From recognition to understanding: enriching visual models through multi-modal semantic integration

Sharifi Noorian, S.

DOI
10.4233/uuid:51fc7a95-fe5f-4519-989c-39646e191148

Publication date
2025

Document Version
Final published version

Citation (APA)
Sharifi Noorian, S. (2025). *From recognition to understanding: enriching visual models through multi-modal semantic integration*. [Dissertation (TU Delft), Delft University of Technology].
https://doi.org/10.4233/uuid:51fc7a95-fe5f-4519-989c-39646e191148

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

FROM RECOGNITION TO UNDERSTANDING: ENRICHING VISUAL MODELS THROUGH MULTI-MODAL SEMANTIC INTEGRATION

FROM RECOGNITION TO UNDERSTANDING: ENRICHING VISUAL MODELS THROUGH MULTI-MODAL SEMANTIC INTEGRATION

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus prof.dr.ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
Monday 10 February 2025 at 15.00 o'clock

by

Shahin SHARIFI NOORIAN

Master of Science in Geo-informatics,
Technische Universität München, Germany
born in Karaj, Iran.

This dissertation has been approved by the promoters.

Prof. dr. ir. G.J.P.M. Houben
Prof. dr. ir. A. Bozzon
Dr. J. Yang

Composition of the doctoral committee:

Rector Magnificus,	Chairman
Prof. dr. ir. G.J.P.M. Houben	Delft University of Technology, Promotor
Prof. dr. ir. A. Bozzon	Delft University of Technology, Promotor
Dr. J. Yang	Delft University of Technology, Copromotor
Independent members:	
Prof. dr. E.M. van Bueren	Delft University of Technology
Prof. dr. F. Casati	University of Trento, Italy
Prof. dr. J. Good	University of Amsterdam
Prof. dr. ir. M.F.W.H.A. Janssen	Delft University of Technology
Prof. dr. K.G. Langendoen	Delft University of Technology, reserve member

SIKS Dissertation Series No. 2025-
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Keywords: Multi-modal Learning, Semantic Reasoning, Visual Understanding, Deep Learning, Human-in-the-Loop

Style: TU Delft House Style, with modifications by Moritz Beller
https://github.com/Inventitech/phd-thesis-template

ISBN:

E-mail: shahin.sharifi.noorian@gmail.com

An electronic version of this dissertation is available at:
http://repository.tudelft.nl/.

What we observe is not nature itself, but nature exposed to our method of questioning.

Werner Heisenberg

Contents

Summary

This thesis addresses the semantic gap in visual understanding, improving visual models with semantic reasoning capabilities so they can handle tasks like image captioning, question-answering, and scene understanding. The main focus is on integrating visual and textual data, leveraging human cognitive insights, and developing a robust multi-modal foundation model. The research starts with the exploration of multi-modal data integration to enhance semantic and contextual reasoning in fine-grained scene recognition. The proposed multi-modal models, which combine visual and textual inputs, outperform traditional models that rely solely on visuals. This is particularly true in complex urban environments where visual ambiguities often occur. This method emphasizes the significance of semantic enrichment through multi-modal integration, which helps resolve visual ambiguities and improve scene understanding.

To advance visual understanding, the thesis uses a human-in-the-loop approach to identify and characterize unknown-unknowns in visual models. The Scalpel-HS framework is introduced, involving humans to compare what models should have learned versus what they actually know. This framework uses human cognitive abilities to detect and address semantic gaps, improving the reliability of visual models, which is critical in applications like medical imaging and autonomous driving. By integrating human insights, the framework further enhances the interpretability and accuracy of visual models. Additionally, the Perspective tool is developed to identify and characterize atypical images through human computation. This tool is crucial in detecting instances that current models fail to recognize, thereby enhancing model performance in diverse and challenging scenarios. The ability to identify atypical instances is important for improving model robustness and ensuring reliable performance in real-world applications. These research works underscore the value of human-computation systems in boosting machine learning models' abilities.

The thesis concludes with the introduction of GraphFusion, a multi-modal graph neural network pre-trained on large-scale, unlabeled datasets. GraphFusion captures long-range dependencies across different modalities, showing improvements in tasks such as cross-modal retrieval and visual question answering. This model demonstrates the potential of using large amounts of unlabeled data to train more comprehensive and versatile models. The success of GraphFusion in various tasks shows its ability to learn unified multi-modal representations, which are essential for advanced visual understanding.

Overall, the contributions of this thesis emphasize the potential of integrating multi-modal data and human cognitive insights into visual understanding models. The research provides valuable methods and tools that enhance the semantic reasoning capabilities of these models, ensuring they are more aligned with human-like understanding. The findings and innovations presented in this thesis pave the way for future research in multi-modal representation learning, highlighting the importance of combining diverse data sources and human expertise to advance the field of computer vision.

Samenvatting

Dit proefschrift behandelt de semantische kloof die bestaat in visueel begrip en verbetert visuele modellen met behulp van semantische redeneervermogens zodat ze taken zoals beeldbijschriften, vraagbeantwoording en scènebegrip aankunnen. De nadruk ligt op de integratie van visuele en tekstuele gegevens, het benutten van menselijke cognitieve inzichten en het ontwikkelen van een robuust multi-modaal basismodel. Het onderzoek begint met de verkenning van multi-modale gegevensintegratie om het semantisch en contextueel redeneervermogen bij fijne scènedetectie te verbeteren. De voorgestelde multi-modale modellen, die visuele en tekstuele inputs combineren, presteren beter dan traditionele modellen die uitsluitend op visuele gegevens vertrouwen, vooral in complexe stedelijke omgevingen waar visuele ambiguïteiten vaak voorkomen. Deze methode benadrukt het belang van semantische verrijking door multi-modale integratie, wat helpt bij het oplossen van visuele ambiguïteiten en het verbeteren van het scènebegrip.

Om het visuele begrip verder te verdiepen, introduceert het proefschrift een 'humanin-the-loop'-benadering om onbekende onbekenden in visuele modellen te identificeren en te karakteriseren. Het Scalpel-HS-raamwerk wordt geïntroduceerd, waarbij mensen worden ingezet om te vergelijken wat modellen zouden moeten hebben geleerd met wat ze daadwerkelijk weten. Dit raamwerk maakt gebruik van menselijke cognitieve vaardigheden om semantische lacunes op te sporen en aan te pakken, waardoor de betrouwbaarheid van visuele modellen wordt vergroot, vooral in kritieke toepassingen zoals medische beeldvorming en autonoom rijden. Door menselijke inzichten te integreren, verbetert het raamwerk de interpreteerbaarheid en nauwkeurigheid van visuele modellen aanzienlijk, wat de weg vrijmaakt voor betrouwbaardere AI-systemen. Bovendien wordt het Perspective instrument ontwikkeld om atypische afbeeldingen te identificeren en te karakteriseren door middel van menselijke computationele inspanningen. Dit instrument is cruciaal voor het detecteren van gevallen die huidige modellen niet herkennen, waardoor de prestaties van modellen in diverse en uitdagende scenario's worden verbeterd. Het vermogen om atypische gevallen te identificeren is essentieel voor het verbeteren van de robuustheid van modellen en het waarborgen van betrouwbare prestaties in realistische toepassingen. Het onderzoek onderstreept het belang van 'human-computation'-systemen om de mogelijkheden van machine learning-modellen aan te vullen bij het effectief omgaan met atypische gegevens.

Het proefschrift mondt uit in de introductie van GraphFusion, een innovatief multimodaal grafisch neuraal netwerk dat is voorgetraind op grootschalige, niet-gelabelde datasets. GraphFusion legt langeafstandsrelaties vast tussen verschillende modaliteiten en toont significante verbeteringen in taken zoals cross-modale retrieval en het beantwoorden van visuele vragen. Dit model demonstreert het potentieel van het gebruik van grote hoeveelheden niet-gelabelde gegevens om uitgebreidere en veelzijdigere modellen te trainen. Het succes van GraphFusion in verschillende taken benadrukt het vermogen om geïntegreerde multimodale representaties te leren, wat essentieel is voor geavanceerd visueel begrip.

Kortom, de bijdragen van dit proefschrift benadrukken het transformerende potentieel van de integratie van multimodale gegevens en menselijke cognitieve inzichten in visuele modellen. Het onderzoek levert waardevolle methoden en instrumenten op die de semantische redeneercapaciteiten van deze modellen verbeteren, waardoor ze beter aansluiten bij het menselijk begrip. De bevindingen en innovaties in dit proefschrift effenen het pad voor toekomstig onderzoek in multimodaal representatie-leren en benadrukken het belang van het combineren van diverse gegevensbronnen en menselijke expertise om het vakgebied van computer vision verder te brengen.

Acknowledgments

Looking back on this PhD journey, I am overwhelmed with gratitude for the countless people who have supported me, inspired me, and walked beside me through this incredible chapter of my life. This dissertation is not just the result of years of research but also of the love, patience, and encouragement of so many wonderful individuals.

First and foremost, my deepest gratitude goes to Sepideh—my best friend, my partner in every sense, and the person who has shared every step of this journey with me. From the moment we sat side by side as bachelor students to working as interns at the same company, and later continuing through our master's and PhD paths together, you've always been there. Our shared experiences, countless laughs, and late-night study sessions have been the glue of our friendship and the joy of my life. Thank you for your endless patience, for always believing in me even when I doubted myself, and for reminding me to laugh and enjoy the small moments, no matter how tough things got. This PhD may be my name on paper, but it's a journey we've traveled together.

To my family, my constant source of strength. My mother, Dayan, your love and encouragement have been the foundation of my confidence and perseverance. My father, Agha Sharifi, who is no longer with us, remains an enduring presence in my life. Your wisdom, kindness, and values continue to guide me every day. To my siblings—Babak, Afshin, and Ghazaleh—thank you for always having my back, no matter what. Afshin, I still remember when everyone had their own opinion about what I should study, but you were the one who suggested computer science. I wasn't sure at the time, but I trusted you, and it turned out to be the best decision I could have made. If I'd taken another path, my life would have been completely different—so thank you for steering me in the right direction. Babak and Ghazaleh, your love and encouragement have been constant, even when I was too busy or distracted to say how much it meant. You three are the best team anyone could ask for. To Ramin, Parisa, and Farah, you've been more than just in-laws—you've been like a second family to me, and your kindness, humor, and support have meant so much. And to my amazing nieces—Ghazal, Baran, and Hasti—you're the brightest sparks in our family, always bringing a smile to my face, no matter the distance.

To my parents-in-law, Mahmoud and Mahnaz, your kindness and encouragement have meant so much to me. To Ali, Negin, Danial, and Danika—being with you is always pure joy. Despite the distance, every moment we share is filled with laughter and unforgettable memories. No matter what I say about your kindness and chivalry, it won't be enough. To Raheleh, your kindness and support have always meant so much. You have a way of making everything feel easier and more manageable, and Adrian's curiosity and charm always brighten the room. And Parham, you've been more than just a brother-in-law (a.k.a. *JARI*)—you're a true so-called "*KA*." The time we spend talking and bouncing ideas around

has made this journey all the more enjoyable.

To my childhood friend, Nima (Zarei Neyestanak)—how crazy is it that two kids who grew up together ended up in the same country after 20 years again? Whether we see each other often or not, just knowing you're around always gives me a sense of relief. And Alalleh, thanks for putting up with both of us!

To the "Sibil" (ᔓᔐ) crew—Parham, Sia, and Kawe—thank you for the years of countless moments of laughter, camaraderie, and truly *NICE!!!* memories that will always stay with me. To my great friends—Nima, Magda, Samira, Jos, Ada, Peyman, Soheyla, Sara, Asal, Shayan, Bas, Shabnam, Majid, and Ghazaleh—all the moments we've shared over the years have been really priceless for me.

To Geert-Jan, thank you for your thoughtful feedback and consistent guidance throughout this journey. Your support and understanding, especially as I navigated the challenges of being a part-time PhD candidate, have been invaluable in helping me stay focused and achieve my goals.

Alessandro, thank you for your encouragement and constructive feedback, which have been instrumental in shaping my work. Your approachable nature and willingness to engage in meaningful discussions made this journey more manageable and enjoyable.

Jie, thank you for the countless brainstorming sessions and for always being ready to dive deeply into ideas. Your sharp insights and ability to connect different perspectives not only helped solve problems but also inspired me to think more creatively.

To the members of my committee, thank you for your valuable feedback and for taking the time to review my work. I am truly grateful for your contributions to this final milestone.

To my incredible colleagues and office mates in the WIS group—Asterios, Christoph, Achilleas, Ujwal, Sihang, Agathe, Petros, Christos, Vasilis, Kyriakos, Giorgos, Oana, Tim, Manuel, Felipe, Ziyu, David, Nirmal, Arthur, Guz, Sara, Andra, Alisa, Peide, Garrett, Lorenzo, Nadia, Daphne, and Ali—thank you for the discussions, laughter, and the sense of community.

Finally, to everyone at WIGeoGIS, you've been so much more than just colleagues—you've been like a second family to me. Having spent over 10 years with you, starting as an intern and growing alongside this incredible team, I've experienced nothing but support, collaboration, and kindness. A special thanks to the management board—Michael, Zoltan, Georg, Wolf, and Markus—for believing in me, encouraging my growth, and making WIGeoGIS a place where I always felt valued. Balancing work and my PhD would have been impossible without your understanding and flexibility, and for that, I'll always be grateful.

Shahin
Den Haag, 2024

1

Introduction

1.1 Background and Motivations

Human cognition relies heavily on visual understanding for recognition, interpretation, and decision-making based on the visual stimuli in our environment [1]. Over recent years, the field of computer vision has witnessed substantial advancements, fueled by progress in deep learning methodologies [2], especially Convolutional Neural Networks (CNNs) [3], and the availability of extensively annotated image datasets such as ImageNet [4]. This progress has led to significant breakthroughs in various vision tasks like object recognition, image classification, and semantic segmentation, subsequently fostering the growth of numerous real-world applications from autonomous vehicles to medical image diagnostics [5].

Despite the remarkable progress in visual recognition tasks, existing models cannot yet effectively cope with the complexity and diversity of real-world scenarios due to the lack of semantic reasoning abilities. High-level semantic reasoning is vital in applications such as autonomous driving systems, which must interpret complex scenes involving objects like pedestrians, other vehicles, and traffic signs while accounting for contextual information, including weather conditions and road infrastructure [6]. Similarly, medical image analysis often necessitates precise diagnosis of diseases and abnormalities contingent on accurately interpreting subtle visual cues, patient history, and other non-visual data [7]. In both cases, attaining human-level semantic reasoning is crucial for achieving high accuracy and robustness in real-world applications.

Nonetheless, existing models often struggle with tasks demanding high-level reasoning, contextual understanding, or the incorporation of multiple modalities [8]. For instance, current models frequently exhibit shortcomings in accurately interpreting complex scenes with occluded objects or when lighting conditions or viewpoint changes lead to significant alterations in object appearances [9]. Such limitations in contemporary visual understanding models primarily stem from the "semantic gap" between low-level visual features and high-level human interpretation [10]. These models often fail to capture the rich semantics and interplay between different modalities, such as textual and auditory information, which humans seamlessly integrate into their visual understanding [11]. Furthermore, models

should be able to reason and generalize beyond the specific instances presented in the training data, adapting to new and previously unseen situations with minimal supervision [12]. This necessitates incorporating semantic reasoning capabilities into learning, empowering models to utilize their prior knowledge to make more informed predictions and decisions.

One promising avenue to overcome these limitations is the development of multi-modal models that combine visual and textual features [13]. By incorporating text-based semantic information, multi-modal models can more effectively capture comprehensive visual content and complex real-world scenes [1, 14], which is particularly crucial for various domain-specific tasks such as storefront recognition or content moderation, where visual features alone might be inadequate to distinguish categories [15]. Nonetheless, efficiently integrating multi-modal semantic information presents several challenges [16]. Fusing visual and textual features is difficult as these modalities have different statistical properties and levels of abstraction[17]. This requires specialized techniques to fuse information while preserving unique characteristics and enabling learning from complementary aspects. Complex architectures are often needed to integrate multi-modal features and learn meaningful representations leveraging each modality's strengths[13]. Moreover, the performance of multi-modal models largely relies on high-quality labeled data, which can be costly and time-consuming to obtain[18]. Acquiring multi-modal data is particularly challenging due to the necessity for consistent and accurate annotation of visual and textual components, demanding substantial human effort and resources.

This thesis endeavors to address the challenges above in pursuing human-level visual understanding. By exploring novel approaches that minimize reliance on extensive labeled data and harness the potential of multi-modal information, we aim to develop models that can better comprehend complex visual concepts and adapt to a vast range of scenarios. Moreover, incorporating human-in-the-loop strategies will enable us to leverage human cognition and expertise to identify and bridge the semantic gaps in current models, ultimately enhancing their performance and interpretability.

Pursuing human-level visual understanding is not only of academic interest but also has far-reaching implications for many real-world applications. Improved visual understanding models can revolutionize our interaction with and benefit from artificial intelligence systems, from autonomous vehicles and robotics to healthcare and content moderation. By addressing the limitations of existing models and pushing the boundaries of visual understanding, this thesis aims to contribute to advancing computer vision research and its practical applications in our increasingly digitalized world.

1.2 Problem Statement

Developing visual understanding models that can perform at a human level necessitates addressing several challenges rooted in real-world visual information's inherent complexity and variability. These challenges encompass aspects such as the subtleties within visual appearances, the importance of context, the identification of feature blind-spots, the proactive detection of atypical instances, and the challenges in multi-modal vision-language representation learning.

Subtleties within Visual Appearances.
In many real-world applications, objects or scenes share similar visual features, making it difficult for models to accurately recognize and distinguish between them [19]. For instance, in storefront recognition, as shown in Figure 1.1, the images of two different business places (pizzeria and bakery) appear very similar. Thus, only textual information can semantically identify the correct class of business places [20]. Addressing this challenge is crucial for achieving accurate scene understanding and decision-making, as well as for the advancement of domain-specific recognition tasks [3].

Figure 1.1: An example of visual ambiguity where the only discriminative feature for recognizing the image class lies in the semantics of scene text.

The Importance of Context.
Current image recognition models effectively learn image representations, which can be applied to high-level image analysis tasks like scene classification [3, 21, 22]. However, a significant difference exists between scene classification and general image classification. General image classification focuses on object-centric images, where each category is closely related to an object in the image. In contrast, scene classification involves recognizing multiple objects and their spatial layout within a scene [23]. As shown in Figure 1.2, Object-centric images typically contain a single object, and classification relies on the object's features. Scene classification, however, requires recognizing key objects and understanding their relationships, necessitating higher-level semantic representation [24]. Context plays a vital role in visual understanding, as an object or scene's meaning often depends on its environment [25]. Current models frequently struggle to capture and incorporate contextual information, leading to misinterpretations and errors in high-level

1

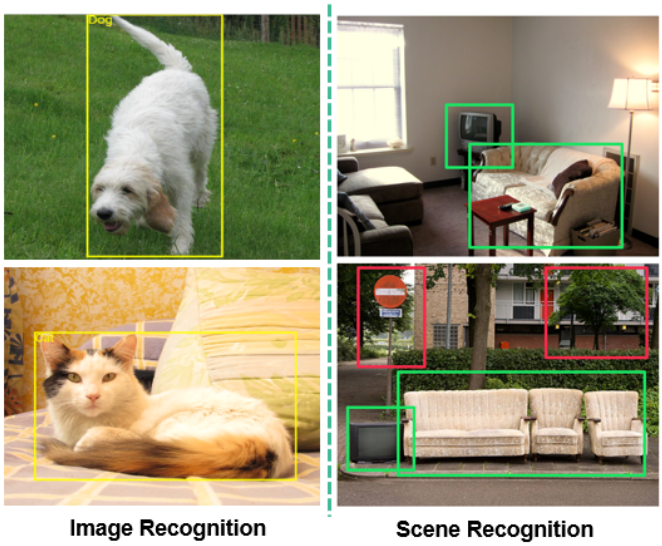


Figure 1.2: The difference between image classification and scene classification task. Images for scene classification are usually not object-centric.

reasoning tasks. For instance, understanding the context of an individual's actions in surveillance systems is essential for differentiating between normal behavior and suspicious activities [26]. In autonomous vehicle navigation, contextual understanding is crucial for safe and efficient operation, as the vehicle must interpret other road users' actions and navigate complex traffic scenarios [27]. Addressing the challenge of contextual understanding is critical for ensuring accurate and reliable visual understanding in complex real-world situations [28].

Identifying and Characterizing Feature Blind-Spots.
Visual understanding models may have feature blind-spots (called unknown-unknowns) that can lead to misclassifications or misinterpretations [29]. The leading cause of these feature blind-spots is usually rooted in an imbalance of training data, which can cause models to fail in real-world scenarios where accurate scene recognition is essential, such as autonomous vehicle navigation, medical image analysis, or surveillance systems [9]. For example, the inability to recognize a specific traffic sign or pedestrian in autonomous vehicle navigation could result in accidents. At the same time, misinterpreting tumor features may lead to incorrect diagnosis or treatment plans in medical image analysis. Similarly, overlooking suspicious activities due to blind-spots may compromise security measures in surveillance systems. By recognizing and characterizing feature blind-spots, researchers can better understand the limitations of visual understanding models and explore opportunities to improve their performance. Leveraging human cognitive abilities may offer valuable insight into addressing these challenges and enhancing the models' overall performance [30].

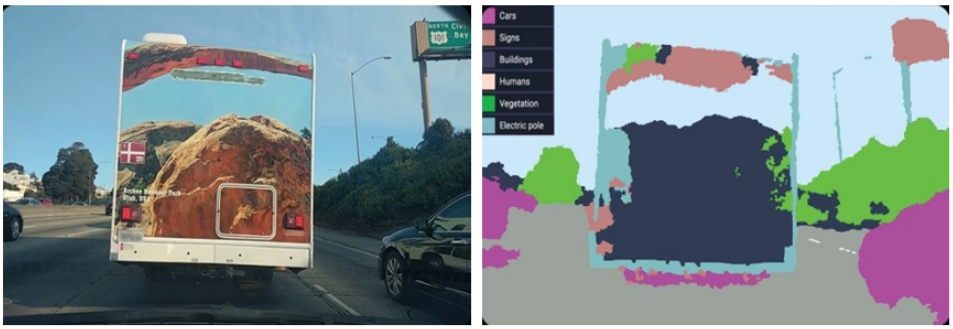


Figure 1.3: Example of an atypical visual concept where current models fail to interpret correctly.

Proactive Detection of Long-tail Visual Concepts.
Real-world data often contain many infrequent categories, referred to as long-tail visual concepts. Traditional models struggle to learn and recognize these concepts due to the scarcity of labeled examples. Developing a scalable human-computational system for proactive detection of atypical instances unrecognizable by visual models is essential for improving the robustness and reliability of these models in real-world scenarios [31]. By identifying such instances, researchers can better understand and address the limitations of visual understanding models when dealing with poor high-level reasoning. For example, detecting atypical instances like a truck's back-side painted to resemble an open road in autonomous vehicle navigation could significantly impact the vehicle's ability to recognize and react appropriately 1.3. In remote sensing applications, detecting atypical instances might be critical for monitoring environmental changes or identifying areas of interest for further investigation [32]. In disaster response scenarios, identifying unusual instances of damage or infrastructure collapse can help allocate resources more effectively and prioritize rescue efforts. Similarly, detecting atypical animal behavior or habitat changes in wildlife conservation can inform conservation strategies and help protect endangered species. By proactively detecting these atypical instances, visual understanding models can better adapt to real-world challenges and provide more accurate and reliable insights across various applications.

Exploiting Web-scale Unlabeled Data for Multi-modal Representation Learning.
Multi-modal vision-language understanding involves the integration of visual and linguistic modalities to enhance the comprehension of complex real-world information [33]. However, existing methodologies struggle to effectively capture the complex interdependencies between these modalities, leading to misalignments in their representations and suboptimal performance in downstream tasks [34]. Additionally, the dependency on high-quality labeled training data can hinder the scalability and applicability of these models. Addressing these challenges, including the effective utilization of large-scale unlabeled datasets, is essential for advancing vision-language understanding and enhancing the performance of multi-modal visual understanding models in various real-world applications, such as image captioning [35], visual question-answering [36], and visual grounding [37].

1

1.3 Research Questions

The primary goal of this thesis is to explore the following main research question:

- **MRQ:** How can we develop visual understanding models that achieve human-like semantic comprehension of complex visual concepts while minimizing the reliance on large-scale, high-quality labeled datasets?

The main research question addresses the challenge of creating advanced semantic-enhanced visual understanding models. We aim to build models capable of capturing intricate concepts and relationships across different modalities, potentially enabling high-level reasoning and comprehension in visual models across various applications and tasks. In pursuit of this goal, we structure our research around the three key aspects that approach the main research question from different perspectives: 1) Multi-Modal Data Integration, 2) Human-in-the-Loop Approach, and 3) Foundation Model for Multi-modal Learning. Subsequent sections will introduce specific sub-research questions that align with these aspects, each delving into the nuances of their respective domains to support the overarching research aim.

Multi-Modal Data Integration
This thesis aims to semantically enhance visual understanding models by effectively integrating visual and textual information within a multi-modal learning framework. This integration is crucial for complex tasks requiring context sensitivity, cross-modal interaction, and high-level reasoning. Examples of these tasks include scene recognition, visual question answering, and multi-modal information retrieval. To this end, we formulate the first sub-research question as follows:

> ***RQ1:*** *How can the integration of multi-modal data, specifically text and visual information, improve the semantic and contextual reasoning abilities of models in fine-grained scene recognition?*

In Chapter 2, we study the effect of merging textual cues with visual data on the efficacy of models in fine-grained scene recognition tasks. The hypothesis is that integrating signals from other modalities can offer more context for correctly interpreting visual scenes. This method involves extracting textual information from visual data and supplementing the visual model with additional semantic information. This helps resolve uncertainties in the visual data and allows the model to differentiate between visually similar but contextually distinct scenes. By investigating this, we hope to refine visual understanding models and make them better at handling complex real-world visual recognition tasks, similar to how humans understand visual information. Chapter 2 is based on the two full conference papers as listed below:

- *Noorian, Shahin Sharifi, Achilleas Psyllidis, and Alessandro Bozzon. "ST-Sem: A Multimodal Method for Points-of-Interest Classification Using Street-Level Imagery." International Conference on Web Engineering. Springer, Cham, 2019. [38]*
- *Sharifi Noorian, Shahin, et al. "Detecting, classifying, and mapping retail storefronts using street-level imagery." Proceedings of the 2020 International Conference on Multimedia Retrieval. 2020. [39]*

Human-In-the-Loop Approach

Enhancing visual understanding models with semantic capabilities involves recognizing and addressing semantic gaps and feature blind spots. These gaps can result in certain errors, referred to as unknown-unknowns, where a model confidently makes incorrect predictions. Identifying these errors is challenging due to the model's overconfidence. Such errors may not be apparent during the model's development and training but can lead to critical issues after deployment in production. This poses significant reliability challenges for visual understanding models, particularly in high-risk tasks like self-driving or healthcare. Recent research highlights the importance of human-in-the-loop methods for interpreting machine learning. Human judgment and cognitive abilities can effectively illuminate what the model has understood and its blind spots - what it should have learned but didn't - for a specific visual-semantic task. To this end, we formulate the second sub-research question as follows:

> ***RQ2:*** *How can we efficiently utilize human cognitive insights to identify and characterize unknown-unknowns in visual models?*

To address **RQ2** in Chapter 3, we explore the utilization of human-in-the-loop approaches to deepen the understanding of 'unknown-unknowns' - errors in machine learning models due to the model's overconfidence in its incorrect predictions. We introduce a framework that employs human semantic analysis to pinpoint and describe the nature of these gaps at scale. By engaging humans to delineate what the machine is expected to know and contrast it with its actual knowledge, we propose a novel method of characterizing unknown-unknowns at the conceptual level. The framework combines information extraction with machine learning interpretability methods and scales human efforts using data partitioning and sampling techniques. The resulting characterization of unknown unknowns is rich and descriptive and paves the way for more efficient detection methods. This research question investigates the effectiveness of systematically utilizing human cognitive abilities to identify and characterize semantic gaps in a structured format. Consequently, it can potentially facilitate the creation of more reliable, robust, and semantically enhanced systems. Chapter 3 is grounded in the following paper on unknown-unknowns characterization in image recognition:

- *Sharifi Noorian, Shahin, et al. "What Should You Know? A Human-In-the-Loop Approach to unknown-unknowns Characterization in Image Recognition." Proceedings of the ACM Web Conference 2022. [40]*

1

Continuing with the theme of the human-in-the-loop approach, we shift our focus from examining the semantic capacity of visual models to investigating the characteristics of atypical data instances with long-tail visual concepts. These instances often cause most visual recognition models to fail due to their distinct or uncommon visual concepts and the models' limited reasoning ability. To this end, our third sub-research question centers on the proactive identification and description of atypical visual instances, as follows:

> ***RQ3:*** *How can we develop a scalable human-computation system to proactively identify and characterize atypical instances that visual models often fail to recognize due to a lack of high-level semantic reasoning?*

In Chapter 4, we aim to address **RQ3** by examining the effectiveness of a system that combines human intuition with algorithmic efficiency to identify and characterize a wide range of atypical data instances. This system provides insights into the nature of atypical visual concepts, thereby enabling the development of benchmarks for evaluating the models' semantic comprehension and reasoning capability. Chapter 4 is based on the following research paper on leveraging human understanding for identifying and characterizing image atypicality:

- *Sharifi Noorian, Shahin, et al. "Perspective: Leveraging Human Understanding for Identifying and Characterizing Image Atypicality." IUI '23: Proceedings of the 28th International Conference on Intelligent User Interfaces. [41]*

Foundation Model for Multi-modal Learning

In our final attempt to address the main research question, we focus on developing a foundation model pre-trained with large-scale multi-modal data, which can be generalized and transfer the learned knowledge to a wide range of downstream cognitive tasks through fine-tuning. Thus, we formulate the fourth sub-research question as follows:

> ***RQ4:*** *How can we develop and pre-train a foundation model capable of cross-modal comprehension and reasoning by leveraging large-scale multi-modal unlabeled datasets?*

To address the **RQ4**, in Chapter 5, we aim to design and train a foundation model capable of capturing complex interactions and long-range feature dependencies within and across different modalities. We demonstrate the efficacy of this approach in enhancing comprehension and high-level visual-semantic reasoning, which can be generalized to various downstream cognitive tasks. Chapter 5 is based on the following publication:

- *Sharifi Noorian, Shahin, et al. "GraphFusion: Unified Vision-Language Representation Learning using Heterogeneous Graph Neural Networks." This work has been submitted to the 2025 International Conference on Learning Representations (ICLR 2025) conference and is currently under review. [41]*

In conclusion, these research questions aim to bridge the gap between current visual understanding models and the nuanced complexities of real-world visual and textual data.

1.4 Original Contribution

In this thesis, we have made several significant original contributions to Visual Understanding models and multi-modal semantic techniques. Our research offers not only valuable theoretical insights but also presents practical implementations in the form of software tools, which have been published on GitHub. The following sections outline the original contributions associated with each thesis chapter.

Enhancing Visual Recognition by Multi-Modal Data Integration (Chapter 2)

We presented a novel multi-modal method for effectively combining visual and textual features in Visual Understanding models, specifically for domain-specific scene recognition tasks. Our approach allowed for integrating textual information from input images to enrich the visual features, leading to improved model performance. We publish two accompanying software for the methods we introduced in Chapter 2. All source codes and scripts corresponding to methods and experiments in Chapter 2 are available on [1].

Identifying Semantic Gaps with Human-Cognitive Integration (Chapter 3)

This chapter explored human-in-the-loop methodologies for identifying and addressing understanding gaps in Visual Understanding models. We developed an innovative approach leveraging human cognitive capabilities to discover the model's blind spot and feature deficiency. Our software implementation, published on [2], demonstrates the practical application of our proposed approach, enabling researchers and practitioners to incorporate human expertise in their model refinement processes.

Decoding Long-tail Visual Concepts Using Human-Computational Approach (Chapter 4)

We presented a scalable human-computational framework designed to detect and characterize atypical instances that current recognition models struggle to recognize due to a lack of high-level reasoning. This framework contributes to enhancing the robustness and adaptability of Visual Understanding models. Our software implementation, available on [3], showcases the efficiency and effectiveness of our framework, providing a valuable resource for researchers and practitioners working on similar problems.

[1] https://doi.org/10.4121/507f9fdd-a38d-449d-9784-a41eec701899
[2] https://doi.org/10.4121/cc6ca9df-ef0d-4af7-8feb-3ee6ecf74d7c
[3] https://doi.org/10.4121/90e68f71-d320-42cb-8f60-bb60aae5eaab

1

A Graph-based Foundation Model For Multi-modal Learning (Chapter 5)
This chapter investigated the potential of leveraging large-scale unlabeled datasets for training semantic-aware visual representation learning models. We developed a method for effectively utilizing unlabeled data and weak supervision sources, reducing reliance on high-quality labeled training data and achieving more cost-efficient training processes in downstream visual understanding tasks. Our software implementation, published on [4], demonstrates the practical application of our proposed approach, offering researchers and practitioners a valuable tool for training better-performing and more interpretable Visual Understanding models.

These original contributions not only advance the field of Visual Understanding models and multi-modal semantic techniques but also provide practical software implementations that can be used by researchers and practitioners alike. By making these resources publicly available, we hope to foster further exploration and development in this area, ultimately creating more robust, interpretable, and efficient Visual Understanding models.

[4] https://github.com/shahinsharifi/GraphFusion

1.5 Additional Contribution

In addition to the core contributions of the thesis, we have also engaged in two research projects in the context of urban analytics. Although these works are not directly related to the main focus of the thesis, they are noteworthy contributions and have been included in the appendix. Both projects have their software implementations and provide valuable insights into the urban analytics domain.

P-median Model for Facility Siting with Live Traffic Data [42]
This research project introduced a novel p-median model for the facility siting problem that considers live traffic data. Our model can provide more accurate and efficient solutions for optimal facility location decisions by incorporating real-time traffic information. This work advances the field of urban analytics and facility location optimization and has practical implications for urban planners and decision-makers. All associated data and source code for reproducing our results are publicly available [5,6], demonstrating the effectiveness of our approach and providing a valuable resource for researchers and practitioners working on similar problems.

Measuring Subjective Perception in Urban Environments using Street-level Imagery[43]
In the second research project, we developed an application to measure the subjective perception of people in urban environments using street-level imagery. By capturing the opinions and perceptions of individuals, this work provides a valuable understanding of the factors that contribute to the quality of urban life. The insights gained from this research can inform urban planners and policymakers in their efforts to create more livable and sustainable urban environments. All materials for reproducing our experiments are published online [7,8], showcasing the functionality of our application and providing a practical tool for researchers and practitioners interested in subjective perception analysis using street-level imagery.

These additional research collaborations demonstrate the versatility of our research interests and contribute to the broader field of urban analytics. By making these resources available on GitHub, we hope to foster further exploration and development in these areas, ultimately contributing to better urban environments and more effective urban planning strategies.

1.6 Thesis Outline

This thesis consists of six chapters, including the current chapter (Chapter 1), which consists of the problem statement, research questions, and the original contributions of this thesis. The remaining chapters are based on full research papers published at various conferences:

[5] https://github.com/shahinsharifi/AGILE2018
[6] https://doi.org/10.4121/507f9fdd-a38d-449d-9784-a41eec701899
[7] https://doi.org/10.17605/osf.io/aqgxr
[8] https://github.com/shahinsharifi/subjectivity.git

- **Chapter 2** is based on two full research papers published at the International Conference on Web Engineering (ICWE 2019) and the International Conference on Multimedia Retrieval (ICMR 2020), respectively.
- **Chapter 3** is based on a full research paper published at the Web Conference (WWW 2022).
- **Chapter 4** is based on a full research paper published at the International Conference on Intelligent User Interfaces (IUI 2023).
- **Chapter 5** is based on a full research paper that will be submitted to the International Conference on Learning Representations (ICLR 2025: Under Review)
- **Chapter 6** summarizes the main findings, outlines the core contributions, and provides an outlook on future works.

2

Enhancing Visual Recognition by Multi-Modal Data Integration

In the field of artificial intelligence and computer vision, a significant challenge remains in bridging the "semantic gap" - the difference between the computational interpretation of visual data and the meaningful, human-like understanding of such images. This chapter explores the complexities of this problem, with a specific focus on the limitations of visual models that rely solely on visual cues for comprehension.

Real-world visual concepts often require more than traditional image processing techniques can offer, highlighting the need for a more comprehensive approach. To address this challenge, our research adopts a multi-modal learning framework. This shift involves integrating textual and visual data to enhance the semantic understanding of visual models. By incorporating text extracted from images, we can achieve a deeper and more contextual comprehension of the visual content, significantly reducing the semantic gap. Our approach does not rely solely on large-scale, high-quality labeled datasets, which is a common constraint in traditional visual recognition models. Instead, it harnesses the synergy of multiple data modalities to improve understanding.

In this work, we showcase the practical application and evaluation of our multi-modal approach by developing an innovative system designed for detecting, recognizing, and automatically geolocating business stores using street-level imagery. Our system demonstrates proficiency in identifying storefronts and distinguishing their types—from bakeries and restaurants to toy shops—by leveraging visual and textual cues extracted from images. We assess the system's performance by its accuracy in these tasks, illustrating the substantial advantages of multi-modal learning in practical scenarios. This chapter thoroughly examines our approach to enhancing the semantic understanding of models, highlighting the significant potential of multi-modal learning to bridge the semantic gap in computer vision. Furthermore, this research lays the groundwork for future studies, suggesting that

our work has broad implications, not just for a specific application such as business store mapping. It also promotes the development of more advanced, context-aware systems that can interpret complex visual environments similar to human perception. The content of this chapter is based on the following papers:

- Noorian, Shahin Sharifi, Achilleas Psyllidis, and Alessandro Bozzon. "ST-Sem: A Multimodal Method for Points-of-Interest Classification Using Street-Level Imagery." Web Engineering: 19th International Conference, ICWE 2019, Daejeon, South Korea, June 11–14, 2019, Proceedings 19. Springer International Publishing, 2019.
- Sharifi Noorian, Shahin, et al. "Detecting, classifying, and mapping retail storefronts using street-level imagery." Proceedings of the 2020 International Conference on Multimedia Retrieval. 2020.

2.1 Introduction

Commercial functions are integral to cities worldwide. These functions' listings are used in mapping and location services, recommender systems, search engines, and social media platforms. For these systems to provide accurate and reliable information to the users, such listings must be kept up to date. One of the most challenging issues is keeping track of the frequent changes that characterize this type of business (e.g., a candy shop turning into a bakery).[1] It is estimated that 10% of establishments go out of business every year, and in some segments of the market, such as the restaurant industry, the rate is as high as 30% [44].

The traditional way of keeping such listings up to date requires lots of manual work and often also entails the integration of several third-party resources (e.g., data from the local chamber of commerce). An opportunity to complement these conventional approaches arises from the recent advent of street-level images available on various platforms (e.g., Google Street View or Mapillary). These frequently updated panoramic views of the urban environment allow us to retrieve pictures of the storefronts at scale. We argue that the information included in the storefronts (e.g., commercial logos, names, text, etc.) could help identify the type of business establishment. Recent studies have used street-level imagery to analyze various aspects of the urban environment [45–47], and automatically detect urban objects [47, 48].

The most challenging aspects of automatically detecting, mapping, and classifying commercial functions from street-level imagery are (1) The high degree of visual variability in storefronts, which limits the accuracy and generalizability of prediction models; (2) Image acquisition factors such as noise, motion blur, occlusions, lighting variations, specular reflections, perspective, and geo-location errors; (3) The need for methods with efficient runtime execution performance, considering the continuous changes and the large number of businesses in a city.

This chapter introduces a multi-modal late-fusion method for a domain-specific visual understanding task. Our proposed method combines visual and textual cues from street-level imagery to make semantically-aware predictions. It can correct semantic ambiguities and incorrect digitization of detected textual information. Our proposed method involves three stages: 1) identifying the physical boundaries of storefronts; 2) recognizing their respective commercial functions (e.g., restaurant, bakery, clothing store, etc.); and 3) estimating their geo-locations. The late-fusion approach makes our storefront-type classification module adaptable, allowing us to utilize various pre-trained models. This feature minimizes the need for training from scratch and broadens the applicability of our method to different street-level imagery datasets from non-English-speaking countries. First, we evaluate each component of our proposed method separately. We compare the detection module with two state-of-the-art methods, Faster R-CNN [49] and SSD [50], which have shown superior performance in several object detection challenges[51, 52]. Results show that, while having higher precision than Faster R-CNN(2 %) and SSD (9 %), our approach is considerably faster than the baselines (up to 60%). Furthermore, we show

[1] See examples at: https://sites.google.com/view/storefrontsmapping

that our proposed recognition method can outperform state-of-the-art visual-only models for POI classification – Places365-CNNs [20] – by 16.86%, and multimodal approaches – Karaoglu et al.[53] by 6.8%. Finally, we ran a crowd-sourcing campaign on Amazon Mechanical Turk. Our proposed approach achieves almost the same precision and recall as a human annotator in detecting and classifying retail storefronts. We also investigate our method's performance in several edge cases to highlight limitations and suggest future directions of improvement.

The remainder of the chapter is organized as follows: Section 2 discusses related work. Section 3 describes the proposed method for updating local business listings from street-level imagery. Section 4 presents the experimental setup and discusses the obtained results. Finally, Section 5 summarizes the conclusions and discusses future lines of research.

2.2 Related work

We discuss related work on knowledge extraction using street-level imagery and fine-grained scene classification.

Knowledge Extraction using Street-level Imagery

Street-level imagery can be a useful data source to extract knowledge about the urban environment [48], especially for tasks requiring high spatial coverage. Recent work shows the feasibility of utilizing street-level imagery in assessing structural changes in urban areas [47], inferring subjective properties of urban areas such as safety, liveliness, and attractiveness [54], mapping urban greenery [55–57], geo-locating high-density urban objects [48], or estimating city-level travel patterns [45]. Other works applied computer vision techniques to Google Street View images for inferring the socioeconomic attributes of neighborhoods in the US [58], finding morphological characteristics to distinguish European cities [59], detection of building entrance in outdoor scenes[60], or detection and classification of traffic signs[61]. Like our work, Yu et al. [62] address the problem of detecting storefronts using street-level imagery. The authors trained a deep learning model on a proprietary dataset(~ 2M annotated images), however, without addressing the classification issue into business-related categories. To the best of our knowledge, our work is the first to address the problem of storefront detection, classification, and geo-localization in an integrated fashion.

Fine-grained Scene Recognition

Deep Convolutional Neural Networks (CNNs) have been successful in various vision-related tasks such as face detection, image segmentation, and scene recognition [21, 63]. However, such breakthroughs in visual understanding do not imply that these models are suitable for fine-grained POI classification based on the visual appearance of storefronts from street-level imagery. This is due to the high degree of intra-class and the low degree of inter-class differences in the appearance of storefronts across business categories [53]. Yan et al. [64] take Spatial Context (i.e., nearby places) into account as complementary information to boost the performance of CNN models for classifying business places. Text in scene images, which frequently appears on shop fronts, road signs, and billboards, usually conveys a large amount of valuable semantic information about the object or the scene in the same image. Regarding the fine-grained classification of storefronts based on their business type,

this textual information is crucial in making more accurate predictions [65]. Most similar to our work is that of Karaoglu et al. [66]. The latter proposed a multimodal approach that combines visual features and textual information from the imagery data in a single feature space as input for an SVM classifier. Our work differs from the existing literature in that we incorporate multilingual word embeddings trained on a large corpus to measure semantic relatedness between spotted textual information in street-level imagery and the candidate types of storefronts. Then, we propose a late fusion approach to leverage the obtained prediction scores of both modalities and generate a final score for each candidate class. We compare *ST-Sem* against the approach of Karaoglu et al. [66] in Section 4, showing improved performance.

2.3 Method

The architecture of our model is depicted in Figure 2.1. It consists of three main modules. The first and most important module is the *storefront detector*, which is designed to extract the physical extent of retail storefronts from street-level imagery. As there is often more than one storefront in an instance of street-level imagery, the detector module outputs a list of bounding boxes. In the second step, detected bounding boxes are iteratively fed into both *classification* and *geo-location estimation* modules. The *classification* module utilizes the bounding box information to crop the original input image and outputs a probability distribution over candidate classes (business types). Simultaneously, the *geo-location estimation* module calculates the actual latitude and longitude of each detected bounding box by using the metadata of street-level imagery. In the following paragraphs, we describe each module in detail.

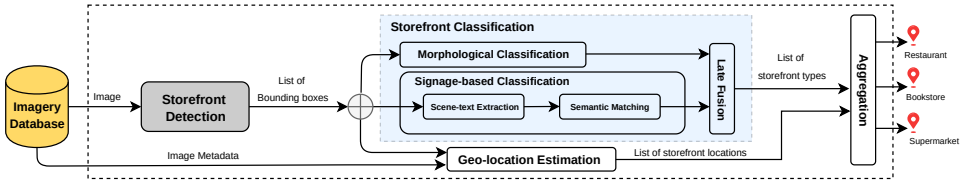


Figure 2.1: The architecture of our end-to-end model

2.3.1 Storefront Detection

For fast storefront detection, we rely on a state-of-the-art one-stage object detector YOLOv3 [67]. The YOLOv3 is the third version of YOLO [68]; while not being the most accurate object detection algorithm, it suits our requirements as it is a very suitable choice for near real-time detection, with limited loss of required accuracy.

Training. We manually annotated 1200 storefront images, which were randomly collected from 5 different countries using Google Street View. We divide the dataset into three parts: training(~ 1000 images), validation(~ 100 images), and test(~ 100 images). We also augment the labeled training data by adding Gaussian noise, varying Brightness, and randomly Rotating images, which results in $1000 \times 5 = 5000$ images in the training set. Due to the scarcity of well-annotated data for business storefront detection, as previous

studies suggested [69, 70], we use a Transfer Learning strategy in order to improve the quality of our detector. As the designer of YOLO has already pre-trained the network using the OpenImages dataset to extract features [71], we immediately applied the pre-trained weight values for further training. As storefront objects are often in the middle or large size, we remove 12 layers from the original YOLOv3 architecture that is responsible for detecting smaller objects. We empirically observed that by removing these layers, the training and inference time decreases by 10%. Additional details on the architecture of the neural network are provided on the companion page.[2]

The input training images are resized to 416×416, and the network has been trained for 5,000 iterations with a batch size of 64. At the end of the training, the loss converges to less than 0.04 on the validation set. As Georgakopoulos et al.[72] suggested, at a general improvement for the training process, we initially set the learning rate to 0.001 for the first 3,000 iterations as we are starting with zero information, and so the learning rate needs to be high. After 3,000 iterations, we decrease the learning rate to a few steps by a factor of 0.1. The YOLO network predicts bounding boxes using dimension clusters called anchor boxes[73]. We calculated anchor boxes for our storefront dataset using the k-Means algorithm and adapted in our output layers.

Inference. At the inference stage, the final output is delivered as a storefront box, paired with its corresponding confidence score. Given a 416×416 image, our storefront detector outputs $((13 \times 13) + (26 \times 26)) \times 3 = 2,535$ bounding boxes. Thus, we must filter boxes based on their objectness score (objectness score reflects how likely the box contains an object[67]) such that boxes having scores below a threshold are eliminated. Furthermore, we perform Soft Non-maximum Suppression [74] to eliminate redundant overlapping boxes with lower confidences.

2.3.2 Storefront Classification

The information captured in street-level imagery is primarily visual. Therefore, storefronts can be described based on the morphological characteristics of their facades, such as height, color, materials, and geometry. Business-related storefronts often have signs or visual labels that display the name, logo, and other relevant information to help people identify the businesses while navigating physical space. These signs can serve as valuable sources of information for classifying retail storefronts. Considering the importance of both visual and textual features, we propose a novel multimodal approach called *ST-Sem* to improve the fine-grained classification of business storefronts. *ST-Sem* leverages visual and textual features extracted from street-level imagery.

The architecture of *ST-Sem* is depicted in Fig 2.2, consisting of three main components. First, the Scene Recognition module predicts the type of storefront at the contextual level based on the common visual characteristics associated with each storefront type. Next, the Scene-text Semantic Recognition module detects textual data in the image, transcribes it into a bag of words, and measures the semantic similarity between the bag of words,

[2]Companion page: https://sites.google.com/view/storefrontsmapping

typically representing the storefront type and each candidate type. Finally, the Class Rank module generates a final score for each candidate class using a Linear Bimodal Fusion (LBF) method. This method combines the prediction scores from the first and second modules. In the following paragraphs, we provide a detailed description of each component.

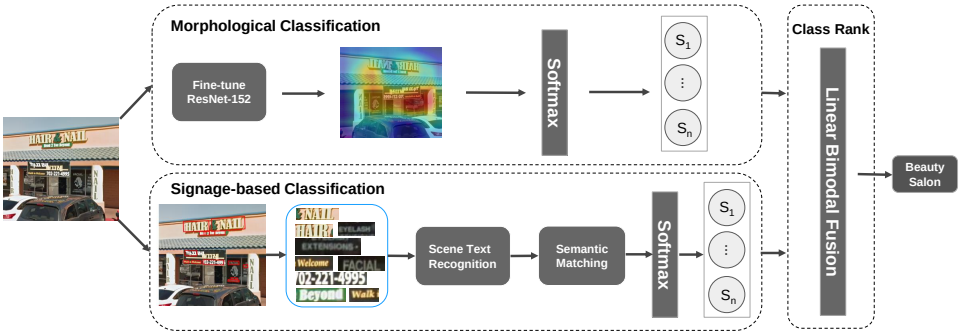


Figure 2.2: The architecture of *ST-Sem*, our proposed method for multi-modal storefront classification.

Morphological Classification

We approach identifying storefront types based on visual information as an image classification problem. For this task, we utilize the Residual Network (ResNet) as our framework, as it has demonstrated excellent performance on ImageNet classification [75]. Specifically, we employ the pre-trained *ResNet152-places365* model provided by [20], which can classify images into 365 classes. However, not all these classes are relevant to our objective of identifying storefront types, such as *cliff* or *coral*. To narrow down the classes, we select 24 business-related place types as our candidate class labels [3]. We then fine-tune the pre-trained *ResNet152-places365* model by removing its last fully-connected layer and replacing it with a new fully-connected layer containing 24 neurons. This is followed by a softmax classifier, which outputs a probability distribution for the 24 storefront types. To initialize the weights of the added fully-connected layers, we randomly generate them from a Gaussian distribution with a mean of zero and a standard deviation of 0.01.

Signage-based Classification

The signage-based classification module is composed of three sub-components. In the following paragraphs, we describe each sub-component in detail.

Scene-text Detection. This sub-module aims to localize and crop text in images as word boxes. Scene-text detection is challenging because scene texts have different sizes, width-height aspect ratios, font styles, lighting, perspective distortion, and orientation. This work incorporates a state-of-the-art method, TextBoxes++ [76], a fast and end-to-end trainable scene-text detector. The reason for choosing TextBoxes++ is that it outperforms

[3]The list of these place types can be found with our code and dataset on the companion page: https://sites.google.com/view/storefrontsmapping

state-of-the-art methods in terms of text localization accuracy and runtime issues of the *IC15* dataset [77] from Challenge 4 of the ICDAR 2015 Robust Reading Competition [4]. The *IC15* dataset comprises 500 test images containing incidental scene text captured by Google Glass. Therefore, it is a good benchmark dataset to evaluate the required scene-text detector for storefront-type classification. We adopt the pre-trained model parameters provided by the authors.

Scene-text Recognition. The task of this sub-module is to transcribe cropped word images into machine-readable character sequences. However, it is considerably difficult to accurately recognize scene texts on street-level imagery because of the varying shapes and distorted patterns of irregular texts. To tackle this problem, we adopt a multi-object rectified attention network (MORAN), proposed by [78]. MORAN consists of a multi-object rectification network (MORN) that rectifies images and an attention-based sequence recognition network (ASRN) that reads the text. Regarding reading rotated, scaled, and stretched characters in different scene texts, this approach outperforms state-of-the-art methods on several standard text recognition benchmarks [78], including the SVT-Perspective dataset [79] which contains 645 cropped images from Google Street View. In training the Scene-text recognition on the MJSynth dataset [80], which is dedicated to Natural Scene Text Recognition, we set the batch size to 64 and the learning rate to 0.01, as suggested by the author. The model is trained for 10 epochs.

Semantic Matching. The semantic matching approach follows the assumption that textual information on the storefront indicates the type of business place. Given this assumption, the goal of the semantic matching module is to predict the type of storefront based on the semantic distance between the words extracted from the image and the standard name of each candidate storefront type, as defined in $ImageNet$ synset[5], such as *cafe*, *bakery* etc. However, not all the words in street-level imagery should necessarily have semantic relations to the place type. Some words may be similar to one of the candidate classes; others may be completely irrelevant. For instance, words such as *hair*, *nail*, or *beauty* on storefront images are likely to be related to a *Beauty Salon*. On the contrary, *OPEN/CLOSE* signs do not give any information about the type of storefront.

The text recognition module could result in some noisy texts, which need to be discarded. Before representing a word spotted by the word vector representation, we use a spell detection tool employing the Levenshtein Distance algorithm[6] to find permutations within an edit distance of 2 from the original word, and therefore remove noisy words. To further remove irrelevant words, we manually curated a blacklist of common – yet irrelevant – words, including verbs like *open*,*close*, *push*,*pull*, etc. After reducing potential noise, we need to detect the language to which the input word belongs. To tackle this problem, we incorporate in our experiments the *polyglot* open source tool [7], which makes language prediction with a corresponding confidence score. If no language can be identified

[4] http://rrc.cvc.uab.es/?ch=4
[5] http://www.image-net.org/synset
[6] https://github.com/barrust/pyspellchecker
[7] https://github.com/aboSamoor/polyglot

for the input word, English will be chosen as the default language.

Once the target language is determined, the recognized word must be transformed into a word vector representation. While there can be many implementations for capturing semantic relatedness[81], previous studies have shown that *word embeddings* [82, 83] perform this task particularly well by measuring the cosine similarity of the word embedding vectors. These vector-based models represent words in a continuous vector space where semantically similar words are embedded close to one another. In our experiments, we adopt FastText [84] to transform recognized texts into a word vector representation. The main reason for incorporating FastText is its promising performance in overcoming the problem of out-of-vocabulary words by representing each word as a bag of character n-grams. We use pre-trained word vectors for two languages (English and German), trained on Common Crawl and Wikipedia [8]. According to the detected language l, the corresponding pre-trained word vector V_l is selected; a pre-trained word vector embedding model encodes each recognized word as v_i. Finally, we use the method proposed by [85] to align the V_l in the same space as the English word vector for multilingual semantic matching. Similarly, each candidate class of storefront type C is represented by a word vector c_j with an English word embedding as a reference. Then, we calculate the cosine similarity between each class label (c_j) and each spotted text (v_i) as follows:

$$\cos\left(\Theta_{ij}\right) = \frac{{v_i}^T c_j}{|v_i||c_j|} \tag{2.1}$$

The probability scores P_i for each candidate storefront type are calculated by averaging the similarity scores of all spotted words:

$$P_j = \frac{\sum_{i=1}^{K} \cos\left(\Theta_{ij}\right)}{K} \tag{2.2}$$

Then, a softmax function is used to normalize the probability scores for each candidate storefront type by the sum of the N candidate ranking scores to sum up to 1. The softmax function can be formulated as follows:

$$\sigma(Z)_j = \frac{e^{Z_j}}{\sum_{n=1}^{N} e^{Z_n}} \tag{2.3}$$

where Z is a vector of probability scores, N is the number of candidate classes, $j = 1, 2, ..., N$ is the index of each probability score in the probability vector Z, and $i = 1, 2, ..., K$ is the index of each spotted text. Similar to the scene recognition module, the scene-text extraction module results in a probability score for each candidate storefront type between 0 and 1.

Class Rank

Inspired by search re-ranking algorithms in information retrieval, we use a Linear Bimodal Fusion (LBF) method (here essentially a 2-component convex combination), which linearly combines the ranking scores provided by the CNN model and the semantic similarity scores

[8] https://fasttext.cc/docs/en/crawl-vectors.html

from the scene-text semantic recognition module, as shown in Equation 4.

$$S_{mixed}(d) = w_v.S_v(d) + (1 - w_v).S_t(d) \tag{2.4}$$

where S_{mixed}, $S_v(d)$, and $S_t(d)$ refer to the final ranking score, visual recognition score, and semantic similarity score for storefront type d respectively, w_v and w_t are the weights for the scene recognition component and scene-text extraction component, and $w_v + w_t = 1$. The weights are determined according to the relative performance of the individual components. Specifically, the weight for the scene recognition module is determined using the following equation:

$$w_v = \frac{acc_v}{acc_v + acc_t} \tag{2.5}$$

where acc_v and acc_t are the measured top@1 accuracy of the scene recognition component and scene-text semantic recognition component, respectively.

2.3.3 Geo-location Estimation and Aggregation

We propose a storefront geo-location estimation algorithm working on the street-level image metadata to geo-locate the storefronts. In previous work [48], the geo-location of an urban object is calculated using the intersection of the central line (symmetry line) of the bounding box and the ground-level horizontal plane (i.e., city ground). We adopt [48] by relying on third-party information about existing buildings, finding which building facade has an intersection with the given bounding box, and then calculating the geo-location of the intersection.

We acquire data on all the building facades in a city from OpenStreetMap (OSM).[9] The map from OSM is composed of nodes and ways, i.e., points and segments. We extract all the segments with the attribute "building" into a set noted as S, representing the collection of all the building facades.

To estimate the geo-location of the storefront from the street-level imagery (Figure 2.3), we trace a ray starting from the location of the camera l_c and going with the heading of the bounding box h, where l_c can be immediately acquired from the metadata of the street-level image and h can be easily calculated according to the position of the bounding box on the image [48]. After that, all the segments (facades) close (not farther than R meters) to the camera location l_c are selected into a set S_c ($S_c \subset S$). Then, we check if a segment s ($s \in S_c$) has an intersection with the ray. If the intersection i_s exists, the distance d_{cs} from the camera l_c to the intersection i_s is calculated and recorded. The segment with the minimum d_{cs} is the facade having the targeted storefront, noted as $\hat{s}$. The location of the corresponding intersection $i_{\hat{s}}$ is the estimated geo-location of the storefront.

[9] https://www.openstreetmap.org/

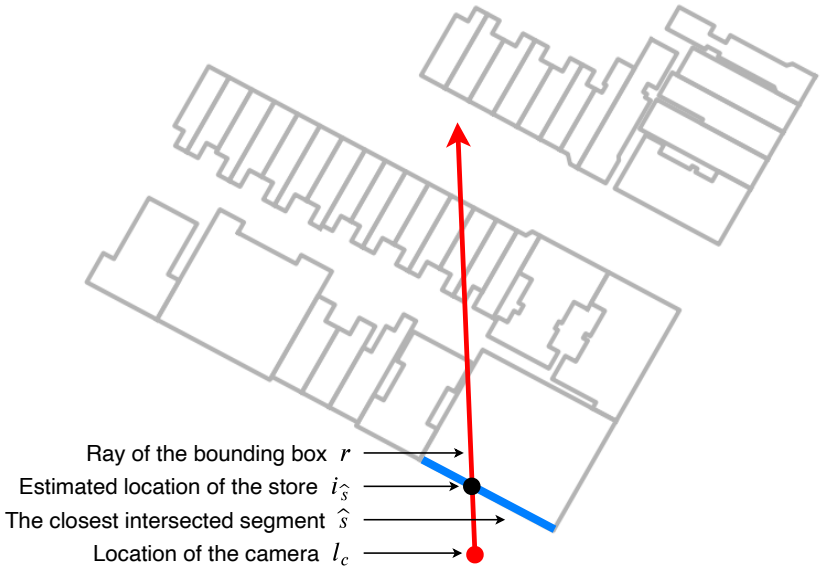


Figure 2.3: Estimation of a storefront's location.

The same storefront might be annotated multiple times from different street-level images or different crowd workers via either an automatic detection method or crowdsourcing. Therefore, "raw" annotations (bounding boxes with labels) produced by automatic detection or crowdsourcing are aggregated to acquire a single estimated annotation for each storefront. We adopt the density-based location aggregation algorithm proposed by [86], which produces one single estimated geo-location from multiple annotations for each storefront. Based on this, the label with the highest confidence score from candidate annotations is selected as the type of storefront.

2.4 Evaluation

In this section, we first describe how datasets are prepared. Then, we separately compare the performance of our detection and classification methods (Shown in Figure 2.1) with 1) State-of-the-art approaches and 2)Human annotators. Finally, we provide a qualitative analysis of the performance of the whole pipeline (detection, classification, and geo-localization).

2.4.1 Dataset

We manually annotated 100 street-level images as a test set for storefront object detection. The dataset comprises 317 storefront bounding boxes(~ 3.2 boxes per image). We refer to this dataset as *Store-Obj*. We also collected a set of single storefront images manually classified into 24 categories. The list of categories comprises 24 top business types,

which are ranked based on their occurrence in popular business listings such as Yelp [10] or Foursquare[11]. We name this dataset as *Store-Scene*. All images in *Store-Scene* only contain a single store, while *Store-Obj* comprises complex panorama images including more than one storefront and many more irrelevant urban objects. Ultimately, we created a small benchmark dataset in Manhattan, New York City. We selected a street about three kilometers long and iteratively collected 150 panoramic images along the street using Google StreetView API. Then, we manually verified that ~ 120 unique businesses exist in the vicinity of the street mentioned above using Google Places API[12]. We observed that the type of collected businesses corresponds to 18 categories of *Store-Scene* dataset. This dataset is used for evaluating the entire pipeline of our model (detection, classification, and geo-localization) in comparison to human performance and for qualitative analysis. We refer to this dataset as Store-location. The properties of each dataset are described in Table 2.1.

Table 2.1: Dataset statistics

Dataset	Problem	#Categories	Training	Testing
Store-Obj	detection	1	1,000	200
Store-Scene	classification	24	-	1,100
Store-Location	detection, classification, geo-localization	18	-	150

2.4.2 Implementation Details

All the training and experiments are conducted on an NVIDIA Tesla K80 GPU. Our method is not trained in an end-to-end manner. Our object detection method is trained using Darknet framework[87] due to its compatibility with the YOLO architecture. To train the other components of our system and fine-tune the compared baselines, we use Tensorflow as a training platform. We perform all experiments using OpenCV, which provides a generic inference module for various Deep Learning models. The source code of our entire pipeline, including the scripts for replicating the results, is available on GitHub, and the link to the repository is provided on the companion page.

2.4.3 Comparison with Object Detectors

We compare our proposed store-front detection approach with two baseline algorithms(Faster R-CNN[49] and Single Shot Detector[50]) in terms of both accuracy and runtime efficiency. Both baseline methods perform superiorly in many general object detection challenges[88]. Therefore, these methods are suitable for evaluating our object detection approach. As evaluation metrics, we adopt precision, recall, F-score, mean average precision over 0.5 IoU threshold, and average inference time per image. We first finetune both baseline methods using the training set of *Store-Obj* dataset. All training images used for tuning baselines methods are resized to (416×416). Then, we perform experiments using *Store-Obj* test set. As shown in Table 2.3, our detection approach outperforms both baseline

[10] https://www.yelp.com/
[11] https://foursquare.com/
[12] https://cloud.google.com/maps-platform/places/

methods in precision(∼+2% & ∼+9%) and mAP@0.5(∼+0.5% & ∼+5%). Regarding recall, Faster R-CNN performs better, but it has a higher inference time (∼+300%) compared to our method. In addition, we present a second variation of our model with a larger input size (608×608). We observe that increasing input size improves precision and recall by ∼+7.8% and ∼+15%, respectively. However, the average inference time also increases by 75%.

Table 2.2: Results of our store-front detector method and the state-of-the-art methods concerning recall (%), precision (%), F1 score, mean average precision over 0.5 IoU threshold (%), and inference time per image (ms). (∗): The second variation of our model is only presented to show the impact of input size on the performance.

Method	Recall	Precision	F1 score	mAP@0.50	Infer. time
SSD	68.29	72.35	70.26	74.3	220
Faster R-CNN	**77.03**	79.33	**78.16**	78.9	325
Ours (yolo-storefront-416)	74	**81**	77	**79.37**	**100**
Ours (yolo-storefront-608)*	89.05	88.22	88	91.35	175

2.4.4 Comparison with Scene Classifiers

As explained in Section 2.3.2, we formulate the identification of store-front type as a fine-grained scene classification problem. First of all, we compare the performance of our approach with two visual-only scene recognition baselines on the *Store-Scene* dataset described in Table 2.1. This comparison mainly aims to show the influence of leveraging textual information from imagery on the classification of business-related storefronts. As shown in Table 2.3, our scene classification approach outperforms both visual-only baselines. Results suggest that it is possible to achieve high performance with limited training data by considering textual information visible on the outdoor appearance of storefronts.

Table 2.3: Results of our proposed storefront classification method in comparison to the state-of-the-art methods concerning top@1 accuracy (%), top@5 accuracy (%), and inference time per image (ms).

Dataset	Method	Top@1 acc.	Top@5 acc.	Infer. time
Store-Scene	GoogLeNet-places365	21.45	55.42	**95**
	ResNet152-places365	28.15	59.45	125
	Karaoglu et al.	38.17	69.56	110
	Ours	**45.01**	**89.44**	205

We also compare the performance of our classification approach with Karaoglu et al.[53], the state-of-the-art multi-modal method that addresses the problem of storefront-type classification by leveraging textual information from images. We fine-tune the CNN models used in this method for visual feature detection, like our morphological classifier. As shown in Table 2.3, our proposed classification approach outperforms the state-of-the-art top@1 from 38.17% to 45.01% (∼+6.8%) on the *Store-Scene* dataset. There is also a remarkable improvement in Top@5 accuracy from 69.5% to 89.4% (∼+20%).

2.4.5 Comparison with Human Annotators

When automatically creating or updating business listings, it is crucial to have a system that performs at human-level accuracy. In other words, the human operator must manually correct any wrong prediction made by such a system, which would cause much additional effort. Therefore, we must ensure that the performance of our system is comparable with the human operator in terms of precision and recall.

We ran our model on the set of collected panorama images and separately conducted a crowd-sourcing experiment through Amazon Mechanical Turk [13]. In the crowdsourcing task, workers are asked to draw a bounding box around every visible storefront on the image and then choose its corresponding category from a given list of 24 business types (Described in Section 2.4.1). We also added an OTHER category which stands for *unknown* or *not-in-the-list* situations. At least three unique human annotators annotate each image. By tracking *workerId*, the back-end system running on our server ensures that each worker submits at most three tasks to avoid biases due to over-repeated participation.

We published 645 HITs, and 318 unique workers executed our tasks. We manually check all the HITs and exclude invalid assessments. The aggregated geo-location of annotations, made by crowd workers, is estimated based on the method explained in section 2.3.3. We run our model on the *Store-Location* dataset, the same street-level images (resolution: 2000×640) used in the crowd-sourcing experiment. Then, we removed duplicate geo-locations from the list of detections, resulting in 97 unique businesses. Each storefront bounding box $\hat{B}$ predicted by our model is considered as True Positive if there are at least two bounding boxes B, obtained from the crowd-sourcing task, where IoU (Intersection over Union) between $\hat{B}$ and B is more significant than 0.5. When $\hat{B}$ is confirmed as True Positive, we compare the result of our storefront classifier with the human categorization. Given L is a set of labels, which are assigned to a storefront bounding box by at least three human annotators. The predicted labels $\hat{L}$ are sorted based on the classifier's confidence. Then, we define the top k prediction set $\hat{L}_k$ as the first k elements in $\hat{L}$, where $k \in \{1, 5\}$. The prediction of business category $\hat{L}_k$ is confirmed as True Positive if one label of $\hat{L}_k$ is agreed by at least two human annotators, represented by L. If the best confidence score of top-k predictions is below 0.4, the label is considered unknown, which is represented by *OTHER* on the list of business categories. As depicted in Table 2.4, our automatic method achieved 83.2% precision on detecting storefronts: it got 39 false positives out of the 232 detections. Then, we manually removed duplicate geo-locations from the list of detections, resulting in 60 in unique businesses. It means a 61.9% recall at 83.2% precision: 60 out of 97 businesses visible on Street View imagery were correctly detected by our automatic system.

Table 2.4: Results of our model in comparison to human assessment results.

	Detection		Classification		Geo-location Estimation	
	Precision	Recall	Top@1 acc.	Top@5 acc.	Precision	Recall
Ours (end-to-end)	83.2%	61.9%	69.1%	92.5%	83.18	61.85

[13] https://mturk.com/

2.4.6 Qualitative Analysis

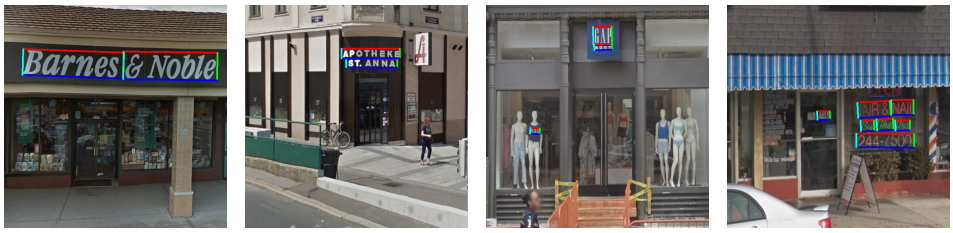
In this section, we discuss examples of real-world scenarios where the proposed approach provides correct and incorrect predictions on *Store-Location*. Figure 2.4 illustrates that our model can detect(~89%) correctly, classify (~78%), and geo-locate (~89%) business-related storefronts, which are visible in the street-level images. In this example, the storefront (i) is predicted correctly, even when there is no word having a direct relation to their types (e.g., $beautysalon$); the proposed semantic matching approach can infer that texts such as $Hair$ or $Nail$, are semantically close to *beauty salon* in the word vector space, thus enabling correct classification.

Figure 2.4: Qualitative results of our integrated approach on detecting, classifying, and mapping storefronts using street-level imagery

Nonetheless, one of the drawbacks of our system is the difficulty in identifying the correct extent of those storefronts, which are divided into different parts. As Figure 2.4 clearly shows, the storefronts f and g are detected separately. However, those bounding boxes belong to the same storefront. As discussed in 2.3.2, due to the high degree of visual variability, it can be very challenging (if not impossible) to correctly classify the business type of storefronts only based on the visual features. As depicted in Figure 2.4, the business types of detected storefronts (c) and (d) are predicted as *Bank*; however, the correct labels are *Optician* and *Bar*, respectively. The reason for the failure of storefront (d) is likely to be that there is a sign of an ATM on the facade of the storefront, which is the only textual feature our model can extract from the image. As the word 'ATM' usually appears in the same context as 'Bank' in the text corpus, our word-vector-based semantic-matching method made a wrong prediction with very high confidence. Similarly, the storefront (c) is predicted incorrectly as Bank since the extracted words ('Tax Service' and 'Income') are semantically related. These failures show an obvious limitation of our method, i.e., that

the textual feature might sometimes be misleading, impacting the overall performance of the proposed approach. Without textual information, the system relies on visual features.

To further analyze the performance of *ST-Sem*, our proposed multi-modal classification approach, we provide additional examples of individual storefront instances where *ST-Sem* provides non-obvious correct (Figure 2.5) and incorrect (Figure 2.6) predictions. As shown in Figure 2.5 (a) and Figure 2.5 (c), *ST-Sem* can recognize the type of storefront, even when there is no word having direct relation to their types (e.g., *book* or *clothes*); the proposed semantic matching module can infer that texts such as *Barnes & Noble* and *GAP* are, respectively, semantically close to *bookstore* and *clothing* in the vector space, thus enabling correct classification. As depicted in Figure 2.5 (b), the proposed method can also measure the semantic similarity between different languages. More specifically, *Apotheke* is recognized as a German scene-text on the image, and then, it is transformed into a multilingual word vector which is semantically similar to *Pharmacy*.



(a) Bookstore (b) Pharmacy (c) Clothing (d) Beauty Salon

Figure 2.5: Examples of correct classifications, with the Storefront dataset's ground-truth label (GT) and probability score. (a) GT: *Bookstore*, Predicted: *Bookstore - 0.71*; (b) GT: *Pharmacy*, Predicted: *Pharmacy - 0.83*; (c) GT: *Clothing*, Predicted: *Clothing - 0.75*, (d) GT: *Beauty Salon*, Predicted: *Beauty Salon - 0.67*

Figure 2.6 shows examples of an incorrect prediction. As shown in Figure 2.6 (a), the scene-text detector failed to detect textual information on the corresponding image due to the uncommon font used in the signs. Therefore, the classification is only based on the visual features. This failure shows an obvious limitation of our method, i.e., the overall performance is highly dependent on the performance of the scene-text detection module. Without textual information, the system relies on visual features. Figure 2.6 (b) shows that the scene-text recognition module recognized two informative words (*pharmacy* and *beauty*) on the image, but the storefront type is not correctly classified. The reason for failure is likely to be that the semantic similarity scores of *pharmacy* and *Beauty Salon* are almost equal for this particular storefront. Therefore, similarly to the previous failure case, classification was only based on the morphological features of the storefront, which can indeed be erroneous.

(a) Toy Shop (b) Pharmacy

Figure 2.6: Examples of incorrect classification results on the Storefront dataset. (a) GT: *Toy Shop*, Predicted: *Bookstore - 0.58* (b) GT: *Pharmacy*, Predicted: *Beauty Salon - 0.42*

2.5 Conclusion

We introduced a novel approach to detect, classify, and geo-locate retail storefronts using street-level imagery. Our approach can detect the physical extent of storefront boundaries even when well-annotated training data is limited. The multi-modal storefront classifier predicts business categories near human-level accuracy by measuring the semantic similarity between detected textual information and the candidate business categories, in addition to morphological characteristics of the storefront's view from the outside. The geo-location aggregation method improves the overall performance of the system by removing false positive predictions. In the future, we plan to incorporate additional semantically rich information, such as contextual information and semantic relationships between objects, which are visible in street-level imagery. Furthermore, to show the scalability of our approach, we plan to extend the scope of our experiments to other cities in non-English-speaking countries.

3

Enhancing Image Recognition with Human-Cognitive Integration

In the rapidly evolving landscape of machine learning and image recognition, the integration of human cognitive capabilities has emerged as a pivotal strategy for overcoming the limitations of automated systems. This chapter delves into the intricacies of harnessing human intelligence within image recognition, mainly focusing on the concept of 'unknown-unknowns' - errors in machine learning models due to the model's overconfidence in its incorrect predictions.

Recent advancements have highlighted the significance of human-in-the-loop approaches in identifying and addressing these unknown-unknowns. By incorporating human insight, we can bridge the gap between what a machine learning model knows and should know. This chapter introduces Scalpel-HS, a novel framework that epitomizes this human-machine collaboration. Scalpel-HS leverages human intelligence for semantic analysis at scale, effectively characterizing unknown-unknowns and enhancing the reliability of image recognition models.

This chapter presents the design and implementation of Scalpel-HS, which includes two essential human computation tasks: outlining the 'Should-Know' elements that indicate what a model should learn, and the 'Really-Knows' elements that reflect the model's current knowledge state. The combined effect of these tasks, bolstered by sophisticated data partitioning and sampling techniques, facilitates a thorough and scalable method for identifying unknown-unknowns.

Through extensive experimentation and analysis, this chapter will demonstrate how Scalpel-HS significantly outperforms existing methods in detecting and characterizing unknown-unknowns. By integrating human cognitive capabilities with machine learning, Scalpel-HS provides a deeper understanding of model failures and paves the way for more reliable and efficient image recognition systems.

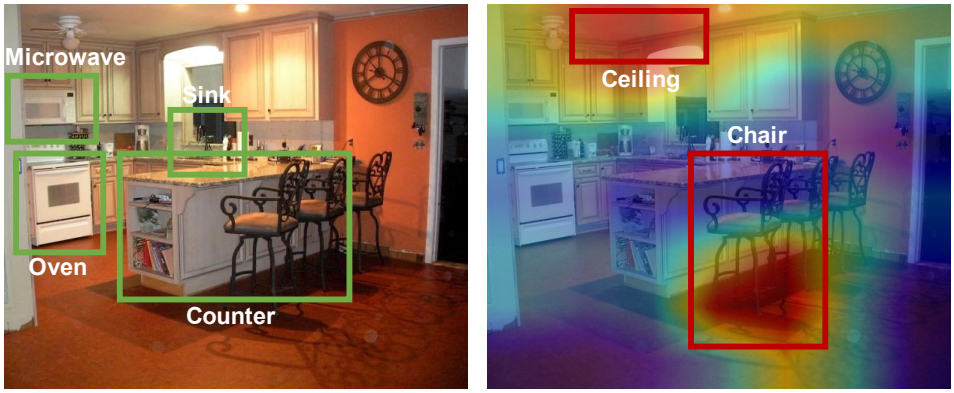
In summary, this chapter presents a novel approach to enhancing image recognition accuracy by embracing the unique strengths of human intelligence. We hope the approach sets the stage for future research and development in human-in-the-loop methodologies, marking a significant step forward in pursuing harmonious human-machine collaboration in artificial intelligence. The content of this chapter is based on the following paper:

- *Sharifi Noorian, Shahin, et al. "What Should You Know? A Human-In-the-Loop Approach to unknown-unknowns Characterization in Image Recognition." Proceedings of the ACM Web Conference 2022. 2022. [40]*

3.1 Introduction

Machine-learned image recognition models are rapidly deployed in many high-stakes contexts [89]. While largely accelerating and aiding the decision-making process, such models suffer from a severe issue of reliability—they can just as easily fail and generate errors that can eventually lead to drastic consequences [90]. Understanding and detecting such errors has become a key demand for both model developers to debug and improve the model [91] and for the users to decide when to trust the model output [92–94].Among image recognition errors, a specific type known as *unknown-unknowns* is of particular interest [95, 96]. Unknown-unknowns refer to the images for which a model is highly confident about its predictions but is wrong. Identifying such errors is challenging due to the overconfidence of the model.

Recent efforts resort to *human-in-the-loop* approaches that ask humans to gather data instances that are potentially difficult for a model to handle [95–97]. An important finding reveals that unknown-unknowns often come with internal consistency, making them particularly suitable to be described by natural language building on top of conceptual knowledge [95, 96, 98]. We bring the notion of *characterizing* unknown-unknowns to allow us to understand better when the model fails. This lies in contrast to previous work that has focused largely on *identifying* unknown-unknowns.



(a) *Should-Know* (b) *Really-Knows*

Figure 3.1: An unknown-unknown example: *Kitchen* image classified as *Conference Room*. The model misses relevant concepts *microwave*, *oven*, *counter*, and *sink* specified in (a) what the model should know, while picking up irrelevant concepts *chair* and *ceiling* shown in (b) what the model really knows (based on the saliency map [99]).

For effective characterization of unknown-unknowns, two types of knowledge are needed: knowledge of what a model *has learned*, that we henceforth refer to as *Really-Knows*, and what a model *should have learned*, referred to as *Should-Know*. Recent work on human-in-the-loop machine learning interpretability [100] has shown the important role of humans as computational agents to describe *Really-Knows* by annotating salient image

areas in image recognition with semantic concepts. In this chapter, we advocate another view of the role of humans as contributors who can shed light on *Should-Know*. We envision that eliciting *Should-Know* from the perspective of human understanding of a given task, can lead to a complete and usable characterization of unknown-unknowns. Consider the example of indoor scene recognition in Figure 3.1, where the model incorrectly classifies a *Kitchen* image as a *Conference Room*: knowing that the model fails by focusing on the *chair* and *ceiling* only tells half the story; knowing that the model should have focused on the *microwave*, *oven*, *counter*, for instance, presents a deeper understanding that would allow further identification of similar errors which the model produces by missing such concepts.

With this in mind, we introduce ***Scalpel-HS***, a human-in-the-loop semantic analysis framework for unknown-unknowns characterization in image recognition. Drawing inspiration from cognitive psychology literature [101–104], ***Scalpel-HS*** is designed with two human computation tasks— for *Should-Know* specification and *Really-Knows* description— that both engage human contributors to operate at the conceptual level. In the *Should-Know* task, human contributors identify a set of objects (with attributes) and relations that are relevant to a given image. In the *Really-Knows* task, human contributors annotate areas of an image shown to be relevant for model prediction with semantic concepts (i.e., visual objects, attributes, and relations). Leveraging the outcome from both *Should-Know* and *Really-Knows* tasks, model unknowns can be characterized by comparing those concepts a model should have learned with what the model actually learned.

Scalpel-HS builds upon a computational pipeline to provide input to the human computation tasks with minimized cognitive load imposed on human contributors. For the *Should-Know* task, we leverage state-of-the-art information extraction techniques to pre-identify objects and relations in the images, allowing human contributors to primarily focus on adjudicating the relevance of concepts to a given scene. This cognitively simplifies the task at hand, in comparison to explicitly synthesizing relevant concepts [105], and results in a more structured vocabulary. For the *Really-Knows* task, we leverage machine learning interpretability methods to highlight important pixels of an image for model prediction [99, 106]. To minimize human effort, ***Scalpel-HS*** employs a semantic data partitioning and sampling method that identifies representative images for human tasks. To do so, ***Scalpel-HS*** starts off by first learning semantically rich image representations.

We demonstrate the effectiveness, informativeness, and cost-efficiency of ***Scalpel-HS*** on several state-of-the-art machine learning models for scene recognition [3, 21]. This task is considered to be complex in image recognition for machines, as well as for humans, as it requires the understanding of context [107, 108]. We show that ***Scalpel-HS*** provides informative, easy-to-understand characterizations of unknown-unknowns that significantly boost state of the art in unknown-unknown detection by *31%*, and can detect *2x to 3x* the sizes of unknown-unknowns compared to the number of annotated images.

In summary, we make the following key contributions:

- We introduce a human-in-the-loop framework that orchestrates both automatic and human computation components for cost-efficient characterization and identification of unknown-unknowns;

- We present the design of human computation tasks for both model should-know and actually-knows descriptions at the conceptual level, with a set of design choices made to account for the cognitive load and fault-tolerance of human work;

- We introduce computational methods for learning semantically rich image representations and for image sampling by partitioning the semantic data space for scaling out human contributions.

3.2 Related Work

Unknown-unknowns. Errors of machine learning fall into two broad categories, namely *known unknowns* and *unknown-unknowns*, denoting low- and high-confidence errors, respectively. Known unknowns have been extensively studied in the literature of active learning [109]. A set of data sampling strategies have been introduced e.g., query-by-committee [110], uncertainty sampling [111], expected error reduction [112]. More recent development concerns with the dynamic selection of optimal strategies in the training process [113–115]. All those strategies rely on information provided by the model and thus are not suitable for the identification of unknown-unknowns that the model is unaware of.

Unknown-unknowns are drawing increasing attention recently due to the criticality for safety and user trust in high-stakes applications. A seminal work by Attenberg *et al.* [95] proposes to ask humans to gather publicly accessible instances that are potentially difficult for a model to handle. This approach has been recently extended by enabling humans access to more information sources to improve the efficiency of unknown-unknowns detection. For example, Lakkaraju *et al.* [97] assume human accessibility to the data and introduce a bandit algorithm to exploit data similarity for faster detection. Vandenhof *et al.* [116] on the other hand assume accessibility to model parameters and propose to engage human contributors to generate instances that contradict model reasoning. Most work so far has focused only on the detection task, with the exception of Liu *et al.* [96] that propose to identify the "pattern" of unknown-unknowns *for detection*, bringing implicitly the task of unknown-unknowns characterization to the horizon. Yet their work does not study characterization on its own—e.g., effectiveness or informativeness. To the best of our knowledge, we are the first to present a focused study on unknown-unknowns characterization, considering the roles of humans in both requirement specification and machine learning behavior interpretation, supported by automatic computational methods for scaling out human contributions.

Automatic and Human Methods. Unknown-unknowns arise from biases in the training data. As such, methods developed for outlier detection are relevant for unknown-unknowns detection. Typical methods can be characterized as either parametric [117, 118] or non-parametric [119–121] i.e., with or without assumptions on the underlying data distributions. In unknown-unknowns detection, outlier detection methods are limited in that 1) they assume the accessibility to the reference data (i.e., training data) which is not necessarily available as in our setting, and 2) they do not take into account what a model has learned thus is limited to identifying model unknowns as we have shown in our experiment.

Another closely related line of work is human-in-the-loop (HItL) machine learning, where human intelligence has been leveraged to address inherent limitations of ML such as reliability and interpretability. Early work in HItL methods mainly focuses on leveraging human intelligence for data labeling [4, 122]. More recent work has investigated the advantage of human computation in debugging ML system components [123] and in identifying biases and noisy labels in the data [124, 125].

HItL approaches are particularly effective in scenarios where model interpretability and reliability are paramount. However, as discussed in works such as De Bruijn *et al.* [126], implementing HItL frameworks requires careful consideration of practical challenges, including variability in human input, resource demands, and the potential amplification of biases inherent in data or human feedback. For example, scalability remains a critical issue, particularly in high-volume contexts where manual oversight becomes impractical. Additionally, achieving consistency in human annotations can be challenging due to cognitive differences or fatigue among contributors. Addressing these aspects requires hybrid frameworks that balance human and machine collaboration, ensuring efficient use of both resources.

The most closely related work, as we discussed, is Lakkaraju *et al.* [97] and Liu *et al.* [96] that use HItL methods for unknown-unknowns detection. Recent work that has directly inspired ours is Balayn *et al.* [100] that propose to use human computation to interpret the behavior of image classifiers by attaching semantic concepts to the saliency maps of classification. We employ this method for unknown-unknowns characterization in image recognition, and take a step further to show that by including human specified requirements of what a model should know, we can significantly improve unknown-unknowns characterization.

3.3 The *Scalpel-HS* Framework

Figure 5.1 presents an overview of ***Scalpel-HS***. Given an image set and a trained image recognition model, it first 1a extracts the scene graphs of the images and 1b the saliency maps for the model classification of the given images. It 2a learns the representation of the images combining both the visual and semantic features and based on that, 2b partitions the image set and sample representative images for the human tasks. The scene graphs and the saliency maps of the sampled images are then respectively fed to the human tasks published in a crowdsourcing platform, 3a the *Should-Know* task, and 3b the *Really-Knows* task, to generate descriptions of what a model should know, and what it actually knows. Output of the two tasks are then aggregated to obtain a characterization of the unknown-unknowns, together with a set of corresponding unknown-unknown images through the 4 aggregation and detection component. In the following, we describe the components in more detail.

Scene Graph Extraction (1a). Understanding a natural scene image usually requires reasoning about the relationship between objects in the image. For example, in the recognition of rooms, a sink next to an oven indicates a kitchen while a sink next to a mirror more likely indicates a bathroom. To help humans specify the required knowledge in scene recognition, i.e., *Should-Know*, we extract scene graphs. A scene graph is a structured representation

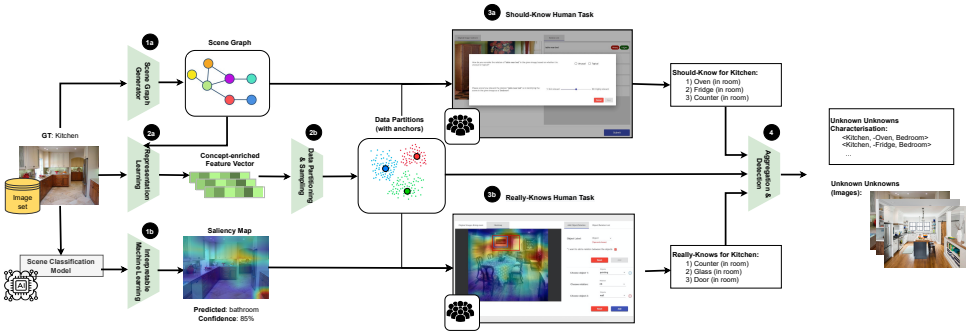


Figure 3.2: The ***Scalpel-HS*** framework. It takes as input an image set and a trained image recognition model; as output, it produces a characterization of unknown-unknowns and identifies the corresponding unknown-unknown images. To do so, it extracts the **(1a)** scene graph and **(1b)** saliency map of model classification of a subset of images—sampled by **(2a)** representation learning and **(2b)** data partitioning—, feeds them to the **(3a)** *Should-Know* and **(3b)** *Really-Knows* human computation tasks published on a crowdsourcing platform, and **(4)** aggregates the output for unknown-unknowns characterization and detection of more corresponding unknown-unknown images.

of objects and the relationships between them present in an image. It consists of a set of relationships, each represented as $[o_i, r_{ij}, o_j]$, where o_i and o_j refer to two objects (image patches usually captured by bounding boxes) in the image, and r_{ij} represents the relation between two objects. Given a scene image, we generate the visual scene graph using state-of-the-art methods Neural Motifs [127].

Saliency Map Extraction (1b). Understanding machine behavior in scene recognition is the machine learning interpretability problem. The most extensively studied interpretability approach for image classification is saliency, a local interpretability post-hoc method that highlights the most important pixels of an image for model decisions in what is called a saliency map [106]. We choose this method to help humans describe what a model *Really-Knows*.

We opted for SmoothGrad [99], which is sensitive to the parameters of a model (thus catering for more accurate capturing of a model behavior) while minimizing noisy results (i.e., highlighting irrelevant pixels). Our framework is though agnostic to the employed local interpretability method. To minimize human effort, we sample a subset of representative images for human annotation. Data sampling is performed via a data partitioning method based on a new type of image representation.

Representation Learning (2a). Due to the complexity of the scene recognition task, representative images should be diverse in terms of both the semantic information contained and the visual appearance. Existing methods generally rely on pre-trained models for visual feature extraction only, which is suboptimal in our context. We propose to fuse in also the semantic information and introduce a self-supervised learning approach for learning semantically rich image representations. We describe the details in Section 3.4.1.

Data Partitioning and Sampling (2b). Prior work has shown that unknown-unknowns are caused by systematic biases in training data, and reside in specific partitions (i.e., blind spots) of the feature space [95, 96]. Following this, we propose a data partitioning method for sampling, *Semantic Space Partitioning (SSP)*, that identifies the optimal subset of representative images in the semantic space. Our method partitions the semantic space and selects candidate images in such a way that (the weighted sum of) cosine distances from the candidate data points to others in the same region are minimized. As the result, semantically similar images will be grouped in partitions, centered around the representative images. Those representative images are then sampled for human annotations. We describe the details of our SSP method in Section 3.4.2.

The *Should-Know* Task (3a). For a given set of valid objects and relations pertaining to a scene, it is crucial to understand the salience of each object and relation in identifying the scene in the given image. For example, from a human perspective, a *bed* when compared to a *carpet* can be deemed to be relatively more salient in identifying the scene as a *bedroom*. In this task, human workers identify the salient objects, their attributes, and the relations between objects for identifying a given scene in an image. We describe details of the task design in Section 3.5.1.

The *Really-Knows* Task (3b). The goal of the task is to find out which objects in the scene influence the prediction of the machine learning model, and whether this is congruent with the human mental model. Human workers identify objects and relations found by the machine and rate their relevance in identifying the scene. Details of the task are described in Section 3.5.2.

Aggregation & Detection (4). The results of the two human tasks are aggregated to characterize unknown-unknowns. Denoting the true class and wrongly predicted class as y and y', respectively, the characterization is represented in the form of the triple $\langle y, (+)c, y' \rangle$ for False Positive (in terms of y') and $\langle y, (-)c, y' \rangle$ for False Negative (in terms of y). For example, $\langle$*Conference Room*, $(+)$*sofa*, *Living Room*$\rangle$ indicates that the model wrongly classifies a conference room image to be a living room because of the focus on the spurious concept sofa; $\langle$*Kitchen*, $(-)$*oven*, *Conference Room*$\rangle$ indicates that the model wrongly classifies a kitchen image to be a conference room by missing the concept of the oven in the kitchen.

Apart from the characterization, this component detects more unknown-unknowns of the same characteristics utilizing the data partitions: images in the same partition as the human-annotated (representative) one are likely to be unknown-unknowns sharing the same missing or spurious concepts. The component, therefore, identifies more images as unknown-unknowns those on which the model confidence is greater than a threshold.

3.4 Image Representation and Sampling

This section describes our methods for semantically rich image representation learning and semantic space partitioning.

3.4.1 Representation Learning

Our representation learning model comprises components for visual feature extraction, semantic feature extraction, multi-modality fusion, and image representation generation. Together those components make a representation learning model that can be trained in an end-to-end fashion. We describe the details in the following.

Visual Feature Extraction. We use the pre-trained model Faster-RCNN [128] to generate feature vectors for nodes and relationships in the generated scene graph. For each object node o_i in the scene graph, a visual feature vector $\mathbf{V}_{o_i}$ is extracted from its corresponding image region. For each relationship node r_{ij}, its visual feature vector $\mathbf{V}_{r_{ij}}$ is extracted from the union region of o_i and o_j on the image.

Semantic Feature Extraction. Each node, either object node or relationship node, has a text description generated by the scene graph generator. From such text description, we obtain the initial semantic features using the pre-trained GloVe embeddings [82]. Those embeddings are trainable parameters in our representation method. We denote the semantic feature of the node and relationship as $\mathbf{E}_{o_i}$ and $\mathbf{E}_{r_{ij}}$, respectively.

Multi-modality Fusion. We fuse the visual and semantic features into a joint multi-modal representation. Inspired by [129], the visual feature vector and label feature vector are first concatenated, then fused as follows:

$$\mathbf{Z}_{o_i} = \tanh(\mathbf{W}_1^T \mathbf{V}_{o_i} + \mathbf{W}_2^T \mathbf{E}_{o_i}) \tag{3.1}$$

$$\mathbf{Z}_{r_{ij}} = \tanh(\mathbf{W}_1^T \mathbf{V}_{r_{ij}} + \mathbf{W}_2^T \mathbf{E}_{r_{ij}}) \tag{3.2}$$

, where $\mathbf{Z}_{o_i}$ and $\mathbf{Z}_{r_{ij}}$ are joint feature vectors for object and relation nodes, respectively. $\mathbf{W}_1$ and $\mathbf{W}_2$ are the shared parameters.

Image Representation Generation. To obtain a single vector representation of an image, we combine the visual-semantic features of all the objects and relationships. Considering the fact that those objects and relations together make a graph, we employ a multi-layer graph convolutional networks (GCN) [130] to capture the graph structure while combining the object and relationship features. Through multiple layers of linear and non-linear transformation layers, GCN generates new node features that contain structural information of the graph. To obtain a global representation of the image, we aggregate the node representations learned by GCN in different layers—denoted as $\mathbf{U}_{o_i}$ and $\mathbf{U}_{r_{ij}}$ for object and relationship nodes respectively—into a graph representation using a graph pooling operation, known as readout [131, 132]. The entire representation learning model contains parameters of embedding, the multi-modality fusion layer, and those of the GCN. We describe the training details in Appendix 3.8.1.

3.4.2 Semantic Space Partitioning

Formally, we denote the distance between the feature vectors f_i, f_j of two images i, j as following:

$$dist(f_i, f_j) = 1 - \frac{f_i{}^T f_j}{\|f_i\| \|f_j\|}. \tag{3.3}$$

Given a budget $\mathcal{B}$, i.e., the number of representative images to be sampled for human tasks, our partitioning method finds the representation images by minimizing the following objective function:

$$\begin{aligned} & min \sum_{i \in \mathcal{D}} \sum_{j \in \mathcal{D}} dist(f_i, f_j) X_{ij} \\ & s.t. \sum_{j \in \mathcal{D}} X_{ij} = 1 \\ & X_{ij} \leqslant Y_j \\ & \sum_{j \in \mathcal{D}} Y_j = \mathcal{B} \end{aligned} \tag{3.4}$$

, where X_{ij} indicates the decision of whether image i is assigned to partition j; Y_j indicates if image j is selected as a representative sample (note that the index j is overloaded to represent both the partition and the representative image of the partition). Due to the large number of possible solutions that are associated with the problem of finding an optimal set of representative samples, it is very challenging to provide a deterministic solution. We employ a meta-heuristic approach based on genetic algorithms (GA). Details of our algorithm are described in Appendix 3.8.2.

3.5 The Human Computation Tasks

We now describe the human computation tasks for specifying what an image recognition model *Should-Know* and what it *Really-Knows*.

3.5.1 The *Should-Know* Task

In this task, human workers are presented with a sampled image and the corresponding scene graph to identify concepts (salient objects, their attributes, and relations) in scene recognition. The procedure is shown in Figure 3.3 (zoomed figures in Appendix 3.8.3). Workers are first asked to (a) validate the automatically generated objects and relations within the scene graph. Erroneous objects and relations can thereby be identified and filtered (Neural Motifs performance in object and relation classification are 33%, 59% respec-

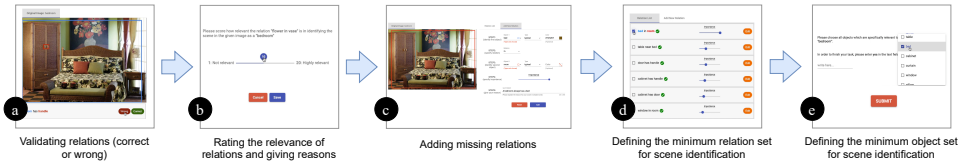


Figure 3.3: The procedure of the *Should-Know* task for specifying what a model should know in scene recognition. (Zoomed in Appendix 3.8.3)

Figure 3.4: The procedure of the *Really-Knows* task for describing what a model really knows in scene recognition. (Zoomed in Appendix 3.8.3)

tively [127]). They are then tasked to ⓑ rate the relevance of concepts in scene recognition using a slider ranging from 1 to 20 (not relevant at all to highly relevant).

To further scope down to the highly relevant concepts, workers are asked to identify the minimum set of concepts that can sufficiently identify the scene. It implicitly requires humans to first ⓒ add missing concepts, and then ⓓ,ⓔ determine the indispensable concepts for identifying the scene in the given image. To add missing concepts, workers need to specify the concept by entering the name and drawing a bounding box in the image. In the case of the concept being an object attribute or a relationship, the worker needs to further specify the relations. Indispensable concepts are selected using checkboxes from the concept list.

3.5.2 The *Really-Knows* Task

While a scene in the human mind is composed of objects possessing clear boundaries and having intelligible locations in space relative to each other [133], to the machine it is the composition of pixels rather than objects. To understand machine behavior, we map the pixels highlighted in warm colors by the saliency map, to actual concepts that humans can understand. The procedure is shown in Figure 3.4 (zoomed figures in Appendix 3.8.3). Workers are asked to ⓐ draw bounding boxes to annotate objects highlighted by the saliency map, ⓑ name the objects and assign the attributes (e.g., color), ⓒ define relations among the objects, and ⓓ add all the objects and relations highlighted by the saliency map to the list.

By comparing to their own mental models in scene recognition, workers then ⓔ rate the relevance these objects/relations are in identifying the scene, using a slider (ranging from 1 to 20, meaning not relevant at all to highly relevant). They are also encouraged to give reasons. Note that with annotations from this task, a characterization of false positive prediction due to the wrong focus on the spurious concept can already be obtained. We compare in our experiments unknown-unknowns detection using only the characterization obtained from this *Really-Knows* task and that from also the *Should-Know* task.

3.6 Experimental Setup and Results

We evaluate the performance of ***Scalpel-HS*** by investigating the following questions[1]: **Q1**: How effective is it in detecting and characterizing unknown unknowns? **Q2**: how

[1]Source code and data are available at https://sites.google.com/view/www22-scalpel-hs

informative are those characterization provided by our framework? and **Q3**: how cost-efficient is our framework under a limited budget?
For these questions, we further evaluate the contribution of the individual components of ***Scalpel-HS*** and compare them to the state of the art whenever possible.

3.6.1 Experimental Setup

Datasets. We use two image datasets: (1) *PLACES*: it contains 10 million images divided into over 400 unique scene classes with 5000 to 30,000 training images and over 100 test images per class. As not all classes are about scenes, we select a subset of data containing nine indoor scene classes. (Details provided in the Appendix 3.8.4.) The subset contains 60000 training and 1000 test images equally distributed across the nine classes.
(2) *MIT67*: this dataset contains 15620 images in 67 indoor classes. Unlike the PLACES dataset, the number of images varies across classes, however, there are at least 100 images per class. We filter images of the same set of scene classes as PLACES and select a subset consisting of 3224 test images with at least 100 images per class. Note that due to the limited number of images in MIT67, there are only 3216 images left after filtering; we therefore only use the training set of PLACES for model training. Test sets from PLACES and MIT67 allow us to experiment with model unknowns exposed in test data of different distributions.

Unknown Unknowns Creation. We consider the unknown unknowns characterization effective in two senses: it exposes the reasoning of an image recognition model in a high-confidence yet wrong prediction, and it allows for the detection of unknown unknowns images of the same type. Note that while the ground truth labels are given in the test set—hence unknown unknowns images are known by comparing the model output to the ground truth labels—the ground truth of the model reasoning is in-transparent. To cope with this issue, inspired by previous work [100], we bias model reasoning by forcing the model to focus on spurious concepts or to miss relevant concepts through data re-sampling. To do so, we create unknown unknowns of False Positive by removing concepts from training images of all classes except those of the class of interest. By doing so, the model will strongly associate the spurious concept with the class of interest and make wrong predictions for test images of other classes. Similarly, we create unknown unknowns of False Negative by removing concepts from the training images of the class of interest (not other classes). To make sure the concepts are distributed in several classes, we select 15 most frequent concepts (objects and relations) and then those that are distributed across at least three classes. A co-occurrence matrix between concepts and scene classes is provided in the companion page. The induced unknown unknowns are summarized in Table 3.1.

Apart from evaluating the effectiveness of our framework in exposing the incorrect reasoning and in detecting unknown unknowns that are manually induced, we further look into the informativeness of the characterization for "natural" unknown unknowns, i.e., those unknown unknown images without the chosen concepts (or missed by the scene graph extractor). Note that while we cannot make sure that the model makes high-confidence errors on all images with the identified characteristics, we are sure about the errors when they occur given the ground truth labels and about model rationales for images annotated by our framework.

Table 3.1: Summary of induced unknown unknowns.

Type	Index	Class of Interest	Concept
False Positive	FP1	*Kindergarden*	*person*
	FP2	*Bedroom*	*bed*
	FP3	*Conference Room*	*chair*
False Negative	FN1	*Kitchen*	*oven*
	FN2	*Bathroom*	*sink*
	FN3	*Dining Room*	*wine glass*
	FN4	*Living Room*	*couch/sofa*
	FN5	*Conference Room*	*woman at table*
	FN6	*Kindergarden*	*boy wearing shirt*

Scene Recognition Models. We conduct our experiment with two state-of-the-art convolution neural networks, ResNet[75] and DenseNet[134], which have shown superior performance on various classification tasks [135]. We train the two scene classifiers for 50 epochs on the biased training data and confirm that biases are successfully injected into the scene classifier by observing the overfitting of the performance metrics during the training phase.

Table 3.2: Performance (P = Precision, R = Recall, and F = F1-score) comparison with baseline methods on detecting unknown unknowns images. We highlight the best performance for each metric in bold.

Type	Comparison Method	ResNet						DenseNet					
		Places			MIT67			Places			MIT67		
		P	R	F	P	R	F	P	R	F	P	R	F
False Positive	Random	0.383	0.187	0.251	0.311	0.158	0.209	0.39	0.161	0.228	0.336	0.120	0.177
	Least Average Similarity	0.558	0.272	0.366	0.318	0.161	0.214	0.558	0.272	0.366	0.663	0.218	0.329
	Least Maximum Similarity	0.379	0.185	0.249	0.232	0.118	0.156	0.379	0.185	0.249	0.616	0.209	0.312
	Most Uncertain	0.348	0.170	0.228	0.44	0.223	0.296	0.351	0.183	0.240	0.53	0.190	0.279
	UUB	0.629	0.378	0.472	0.755	0.383	0.509	0.617	0.394	0.480	0.702	0.282	0.402
	Scalpel-HS	**0.855**	**0.522**	**0.648**	**0.915**	**0.465**	**0.616**	**0.874**	**0.716**	**0.787**	**0.766**	**0.521**	**0.620**
False Negative	Random	0.21	0.08	0.11	0.28	0.152	0.197	0.293	0.271	0.281	0.33	0.09	0.141
	Least Average Similarity	0.542	0.279	0.368	0.663	0.218	0.329	0.542	0.279	0.368	0.589	0.155	0.246
	Least Maximum Similarity	0.372	0.185	0.247	0.616	0.209	0.312	0.372	0.185	0.247	0.495	0.135	0.212
	Most Uncertain	0.585	0.219	0.319	0.452	0.246	0.319	0.549	0.237	0.331	0.44	0.12	0.188
	UUB	0.551	**0.634**	0.589	0.480	0.376	0.422	0.553	0.649	0.597	0.456	0.271	0.340
	Scalpel-HS	**0.711**	0.525	**0.604**	**0.653**	**0.435**	**0.522**	**0.577**	**0.678**	**0.624**	**0.704**	**0.364**	**0.480**

Baseline Methods. Following previous work [96, 97], we compare the performance of our pipeline against the following methods: 1) Random Sampling: Randomly selects instances from the test data to be queried with humans. 2) Least Average Similarity [136]: Computes the average Euclidean distance for each test instance to all training instances, and chooses the instances with the highest distances. 3) Least Maximum Similarity [136]: Computes the minimum Euclidean distance of test data instances to all training data instances and chooses instances with the highest distances. 4) Most Uncertain [109]: Ranks the instances in the test dataset by increasing order of the prediction confidence as assigned by the scene classification model. 5) UUB [97]: Combines clustering and the bandit algorithm to query an Oracle. Least Average Similarity and Least Maximum Similarity are popular outlier

detection methods; Most Uncertain is similar to the uncertain sampling strategy used in active learning; and UUB is the state of the art unknown unknowns detection method. Apart from those, we further compare with variations of our own framework considering only output from the *Really-Knows* task, and those with other baseline representation learning and data sampling methods.

Evaluation Metrics. For effectiveness evaluation, we use Precision and Recall to measure the performance on unknown unknowns identification. We consider detection performance in both cases when we are sure the unknown unknowns happen due to the injected data biases and in the entire test set.

Crowdsourcing. We crowdsourced 300 images from each dataset, selected by our feature space partitioning method. We present the same set of images for both the *Should-Know* and *Really-Knows* tasks. For each task, we recruited 300 workers on Prolific[2]. For quality control, only workers whose approval rate was greater than 90% were considered as qualified; to avoid learning bias between the two tasks, each worker is allowed to perform only a single task throughout the entire experiment. The authors manually examined the quality of worker annotations on a random sample, which was found satisfying. Each worker was paid 1.15 USD (0.8 GBP) for participating in our study, translating to an average hourly reward of 10.25 USD (7.41 GBP).

3.6.2 Scalpel-HS Performance

Effectiveness. Table 3.2 reports the performance of comparison methods on detecting unknown-unknown images. Among the baselines, Most Uncertain, widely used in detecting known unknowns, yields low performance (similar to Random), providing evidence for the important difference between the problems of detecting known unknowns and unknown-unknowns. Among the two outlier detection methods, Least Average Similarity generally outperforms Least Maximum Similarity, indicating that unknown-unknowns are distant to the general image population in the feature space. UUB, which considers model confidence, gives better performance than all the other baseline methods, showing that unknown-unknowns are related to not only the data but also what models have learned from the data. Most importantly, our proposed framework ***Scalpel-HS*** achieves the best performance across all settings (unknown-unknown types, datasets, and metrics), outperforming UUB by a significant margin of 31% in F1-score, strong evidence demonstrating the effectiveness of our framework in unknown-unknowns detection. The relative detection performance of ***Scalpel-HS*** on the two types of unknown-unknowns (False Positive vs. False Negative) is consistent across datasets given the same model; similarly, it is consistent across models given the same dataset. Those results show the robustness of our framework in unknown-unknowns detection.

Informativeness. To gain a deeper understanding of the informativeness of unknown-unknowns characterization provided by ***Scalpel-HS***, we report in Table 3.3 its performance on uncovering the exact reasoning of the model on unknown-unknown images. Our framework successfully exposes the characteristics of all manually created unknown-unknowns

[2]https://www.prolific.co

Table 3.3: ***Scalpel-HS*** Performance in uncovering ResNet reasoning on unknown-unknown images (FP = False Positive, FN = False Negative, # = the number of corresponding unknown-unknown).

Type	Index	PLACES				MIT67			
		#	P	R	F	#	P	R	F
FP	FP1	488	0.896	0.588	0.710	158	0.769	0.443	0.562
	FP2	618	0.939	0.629	0.753	545	0.636	0.366	0.465
	FP3	9	0.16	0.111	0.13	13	0.0	0.0	0.0
	All	1115	0.914	0.607	0.729	716	0.666	0.384	0.487
FN	FN1	45	0.531	0.377	0.441	330	0.628	0.472	0.539
	FN2	60	0.275	0.5	0.355	90	0.755	0.453	0.566
	FN3	162	0.721	0.623	0.668	166	1	0.282	0.440
	FN4	98	1	1	1	76	0.709	0.549	0.619
	FN5	16	0.086	0.25	0.129	207	0.166	0.052	0.08
	FN6	47	0.095	0.148	0.116	66	0.755	0.486	0.592
	All	428	0.516	0.600	0.555	880	0.672	0.455	0.543

Table 3.4: Examples of "natural" unknown-unknowns identified by ***Scalpel-HS*** for ResNet. Note that False Positive is defined w.r.t. Predicted Class and False Negative is defined w.r.t. True Class.)

Type	True Class	Concept	Predicted Class
False Positive	*Hospital Room*	(+)*sink*, (+)*counter*	*Bathroom*
False Negative	*Kitchen*	(-)*oven*, (-)*counter*	*Bathroom*
	Conference Room	(-)*chair at table*	*Kitchen*

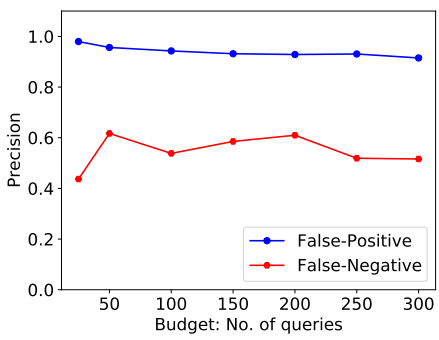
(except FP3 on MIT67, which corresponds to 13 images only), showing the strong characterizing power of our framework for unknown-unknowns. We find a large variability of the performance in detecting unknown-unknown images with different characteristics, showing the specificity of unknown characteristics for detection. As a remark, we note a discrepancy between the overall performance in Table 3.3 and Table 3.2 due to the presence of "natural" unknown-unknowns that are not manually induced.

In Table 3.4, we show a few of those additional unknown-unknowns exposed by our framework, i.e., those that are not manually induced. Those characterizations provide easy-to-understand reasons for model failures in unknown-unknowns and are thus highly useful for identifying similar errors. In our experiment, they allow us to detect 19% extra False Positive natural unknown-unknowns, and 38% extra False Negative natural unknown-unknowns.

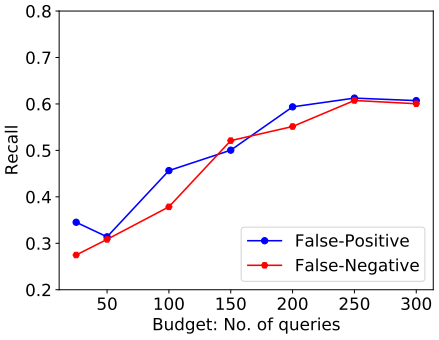
We show examples of manually induced unknown-unknowns detected by our framework in Appendix 3.8.5 and more (including natural ones) on the companion page.

Cost-Efficiency. Figure 3.5 depicts the performance of ***Scalpel-HS*** under different budgets. As expected, precision decreases and recall increases when the budget increases; however, we observe that the decrease in precision is much slower than the increase in recall. With 300 images annotated, accounting for 3% of the overall test images in PLACES and 20% of

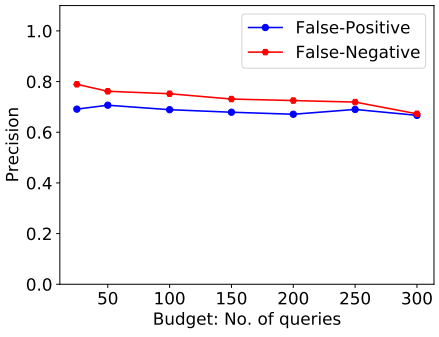
the unknown-unknowns with the identified characteristics, we reach a recall of over 60% of all the unknown-unknowns; on MIT67, the 300 annotated images account for 9% of the overall test images and 19% of the unknown-unknowns with the identified characteristics, we reach a recall of 42%. Those results show that our framework allows detecting **2x** to **3x** unknown-unknowns w.r.t. a given budget, demonstrating that our framework is highly cost-efficient.



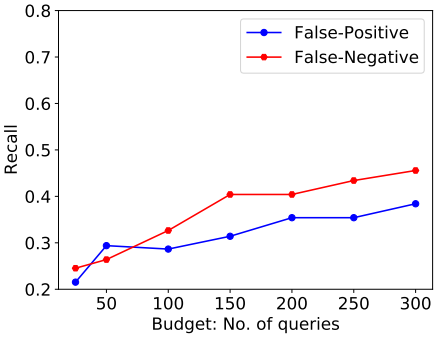
(a) PLACES - Precision



(b) PLACES - Recall



(c) MIT67 - Precision



(d) MIT67 - Recall

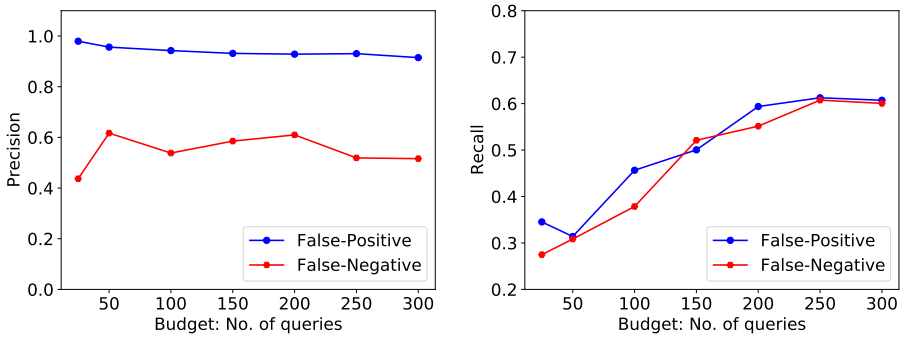
Figure 3.5: Performance of our framework on unknown-unknowns detection for ResNet under different budgets.

3.6.3 Contribution by Automatic Components

We now evaluate the contribution of our representation learning and image sampling methods. Figures 3.6 (a,b) compare the performance of our framework with our proposed semantically-rich representation learning method and the method with visual features only for representation learning (ResNet152 pre-trained on ImageNet and fine-tuned on our dataset, the rest components of the framework kept the same). The result shows that our proposed representation learning method is a better approach in both precision and recall across almost all budgets, signifying the utility of semantic features in the images for

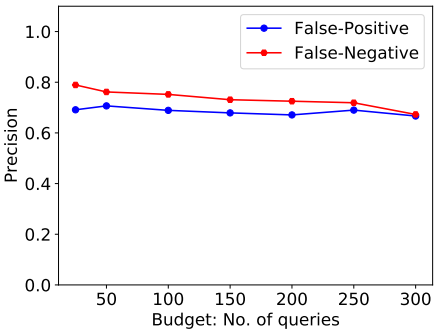
image sampling and ultimately for unknown-unknowns characterization and detection.

Figures 3.6 (c,d) compare the performance of our framework with our proposed semantic space partitioning (SSP) and other baseline data partitioning or sampling methods. These include 1) Random Sampling; 2) DSP [97]: optimizes the overall distances within data partitions (minimization) and across partitions (maximization), and then randomly samples representative images. 3) K-means: generates clusters, and the nearest instance to the center of each cluster (mean) is selected as the anchor of that partition. We observe that SSP achieves much higher precision with comparable recall (higher when the budget increases), showing the superiority of our partitioning methods in sampling representative unknown-unknowns and the effectiveness of joint partitioning and sampling.

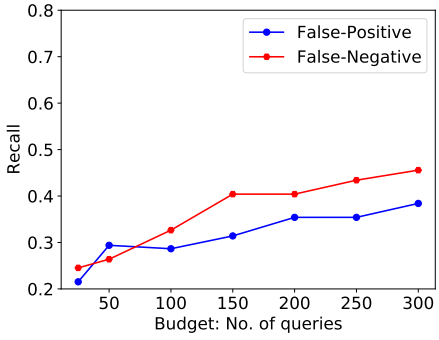


(a) RL - Precision

(b) RL - Recall



(c) SSP - Precision



(d) SSP - Recall

Figure 3.6: Comparing the effects of our proposed representation learning (RL) and data sampling (SSP) with baseline methods on the performance of our framework under different budgets. Results are obtained on ResNet using the PLACES dataset.

3.6.4 Impacts of *Should-Know* and *Really-Knows*

We evaluate the effect of including human annotations, and in particular, human specification of what a model should know, by comparing the following configurations of our framework: 1) no human task, 2) including only the *Really-Knows* task, and 3) including both the *Really-Knows* and *Should-Know* tasks.

Figure 3.7 compares the performance of those configurations under different budgets. We observe that involving human annotations significantly impacts the precision under any budget. In addition, we observe that integrating human annotations significantly impacts recall as the budget increases. Compared to the version of no human tasks, our framework improves by 26% and 10% in precision and recall, respectively (budget=300). Compared to *Really-Knows* only, with *Should-Know*, the performance of ***Scalpel-HS*** improves by 5% in both precision and recall. We also notice that the *Should-Know* task is beneficial for precisely identifying unknown-unknowns when the budget is low.

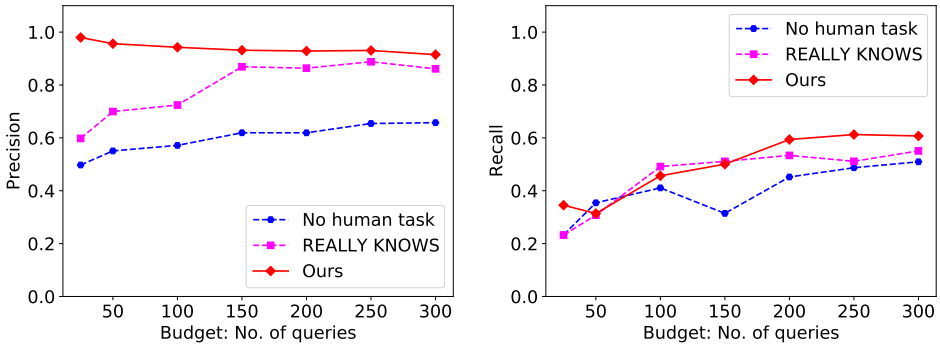


Figure 3.7: Impacts of human tasks on the performance of our framework under different budgets.

3.7 Conclusion

We presented ***Scalpel-HS***, a human-in-the-loop, semantic analysis framework for characterizing and detecting unknown-unknowns of image recognition models. It involves human contributors to specify both what the model should know and describe what it really knows, while minimizing the cognitive load of the tasks leveraging scene graph extraction and machine learning interpretability techniques. It scales out human contributions through both a new semantically-rich representation learning and data sampling method. Our extensive evaluation on multiple models and datasets with different types of unknown-unknowns demonstrates that characterizations provided by ***Scalpel-HS*** are not only informative but also highly effective and cost-efficient for unknown-unknowns detection.

3.8 Appendix

3.8.1 Learning Image Representations

To obtain graph-level representations, we train the graph encoder end-to-end using the approach proposed by Sun et al. [137], by maximizing the mutual information between graph-level and patch-level representations. The patch-level representation refers to the representation of nodes learned by aggregating the features of their neighborhood nodes, at the last GCN layer, while the graph-level representation is a fixed length vector obtained by pooling all patch-level representations. To train our graph encoder, we define the following objective function:

$$max \frac{1}{|G|} \sum_{g \in G} \left[\frac{1}{|g|} \sum_{i=1}^{|g|} D\left(\vec{U}_i, \vec{U}_g\right) \right]$$

where $|G|$ is the number of graphs in train set, $|g|$ is the number of nodes in graph g, and $\vec{U}_i$, $\vec{U}_g$ are representations of node i and graph g, respectively. D denotes a mutual information estimator which is modeled as a discriminator to score the agreement between patch-level and graph-level representations. The agreement score is obtained by simply computing the dot product between two representations.

Hyperparameter Setting. The number of GNN layers are chosen from $\{4, 8, 12\}$. Initial learning rate is chosen from the set $\{0.01, 0.001, 0.0001\}$. We set the batch size to 64. The number of epochs are chosen from $\{30, 60, 90\}$.

3.8.2 Semantic Space Partitioning Algorithm

The input of our GA-based method is a set of data points $\mathcal{D} = \{d_1, d_2, ..., \}$ where each d_i consists of (feature vector, weighted concept) pairs, and the number of representative samples $\mathcal{B}$ to be found. Note that the number of generated partitions is equal to the number of representative samples. We initialize the genetic algorithm by constructing a population of random chromosomes $\mathcal{P} = \{p_1, p_2, ..., \}$, where each chromosome p_i consists of $\mathcal{B}$ candidate samples. Algorithm 1 implements the fitness function, which corresponds to the proposed objective function (See Eq.4.1). The fitness function guides the exploration through the search space towards an optimal solution.

GAs are prone to premature convergence to local optima. Inspired by [138], we address this problem by adjusting the mutation rate (Pm) while the algorithm explores the search space. To avoid generating invalid solutions and improve the GA's performance, we use a greedy approach to the mutation process, which mutates the offspring only if the mutated solution gains a lower fitness value.

Hyperparameter Setting. Population, elitism, mutation rate, and cross-over rate are regarded as hyperparameters and, therefore, can be found via grid search ($N \in \{50, 100, 150, 200\}$; elitism $E \in \{5, 10, 15, 20, 30\}$; mutation rate $P_m \in \{0.0005, 0.001, 0.005, 0.01, 0.05\}$; and cross-over rate $P_c \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$). We set $N = 100; E = 20\%; P_m = 0.005; and P_c = 0.7$. We used a stagnation-based termination criterion; following [139], we terminate the algorithm

Algorithm 1 Fitness algorithm

```
1: procedure Fitness(R, D, F, W)
2:     fitness ← 0
3:     for d ∈ D do
4:         max_similarity ← ∞
5:         for r ∈ R do
6:             f_r ← GetWeightedFeatures(F, W, r)
7:             f_d ← GetFeatures(F, d)
8:             similarity ← Cosine(f_r, f_d)
9:             if similarity ≤ max_similarity then
10:                max_similarity ← similarity
11:            end if
12:        end for
13:        fitness ← fitness + max_similarity
14:    end for
15:    return fitness
16: end procedure
```

after $\sqrt[D]{\mathcal{B}}$ generations, where $\mathcal{B}$ denotes the number of representative samples to be found, and D is the number of data points.

3.8.3 Annotation Workflow in Large Figures

Figures 3.9 and 3.10 show the zoomed version of our *Should-Know* and *Really-Knows* human computation tasks.

3.8.4 Experimental Setup Details

We filter the two datasets by keeping images of the following scene classes: *dining room, bathroom, conference room, kindergarten, hospital room, kitchen, living room, bedroom, dorm room.*

3.8.5 Examples of Unknown-unknowns Identified by *Scalpel-HS*

We show a few examples of unknown-unknown images ***Scalpel-HS*** identified and characterized for both False Positive and False Negative. Figure 3.8 illustrates two characterizations of each type for both the sampled images (i.e., anchors) and two similar ones detected by our framework.

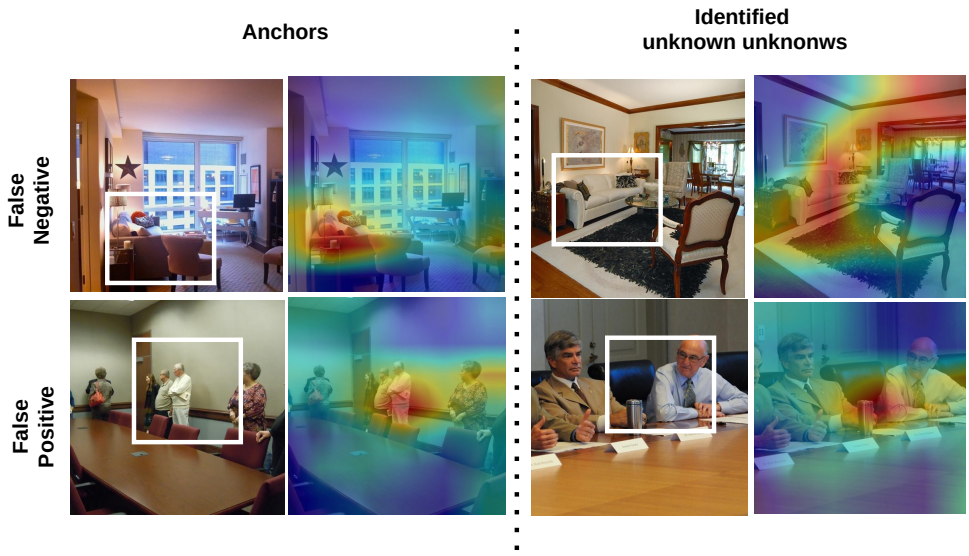
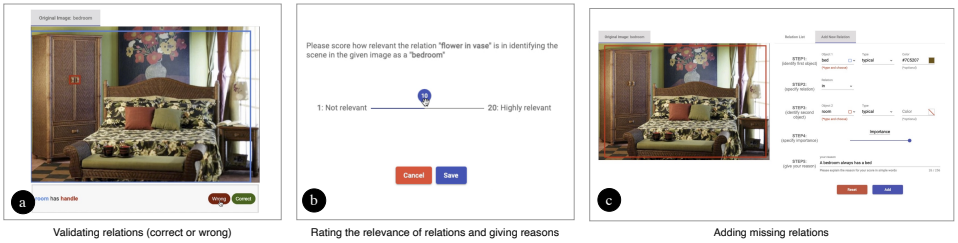
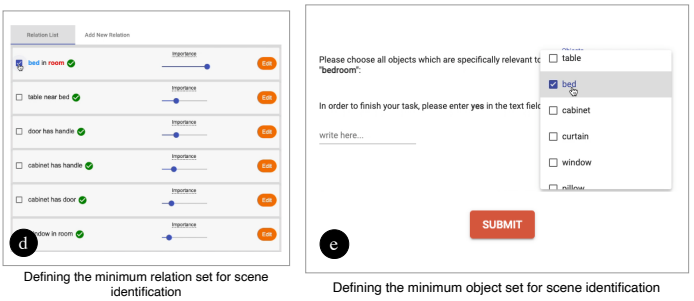


Figure 3.8: Example of unknown-unknowns characterized and detected by ***Scalpel-HS***: (upper-row) False Negative <*Living room*, (-)*sofa*, *Dorm room*> and (lower-row) False Positive <*Conference room*, (+)*person*, *Kindergarden*>. For each case, we show the sampled representative image with relevant concepts on the left and an additional similar unknown-unknown image on the right. All images are shown together with a corresponding saliency map showing where the model is attending to in making the incorrect prediction. Note that in the False Negative case, the sofa leads to False Negative w.r.t. *Living Room* yet False Positive w.r.t. *Dorm Room*.



Validating relations (correct or wrong) | Rating the relevance of relations and giving reasons | Adding missing relations

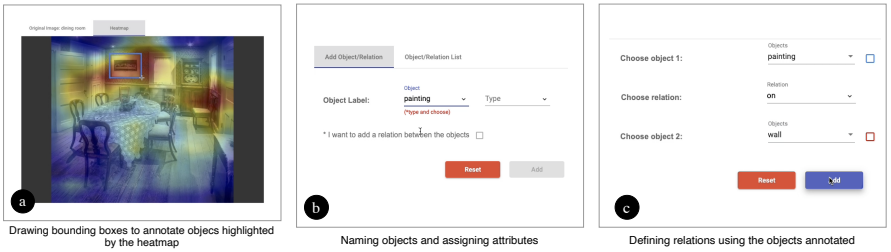
(a) (b) (c)



Defining the minimum relation set for scene identification | Defining the minimum object set for scene identification

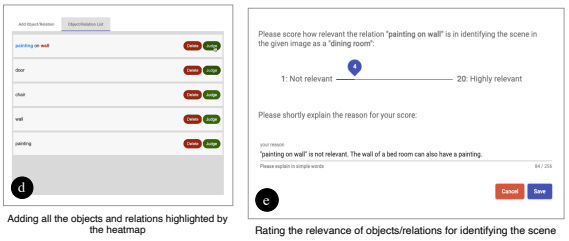
(d) (e)

Figure 3.9: The procedure of the *Should-Know* task zoomed.



Drawing bounding boxes to annotate objecs highlighted by the heatmap | Naming objects and assigning attributes | Defining relations using the objects annotated

(a) (b) (c)



Adding all the objects and relations highlighted by the heatmap | Rating the relevance of objects/relations for identifying the scene

(d) (e)

Figure 3.10: The procedure of the *Really-Knows* task zoomed.

4

Decoding Long-tail Visual Concepts Using Human-Computational Approach

In the evolving landscape of image classification, the crux of reliability and precision often pivots on the quality and diversity of the dataset used. The burgeoning field of machine learning and computer vision has made significant strides, yet the challenge of correctly classifying atypical images - those that deviate from standard representations - persists. This chapter introduces a novel paradigm that blends human intuition with computational efficiency to address this challenge.

Image classification models, despite their advanced algorithms, falter when confronted with atypical instances - images that diverge from the norm due to unique attributes or contexts. The traditional datasets used to train these models often lack the diversity necessary to encompass the wide spectrum of real-world variability. This gap in training leads to models that are adept in handling typical instances but are inept when faced with atypicality. It raises a pivotal question: How can we enhance the ability of these models to recognize and correctly classify atypical instances?

Our approach leverages human computational skills to identify and characterize these atypical instances. Humans, with their innate ability to perceive and process complex visual cues, are adept at recognizing atypicality in images. By incorporating human judgment, we aim to create a more robust and comprehensive dataset that includes a diverse range of atypical instances. This human-in-the-loop method is not only more inclusive but also paves the way for developing more reliable and accurate image classification models.

In this chapter, we delineate a framework that systematically integrates human perception with algorithmic processes. The primary objective is to enrich the training datasets with atypical instances accurately identified and characterized by human annotators. This enrichment promises to bolster the model's ability to handle real-world variability, thereby enhancing its overall accuracy and reliability. The content of this chapter is based on the following paper:

- *Sharifi Noorian, Shahin, et al. "Perspective: Leveraging Human Understanding for Identifying and Characterizing Image Atypicality." IUI '23: Proceedings of the 28th International Conference on Intelligent User Interfaces. [41]*

4.1 Introduction

Data quality is a key factor in the success of image classification systems. Despite their impressive performance, image classification models remain largely unreliable, especially in situations slightly different from those captured in their training phase [140, 141]. As an implication, lack of reliability can lead to negative and sometimes damaging effects, particularly in critical domains such as transport, finance, or medicine.

Among image recognition errors, a specific type known as unknown unknowns has gained particular interest [95]. Unknown unknowns refer to the images for which a model is highly confident about its predictions but is wrong. Unknown unknowns are often discovered after deployment since identifying such errors is challenging due to the over-confidence of the model. Thus, high-quality test data has become vital for understanding and proactively uncovering vulnerabilities in image classification models, as partly demonstrated by recent efforts from both academia and industry [90, 142], e.g., the Dynabench platform by Facebook[1] and the CATS4ML data challenge by Google[2].

A promise of these efforts is the creation of a feedback loop in the lifecycle of an image classification model, thereby enabling a never-ending learning scenario where model performance can continuously improve. Existing methods generally consider both a model-and-human-in-the-loop approach, where human workers identify adversarial instances that are challenging for certain specific image classification models [95–97]. Those methods, mainly contributed by Human Computation studies, are concordant with findings from Computer Vision showing that human visual systems are more robust than machines [143–145].

Little work, however, has addressed deeper questions pertaining to *i)* the characteristics of images that lead to difficulty in their classification from a human perspective and *ii)* whether such human understanding is aligned with the distribution of data that the machine perception is built upon [107].

An instrument that can allow us to gain insights into the difficulty of both human and machine classification of images is the notion of atypicality [102, 103], defined as "the strength of association between observable properties and concepts". From the human perspective, the difficulty in classification has been explained through the difficulty in recognizing components of the image [146]. When such components deviate from the norm (either due to their unusual representation, attributes that deviate from our mental models, the unfamiliar context they are presented, or partial or complete occlusion), we experience difficulty in the image classification task. From the machine learning point of view, models that fail in image classification generally learn incorrect or spurious associations (or correlations) between an image class and the components, arising from incompleteness, imbalances, or undesired biases in the data. This has mainly been found to be due to the under-representation of atypical images in the data [40, 97].

To utilize human understanding for identifying and characterizing atypical images in the context of image classification, a straightforward design for such a task would be gathering responses about the atypicality of a given image from human annotators on a subjective rating scale. However, in the context of image classification, the quality of the resulting insights would depend not only on the cognitive capability of the human

[1] https://dynabench.org
[2] https://cats4ml.humancomputation.com

annotators (and their open world view) but also on their ability to envision the perceived atypical images with respect to the distribution of data. In other words, the perceived atypical images from the human point of view may not necessarily represent a rare concept that the classification model has not encountered during the training phase.

Moreover, cost-efficiency represents another important challenge in real-world settings of image classification in the wild, where stakeholders (e.g., developers and users) have access to a large number of images without knowing the model performance on such images. In such a setting, reducing the number of images for human annotation is of critical importance to save human effort and hence cost. This chapter, therefore, seeks to answer the following research questions:

RQ: How to support humans identify and characterize image atypicality in a cost-efficient manner effectively?

Given the research question, we developed ***Perspective***, an annotation tool that supports effective and scalable human computation for proactive identification and characterization of atypical images. Given an image for annotation, ***Perspective*** presents users with both a global view of images in the class of interest – including both random and visually diverse samples) and a local view of visually similar images in the dataset (possibly from multiple classes) to support human annotation of atypically. ***Perspective*** employs a data sampling method that accounts for both the atypicality and the redundancy of visual and semantic information in the sampled images, thereby narrowing down the most likely atypical images that can be passed along for human annotation. Through controlled crowdsourcing experiments, we demonstrate that our annotation tool can significantly improve worker performance in terms of accuracy and speed in atypicality annotation and that the sampling method is effective in filtering atypical images for annotation.

Through several iterations of annotation (including crowd workers) on 10K images, we present a coding scheme of 20 distinct characterizations of image atypicality, ranging from atypicality with respect to the semantic content (e.g., unusual objects or objects presented in an unusual context), the visibility of objects (e.g., occlusion), to the image quality (e.g., resolution and lighting) and formation (e.g., vantage point, out of focus). The coding scheme reveals the diversity of image atypicality characteristics; particularly, atypical semantic content constitutes the largest category of image atypicality, indicating the heavy skewness of image atypicality due to the unusual content.

To demonstrate the utility of image atypicality annotation, we test the performance of several vision APIs from the industry against our identified atypical images. Results show that atypical images present a strong challenge to state-of-the-art image classification services. To gain a deeper understanding of the alignment between human and machine perception of atypicality, we further fine-tune several image classification models locally, and manually compare the model rationales (using interpretable machine learning techniques) with human annotations. Our analysis shows that model rationales match human-annotated image atypicality to a large extent.

This highlights the potential of ***Perspective*** for not only collecting atypical images to expose model errors but also for identifying reasons which have important implications for developing and deploying reliable image classification models. For example, the identified atypical images can be used to augment the training data in order to improve model performance; the reasons for atypicality can also be used to defer atypical images where models are more likely to fail for human takeover in a hybrid human-AI setting [147, 148].

In summary, we make the following key contributions:

- We introduce a scalable human-in-the-loop framework that orchestrates automatic and human computation components for efficient and effective identification and characterization of image atypicality.
- We identify 10K atypical images and provide a set of structured characterizations (code schemes) of atypicality across four atypicality categories: semantic content, object visibility, image quality, and formation.
- We demonstrate the utility of the collected atypical images by testing against state-of-the-art computer vision models and exposing their weaknesses through atypicality characteristics.
- We provide a set of insights into the need for support for data exploration and navigation in human annotation and the alignment between human and machine perception of image atypicality.

4.2 Related Work

We discuss related works pertaining to data quality issues and their implications, and others present methods that tackle concomitant problems from both algorithmic and human computation angles. The term "data quality" in the context of machine learning usually refers to the coverage or representativeness of data distribution in terms of relevant attributes, e.g., demographics [149] and location [150], or to the correctness of the label [124].

In image classification, it has been found that state-of-the-art models fail when the objects are in strange positions [141] or even exhibit slight changes in position [151] not captured in the training phase. Such a problem remains even with big training data. For example, studies have shown those image classification models trained on the ImageNet dataset exhibit misclassifications consistent with racial stereotypes [149], biases towards textures [19], and limited generalizability to under-representative geo-locations [150]. Those problems are mainly attributed to the inequality of representation in the images within concepts, hence atypicality [152]. Technically, the coverage or distributional representativeness issue in the training data can lead to incomplete models that are prone to generate high-confidence errors, referred to as unknown unknowns [40, 95, 153]. Due to the high confidence, such errors are hard to detect and consequently, imply an ever-big challenge in high-stakes domains with safety, trust, or ethical requirements.

The problem of data quality has been addressed from different perspectives. A large body of work has focused on reducing undesired bias through data preprocessing or

posing additional regularization in model training or inference [154, 155]. Work can also be found on calibrating prediction confidence such that model confidence can become a reliable signal of error risks [156, 157]. Those ideas, while helping to alleviate the issue, are suboptimal by ignoring underrepresented instances or by trading off accuracy for fairness or confidence. We are, instead, more interested in methods that augment the data with adversarial instances.

A closely related line of research in machine learning is adversarial training, referring to the class of methods that automatically generate adversarial instances [158, 159]. The observation mainly drives the idea that *human-imperceptible* differences in the processed data can lead to prediction failures. Adversarial training methods are, therefore, often designed to generate instances similar to existing training instances (with imperceptible differences) while coming with different ground truth labels. As an implication, those methods cannot "naturally" generate images with significant deviation from the training data (e.g., recognizable by humans, often due to the different objects or contexts), rendering them strongly limited in the types of adversarial instances can be generated.

Human-in-the-loop methods have been developed mainly to address model errors. Unlike machines that fully rely on knowledge explicitly encoded in predefined training data, humans excel at leveraging broad, tacit, and contextual knowledge in decision-making and justification. Human computation has, therefore, emerged as a new, promising approach to detecting model errors. A seminal work by Attenberg *et al.* ([95]) proposed to ask humans to gather publicly accessible instances that are potentially difficult for the model to handle. Lakkaraju *et al.* ([97]) introduce a data partitioning technique that first organizes the data into multiple partitions based on feature similarity and then uses an explore-exploit strategy to search for difficult instances across these partitions. An important finding in human computation studies reveals that model errors often come with internal consistency, making them particularly suitable to be described by human language building on top of concepts and properties [96].

The potential of human computation is also verified by studies in Computer Vision, where findings have shown that human visual systems are more robust than machines, especially to distributional shift [143–145], making humans a promising computational means to identifying challenging images for machines. Existing human-in-the-loop methods, however, are specific to address errors of individual models and are hard to generalize to different tasks. Most importantly, we lack a general understanding of the characteristics of an image which leads to the difficulty in classifying it from the human perspective. We set out to fill this gap through our work in this chapter.

In terms of interface design, several studies show the effectiveness of visual analytics tools in discovering model errors. For instance, DriftVis by [160] addresses concept drift in data streams, combining a drift detection method and a streaming scatterplot visualization. ConceptExplorer by [161] detects and analyzes concept drift in multi-sourced time-series data, with visual detection based on prediction models, drift level index, and consistency judgment. [162] presents a system for processing drift detection and visualization in business process event logs. [163] introduces a visual approach for identifying and explaining out-of-distribution samples that cause degradation in predictive models.

It uses an improved ensemble detection method and a grid-based visualization with a novel kNN-based layout algorithm for better context analysis. Our work complements this work by introducing a human annotation tool for identifying and characterizing image atypicality.

4.3 *Perspective*

This section describes ***Perspective***, our proposed annotation tool for identifying and characterizing image atypicality. We first present an overview of the tool and then describe in more detail its components.

Overview.
Perspective takes as input a set of images, samples a subset of images, and feeds them to an annotation interface for human workers to annotate, concerning atypicality rating and rationale. Figure 4.1 presents the overall workflow of the tool. It contains four components:

1. *Image Representation Learning*, to obtain a low-dimensional vector representation of every image in the input dataset for the following sampling components;
2. *Target Images Sampling*, to sample a diverse set of potentially atypical images for atypicality annotation;
3. *Auxiliary Images Sampling*, to sample visually similar images as well as images representative of the class of interest for a given target image and class;
4. *Atypicality Annotation*, to engage human workers in rating the atypicality and providing rationales for the sampled target images, by referring to the auxiliary images.

For ease of understanding, we introduce the components in backward order.[3]

4.3.1 Image Atypicality Annotation

At the front end of our approach is the atypicality annotation task, which is used to both develop the codes of atypicality by trained annotators and to annotate images at scale by crowd workers.

Task Definition & Design.
We consider two types of target image classes, namely object and scene images. Object images are those that contain a given object, e.g., "bird", or "muffin"; scene images, on the other hand, are those that contain multiple objects that together describe a theme, often being an activity or event, e.g., "graduation", "thanksgiving". Atypicality generally means that to the human perception, the object of interest shows an unusual appearance, in an unusual context, or the scene of interest contains unusual objects. In image classification, we further emphasize atypicality with a relative meaning, that is, we consider an object image to be atypical if the object of interest is present in a context *more similar* to the

[3]Implementation details of all methods in the four components are provided in the companion page: https://sites.google.com/view/iui23perspective.

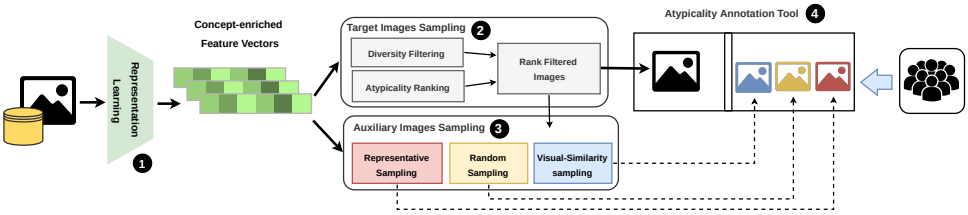


Figure 4.1: Workflow of ***Perspective***. Images from a given dataset are fed to the 1) Representation Learning module to obtain image representations, which are then fed to both the 2) Target Images Sampling module to sample images for annotation and the 3) Auxiliary Images Sampling module to sample three types of auxiliary images, i.e., representative images of a class, random images of a class, and visually similar images, to assist the annotation. The selected target image, together with the auxiliary images are sent to the 4) Atypicality Annotation Interface for human annotation. Note that within Target Images Sampling, images are first filtered by Diversity Filtering, and then ranked by Atyplicaty Ranking. The top-ranking images are sent to the Auxiliary Images Sampling module for sampling images visually similar to the target image.

context of any other classes, or a scene of interest is unusual due to the contained objects being *more similar* to the context of any other classes.

In atypicality rating, we consider two types of errors that annotators can potentially make, namely wrong recognition of typical images to be atypical, i.e., type I error, and the other way around, i.e., type II error. To reduce type I error, human workers need to have insight into a set of typical images representative of the entire class. To reduce type II errors, it is useful to show to the worker which classes visually similar images belong to. We note that both types of auxiliary images are selected from a given dataset that, while coming with its own limit in terms of coverage, is often available at a large size (e.g., publicly available training set or test data in the wild). Methods for sampling those images will be introduced in the next subsections and validated in our experiments.

Task Interface.

Figure 4.2 shows the task interface. It contains three parts: 1) the target image (left), 2) the auxiliary images (right), and 3) the questions that workers answer (bottom).

The auxiliary images are organized in different tabs, each displaying one type of the auxiliary image. In addition to the visually similar images and representative images, we show a random set of images from a given class to help workers gain an idea of the general distribution of visual information in a class.

The task starts by confirming if the target image contains a certain object or is about a scene (initial labels can be obtained from any given image classifiers). Workers are asked to judge and characterize image atypicality for the given image and the associated label by analyzing the target image and comparing that to the different types of auxiliary images. They are asked to rate the level of atypicality using a 7-point Likert-scale (from Highly Typical to Highly Atypical) and when the judgment is atypical (rating bigger than the threshold 4), workers are asked to enter their rationales by selecting from a drop-down list our developed codes of image atypicality (described in Section 4.4).

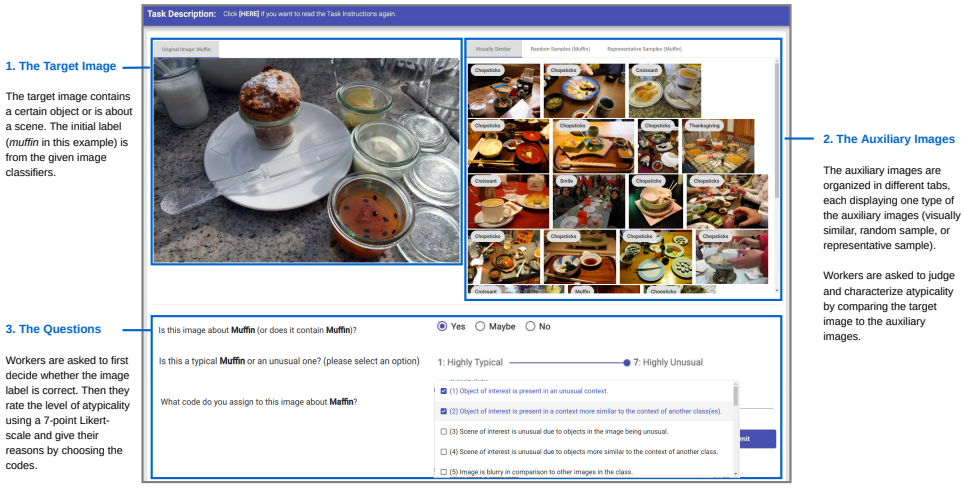


Figure 4.2: A screenshot of the worker interface while using the ***Perspective*** annotation tool.

4.3.2 Sampling Target Images

We now describe our method for sampling the target images for annotation, in order to reach a high cost-efficiency for annotation. To design the sampling method, we consider two requirements of the sampled images: 1) atypicality, i.e., the set should contain as many as possible the indeed atypical ones; and 2) diversity, i.e., the images show a variety of the atypicality characteristics. To this end, we introduce a two-stage method that first filters a subset of images with high visual diversity, and then ranks them according to an atypicality measure we derived from visual features.

Diversity Filtering.

To sample a subset of diverse images, we use the recently proposed adversarial filtering method AFLite [164]. The goal of AFLite is to remove "spurious artifacts in data beyond what humans can intuitively recognize, but those which are exploited by powerful models." For that purpose, the method is designed to reduce the bias in the training data by selecting only a subset of data samples that are the most diverse possible (to avoid spurious artifacts). Consequently, filtered samples by AFLite contain a rather equal distribution of both highly typical samples (if it did not contain any, the model would not be able to learn the typical representations of the target classes of the model) –that we need to exclude through annotation–, and rarer samples –the ones that are indeed atypical.

AFLite works in an iterative process consisting of model training and evaluation. At each iteration, the available dataset is randomly partitioned into two subsets for training and test sets, respectively. The partition is performed m times, and a linear classifier is trained and evaluated independently on each partition. Note that the linear classifier uses the image representation vector as features, which we introduce in the next subsection "Representation Learning" –this allows the sampling to consider visually meaningful features as compared to the low-level, pixel-based features. The evaluations on the m test sets are aggregated into a *predictability score* per sample in the dataset, representing

the ratio of the number of times the sample received a correct prediction over the total number of predictions. In the case of no ground truth labels available, we approximate the predictability score with the agreement among predictions from the linear classifiers. The top k samples with the highest predictability scores are then removed from the dataset, and then we proceed to the next iteration. The stopping criteria is defined over the number of samples remaining in the dataset, and over the number of samples that have a predictability score higher than a pre-defined threshold.

Atypicality Ranking.
Atypical images are a type of outliers, and can be detected through a relevant distribution of image representations: images that are deviated from the mean/medium of a relevant distribution are considered atypical. A general approach for outlier detection in non-parametric distributions is item ranking. In our scenario, we consider leveraging the distribution of images by model-based image representations [165], i.e., the activation of the neurons in a given layer of a neural network model. Using the model-based image representation is favored over the original image representation as it accounts for the varying contribution of different visual features in classification.

Specifically, for a given image i our goal is to find the rank of the image in a subset of images $\mathcal{V}$ randomly sampled from the large dataset (such that the subset keeps the same distribution as the large image set in the wild). To do so, for each image represented by the feature vector (introduced in the next subsection), we run it through an independent multi-layer perceptron model for image classification, record the activation values of neurons in the last layer before the classification layer, and use that as a new representation of the image. Images in $\mathcal{V}$ are then ordered based on the activation values – multiple orderings corresponding to multiple neurons are aggregated into one order. We then obtain a similar representation of image i and find its ranking position in the ordered list of $\mathcal{V}$.

The ranking effectiveness is, to a large extent, dependent on the neural network model. We can start with a given deployed model when available as the initial model. To best leverage human annotations, we progressively train the model following the active learning process [166]. Active learning is a way of training a machine learning model using an optimal subset of the training data, by selecting the most informative instances from the given dataset in multiple iterations. In each iteration, the model is retrained with the newly selected instances combined with the existing ones. Informativeness has various forms that are modeled in different sampling strategies, e.g., uncertainty sampling measures the informativeness of an instance by the uncertainty of model prediction [167]. In our scenario, we replace the informativeness criteria with our atypicality (ranking) measure for sampling.

4.3.3 Sampling Auxiliary Images

Sampling for the auxiliary images is straightforward for random and visually similar images: when the image representations are available, visually similar images are found through a nearest neighbor search. For representative image sampling, we develop an optimization-based approach to ensure that we present the whole spectrum of the visual appearance of a given class to human annotators. We convert the problem into a data partitioning problem, where the goal is to split the images of a given class into partitions such that

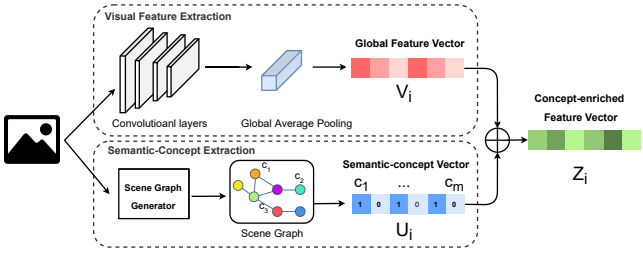


Figure 4.3: The image representation learning model.

images of the same partition are visually similar and those from different partitions are not; representative images can then be sampled from each of the partitions. Given a budget $\mathcal{B}$ (the number of representative images that can be sampled), we solve the following objective function for sampling:

$$\begin{aligned} min & \sum_{i \in C} \sum_{j \in C} D(Z_i, Z_j) X_{ij} \\ s.t. & \sum_{j \in C} X_{ij} = 1 \\ & X_{ij} \leqslant Y_j \\ & \sum_{j \in C} Y_j = \mathcal{B} \end{aligned} \tag{4.1}$$

, where D represents the cosine distance between the feature vectors of two images, i.e., Z_i, Z_j; X_{ij} indicates the decision of whether image i is assigned to partition j; Y_j indicates if the image j is selected as a representative sample (note that the index j is overloaded to represent both the partition and the representative image of the partition). Due to the large number of possible solutions that are associated with the problem of finding an optimal set of representative samples, it is very challenging to provide a deterministic solution. We employ a meta-heuristic approach based on a genetic algorithm which has proven to be effective in finding an optimal solution for such partitioning problems [42].

4.3.4 Representation Learning

We now present our approach for generating the image representation that supports all the previously introduced components. Considering the fact that the atypicality definition is especially relevant about the *content* of an image, i.e., objects and contexts, we aim to generate image representations that not only capture the visual features but also the semantic concepts in the image. Our representation learning approach is depicted in Figure 4.3 that extracts both the visual and semantic features, and concatenates them as the image representation.

Visual Feature Extraction.
We use the convolutional network ResNet-152 [75] to generate a feature vector of the input image. To be specific, we feed the image to a pre-trained model until the final max-pooling

layer (prior to the fully-connected layers), and extract the activations at that layer. Then we flatten the output of the max-pooling layer to obtain a feature vector $V_i : \mathbb{R}^{1\times 2048}$ for each input image i.

Semantic Feature Extraction.
To extract semantic concepts, we use scene graph, a structured representation of objects and their relationships in an image. It consists of a set of relationships, which are represented as $\langle o_1, r, o_2\rangle$, where o_1 and o_2 refer to two objects in the image, and r represents their relation. We generate scene graphs using the state-of-the-art scene graph generation method Neural Motifs [127]. After obtaining the scene graphs for a given set of N images, we extract a set of unique objects and relations. Then, for each input image i, we construct a fixed-length concept vector $U_i : \mathbb{R}^{1\times M}$, where M corresponds to the number of unique concepts. Given $u_i^c \in U_i$ as the c-th concept in the concept vector, we set u_i^c to 1 if the concept c appears in image i, otherwise 0. Finally, for each input image i, we concatenate U_i to the visual feature vector V_i, resulting an enriched image representation Z_i.

4.4 Annotation, Evaluation, and Experimentation Setup

We conduct our annotation and experiments on Open Images [71], a dataset of 9.2M images with 30.1M image-level labels for 19.8K concepts. Following the CATS4ML data challenge, we use a subset of 117K images of 23 classes, including Canoe, Lipstick, Bird, Firefighter, Graduation, etc.[4]

We apply diversity filtering and pick the top 10K most atypical images through our atypicality ranking method. We asked six researchers of our group to act as trusted annotators and manually annotate the 10K images based on their perceived degree of atypicality. Each author independently annotated 1K images, and the remaining 4K images were annotated by crowd workers using the ***Perspective*** interface. As a result of this process, 1925 images were identified as atypical.[5]

4.4.1 Developing a Coding Scheme

To characterize image atypicality, we follow the open coding method rooted in grounded theory [168] and use thematic analysis to develop insights from the images [169]. As a first step, six authors independently assessed a random subset of 46 atypical images (sampled from the set of 1925 atypical images, two for each class) using the interface shown in Figure 4.2. In this round, authors provided detailed explanations for characterizing given images as atypical based on their understanding of the image class in general and the distribution of images in the dataset. Authors then iteratively identified different rationales from their explanations for characterizing image atypicality and assigned codes to represent them. Next, to refine the coding scheme and resolve any disagreement, all authors discussed each of the 46 atypical images and iteratively identified and assigned codes to characterize

[4] see https://cats4ml.humancomputation.com for the full list.
[5] Our annotated dataset will be released on the companion page.

image atypicality.[6] For the sake of completeness and to ensure that the resulting coding scheme can be used by the community for further research, complementary codes which went beyond what was observed in this sample were added.

4.4.2 Evaluating the *Perspective* Annotation Tool

Auxiliary Images Sampling.

We conduct a controlled crowdsourcing experiment to evaluate the effectiveness of the different types of auxiliary images for annotation. We design a between-subject study across four experimental conditions of the auxiliary images: 1) **Random** samples from the dataset; 2) **Representative** samples; 3) **Visually Similar** samples; and 4) **Combined**, which combines all the three above types of auxiliary images.

We randomly selected equal splits of typical and atypical images annotated by the authors, resulting in 300 images and at least ten samples per class label. We ensure that 180 common images are used for all four conditions.

We recruited 50 workers on the Prolific crowdsourcing platform for each condition.[7] Only workers whose approval rate was greater than 90% were considered qualified. During the task execution, each worker is asked to annotate six images, three atypical images, and three typical images randomly selected. To avoid learning bias, each worker can perform only one task throughout the experiment. All the tasks across four conditions were published and completed within the same four-hour period on Prolific to reduce the bias of worker availability. Each worker was paid 0.90 USD (0.65 GBP) for participating in our study. According to Prolific, the actual average hourly reward of our experiment that workers received was 11.75 USD (8.59 GBP).

We measure worker performance in terms of both annotation accuracy and speed. Precisely, accuracy is calculated using the metrics of precision and recall with respect to both typical and atypical images (indicated by author annotations). We measure the speed of worker annotation by the average time spent on identifying each atypical and typical image. .

Target Images Sampling.

We evaluate the effectiveness of our target image sampling techniques, i.e., diversity filtering and atypicality ranking. To do so, we compare the precision of the sampling by diversity filtering alone, atypicality ranking alone, and combined. Precision is measured by the fraction of truly atypical images sampled. To evaluate diversity filtering, we pick 1K images out of 12K obtained filtered images and measure the precision. We repeat the experiment ten times and report the average precision. For a fair comparison, we rank the whole data set using atypicality ranking techniques and select the top 1K images from the ranked list. Finally, we combine both techniques by ranking the 12K images obtained by diversity filtering, using the ranking score obtained from the atypicality ranking. Then, we pick the top 1K instances to calculate the precision.

[6] We do not report inter-rater reliability, as the disagreement between the researchers was resolved through detailed discussions and critical reflections through multiple rounds of iterative coding [170].

[7] Note this group of workers is recruited only for evaluating our task design; workers annotating the 4K images are recruited separately.

4.4.3 Human vs. Machine Perception

We apply our identified atypical images to evaluate the performance of state-of-the-art image classifiers on those images. We first test three industrial APIs: Google Vision API[8], Amazon Rekognition API[9], Microsoft Azure Vision API[10]. To gain a deeper understanding of the alignment between human and machine perceived atypicality, we locally fine-tune three models pre-trained on ImageNet, namely InceptionV3 [156], a VGG19 [171], and a DenseNet121 [172], onto a subset of the images in the Open Images dataset, corresponding to 13 classes containing the largest number of images.[11] To be able to compare the rationales of model predictions to the human characterization of atypicality, we extract saliency maps from the three models using SmoothGrad [99] and manually interpret the visual elements the models highlight.

4

4.5 Results

4.5.1 Characterizing Atypical Images

The resulting coding scheme from the six authors is presented in Table 4.1. We group the codes into four categories, namely, *Semantic Content*, *Image Medium Quality*, *Object Visibility*, and *Formation*. *Semantic Content* contains codes that describe the atypicality of an object in an unusual context (for object images) or a context with unusual objects presented in (for scene images). This category is different from *Image Medium Quality* that describes the atypicality in terms of image resolution, lighting, and color scheme, and from *Formation* that describes atypicality on how the image was formed (e.g., photographed) in terms of the type of medium, vantage point, and focus point. *Semantic Content* is also different from *Object Visibility* in that the former describes the atypicality of the object or scene itself, e.g., an unusual type of Pizza, whereas the latter concerns the appearance of typical object or scene, e.g., a normal Pizza partly occluded in the image. As a remark, we note that many of the codes represent the human perspective of an image while considering the perceived distribution of other images in the class, as well as other classes in the dataset.

Atypicality Distribution

To understand the distribution of atypical images with varying characteristics in our dataset (i.e., corresponding to different codes), we considered a uniformly random sample of 220 atypical images and coded them using the coding scheme. Figure 4.4 presents the distribution of codes that were observed as a result.

Semantic Content is the largest atypicality category: 60% of the images were assigned *Code#1*, suggesting that the most frequent characterization of an atypical image in our sample corresponded to the object of interest being present in an unusual context; 29% images were assigned *Code#2*, i.e., atypical images where the object of interest is presented in an atypical context, in a relative sense (i.e., the context being more similar to that of other classes). *Object Visibility* is another category of atypicality many images are assigned, especially *Code#17* that describes image atypicality from object appearance in terms of shape or other attributes. Interestingly, *Code#18* was assigned to 36% of the atypical images,

[8] https://cloud.google.com/vision
[9] https://aws.amazon.com/rekognition/
[10] https://azure.microsoft.com/services/cognitive-services/computer-vision/
[11] Fine-tuning details are provided on the companion page.

Type	Code	Description
Semantic Content	**1**	Object of interest is present in an unusual context in comparison to other images in the class.
	2	Object of interest is present in a context more similar to the context of one or more other classes.
	3	Scene of interest is unusual due to objects in the image being unusual.
	4	Scene of interest is unusual due to objects which are more similar to the context of another class.
Image Medium Quality	**5**	Image is blurry in comparison to other images in the class.
	6	Image is blurry, making it more similar to images of another class.
	7	Lighting in (a portion of) the image is too dark in comparison to other images in the class.
	8	Lighting in (a portion of) the image is too dark, making it more similar to images of another class.
	9	Lighting in (a portion of) the image is too bright in comparison to other images in the class.
	10	Lighting in (a portion of) the image is too bright, making it more similar to images of another class.
	11	Color scheme of the image is inconsistent with other images in the class.
	12	Color scheme of the image is more similar to that of images in other class(es).
Object Visibility	**13**	Aspect ratio of the object of interest is smaller than other images in the class.
	14	Aspect ratio of the object of interest is larger than other images in the class.
	15	The majority of the object(s) of interest in comparison to other images in the class is(are) occluded.
	16	Dominant object in the image belongs to another class(es).
	17	The shape or other attributes of the object of interest look unusual with respect to other images in the class.
Formation	**18**	The type of medium for representing the object of interest is inconsistent with other images in the class.
	19	The vantage point of the image is inconsistent with other images in the class.
	20	The object of interest is out of focus in comparison to other images in the class.

Table 4.1: Coding scheme to characterize atypical images (in a given dataset). Multiple codes can be assigned to a single image.

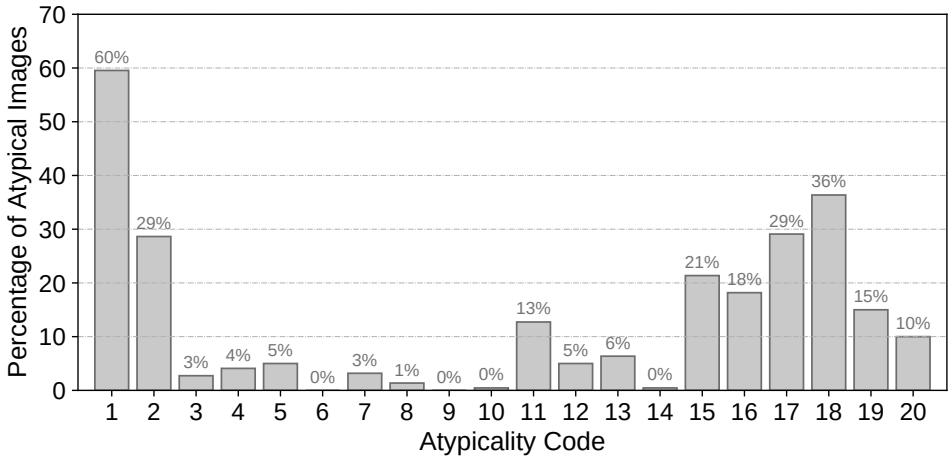


Figure 4.4: Distribution of (a random sample of 220) atypical images as characterized using the coding scheme for image atypicality.

indicating a different medium of representing the object of interest in comparison to other images of the class. As mentioned earlier, certain complementary codes were added to the coding scheme for the sake of completeness. Either a small fraction of atypical images or none were found to correspond to such complementary characterizations (e.g., *Code#6*, *Code#8*, *Code#9*, *Code#10*).

Examples of Atypical Images

4.5 presents examples of atypical images and the corresponding codes assigned to them. 4.5 (a) shows an image with the class label **Muffin**, characterized as being atypical due to *Code#1* the unusual context of the muffin, i.e., presented in a glass, as well as *Code#2* since the surrounding context is most similar to the class of **Chopsticks** where there are multiple dips nearby the main plate.

Figure 4.5 (b) shows an image with the class label **Pizza**, characterized as being atypical due to *Codes #9 & #10* the bright lighting, and importantly, the presence of the pizza in an *Code#17* unusual shape due to the close-up angle, which is also related to *Code#19* the inconsistent vantage point.

Figure 4.5 (c) shows a particularly interesting example of an atypical image corresponding to the class label of **Bird**, which is characterized by a range of codes. We can see that a **Bird** in the image is occluded by two children on the photograph held in a person's hands, thereby characterizing the atypicality of this image on several different fronts.

(a) *Codes: #1,#2* (b) *Codes: #9,#10,#17,#19* (c) *Codes: #1,#15,#16,#18,#20*

Figure 4.5: Performance of industrial vision APIs on the typical and atypical images.

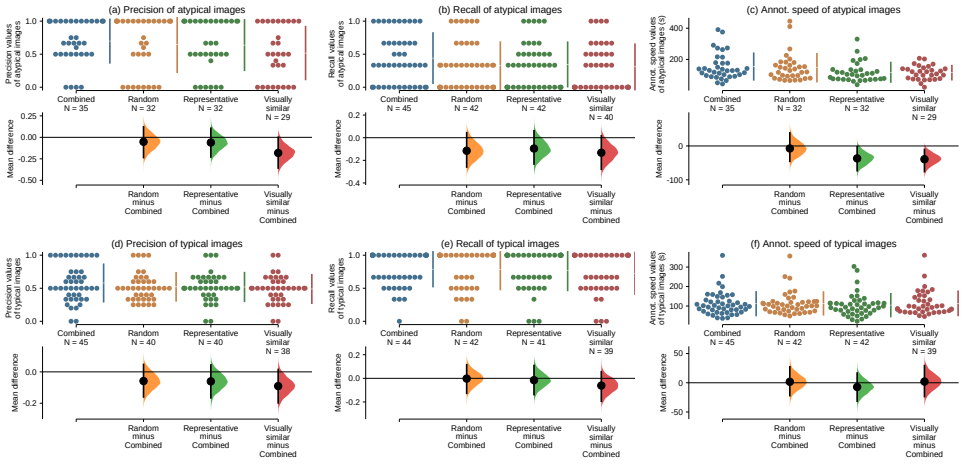


Figure 4.6: Estimation plots of worker precision, recall, and annotating speed on atypical and typical images, respectively, wherein the jitter plots each point represents a worker's performance value (precision, recall, or speed).

4.5.2 Effectiveness of *Perspective*

Auxiliary Images Sampling.

Figure 4.6 shows estimation plots of worker performance (precision, recall, and annotating speed for atypical and typical images) [173]. In this figure, jitter plots show all the measures and how they distribute across the four experimental conditions. The estimation plots also show the effect size by displaying the resampling distributions of the mean difference. We found that the resulting data on all the measures do not follow normal distributions (Shapiro-Wilk tests).

Precision & Recall. In terms of precision, the Combined condition outperforms the other three conditions on both atypical images and typical images (Figures 4.6 (a) and (d)). We can observe comparatively large effect sizes of the differences between the Combined condition and the Visually Similar condition. Regarding recall, the Combined condition also achieves higher performance on atypical images than the other conditions, with relatively large effect sizes. For typical images, the mean recall of Combined, Random, and

Method	Diversity Filtering	Atypicality Ranking	Combined
Precision	25.06	21.98	29.1

Table 4.2: Precision (%) of different target image sampling methods in identifying atypical images.

Representative conditions are almost equal, while that of the Visually Similar condition is relatively lower. Note that we found no significant difference in worker precision or recall ($p > 0.05$, Kruskal–Wallis tests), which is likely due to the low number of images (six) annotated by each worker.

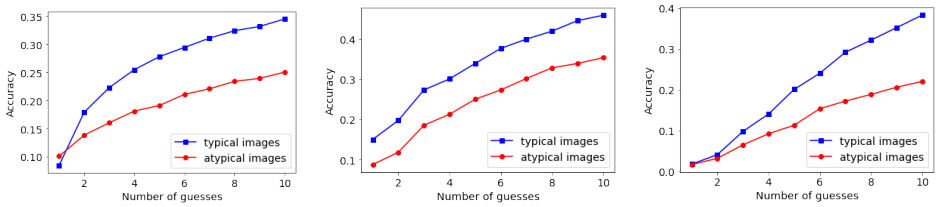
Annotation Speed. We observe that workers in Representative and Visually Similar conditions annotated atypical images faster, as shown in Figure 4.6 (c). In the annotation of typical images, workers across all the conditions exhibited comparable annotation speeds.

Summary. The Combined condition is most effective for accurate annotation (precision and recall), especially for identifying atypical images. The result signifies the advantage of displaying different types of auxiliary images for annotation accuracy. This, however, comes with the trade-off of longer annotation times on average. The Representative condition results in relatively high-quality annotations while enabling fast annotation, especially on typical images, as compared to other conditions. When presented alone, visually similar images do not allow workers to deliver high-quality annotations, possibly due to the lack of a global view of the image class. By analyzing worker activity logs in the Combined condition, we noticed that all the workers who had switched the tabs (32 out of 50) clicked on visually similar images in the annotation. This suggests the perceived utility of visually similar images in informing atypicality identification in the Combined condition.

Target Images Sampling.
Table 4.2 reports the precision of our sampling methods in identifying target atypical images. When diversity filtering and atypicality ranking are used together, we observe a significant improvement in precision.

4.5.3 Human vs. Machine Perception

Industrial APIs.
Figure 4.7 shows the performance of the three image classification APIs on our identified typical and atypical images. These APIs classify an image with multiple labels along with their confidence scores; we, therefore, evaluate the accuracy with respect to the number of guesses allowed – classification is considered correct if one of the guesses is correct. Note that the comparison between different APIs is *not fair* due to the different sizes and vocabularies of image classes they cover. We resample the typical images according to the distribution of atypical images, such that the results on typical and atypical images are comparable. We observe from the figures that the APIs consistently show higher accuracy on typical images than on atypical ones. In particular, when the number of guesses is five, the average accuracy on atypical images is 18%, as compared to 27% on typical ones.



(a) Google Vision API (b) Amazon Rekognition API (c) Microsoft Azure Vision API

Figure 4.7: Performance of industrial vision APIs on the typical and atypical images.

Atypicality	DenseNet121	VGG19	InceptionV3
Typical	63.90	55.69	57.65
Atypical	42.77	22.96	39.04

Table 4.3: Percentages (%) of correct predictions of the fine-tuned models on typical and atypical images.

Local Models.
Locally fine-tuned models also consistently show a higher rate of correct predictions on the typical images than on the atypical ones (see Table 4.3). Due to the high bar of our atypicality judgment, specific samples annotated as typical with incorrect predictions might actually be atypical for more lenient characterizations of atypicality, which can further reinforce the above observations. Those results indicate that, statistically, challenges in model predictions are generally aligned with human judgments of atypicality.

To gain a deeper understanding of the alignment of atypicality perceived by humans and machines, we look into the saliency maps as the rationales of model classifications and compare those to image atypicality characterized by humans.

Typical Images Correctly Classified. The models do not always use a correct rationale for correctly predicting the labels of typical images. Typicality does not mean that the models can easily learn the correct reasoning. Potential spurious biases across the typical images of a training dataset can lead the model to pick up on simpler, incorrect reasons. For instance, for the class **Canoe** (see Table 4.4 (1)), the DenseNet121 model has learned to look at the presence of water when canoes are in the water, but at the presence of a canoe itself when no water is present in the image.

Atypical Images Incorrectly Classified. The rationale of the models is more frequently aligned with human judgments for atypical images that the model predicts incorrectly. It is especially the case when the atypicality code relates to another class. For instance, an image of an Athlete surfing on the water is predicted as **Canoe** by the model due to the presence of water (see Table 4.4 (4)), which follows *Code#2* indicating that the background of this image is more similar to the ones of **Canoe** images in the dataset.

Typical Images Incorrectly Classified. Only a few typical images receive a wrong pre-

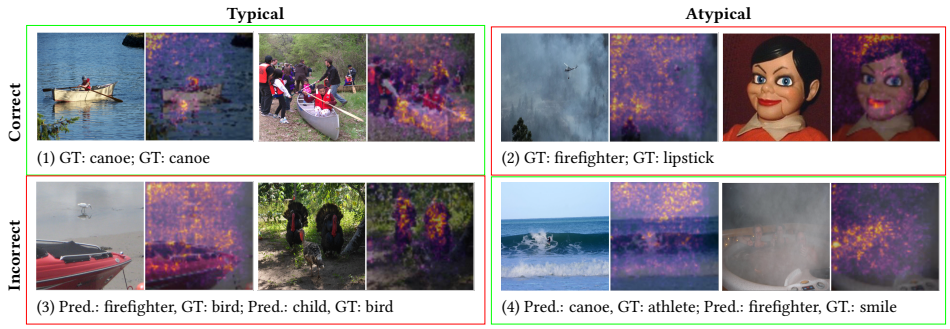


Table 4.4: Example images that received correct or incorrect predictions from the DenseNet121 model and judged as typical or atypical by humans (in green: alignment between human and machine reasoning, in red otherwise). The images are associated with their saliency maps on the right and the predicted (Pred.) and ground truth (GT) labels underneath.

diction from the models. Such misalignment between the model and human reasoning is the most complex to interpret the images and saliency maps. The most obvious cases are when the image contains two rather dominant visual cues hinting at two different classes: one referring to the expected but wrongly predicted class and one potentially referring to another class. For instance, an image of a **Bird** on the beach with a red boat is associated by the model to **Firefighter** probably due to the red color (see Table 4.4 (3)).

Atypical Images Correctly Classified. Part of the images which received correct predictions while marked atypical merit their atypicality judgment to be reviewed once the human judges have further understood the model rationale by analyzing multiple images and saliency maps. As an example, a **Firefighter** image showing a small helicopter in a background of smoke (see Table 4.4 (2)) is marked atypical as firefighter images instead usually contain a firetruck or individual firefighters. Yet, the model learned to use the smoke to predict the label **Firefighter** (e.g., see the image in Table 4.4 (4)). Another image presents a plastic doll with red lips (see Table 4.4 (2)), that was coded as atypical due to the medium of representation (doll with simple facial features) that is unusual for **Lipstick**, yet the model still correctly focuses on the lipstick to make its predictions. These cases show that it is not always sufficient to use the atypicality codes to estimate whether a model prediction will be correct. Still, it also requires an understanding of how important a given atypicality characterization is concerning other potentially more typical characteristics of the image that are less obvious from an open-world human perspective (e.g., the smoke for the firefighter instead of the individual firefighter). This hints at new opportunities in the coding process: a sequential coding procedure could potentially first allow the judges to build an understanding of model reasoning by visualizing saliency maps, ground truth, and predictions and then ask them to characterize image atypicality based on such understanding of the reasoning.

4.6 Discussion

4.6.1 Importance of Context Expansion

Results from our controlled crowdsourcing experiments show that workers, when presented with representative images of a given class and with visually similar images to the target image, perform significantly better in terms of both annotation quality and speed. These results verify our initial assumption that human perspectives that rely on annotator experiences can be limited in envisioning image atypicality; for that, being able to see image distributions in the dataset (images in the same class but also the other classes). This is confirmed further by our results from the coding exercise, where many of the codes represent not only the human perspective but also such perspective conditioned on the distributions of the images. These results therefore, pose new research questions on what impacts human perspective, and especially how new experiences gained from human interactions with new environments (objects, scenes) shape the development of human perspectives. Such questions are related to the literature on cognitive science and creativity especially. In this literature, it has been shown that collecting and navigating through information is an important phase in the creative process, which expands the current context of the topic (and fosters associative and inspirational learning) [174, 175]. While partly answering the research questions, more research is needed to cross-check the exact influence of human experience on the perception of atypicality.

4.6.2 Need for Collaboration and Interaction Tools

From the tooling perspective, the results imply that providing adequate support for human annotation is an important and perhaps indispensable part of human annotation. In our work, we have mainly explored methods and interfaces for sampling and visualizing images from certain distributions, while much is left for future studies. An important aspect to be considered in developing new tools would be to consider the cooperation among human workers. In our specific task of image atypicality identification and characterization, being able to communicate with other workers allow further expanding the current context of an individual worker as constrained by what they observe and their own memory, making it possible to connect to the new contexts other workers are experiencing. When developing support for context expansion from either extra information or communication, an essential type of atypicality that needs to be accounted for is the semantic content atypicality, namely unusual content and context. This type of atypicality makes the majority and is perhaps the most complex type given the diversity of objects and scenes. Future work in this direction can benefit from cognitive science but also more technical domains such as knowledge management, to link the annotation interface to knowledge bases in the backend that can offer in real-time new concepts related to the running context. This also calls for new research on interaction techniques, namely, how to display the increasing amount of information to workers while not significantly increasing their cognitive load.

4.6.3 Response and Data Sampling Biases

One common issue in most crowdsourced image annotation tasks is "response bias", where annotators may tend to complete tasks quickly to earn rewards, leading them to choose the simplest answer without considering answer quality. In our task, workers might have

the tendency to label images as atypical which does not require characterizing atypicality. To reduce such a bias, we have employed several approaches such as using "gold standard" images to filter out unreliable annotations and soliciting multiple annotations per image. Another potential source of bias in our approach can be the data sampling bias. This has been a major consideration in the design of our approach, which takes into account image diversity in addition to atypicality. ***Perspective*** however, can potentially be further improved by integrating alternative data sampling methods, e.g., combined with out-of-distribution data detection [163].

4.6.4 Implications for Machine Learning and Interdisciplinary Research

As for the application domain, findings from our study have several implications on machine learning (computer vision specifically). An important one is the notion of atypicality as the proxy of data quality. We showed how easy it is for state-of-the-art machine learning systems to fail in dealing with atypical images. A direct implication would be the need to consider atypical images not only in model development but also in model deployment: it is nearly impossible to collect the perfect set of images covering all possible scenarios in one shot, yet this can be compensated by the incremental discovery of atypical images in model deployment, from which the data quality can be gradually improved by integrating such images. This implication aligns with the current discussions on data-centric AI, which stresses more the importance of developing higher-quality datasets than models [176]. Our contribution in this sense is showing how human perspectives can be leveraged in the process of image atypicality identification and characterization. A further implication from this study is, therefore, the need to better bridge machine learning research and data science, with the IUI and HCI communities.

There are multiple ways of using ***Perspective*** to improve the reliability of image classification. One approach is augmenting the training data with the identified atypical images to retrain the model, in an active learning setting where model performance can be continuously improved with new images [177–180]. Another approach one can also consider to use ***Perspective*** for reliable image classification is a hybrid human-AI setting, where humans can take over decision-making when model decisions are unreliable. In such a scenario, the human-identified reasons for image atypicality can be used to build a decision deferral mechanism that filters images for decision handover [147, 148].

4.7 Conclusions and Future Work

In this chapter, we have presented a study on image atypicality identification and characterization through human annotation. We introduced ***Perspective***, an annotation tool that increases annotation accuracy and speed by presenting several distinct sets of auxiliary images, and that increases cost-efficiency by carefully designed sampling techniques. Iterative coding resulted in a coding scheme for image atypicality, comprising 20 distinct characterizations of image atypicality. Trusted and crowdsourced annotation resulted in 10K images with atypicality judgments. Experiments show that the identified atypical images present a strong challenge to state-of-the-art image classification services and models, and that atypical characteristics can well explain model rationales in instances of

incorrect classification.

In the imminent future, we will improve the annotation tool to account for model behavior, explore the integration of model interpretability methods, and study further the utility of atypical images for improving system performance by, e.g., augmenting the training data.

5

A Graph-based Foundation Model For Multi-modal Learning

The rapid advancements in artificial intelligence (AI) aim to bridge the gap between human cognitive capabilities and machine intelligence. In this pursuit, multi-modal learning emerges as a crucial step, enabling systems to process and understand various forms of data simultaneously, much like the human brain. In this chapter, we introduce a novel graph-based foundation model named GraphFusion, designed to harness the power of multi-modal learning by integrating diverse data types, specifically images and text, through a unified graph-based framework. Our approach leverages the inherent strengths of graphs to model complex relationships and dependencies across different modalities, thereby facilitating a deeper understanding and semantically enriched representation of multi-modal data.

Using self-supervised learning techniques, the proposed foundation model pre-trains on a large-scale, multi-modal dataset with weak semantic correlations. This approach helps the model understand the subtle interactions between different data types. The pre-training enables the model to comprehensively understand the data, which can be fine-tuned for various tasks in natural language processing and computer vision. Our proposed model integrates various inductive biases inherent in different data modalities, effectively handling the heterogeneity and complexity of the datasets. Our results showcase significant performance improvements across multiple benchmarks, highlighting the model's ability to comprehend and handle multi-modal data cohesively. The proposed graph-based foundation model emphasizes the benefits of multi-modal learning and demonstrates the efficiency and adaptability of graph-based architectures. It provides a framework for creating models proficient in multi-modal representation learning with graphs.

5.1 Introduction

The advancement of artificial intelligence toward bridging the gap between human cognitive capabilities and machine intelligence has identified the integration of visual and textual data as a pivotal challenge [181–183]. Achieving sophisticated multi-modal understanding, where systems can process and interpret multiple forms of data simultaneously, is crucial for various applications. These applications include but are not limited to image captioning [181, 184], visual question answering [184–186], and more. The importance of this integration is especially underscored by advancements in technologies that enable richer interactions between visual elements and textual information [187, 188].

A multitude of strategies have been proposed to address the challenge. Early methodologies relied on hand-crafted features and shallow learning models, but these were quickly found to be insufficient for handling the complexity of the problem [189, 190]. More recent methodologies have turned to deep learning, using architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to learn representations [191, 192]. However, these methodologies often struggle to effectively model the interactions between the modalities and align their representations [193, 194]. Given these challenges, attention has turned towards more sophisticated models that can better capture the complex relationships between visual and linguistic modalities. For instance, attention-based models have been proposed as a solution, allowing the model to focus on relevant parts of the input when generating output [195, 196]. These models have shown promise in tasks such as image captioning, where the model must focus on different parts of the image when generating each caption word. However, while attention mechanisms have improved performance, they do not fully solve the problem of aligning visual and linguistic representations [197].

Addressing this gap, we present ***GraphFusion***, a novel graph-based foundation model explicitly designed for multi-modal learning. Unlike traditional approaches that struggle with the heterogeneity and complexity of multi-modal data, ***GraphFusion*** leverages a unified graph-based framework to model the intricate relationships and dependencies across different modalities, particularly images and text [198, 199]. We use self-supervised learning techniques to pre-train the ***GraphFusion*** model, leveraging large-scale multi-modal datasets from diverse social media platforms [200, 201]. Each image in these datasets is paired with several textual descriptions provided by human annotators [202]. This varied dataset is essential for training ***GraphFusion*** to effectively handle and integrate information from different sources. Our self-supervised approach enables the model to identify and utilize complex visual and textual data interactions independently [203]. Consequently, ***GraphFusion*** develops a sophisticated understanding of the interplay between these modalities, improving its capability to analyze and interpret complex multi-modal inputs. This is particularly important for applications that demand advanced comprehension of intertwined visual and textual content, such as content recommendation engines or automated content moderation systems [204].

Our results underscore the model's robust performance across multiple benchmarks, illustrating its capability to provide a cohesive understanding and handling of multi-modal data [205]. ***GraphFusion*** demonstrates the tangible benefits of multi-modal learning and

highlights the efficiency and adaptability of graph-based architectures in creating proficient models for multi-modal representation learning [206]. Through this work, we establish a compelling framework for future advancements in the field, emphasizing the potential of graph-based approaches in enhancing AI's multi-modal learning capabilities, paving the way for its application across various tasks in natural language processing and computer vision [207].

In summary, the primary contributions of this work are as follows:

- We introduce ***GraphFusion***, a generic Vision-Language model based on Heterogeneous Graph Neural Networks. ***GraphFusion*** allows for the learning of unified multi-modal representations that encapsulate the semantics of both modalities by representing visual and linguistic data as nodes in a heterogeneous graph.

- We present a novel set of techniques for pretraining the ***GraphFusion*** on large-scale image-text datasets. This design choice fosters alignment and ensures that the learned representations are semantically rich and well-aligned across the two modalities.

- We validate the effectiveness of our approach by conducting comprehensive experiments on cross-modal retrieval tasks using various standard benchmark datasets.

5.2 Related Work

5.2.1 Vision-language Pretraining

Self-supervised learning has recently been recognized as a powerful strategy for pretraining deep neural networks. Within this domain, Vision-language pretraining has garnered considerable attention. It involves training a model on a substantial dataset of image-text pairs without manual annotations, enabling it to learn significant visual and textual features and to understand their interconnections.

One pioneering effort in Vision-Language Pretraining was CLIP [192], which utilized a contrastive loss function to align semantically related image-text pairs. The studies demonstrated that, with an adequately large dataset, CLIP could match the performance of fully supervised models across various vision-language tasks.

Subsequent research has explored numerous facets of Vision-Language Pretraining. For example, ALIGN [208] enhanced the training methodology by incorporating noisy image and alt-text pair data to improve the model's robustness to real-world variability. Other researchers, such as Li et al. [209] and Chen et al. [210], have employed pre-trained object detectors to identify visual concepts that aid the training of multi-modal transformers, thus focusing the model's attention on more prominent visual features.

Recent developments have also emphasized cross-modal interactions. Models like ALBEF [211], TCL [212], FLAVA [213], and Florence [214] have integrated multi-modal fusion layers atop modality-specific transformer encoders, showing substantial efficacy on diverse vision-language tasks. Additionally, advances in generative modeling for tasks such as image captioning have been pursued by researchers like Mokady et al. [215], further enhancing performance on complex tasks like visual question answering.

In summary, Vision-Language Pretraining is a rapidly evolving research area, producing significant breakthroughs that leverage the synergy between images and text to develop more robust and generalizable model representations useful across various applications.

5.2.2 Heterogeneous Graph Neural Networks

Heterogeneous Graph Neural Networks (HGNNs) are increasingly used to model complex interactions across different domains. Notable examples include R-GCN [216] and HAN [217], which are designed to manage graphs with diverse node and edge types. In the vision-language context, Zhou et al. [218] proposed a unified graph neural network that combines visual and linguistic features within a single graph representation. However, their approach does not explicitly model the interconnections between visual and linguistic elements as edges, potentially limiting effective modality alignment. Our HGNN model addresses this by treating visual and textual data as nodes and explicitly modeling their interrelations as edges within the heterogeneous graph, enhancing modality alignment.

In summary, our graph-based approach leverages advancements in vision-language pretraining and heterogeneous graph neural networks to effectively model and align visual and linguistic modalities. By structuring data as nodes and their relationships as edges in a heterogeneous graph, our method promotes information transfer and learning of joint representations that capture both modalities' semantics. Our empirical results on various downstream tasks, including image captioning, visual question answering, and visual grounding, highlight our method's capability to tackle these challenges and its potential to push forward vision-language understanding.

5.3 The GraphFusion Framework

The overall architecture of our proposed approach is illustrated in Figure 5.1. This section presents the methodology and mathematical formulations of our proposed Heterogeneous Graph Neural Network (HGNN) approach for multimodal vision-language understanding. The method consists of the following main components: (1) Graph construction, (2) Graph Neural Network (GNN), (3) Hierarchical Pooling, and (4) Pretraining loss functions, including Masked Language Modeling, Image-text Matching, and Masked-Image-Modeling.

5.3.1 Problem Formulation

We formulate the problem of multimodal vision-language understanding as learning a joint representation space $\mathcal{S}$ that effectively fuses information from both the visual and linguistic modalities. Given a set of images $\mathcal{I} = i_1, i_2, ..., i_N$ and associated textual descriptions $\mathcal{T} = t_1, t_2, ..., t_N$, our goal is to learn a mapping function f (Equation 5.1) such that the learned joint representations $s \in \mathcal{S}$ capture the semantic relationships between the modalities $\mathcal{I}$ and $\mathcal{T}$.

$$f : \mathcal{I} \times \mathcal{T} \rightarrow \mathcal{S} \tag{5.1}$$

The representation space $\mathcal{S}$ should be semantically coherent and well-aligned across modalities, enabling downstream vision-language tasks that require correlating and amalgamating visual and textual data. For instance, in image captioning, s_i can generate a textual

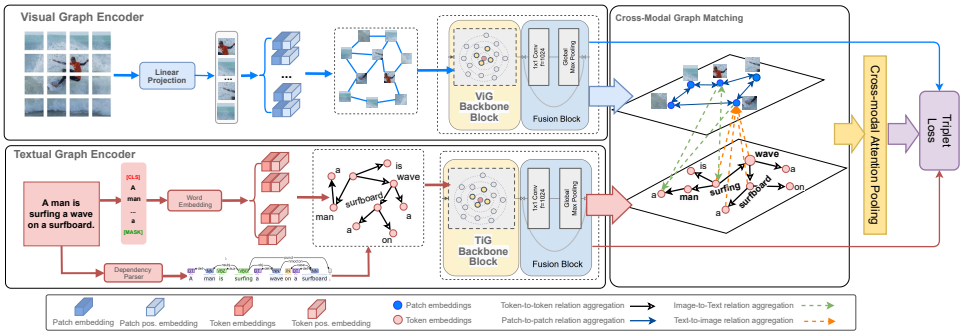


Figure 5.1: The input of the network is a pair of image-text. Image and text are separately transformed into unimodal graphs and passed to separated Graph Neural Networks simultaneously. A graph-matching layer is adopted for cross-modal context modeling and multi-modal fusion. Masked attention pooling encodes aligned image and text data into a multi-modal embedding space. Then, a triplet loss function is fed to calculate the loss in the training phase. The joint embedding is passed to a 2-way prediction head for inference to calculate the matching score.

description t_i for image i_i, by conditioning a language generation model on s_i. Similarly, in visual question answering, s_i contains information from both the input image i_i and question q_i to predict an answer a_i. We propose GraphFusion, a novel Vision-Language model based on Heterogeneous Graph Neural Networks. GraphFusion represents visual and linguistic elements as nodes in a heterogeneous graph, with edges capturing relationships between them. This graph-based representation allows for modeling the complex interdependencies between modalities and enables information propagation across visual and textual nodes. Through end-to-end pretraining on large-scale datasets, GraphFusion can learn fused joint representations $s \in \mathcal{S}$ that are semantically rich, well-aligned, and suitable for many downstream tasks. In what follows, we provide a detailed explanation of the GraphFusion model architecture and training techniques.

The goal of Vision-Language pretraining is to learn joint representations that effectively fuse visual and textual modalities. To this end, we propose an HGNN model which determines relevant relationships between images $\mathcal{I}$ and text $\mathcal{T}$ by learning a parameterized mapping function:

$$f_\theta : \mathcal{I} \times \mathcal{T} \rightarrow \mathcal{G} \tag{5.2}$$

where $\mathcal{G} = (V, E)$ is a heterogeneous graph comprising: 1) Visual nodes $V^I \in V$ representing images $\mathcal{I}$ 2) Textual nodes $V^T \in V$ representing text $\mathcal{T}$ 3) Edges E connecting cross-modal nodes. The parameters θ of f_θ are learned by optimizing a pretraining objective (e.g. contrastive loss) over large datasets of image-text pairs. By maximizing information flow across the graph, the HGNN determines relationships that align meaning at an abstract level and fuse the modalities. For example, given an image i_j and caption t_k, the HGNN may construct edges:

$$f_\theta : i_j \times t_k \rightarrow (e_{i_j,v^k}, e_{v^j,t_k}) \tag{5.3}$$

, where e_{i_j,v^k} connects the visual node for image i_j to a high-level semantic node v^k (representing e.g. objects/events), and e_{v^j,t_k} relates v^k to the textual node for caption t_k. Through Vision-Language pretraining, the HGNN learns these cross-modal relationships across datasets, developing an abstract understanding necessary to align and fuse visual and textual modalities into joint representations $\mathcal{S} = \{s_1, s_2, ..., s_N\}$. These representations encode high-level semantic concepts that cut across both modalities while retaining modality-specific attributes. By maximizing information flow across modalities and optimizing an end-to-end pretraining objective, our methodology determines the relationships required to solve various downstream vision-language tasks.

5.3.2 Graph Construction

To construct the hierarchical heterogeneous graph, we define the visual and linguistic graphs at different levels of abstraction.

Visual Graph Construction

We incorporate superpixels of the input image to construct the visual graphs. Each superpixel is associated with a node in the RAG, and edges connect nodes if the corresponding superpixels are adjacent in the image. The similarity between the connected superpixels determines the weight of each edge. Let $G = (V, E)$ be the RAG, where $V = v_1, v_2, ..., v_n$ represents the set of nodes corresponding to superpixels, and E represents the set of edges connecting adjacent superpixels. The similarity between two adjacent superpixels, v_i and v_j, can be calculated using various features such as color, texture, and location. Let $\mathbf{F}(v_i)$ and $\mathbf{F}(v_j)$ denote the feature vectors of nodes v_i and v_j, respectively. The weight of the edge $(v_i, v_j) \in E$ can then be defined as:

$$w_{i,j} = \exp\left(-\frac{d(\mathcal{F}(v_i), \mathcal{F}(v_j))}{\sigma^2}\right) \tag{5.4}$$

, where $d(\mathbf{F}_i, \mathbf{F}_j)$ represents the Euclidean distance between the feature vectors $\mathbf{F}_i$ and $\mathbf{F}_j$. The parameter σ controls the scale of the exponential kernel. The RAG captures both the image topology and the similarity between regions, making it useful for various computer vision tasks such as segmentation and object detection.

Textual Graph Construction

For sentence texts, we follow the recent trends in the community of Natural Language Processing and utilize the pre-trained BERT [219] model to extract word-level textual representations. Similar to visual features processing, we also utilize FC layers to project the extracted word features into a D_t-dimensional space, denoted as $T = [t_1, t_2, \cdots, t_N], t_j \in \mathbb{R}^{D_t}$, with length N.

To facilitate cross-modal interaction and embedding space consistency, the projected dimensions are the same ($D_v = D_t$) for visual and textual representations. For subsequent local-global (image-word/sentence-object) intermodal interaction and final cross-modal similarity calculation, we use the average-pooling operation to obtain the global image feature $\bar{V}$ for sentence-to-image and the global sentence feature $\bar{T}$ for image-to-sentence.

Implicit textual graph building and embedding. In contrast to the approaches [38, 29, 5, 26] of explicitly modeling inter-word dependencies, we construct a fully connected graph for each sentence, where the semantic features T of the words serve as nodes and the semantic similarities A^t between words serve as edges. We argue that explicit modeling of sentences tends to focus only on the words of object and relation and loses the benefit of many attribute descriptions. Similar to the visual enhancement process, as shown in Figure 2 (2), we apply GCNs [21, 22] with residuals to reason and get the final textual representations T^f with the relationship enhanced, as follows:

$$A^t = \left(W_\varphi^t T\right)^T \left(W_\phi^t T\right) \tag{5.5}$$

$$T^f = \left(A^t T W_g^t\right) W_{r_t} + T \tag{5.6}$$

, where W_φ^t and W_ϕ^t denote the mapping parameters, W_{r_t} is the residual weights, W_g^t is the weight matrix of the GCN layer.

5.3.3 Multi-scale Heterogeneous Representation Learning

In this section, we present a novel approach for learning multi-scale heterogeneous representations that can effectively capture complex patterns in data with diverse characteristics. Our method incorporates various representation learning techniques to generate a robust and comprehensive feature set for cross-modal tasks.

Hierarchical Feature Extraction. We first extract hierarchical features from the data using a series of convolutional and pooling layers. For each data modality, we obtain a set of multi-scale feature maps $\mathcal{F}i = Fi^{(1)}, F_i^{(2)}, \cdots, F_i^{(M)}$, where $i \in v, t$ denotes the modality (visual or textual), and M represents the number of scales. Each feature map $F_i^{(m)}$ corresponds to a specific scale and is represented as a matrix in $\mathbb{R}^{H_m \times W_m \times D_i}$, where H_m and W_m denote the height and width of the map, and D_i is the feature dimension.

Multi-scale Fusion We employ a fusion strategy to integrate the multi-scale features that combine the feature maps at different scales. The fused feature map F_i is computed as follows:

$$F_i = \sum_{m=1}^{M} \alpha_i^{(m)} F_i^{(m)}, \tag{5.7}$$

, where $\alpha_i^{(m)}$ is a learnable weight parameter that controls the contribution of each scale in the fused representation. The weights are normalized to sum to one, i.e., $\sum_{m=1}^{M} \alpha_i^{(m)} = 1$.

Heterogeneous Feature Embedding We apply separate, fully connected (FC) layers for each modality to embed the fused multi-scale features into a common space. These layers project the features into a D-dimensional space, where D is the desired embedding dimension:

$$E_i = W_i F_i^* + b_i, \tag{5.8}$$

, where W_i and b_i are the weight matrix and bias term for the FC layer corresponding to modality i.

Hierarchical Pooling After the GNN message passing, we obtain the updated node features $H_v = h_v^{(T)} | v \in V$ and $H_l = h_l^{(T)} | l \in L$ for visual and linguistic nodes, respectively, where T is the total number of GNN iterations. To generate a fixed-size joint representation, we apply a hierarchical pooling strategy on the node features:

$$\mathcal{H} = \text{Pool}(\text{Pool}(H_v) \oplus \text{Pool}(H_l)) \tag{5.9}$$

„ where Pool$(\cdot)$ is a pooling function, such as mean or max pooling, and $\oplus$ denotes element-wise addition.

5.3.4 Pretraining Loss

Intra-Modal Contrastive (IMC). We define Intra-Modal Contrastive loss to learn the semantic difference between positive and negative samples within the same modality. Inspired by [220], for each new graph $X_v^{(GM)}$ and $X_l^{(GM)}$, we consider two random "views" of the same graph as a positive pair. We maximize agreement between visual sub-graphs (I_1, I_2) by using the contrastive loss $\mathcal{L}_{\text{nce}}\left(I_1, I_2, \tilde{I}\right)$. Similarly, the agreement between textual sub-graphs are maximized by $\mathcal{L}_{\text{nce}}\left(T, T_+, \tilde{T}\right)$. Overall, we minimize the following objective to guarantee reasonable intra-modal representation learning.

$$\mathcal{L}_{\text{imc}} = \frac{1}{2}\left[\mathcal{L}_{nce}\left(T, T_+, \tilde{T}\right) + \mathcal{L}_{nce}\left(I_1, I_2, \tilde{I}\right)\right] \tag{5.10}$$

The core idea behind the $\mathcal{L}_{imc}$ is to enforce the uniformity of the whole representation space of image and text such that the embeddings are uniformly distributed. Consequently, it improves the quality of the learned representations in each modality, further facilitating joint multi-modal learning.

Image-Text Matching (ITM). We adopt ITM, widely used in previous VLP studies, to fuse vision and language representations. Given an image-text pair, ITM predicts whether they are matched (positive examples) or not (negative examples), which can be regarded as a binary classification problem. We use a cross-attention pooling module to generate the joint representation of the two enriched graphs returned by the Graph Matching module. Then,

we fed the joint embedding into a fully-connected layer to predict the matching probability $\phi(I,T)$. We assume that each image-text pair (I,T) sampled from the pre-training datasets is a positive example (with label 1) and construct negative examples (with label 0) through batch sampling. The ITM loss is defined as:

$$\mathcal{L}_{itm} = \mathbb{E}_{p(I,T)} H\left(\phi(I,T), y^{(I,T)}\right) \tag{5.11}$$

, where $H(;)$ is the cross-entropy, $y^{(I,T)}$ denotes the label. The overall training objective of our model is computed as follows:

$$\mathcal{L} = \mathcal{L}_{itm} + \mathcal{L}_{\text{imc}} \tag{5.12}$$

5.4 Experiments and Resuts

In this section, we report the results of our experiments to evaluate the proposed approach, ***GraphFusion***. We will introduce the dataset and experimental settings first. Then, we compare the performance of ***GraphFusion*** with state-of-the-art image-text retrieval approaches quantitatively. In addition, we conduct ablation studies to investigate the effectiveness of each component of our model. Finally, we provide some qualitative analysis of the results.

5.4.1 Pre-training Datasets

Following previous experimental protocols [208, 221], we use COCO [52], Visual Genome (VG) [222], Conceptual Captions (CC) [223], and SBU Captions [224] as the pre-training dataset in our study, where a total of 4.0M unique images and 5.1M image-text pairs are covered. We term this dataset a 4M dataset in our study. To prove that our method can be applied to large-scale datasets, we further use CC12M [225]. Together with the 4M dataset, we reach large-scale pre-training data with 14.97M unique images and 16M image-text pairs.

5.4.2 Downstream Tasks

Image-Text Retrieval includes two tasks: (1) image as query and text as targets (TR); (2) text as query and image as targets (IR). The pre-trained model is evaluated on Flickr30K [37] and COCO [52] by following fine-tuning and zero-shot settings. For the fine-tuning setting, the pre-trained model is fine-tuned on the training data and evaluated on the validation/test data. The pre-trained model is directly evaluated on the test data for the zero-shot setting. In particular, for zero-shot retrieval on Flickr30K, we follow [208] to evaluate the model fine-tuned on COCO. Visual Question Answering (VQA) [36] aims to predict the answer given an image and a question (in text format), which requires an understanding of vision, language, and common-sense knowledge to answer. This task is a generation problem following the same setting [208]. Specifically, an answer decoder is fine-tuned to generate the answer from the 3,192 candidates. Visual Entailment (SNLI-VE) [226] predicts whether a given image semantically entails a given text, a three-class

classification problem. Specifically, the class or relationship between any given image-text pair can be entailment, neutral, or contradictory. Compared with VQA, this task requires fine-grained reasoning. Visual Reasoning (NLVR^2) [227] determines whether a natural language caption is true about a pair of photographs. We evaluate our model on NLVR^2 dataset, which contains 107,292 examples of human-written English sentences paired with web photographs. Since this task takes the text and two images as input, we extend our model by following [208].

5.4.3 Implementation Details

Our experiments are performed on 8 NVIDIA A100 GPUs with the PyTorch framework [36]. Our vision encoder is implemented by ViT-B/16 with 12 layers and 85.8Mpa− rameters. A 6-layer transformer implements both the text and the fusion encoder. They are initialized by the first 6 layers and the last 6 layers of BERT $_{\text{base}}$ (123.7M parameters), respectively. We set $K = 65,536$ and $m = 0.995$. The model is trained for 30 epochs for the pre-training stage with a batch size of 512. We use a mini-batch AdamW optimizer [31] with a weight decay of 0.02. The learning rate is initialized as $1e-5$ and is warmed up to $1e-4$ after 2,000 training iterations. We then decrease it by the cosine decay strategy to $1e-5$. For data augmentation, a 256×256-pixel crop is taken from a randomly resized image and then undergoes random color jittering, random grayscale conversion, random Gaussian Blur, random horizontal flip, and RandAugment [9]. During the fine-tuning stage, the image resolution is increased to 384×384 and the positional encoding is interpolated according to the number of image patches.

5.4.4 Image-Text Retrieval

Table 5.1 compares our model with state-of-the-art methods on image-text retrieval benchmarks. Although models with modality-specific encoders usually perform better due to more parameters and architectural biases, our one-tower model, OneR, achieves the best zero-shot performance among similar baselines. Notably, ***GraphFusion*** achieves competitive results without pre-training or initialization, indicating that once the intermodality gap is addressed, both modalities can be effectively encoded within a single representation space with minimal bias. We evaluated the generalizability of our model by transferring it to downstream tasks without fine-tuning, consistently outperforming state-of-the-art models with an average improvement of +9.5% on COCO and +12.2% on Flickr30K compared to ViLT. Our approach, similar to ALBEF, shows superior performance with a +2.7% TR/R@1 and +3.4% IR/R@1 boost on the MSCOCO(5K) subset, likely due to our inclusion of intra-modal supervision. Our model also surpasses ALIGN, which uses extensive pre-training on 1.8 billion image-text pairs, with a mean of 79.5% vs. 70.9% on COCO and 94.0% vs. 92.2% on Flickr30K, indicating more general and transferable representations. Furthermore, we outperform other baselines on the Flickr30K dataset and surpass ALBEF on the MSCOCO(5K) subset by 2.5% absolute TR/R@1 and 2.2% absolute IR/R@1. Although ALIGN slightly outperforms our model by +0.48% on average across COCO and Flickr30K, our method's performance is expected to improve significantly with larger pre-training datasets.

Table 5.1: Performance comparison of zero-shot and fine-tuned image-text retrieval on Flickr30K and COCO datasets. For text retrieval (TR) and image retrieval (IR), we report the average of R@1, R@5, and R@10.

Method	#Images	Flickr30K (1K test set)						Fine-tuned MSCOCO-5K					
		Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
$\text{UNITER}_{\text{large}}$	4M	87.3	98.0	99.2	75.6	94.1	96.8	65.7	88.6	93.8	52.9	79.9	88.0
METER-Swin	4M	92.4	99.0	99.5	79.0	95.6	98.0	73.0	92.0	96.3	54.9	81.4	89.3
ALBEF	4M	94.3	99.4	99.8	82.8	96.7	98.4	73.1	91.4	96.0	56.8	81.5	89.2
METER-CLIP	4M	94.3	99.6	99.9	82.2	96.3	98.4	76.2	93.2	96.6	57.1	82.7	90.1
$\text{VinVL}_{\text{large}}$	5.6M	-	-	-	-	-	-	75.4	92.9	96.2	58.8	83.7	90.5
ALIGN	1.8B	95.3	99.8	100.0	84.9	97.4	98.6	77.0	93.5	96.6	59.9	83.9	89.7
ALBEF	14M	95.9	99.7	99.9	85.6	97.5	98.9	77.4	94.3	97.2	60.7	84.9	90.5
GraphFusion	4M	95.91	99.8	99.9	85.77	96.84	98.22	78.4	94.72	98.3	62.88	85.1	91.06

5.4.5 VQA, VE, and NLVR^2

Table 5.2 shows the performance comparison on VQA, VE, and NLVR^2, which requires image+text as inputs. In other words, to be successful in these tasks, the model is supposed to have the capability to learn joint multi-modal embeddings. Among five out of six criteria, we deliver competitive results, suggesting that explicitly considering cross-modal alignment and intra-modal supervision contributes positively to feature fusion. Notably, VinVL [228] demonstrates superior performance compared to our method. Its pre-training corpus mainly contains visual QA datasets, including GQA [229], VQA [8], and VG-QAs [230].

Table 5.2: Performance comparison on vision+language tasks.

Method	#Images	VQA		NLVR^2		SNLI-VE	
		test-dev	test-std	dev	test-P	val	test
OSCAR [209]	4M	73.16	73.44	78.07	78.36	74.02	74.02
UNITER [231]	4M	72.70	72.91	77.18	77.85	78.59	78.28
ViLT [232]	4M	71.26	74.02	75.70	76.13	74.02	74.02
UNIMO [233]	4M	73.29	74.02	74.02	74.02	80.0	79.1
VILLA [234]	4M	73.59	73.67	78.39	79.30	79.47	79.03
ALBEF [211]	4M	74.54	74.70	80.24	80.50	80.14	80.51
VinVL [228]	6M	75.95	76.12	82.05	83.08	74.02	74.02
GraphFusion	4M	77.37	76.92	80.88	81.07	80.79	80.66

5.4.6 Ablation Study

The impact of components.

We evaluate the effect of each component in our proposed method by switching off the corresponding module. This ablation analysis involves the following components: *ViG*: intra-modal visual-relation aggregation; *TiG*: intra-modal textual-relation aggregation; *HiG*: cross-modal relation aggregation; *T2V*: message passing from textual to visual modality; *V2T*: message passing from visual to textual modality; *MH*: multi-head attention mechanism; The ablation results tested on the Flickr30k dataset are shown in Table 5.3. Removing visual-relation aggregation significantly degrades image–text retrieval performance, while deleting textual-relation aggregation also hurts retrieval performance. Graph representa-

tions are beneficial for modeling objects and relationships efficiently, promoting image–text retrieval performance. Compared with w/o cross-graph aggregation, the proposed ***GraphFusion*** model with cross-graph relation aggregation has gained 3.7% and 4.5% improvement of R@1 score, respectively, evaluated on the text retrieval and image retrieval tasks. This indicates that integrating cross-graph relations can provide complementary correlation for fine-grained correspondence learning, further strengthening the semantic interaction between different modalities and boosting image–text retrieval performances.

Table 5.3: The ablation study on Flickr30K investigated the effect of different relation-aggregation.

Method	Image to Text			Text To Image		
	R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
w/o ViG	71.05	87.27	90.12	56.56	77.71	82.86
w/o TiG	74.84	88.56	91.46	57.39	78.86	84.08
w/o HiG	76.28	90.26	93.21	58.49	80.37	85.69
w/o V2T	79.26	93.79	96.85	60.78	83.51	89.05
w/o T2V	79.16	93.67	96.73	60.7	83.4	88.93
w/o MH	79.13	93.63	96.69	60.68	83.37	88.9
GraphFusion	**80.21**	**94.88**	**97.95**	**60.49**	**84.54**	**90.11**

The number of neighbors.
In constructing the graph, the number of neighbor nodes K is a hyper-parameter specifying the feature aggregation range. The optimal value for K depends on the task and the average size of the input graphs. However, generally, a very high value of K (too many neighbors) will lead to over-smoothing. On the other hand, very low K (very few neighbors) will cause a deficiency in exchanging information. We fine-tuned the number of neighbors K_v in our visual graph encoder ViG from 3 to 20. Similarly, we finetuned K_c, from 3 to 15, for the cross-modal graph encoder HiG. We observed the optimal number of neighbor nodes between 9 to 15 for Vig, and between 12 and 15 for HiG, respectively. To this end, instead of specifying a fixed number for K, we integrated a range of K, which gradually increases as the layer goes more profound into the network. We present the effect of the number of neighbors in Table 5.4.

Through an ablation study, we demonstrate that each component of ***GraphFusion*** remarkably contributes to an effective aggregation of intra-modal semantic information and, subsequently, more precise cross-modal alignment between image and text data.

5.4.7 Qualitative Results

Cross-attention Visualization.
We visualize the cross-attention maps using Grad-CAM [235] to provide a qualitative assessment of ***GraphFusion***. Figure 5.2 shows that ***GraphFusion*** can associate language with "regions of interest" by attending to meaningful objects and locations, visually reflecting the quality of our model in multi-modal alignment. For instance, the model attends to the

Module	K	Image to Text			Text To Image		
		R@1	R@5	R@10	R@1	R@5	R@10
ViG	3	57.56	82.19	88.3	41.52	69.49	79.04
	6	58.21	83.11	89.29	41.99	70.27	79.92
	9	58.43	83.43	89.64	42.15	70.55	80.24
	12	58.98	84.21	90.47	42.54	71.2	80.98
	15	59.18	84.5	90.79	42.69	71.45	81.27
	20	58.32	83.27	89.46	42.07	70.41	80.08
	9-15	**60.34**	**86.12**	**92.48**	**43.51**	**72.78**	**82.79**
HiG	3	59.31	84.69	90.98	42.79	71.61	81.44
	6	59.56	85.04	91.36	42.96	71.9	81.78
	9	59.52	84.99	91.31	42.94	71.86	81.73
	12	59.48	84.93	91.25	42.91	71.81	81.68
	15	58.68	83.79	90.02	42.33	70.85	80.58
	6-10	**60.34**	**86.12**	**92.48**	**43.51**	**72.78**	**82.79**

Table 5.4: The ablation study on MSCOCO-5K to observe the impact of the different numbers of neighbors (K) in the visual graph (*ViG*) and cross-modal graph (*HiG*).

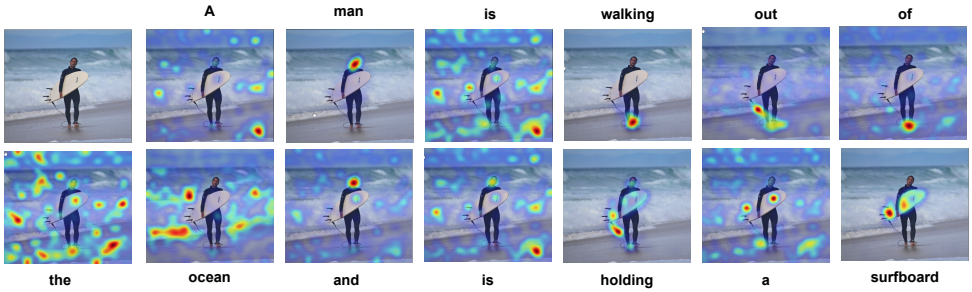


Figure 5.2: Grad-CAM visualization on the cross-attention maps corresponding to individual words.

face of the person when the word "*man*" is given, while for "*walking*" and "*holding*", the model performs surprisingly well, by moving the attention to men's "*feet*" and his "*hands*", respectively. In this experiment, we interestingly see that the model switches its attention from the man's upper body to the conjunction of his feet and the ground when the word changes are "*walking*." It demonstrates the capability of ***GraphFusion*** in understanding the semantic relations between image and text.

Image-text Matching.
To verify the superiority of the proposed model, we further visualize some representative image–text retrieval results on the Flickr30k dataset. Figure 5.3 displays top-5 ranked image-to-text. The proposed ***GraphFusion*** has indexed the semantically relevant textual results for the first image query. Similarly, for the second image query, ***GraphFusion*** can capture the fine-grained correspondence of the objects and their potential relations, such

as "a group of people" and "eating food," and three indexed textual results are semantically relevant to the image query. We can also see that the two textural do not exactly match the image query. However, the main semantic descriptions, such as "many people" and "a crowd of people," match correctly. As for the T2I retrieval, Figure 5.4 shows that the top 3 retrieved images mark the correct results with green boxes. The top-1 image is the ground truth; all other results are close to the sentence's semantics. These results demonstrate that our model can precisely learn the fine-grained semantic correspondence between different modalities, improving retrieval performance.

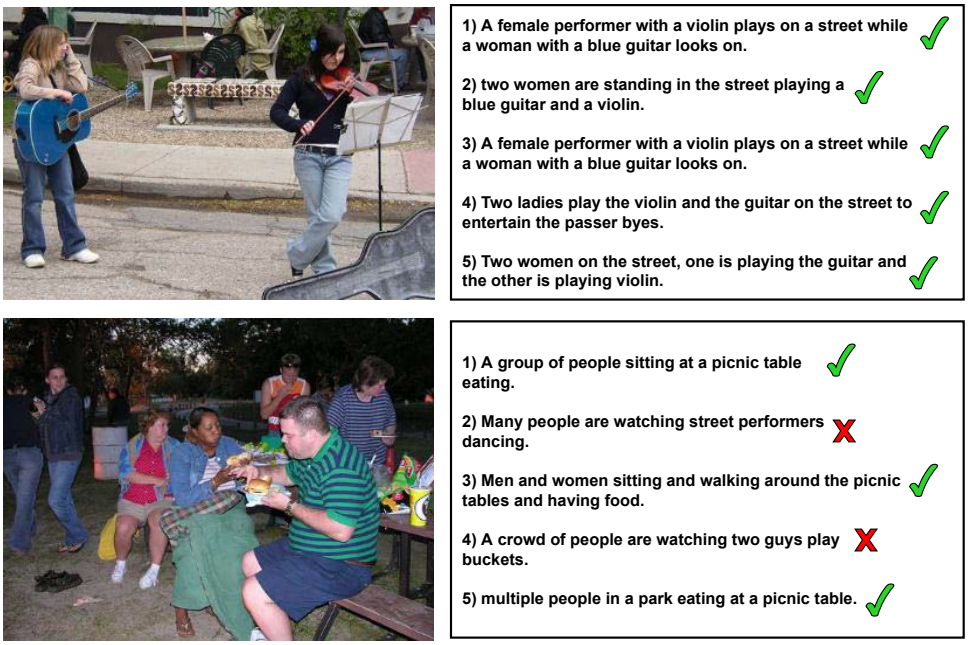


Figure 5.3: Visualization of the image retrieval result. The top 3 images are retrieved for each text. Our approach always retrieves the ground truth in the Top 1 rank.

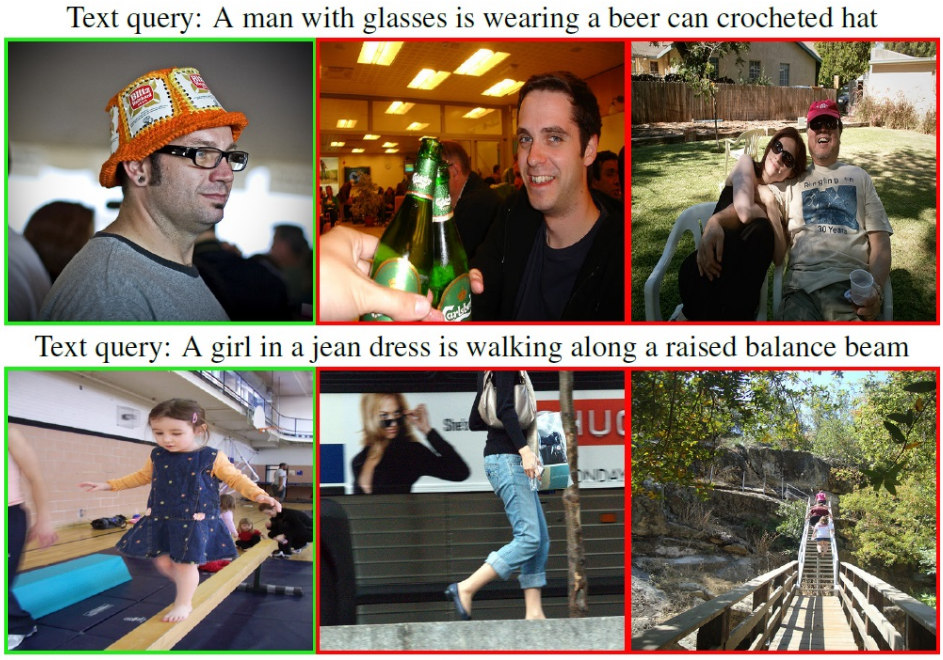


Figure 5.4: Visualization of the image retrieval result. The top 3 images are retrieved for each text. Our approach always retrieves the ground truth in the Top 1 rank.

5.5 Conclusion

In this chapter, we present ***GraphFusion***, a novel methodology for multimodal vision-language understanding, leveraging the capabilities of Heterogeneous Graph Neural Networks (HGNN). By representing visual and linguistic data as nodes in a heterogeneous graph, our proposed model, ***GraphFusion***, can learn joint representations that encapsulate the semantics of both modalities. This graph-based representation allows for efficient information propagation across different types of nodes, enabling the model to capture the intricate interdependencies between visual and textual information. Furthermore, we have proposed a set of techniques for pretraining the ***GraphFusion*** on large-scale image-text datasets, fostering alignment and ensuring that the learned representations are semantically rich and well-aligned across the two modalities.

While our methodology has shown promising results, there are several avenues for future work. One such direction is to explore other graph-based techniques that can further enhance the alignment and representation learning of visual and linguistic modalities. In addition, it would be worthwhile to investigate how to more effectively incorporate attention mechanisms within the heterogeneous graph framework to better focus on relevant parts of the input when generating output. Finally, applying our methodology to a wider range of multimodal tasks, such as visual question answering, image captioning, and visual grounding, could provide additional insights into the generalizability and potential of our approach in addressing the challenges of multimodal vision-language understanding.

In conclusion, the proposed ***GraphFusion*** model offers a promising step forward in addressing the challenges of encapsulating the intricate interdependencies between visual and linguistic modalities and aligning their representations. By explicitly modeling these relationships using Heterogeneous Graph Neural Networks, our methodology has the potential to facilitate advancements in the field of artificial intelligence and contribute to the development of systems capable of advanced multimodal comprehension.

6

Conclusion

Computer vision aims to enable computers to interpret visual content, represented by pixels, at a high level. This interpretation is a crucial step in all types of computer vision problems as it requires the ability to decipher the semantics conveyed through raw pixels. The long-acknowledged gap between low-level features and the semantic meanings of images is known as the semantic gap. Bridging this gap is essential for developing computer vision systems that mirror human perception in understanding visual content.

In this thesis, we explored various challenges and our methods for enhancing visual models with semantic information. Our goal is to improve the reliability of real-world applications in numerous visual understanding tasks. This chapter first summarizes the main contributions and then discusses the limitations of our work. Lastly, we suggest future research directions for multi-modal representation learning in visual understanding.

6.1 Summary of Contributions

This thesis aims to advance visual models by focusing on three key areas to enhance their semantic understanding capabilities. Firstly, we explore the integration of visual and textual inputs within multi-modal learning frameworks to improve performance on tasks involving complex visual-semantic relationships. This approach seeks to refine the models' abilities to interpret intricate visual information by acknowledging the complexity and subtleties of such tasks. Secondly, we investigate how combining human cognitive insights with algorithmic techniques can help identify and characterize shortcomings in visual understanding models, making them more aligned with human perception. Finally, building on insights from our earlier work, we utilize large, unlabeled datasets to pre-train a foundation multi-modal model. We aim to develop a model capable of representing multi-modal data into a unified feature space and capturing long-range dependencies across different modalities, thereby facilitating a wide range of downstream applications that demand advanced comprehension and long-range reasoning capabilities. Each focus area addresses specific research questions discussed in Chapters 2 to 5.

6.1.1 Multi-Modal Data Integration

In Chapter 2, we explore the potential of integrating multi-modal data by formulating the following research question:

> ***RQ1:*** *How can the integration of multi-modal data, specifically text and visual information, improve the semantic and contextual reasoning abilities of models in fine-grained scene recognition?*

To address **RQ1**, we focus on overcoming the limitations inherent in current visual understanding models, particularly in handling visual ambiguities and nuanced differences in appearance. Our solution is a multi-modal late fusion strategy that merges visual and textual inputs, thereby improving scene recognition with a special focus on resolving semantic ambiguities and the challenges in converting textual labels from visuals into digital formats. The cornerstone of our contribution is the Cross-Modal Semantic-Enhanced Visual Understanding Model, designed to enhance scene class categorization in complex environments, surpassing the performance of visual-only models in extensive testing. Additionally, we present the Scene-Text Semantics (ST-Sem) technique to classify Points of Interest (POI) in commercial frontages, utilizing visual labels extracted from street-level images. This technique comprises three modules: scene recognition, scene-text semantic recognition, and a class rank module that merges prediction scores. Our novel approach in data fusion, which integrates visual and textual information after semantic clarification and digitization error correction, provides enhanced flexibility and compatibility across diverse datasets, requiring minimal retraining. Experimental outcomes reveal that our model significantly outperforms both visual-only and other multi-modal methods in POI classification, particularly in urban areas with intricate scenes. The results include both qualitative and quantitative evaluations, confirming the model's efficiency.

Ultimately, our method stands out in its ability to distinguish between similar objects or scenes by incorporating textual data, proving especially useful in scenarios with high visual ambiguity. This research marks a significant contribution to the field of visual understanding and paves the way for future exploration in refining the differentiation of similar entities in real-world settings.

6.1.2 Human-In-the-Loop Approach

In Chapter 3, as our first attempt to leverage human cognitive capabilities to narrow down semantic gaps in the visual models, we formulate the following research question:

> ***RQ2:*** *How can we efficiently utilize human cognitive insights to identify and characterize unknown-unknowns in visual models?*

To address **RQ2**, we introduce the Scalpel-HS framework. This human-in-the-loop semantic analysis framework engages humans in the **Should-Know** and **Really-Knows** tasks to understand what the model should have learned versus what it has learned. We emphasize characterizing the "unknown-unknowns" in image recognition by comparing these two aspects. The framework integrates various components like scene graphs, saliency map extraction, and representation learning. Our experimental results showcase the effectiveness and cost-efficiency of Scalpel-HS, particularly in understanding and mitigating model limitations. This approach highlights the importance of human cognitive abilities in AI, offering significant implications for developing trustworthy AI systems, especially in critical areas like medical imaging and autonomous vehicles.

To further advance our research on bridging the semantic gap using human-in-the-loop techniques, we formulate our third research question as the following:

> ***RQ3:*** *How can we develop a scalable system that combines human and computer capabilities to proactively identify instances that visual models fail to recognize due to insufficient high-level reasoning?*

In Chapter 4, we address **RQ3** by proposing the 'Perspective' annotation tool. This tool aids in identifying and characterizing atypical images through human computation. Our approach involves a two-step image atypicality annotation and sampling process supported by a novel coding scheme based on 10,000 human-annotated images. We tested the identified atypical images against leading image classification services, revealing insights into the alignment between human and machine perception of atypicality. The findings significantly impact the enhancement of image classification systems, suggesting the inclusion of atypical images in training data and the potential for hybrid human-AI systems. Our contributions extend to a large dataset of atypical images with structured characterizations, highlighting the gap between human and machine perception in image classification.

6.1.3 Multi-modal Foundation Model

In Chapter 5, we explore the possible advantages of building a multi-modal foundation model by pre-training on large-scale image-text datasets, such as image-text pairings found on social media platforms. To this end, we formulate our fourth research question posed as:

> ***RQ4:*** *How can we develop and pre-train a foundation model capable of cross-modal comprehension and reasoning by leveraging large-scale multi-modal unlabeled datasets?*

To address **RQ4**, in Chapter 5, we introduce a novel multi-modal graph neural network called GraphFusion, designed for capturing the nuanced dependencies within and across different modalities. Inspired by recent advancements in the knowledge graph domain, GraphFusion first constructs a heterogeneous graph by encoding low-level features of each modality (e.g., superpixels for an image and word tokens for text) as nodes of different types and encoding inductive biases within each modality as edges (e.g., the spatial relationship between pixels and syntactic dependencies between words). This graph-based architecture is crucial in facilitating the flow of information across different modalities (different node types) and within each modality (same node types) through different weighted edges, thereby capturing the nuanced inter-dependencies between visual and textual features. Furthermore, our proposed graph neural network architecture effectively aligns multi-modal features and learns multi-scale joint embeddings by hierarchically aggregating features at every layer. We pretrain GraphFusion on extensive, unlabeled image-text datasets publicly available online. Training our model on such diverse and comprehensive datasets enriches its semantic information, enabling it to generalize across various visual understanding tasks. The practical effectiveness of GraphFusion is further demonstrated through rigorous testing on various downstream tasks, notably in cross-modal retrieval and visual question answering. These experiments show the model's ability to utilize multi-modal data effectively. The success of GraphFusion in these tasks is a testament to its ability to learn unified multi-modal representations, which are essential for tasks requiring a comprehensive understanding of visual and linguistic elements. By introducing GraphFusion, we take a notable step towards narrowing the semantic gap in the visual understating model and moving it closer to practical, real-world applicability. In addition, it opens new avenues for future research in Graph-based Multi-modal Representation Learning.

Our collective efforts to enhance the semantics of visual models provide several methodological and practical insights. We've introduced strategies highlighting the benefits of integrating multi-modal data, incorporating human cognitive abilities, developing a graph-based multi-modal architecture, and utilizing web-scale unlabeled data. These techniques help construct a semantic-aware multi-modal foundation model while ensuring efficiency, reliability, and explainability.

6.2 Limitations and Future Directions

While our research has made notable contributions, some limitations guide future research directions and inform the societal and practical considerations necessary for the responsible

development and application of AI technologies.

6.2.1 Mitigating Bias and Enhancing Dataset Diversity

In machine learning and artificial intelligence, the quality of the data used to train models is paramount to their output. However, biases present in these training datasets can unintentionally shape and influence the resultant model's predictions and behavior. This is a significant concern as it might lead to outcomes that unknowingly reinforce pre-existing societal disparities and biases. Therefore, a keen eye must be kept on the data used to train these models to ensure they represent a fair and unbiased worldview.

Development of Bias Mitigation Algorithms. Investing in the research and development of innovative algorithms is of extreme importance. These algorithms should be designed and tailored to identify potential biases within datasets and then take the necessary steps to mitigate them. Identifying biases could involve using sophisticated techniques capable of detecting skewed data distributions. Once identified, these techniques would work towards automatically balancing them, thus ensuring the integrity of the data. Moreover, there is also a need to develop advanced algorithms that can adjust and fine-tune model training processes. This would aim to minimize and ideally eliminate the influence of biased data. This comprehensive approach to research and development in algorithm design not only enhances the accuracy of data analysis but also upholds the principles of fairness and objectivity.

Creation of More Diverse and Inclusive Datasets. One key objective is to work with an extensive array of stakeholders. These may include industry professionals, academic researchers, community members, and other interested parties. The goal is to gather and annotate datasets that more accurately embody the diversity of real-world scenarios and populations across the globe. This collaborative effort should concentrate on collecting a broad spectrum of visual and textual data, which has been sourced from a diverse range of cultures, languages, and geographical locations. It's not just about the variety of data but also about the quality and relevance of the data to different groups and scenarios. It's crucial to ensure that the models we train are based on data representative of the diversity we see in the real world. By doing so, we can help avoid biases in AI systems and contribute to developing more fair, equitable, and inclusive AI technologies that respect and understand a broad range of human experiences and perspectives.

6.2.2 Addressing Resource and Complexity Issues

Foundation models, such as the GraphFusion, demand significant computational resources. The training phase of these models, in particular, requires considerable processing power and storage capacity. This ongoing need for high computational resources can be a significant hurdle for smaller organizations, startups, or any entity with limited resources. The cost-effectiveness of implementing such models needs to be carefully evaluated against the potential benefits. Furthermore, the complexity of these models adds another challenge in understanding, implementing, and maintaining them, which may deter adopting such

models despite their high performance. Therefore, foundation models like GraphFusion offer impressive results, but their resource requirements and complexity might present considerable obstacles to widespread adoption.

Optimization of Model Architectures. The focus here is on developing more efficient model architectures. The goal is to create designs that require less computational power, thus making them more cost-effective and environmentally friendly while not sacrificing their performance or accuracy. This complex task requires a deep understanding of both the theoretical and practical aspects of machine learning. Techniques such as model pruning, quantization, and knowledge distillation could be explored in this process. Model pruning is a method of reducing the size of a machine-learning model by eliminating unnecessary parts.

In contrast, quantization can reduce the precision of the computations, thus reducing the computational requirements. On the other hand, knowledge distillation is a technique that involves training a smaller model on the output of a larger model, effectively "distilling" the knowledge from the larger model into the smaller one. These techniques could be instrumental in reducing the size and complexity of models, thereby making them more accessible for use in a more comprehensive range of applications and in a broader range of devices, especially those with limited computational resources.

6.2.3 Improving Human-in-the-Loop Scalability

Incorporating human feedback into artificial intelligence systems can result in many issues, particularly scalability and consistency. The inherent variability of human judgment, with its unique subjectivity and often unpredictable nature, can lead to a lack of uniformity in the responses and decisions made by the AI system. Furthermore, significant logistical challenges are involved in integrating human feedback into large-scale systems. These challenges range from the practical aspects of managing and processing vast amounts of data to the technical difficulties of designing and implementing systems that can effectively incorporate and utilize such feedback. As a result, these factors, both individually and collectively, can severely limit the feasibility and effectiveness of methods involving human feedback in AI systems. De Bruijn *et al.* [126] further emphasize the need for socio-technical strategies that blend human expertise with algorithmic processing to ensure trust and efficiency in decision-making. Such approaches must carefully balance the potential benefits of human feedback against challenges of resource constraints and variability in human input.

Hybrid Feedback Mechanisms. Future research must develop hybrid approaches that combine human oversight with automated mechanisms. For instance, synthetic data generation and semi-supervised learning could reduce the burden on human annotators. These mechanisms should aim to emulate human judgment effectively while minimizing the cognitive load on contributors, thus ensuring both scalability and reliability.

Crowdsourcing and Gamification. The fascinating world of crowdsourcing and gamification is a practical strategy for gathering human insights on a grand scale. By actively

engaging a broad and diverse community in the integral processes of data annotation and model evaluation, it becomes feasible to obtain an array of human feedback that is diverse in its range and scalable in its volume. This not only broadens the scope of data that is being evaluated but also enriches the quality of the insights obtained. With the introduction of gamification methods, participants are more likely to stay engaged and motivated, further enhancing the quality and diversity of input. However, to maintain the integrity and reliability of this approach, implementing robust quality control mechanisms becomes imperative. It is equally essential to establish effective incentive structures. These structures would motivate and reward participants, encouraging continued and sustained engagement. In doing so, this ensures the reliability of the crowdsourcing and gamification approach and enhances its overall efficacy and success.

6.2.4 Enhancing Domain Generalization

While our proposed foundation model has shown considerable performance in various downstream tasks, generalizing across various domains remains complex and challenging. This is due to the inherently unique nature of each domain, which often requires specific adjustments to our models. As such, there is a critical need for future work in this area to focus on increasing the adaptability of these models. This includes not only their ability to handle different types of data but also their capability to adjust to different domain-specific requirements and conditions. By doing so, we can ensure that these models maintain relevance and usefulness in various real-world situations, thereby broadening their applicability and potential impact.

The Development of Domain-Agnostic Models. A significant area for future research is the development of domain-agnostic models. These advanced models are designed to perform optimally across diverse tasks and data types. The primary focus is extensive research on foundational model architectures. This research aims to pinpoint and develop architectures to learn and comprehend universal representations of various information types beyond visual and textual. This approach is anticipated to boost our models' overall efficiency and effectiveness, thus advancing the field.

By pursuing these future directions, the research community can address the current limitations and move towards developing more robust, fair, and universally applicable visual understanding models.

Bibliography

References

[1] Yipin Zhou, Zhaowen Wang, Chris Krafft, Qifeng Chen, Ying Xu, and Vladlen Koltun. Visual to sound: Generating natural sound for videos in the wild. *arXiv preprint arXiv:1812.01767*, 2018.

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. pages 1097–1105, 2012.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. pages 248–255, 2009.

[5] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen AWM van der Laak, Bram van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[6] Mariusz Bojarski, Davide Del Testa, David Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Larry D. Jackel, Mathew Monfort, Urs Müller, Jiakai Zhang, et al. End-to-end deep learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[7] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen van der Laak, Bram van Ginneken, and Clara I Sánchez. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19:221–248, 2017.

[8] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *arXiv preprint arXiv:1612.00837*, 2017.

[9] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020.

[10] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1349–1380, 2000.

[11] Quan Wang, Zhendong Zhang, and Haizhou Li. Visual-textual knowledge graphs: A survey. *arXiv preprint arXiv:2102.02302*, 2021.

[12] Gary Marcus. The limits of deep learning. *New York Times*, 2018.

[13] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[14] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. In *Transactions of the Association for Computational Linguistics*, volume 2, pages 453–465, 2014.

[15] Pau Rodriguez, Guillem Cucurull, and Jordi Gonàlez. Multimodal neural networks: A survey of current methods and performances. In *arXiv preprint arXiv:1901.00613*, 2019.

[16] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A"'bibtex survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.

[17] Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012.

[18] Alexander J Ratner, Henry R Ehrenberg, Zeshan Hussain, Jared Dunnmon, and Christopher Ré. Learning to compose domain-specific transformations for data augmentation. *arXiv preprint arXiv:1709.01643*, 2017.

[19] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wiel Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[20] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. In *IEEE transactions on pattern analysis and machine intelligence*, volume 40, pages 1452–1464. IEEE, 2018.

[21] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495, 2014.

[22] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[23] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Adaptive object recognition using adjacency and zoom prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[24] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A Shamma, Michael S Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.

[25] Carolina Galleguillos and Serge Belongie. Context based object categorization: A critical survey. *Computer vision and image understanding*, 114(6):712–722, 2008.

[26] Ruoyi Wang, Heng Wang, Qian Wang, and Xiaomin Wang. Deep learning for smart city applications based on urban big data: A survey. *IEEE Communications Surveys & Tutorials*, 21(3):2094–2117, 2019.

[27] Yunseok Choi, Sangmin Lee, Seonguk Lee, Jaeyoung Lee, and Sungroh Yoon. Attention-based context-aware reasoning for predictive visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6743–6752, 2019.

[28] Limin Zhang, Yufan Zhou, Jianmin Li, and Qi Tian. Context-aware deep learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[29] Stephan R Richter, Ziyu Hayder, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European Conference on Computer Vision*, pages 102–118. Springer, 2018.

[30] Ankan Bansal, Yining Ma, Deva Ramanan, and Andrea Vedaldi. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5769–5778, 2018.

[31] Steve Branson. From pixels to regions: There's plenty of bottom-up in top-down attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3710–3718, 2014.

[32] Lichao Xu, Wen Li, Zehui Lai, Chengqi Zhang, and Li Yang. Deep learning in remote sensing applications: A meta-analysis and review. In *ISPRS Journal of Photogrammetry and Remote Sensing*, volume 152, pages 166–177. Elsevier, 2018.

[33] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[34] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5099–5110, 2019.

[35] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[36] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. *arXiv preprint arXiv:1505.00468*, 2015.

[37] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan Carlos Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[38] Shahin Sharifi Noorian, Achilleas Psyllidis, and Alessandro Bozzon. St-sem: A multimodal method for points-of-interest classification using street-level imagery. In *International Conference on Web Engineering*, pages 32–46. Springer International Publishing Cham, 2019.

[39] Shahin Sharifi Noorian, Sihang Qiu, Achilleas Psyllidis, Alessandro Bozzon, and Geert-Jan Houben. Detecting, classifying, and mapping retail storefronts using street-level imagery. In *Proceedings of the 2020 international conference on multimedia retrieval*, pages 495–501, 2020.

[40] Shahin Sharifi Noorian, Sihang Qiu, Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. What should you know? a human-in-the-loop approach to unknown unknowns characterization in image recognition. In *Proceedings of the ACM Web Conference 2022*, pages 882–892, 2022.

[41] Shahin Sharifi Noorian, Sihang Qiu, Burcu Sayin, Agathe Balayn, Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. Perspective: Leveraging human understanding for identifying and characterizing image atypicality. In *IUI '23: Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages Pages–650, 2023.

[42] Alessandro Bozzon Shahin Sharifi Noorian, Achilleas Psyllidis. A time-varying p-median model for location-allocation analysis. In *21st Conference on Geo-information Science (AGILE 2018)*, 2018.

[43] Vasileios Milias, Shahin Sharifi Noorian, Alessandro Bozzon, and Achilleas Psyllidis. Is it safe to be attractive? disentangling the influence of streetscape features on the perceived safety and attractiveness of city streets. volume 4, page 8. Copernicus Publications Göttingen, Germany, 2023.

[44] HG Parsa, Jean-Pierre I van der Rest, Scott R Smith, Rahul A Parsa, and Milos Bujisic. Why restaurants fail? part iv: The relationship between restaurant failures and demographic factors. *Cornell Hospitality Quarterly*, 56(1):80–90, 2015.

[45] Rahul Goel, Leandro MT Garcia, Anna Goodman, Rob Johnson, Rachel Aldred, Manoradhan Murugesan, Soren Brage, Kavi Bhalla, and James Woodcock. Estimating city-level travel patterns using street imagery: A case study of using google street view in britain. *PloS one*, 13(5):e0196521, 2018.

[46] Yi Zhu, Xueqing Deng, and Shawn Newsam. Fine-grained land use classification at the city scale using ground-level images. *IEEE Transactions on Multimedia*, 2019.

[47] Pablo F Alcantarilla, Simon Stent, German Ros, Roberto Arroyo, and Riccardo Gherardi. Street-view change detection with deconvolutional networks. *Autonomous Robots*, 42(7):1301–1322, 2018.

[48] Sihang Qiu, Achilleas Psyllidis, Alessandro Bozzon, and Geert-Jan Houben. Crowd-mapping urban objects from street-level imagery. In *The World Wide Web Conference*, pages 1521–1531. ACM, 2019.

[49] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[50] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[51] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4990–4999, 2017.

[52] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[53] Sezer Karaoglu, Ran Tao, Theo Gevers, and Arnold WM Smeulders. Words matter: Scene text for image classification and retrieval. *IEEE Transactions on Multimedia*, 19(5):1063–1076, 2017.

[54] Kaiqun Fu, Zhiqian Chen, and Chang-Tien Lu. Streetnet: preference learning with convolutional neural network on urban crime perception. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 269–278. ACM, 2018.

[55] Xiaojiang Li, Carlo Ratti, and Ian Seiferling. Mapping urban landscapes along streets using google street view. In *International Cartographic Conference*, pages 341–356. Springer, 2017.

[56] Xiaojiang Li and Carlo Ratti. Mapping the spatial distribution of shade provision of street trees in boston using google street view panoramas. *Urban Forestry & Urban Greening*, 31:109–119, 2018.

[57] Xiaojiang Li, Chuanrong Zhang, Weidong Li, Robert Ricard, Qingyan Meng, and Weixing Zhang. Assessing street-level urban greenery using google street view and a modified green view index. *Urban Forestry & Urban Greening*, 14(3):675–685, 2015.

[58] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. Using deep learning and google street view to estimate the demographic makeup of the us. *arXiv preprint arXiv:1702.06683*, 2017.

[59] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei Efros. What makes paris look like paris? *ACM Transactions on Graphics*, 31(4), 2012.

[60] Mehdi Talebi, Abbas Vafaei, and Amirhassan Monadjemi. Vision-based entrance detection in outdoor scenes. *Multimedia Tools and Applications*, 77(20):26219–26238, 2018.

[61] Vahid Balali, Armin Ashouri Rad, and Mani Golparvar-Fard. Detection, classification, and mapping of us traffic signs using google street view images for roadway inventory management. *Visualization in Engineering*, 3(1):15, 2015.

[62] Qian Yu, Christian Szegedy, Martin C Stumpe, Liron Yatziv, Vinay Shet, Julian Ibarz, and Sacha Arnoud. Large scale business discovery from street level imagery. *arXiv preprint arXiv:1512.05430*, 2015.

[63] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

[64] Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Rui Zhu. xnet+ sc: Classifying places based on images by incorporating spatial contexts. In *10th International Conference on Geographic Information Science (GIScience 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[65] Yair Movshovitz-Attias, Qian Yu, Martin C Stumpe, Vinay Shet, Sacha Arnoud, and Liron Yatziv. Ontological supervision for fine grained classification of street view storefronts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1693–1702, 2015.

[66] Sezer Karaoglu, Ran Tao, Jan C van Gemert, and Theo Gevers. Con-text: Text detection for fine-grained object classification. *IEEE Transactions on Image Processing*, 26(8):3965–3980, 2017.

[67] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[68] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[69] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[70] Jonti Talukdar, Sanchit Gupta, PS Rajpura, and Ravi S Hegde. Transfer learning for object detection using state-of-the-art deep neural networks. In *2018 5th International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 78–83. IEEE, 2018.

[71] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018.

[72] Spiros V Georgakopoulos and Vassilis P Plagianakos. A novel adaptive learning rate algorithm for convolutional neural network training. In *International Conference on Engineering Applications of Neural Networks*, pages 327–336. Springer, 2017.

[73] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[74] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pages 5561–5569, 2017.

[75] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[76] Minghui Liao, Baoguang Shi, and Xiang Bai. Textboxes++: A single-shot oriented scene text detector. *IEEE Transactions on Image Processing*, 27(8):3676–3690, 2018.

[77] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 1156–1160. IEEE, 2015.

[78] Canjie Luo, Lianwen Jin, and Zenghui Sun. Moran: A multi-object rectified attention network for scene text recognition. *Pattern Recognition*, 2019.

[79] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 569–576, 2013.

[80] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, 116(1):1–20, 2016.

[81] Christoph Lofi. Measuring semantic similarity and relatedness with distributional and knowledge-based approaches. *Information and Media Technologies*, 10(3):493–501, 2015.

[82] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[83] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[84] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[85] Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*, 2017.

[86] Alex Rodriguez and Alessandro Laio. Machine learning. Clustering by fast search and find of density peaks. *Science (New York, N.Y.)*, 344(6191):1492–6, jun 2014.

[87] Joseph Redmon. Darknet: Open source neural networks in c. http://pjreddie.com/darknet/, 2013–2016.

[88] Licheng Jiao, Fan Zhang, Fang Liu, Shuyuan Yang, Lingling Li, Zhixi Feng, and Rong Qu. A survey of deep learning-based object detection. *IEEE Access*, 7:128837–128868, 2019.

[89] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, et al. Machine behaviour. *Nature*, 568(7753):477–486, 2019.

[90] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Kumar Paritosh, and Lora Mois Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. 2021.

[91] Saleema Amershi, Max Chickering, Steven M Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. Modeltracker: Redesigning performance analysis tools for machine learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 337–346, 2015.

[92] Mengnan Du, Ninghao Liu, and Xia Hu. Techniques for interpretable machine learning. *Communications of the ACM*, 63(1):68–77, 2019.

[93] Xin Wang, Wenhu Huang, Fuli Wu, Peng Qi, Alan L Huang, and Fuxin Jiang. Learning semantic concepts and order for image and sentence matching. *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[94] Alexander Erlei, Franck Nekdem, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. Impact of algorithmic decision making on human behavior: Evidence from ultimatum bargaining. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 43–52, 2020.

[95] Josh M Attenberg, Pagagiotis G Ipeirotis, and Foster Provost. Beat the machine: Challenging workers to find the unknown unknowns. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[96] Anthony Liu, Santiago Guerra, Isaac Fung, Gabriel Matute, Ece Kamar, and Walter Lasecki. Towards hybrid human-ai workflows for unknown unknown detection. In *Proceedings of The Web Conference 2020*, pages 2432–2442, 2020.

[97] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. Identifying unknown unknowns in the open world: representations and policies for guided exploration. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 2124–2132, 2017.

[98] Ujwal Gadiraju and Jie Yang. What can crowd computing do for the next generation of ai systems? In *2020 Crowd Science Workshop: Remoteness, Fairness, and Mechanisms as Challenges of Data Supply by Humans for Automation*, pages 7–13, 2020.

[99] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv:1706.03825*, 2017.

[100] Agathe Balayn, Panagiotis Soilis, Christoph Lofi, Jie Yang, and Alessandro Bozzon. What do you mean? interpreting image classification with crowdsourced concept extraction and analysis. In *Proceedings of the Web Conference 2021*, pages 1937–1948, 2021.

[101] Chris Fields. How humans recognize objects: Segmentation, categorization and individual identification. *Frontiers in psychology*, 7:400, 2016.

[102] Eric Margolis and Stephen Laurence. The ontology of concepts-abstract objects or mental representations? *Noûs*, 41(4):561–593, 2007.

[103] Bing Ran and P Robert Duimering. Conceptual combination: Models, theories and controversies. *International Journal of Cognitive Linguistics*, 1(1):65–90, 2010.

[104] Martin N Hebart, Charles Y Zheng, Francisco Pereira, and Chris I Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, 4(11), 2020.

[105] David R Krathwohl. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218, 2002.

[106] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[107] Zijian Zhang, Jaspreet Singh, Ujwal Gadiraju, and Avishek Anand. Dissonance between human and machine understanding. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

[108] Sina Mohseni, Jeremy E Block, and Eric D Ragan. Quantitative evaluation of machine learning explanations: A human-grounded benchmark. 2021.

[109] Burr Settles. Active learning. *Synthesis lectures on artificial intelligence and machine learning*, 6(1):1–114, 2012.

[110] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 287–294, 1992.

[111] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, 1994.

[112] Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. In *ICML*, pages 894–905, 2001.

[113] Yoram Baram, Ran El-Yaniv, and Kobi Luz. Online choice of active learning algorithms. *Journal of Machine Learning Research*, 5:255–291, December 2004.

[114] Wei-Ning Hsu and Hsuan-Tien Lin. Active learning by learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[115] Hong-Min Chu and Hsuan-Tien Lin. Can active learning experience be transferred? *IEEE 16th International Conference on Data Mining*, pages 841–846, 2016.

[116] Colin Vandenhof. A hybrid approach to identifying unknown unknowns of predictive models. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 180–187, 2019.

[117] Deepak Agarwal. Detecting anomalies in cross-classified streams: a bayesian approach. *Knowledge and information systems*, 11(1):29–44, 2007.

[118] Bovas Abraham and George EP Box. Bayesian analysis of some outlier problems in time series. *Biometrika*, 66(2):229–236, 1979.

[119] Eleazar Eskin. Anomaly detection over noisy data using learned probability distributions. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 255–262, 2000.

[120] Tom Fawcett and Foster Provost. Adaptive fraud detection. *Data mining and knowledge discovery*, 1(3):291–316, 1997.

[121] Eleazar Eskin, Andrew Arnold, Michael Prerau, Leonid Portnoy, and Sal Stolfo. A geometric framework for unsupervised anomaly detection. In *Applications of data mining in computer security*, pages 77–101. Springer, 2002.

[122] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(4), 2010.

[123] Besmira Nushi, Ece Kamar, Eric Horvitz, and Donald Kossmann. On human intellect and machine failures: Troubleshooting integrative machine learning systems. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[124] Jie Yang, Alisa Smirnova, Dingqi Yang, Gianluca Demartini, Yuan Lu, and Philippe Cudré-Mauroux. Scalpel-cd: leveraging crowdsourcing and deep probabilistic modeling for debugging noisy training data. In *WWW*, pages 2158–2168, 2019.

[125] Xiao Hu, Haobo Wang, Anirudh Vegesana, Somesh Dube, Kaiwen Yu, Gore Kao, Shuo-Han Chen, Yung-Hsiang Lu, George K Thiruvathukal, and Ming Yin. Crowdsourcing detection of sampling biases in image datasets. In *Proceedings of The Web Conference 2020*, pages 2955–2961, 2020.

[126] Hans de Bruijn, Martijn Warnier, and Marijn Janssen. The perils and pitfalls of explainable ai: Strategies for explaining algorithmic decision-making. *Government information quarterly*, 39(2):101666, 2022.

[127] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Conference on Computer Vision and Pattern Recognition*, 2018.

[128] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[129] Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019.

[130] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Intl. Conf. on Learning Representations*, 2017.

[131] Nicolò Navarin, Dinh Van Tran, and Alessandro Sperduti. Universal readout for graph convolutional neural networks. In *2019 International Joint Conference on Neural Networks*, pages 1–7, 2019.

[132] Rex Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. *arXiv preprint arXiv:1806.08804*, 2018.

[133] Herbert A. Simon. *Models of thought*, volume 352. Yale university press, 1979.

[134] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[135] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[136] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Outlier detection: A survey. *ACM Computing Surveys*, 14:15, 2007.

[137] Fan-Yun Sun, Jordan Hoffman, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *International Conference on Learning Representations*, 2019.

[138] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8732–8740, 2020.

[139] Miltiadis Allamanis, Pankajan Chanthirasegaran, Pushmeet Kohli, and Charles Sutton. Learning continuous semantic representations of symbolic expressions. In *International Conference on Machine Learning*, pages 80–88. PMLR, 2017.

[140] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*, 2018.

[141] Michael A Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4845–4854, 2019.

[142] Andrew NG. Mlops: From model-centric to data-centric ai. https://www.deeplearning.ai/wp-content/uploads/2021/06/MLOps-From-Model-centric-to-Data-centric-AI.pdf, 2021. Deeplearning.ai [Online; posted: June-2021].

[143] Robert Geirhos, Carlos R Medina Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *arXiv preprint arXiv:1808.08750*, 2018.

[144] Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9617–9626, 2019.

[145] Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning*, pages 8634–8644. PMLR, 2020.

[146] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987.

[147] Burcu Sayin, Jie Yang, Andrea Passerini, and Fabio Casati. The science of rejection: a research area for human computation. *arXiv preprint arXiv:2111.06736*, 2021.

[148] Burcu Sayin, Fabio Casati, Andrea Passerini, Jie Yang, and Xinyue Chen. Rethinking and recomputing the value of ml models. *arXiv preprint arXiv:2209.15157*, 2022.

[149] Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 498–512, 2018.

[150] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017.

[151] Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

[152] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 547–558, 2020.

[153] Thomas G Dietterich. Steps toward robust artificial intelligence. *AI Magazine*, 38(3):3–24, 2017.

[154] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6. IEEE, 2009.

[155] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*, 2016.

[156] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[157] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, pages 4694–4703, 2019.

[158] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

[159] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.

[160] Weikai Yang, Zhen Li, Mengchen Liu, Yafeng Lu, Kelei Cao, Ross Maciejewski, and Shixia Liu. Diagnosing concept drift with visual analytics. In *2020 IEEE conference on visual analytics science and technology (VAST)*, pages 12–23. IEEE, 2020.

[161] Xumeng Wang, Wei Chen, Jiazhi Xia, Zexian Chen, Dongshi Xu, Xiangyang Wu, Mingliang Xu, and Tobias Schreck. Conceptexplorer: Visual analysis of concept drifts in multi-source time-series data. In *2020 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 1–11. IEEE, 2020.

[162] Anton Yeshchenko, Claudio Di Ciccio, Jan Mendling, and Artem Polyvyanyy. Visual drift detection for event sequence data of business processes. *IEEE Transactions on Visualization and Computer Graphics*, 28(8):3050–3068, 2021.

[163] Changjian Chen, Jun Yuan, Yafeng Lu, Yang Liu, Hang Su, Songtao Yuan, and Shixia Liu. Oodanalyzer: Interactive analysis of out-of-distribution samples. *IEEE Transactions on Visualization and Computer Graphics*, 27(7):3335–3349, 2021.

[164] Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. pages 1–14, 2020.

[165] Robert Munro. *Human-in-the-Loop Machine Learning*. Manning Publications, 2021.

[166] David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4(1):129–145, March 1996.

[167] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, 1994.

[168] Barney G Glaser and Anselm L Strauss. *Discovery of grounded theory: Strategies for qualitative research*. Routledge, 2017.

[169] Virginia Braun and Victoria Clarke. *Thematic analysis*. American Psychological Association, 2012.

[170] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

[171] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[172] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269. IEEE Computer Society, 2017.

[173] Joses Ho, Tayfun Tumkaya, Sameer Aryal, Hyungwon Choi, and Adam Claridge-Chang. Moving beyond p values: data analysis with estimation graphics. *Nature methods*, 16(7):565–566, 2019.

[174] Thomas Binder, Giorgio De Michelis, Pelle Ehn, Giulio Jacucci, and Per Linde. *Design things*. MIT press, 2011.

[175] Ben Shneiderman. Creativity support tools. *Communications of the ACM*, 45(10):116–120, 2002.

[176] Neoklis Polyzotis and Matei Zaharia. What can data-centric ai learn from data and ml engineering? *arXiv preprint arXiv:2112.06439*, 2021.

[177] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. Data augmentation can improve robustness. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[178] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[179] Sanglee Park and Jungmin So. On the effectiveness of adversarial training in defending against adversarial example attacks for image classification. *Applied Sciences*, 10(22):8079, 2020.

[180] Lin Li and Michael Spratling. Data augmentation alone can improve adversarial training. *arXiv preprint arXiv:2301.09879*, 2023.

[181] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020.

[182] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Neural Information Processing Systems*, 32:13–23, 2019.

[183] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13041–13049, 2020.

[184] Peter Anderson, Qi Wu, Damien Teney, Jacob Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2020.

[185] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10267–10276, 2020.

[186] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. pages 6904–6913, 2017.

[187] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. pages 121–137, 2020.

[188] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[189] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[190] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[191] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021.

[192] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 2021.

[193] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. *International Conference on Computer Vision*, 2021.

[194] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021.

[195] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. *International Conference on Machine Learning*, 2021.

[196] Zirui Wang, Jiahui Liu, and Jianfeng Gao. Simvlm: Simple visual language model pre-training with weak supervision. *International Conference on Machine Learning*, 2022.

[197] Yuhao Jiang, Richard Zhang, and Lei Wang. Translating math formula images to latex sequences using deep neural networks with sequence-to-sequence model. *Artificial Intelligence and Statistics*, 2021.

[198] Shuiwang Ji and Le Wang. Graph-based neural networks for natural language processing and computer vision: A review. *IEEE Access*, 9:123456–123467, 2021.

[199] Fei Wang and Jian Sun. Unified graph-based multi-modal learning. *Artificial Intelligence Review*, 55:785–795, 2022.

[200] Ming Zhou and Yue Zhang. Datasets for multi-modal learning in social media contexts. *Journal of Machine Learning Research*, 22:2021–2045, 2021.

[201] Youngjin Kim and Hyun Lee. Self-supervised learning for multi-modal integration in retrieval systems. *Neural Processing Letters*, 53:3317–3330, 2021.

[202] Xiaojun Chen and Yuheng Wang. Paired multi-modal data for machine learning. *Data Science and Engineering*, 6:456–470, 2021.

[203] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Aditya Goyal, Nick Kanwisher, Benjamin Recht, and F. N. Iandola. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[204] Chang Liu and Mei Zhao. Applications of multi-modal learning in automated content moderation. In *Proceedings of the International Conference on Learning Representations*, 2021.

[205] Jordan A. Smith and Pei Zhang. Benchmarking multi-modal models in machine learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:337–352, 2022.

[206] Jing Wang and Bin Liu. Adaptability of graph-based models for multi-modal learning. *Pattern Recognition*, 120:108911, 2022.

[207] Li Zhang and Yu Wei. The future of ai: Graph-based models for advanced representation learning. *Journal of Artificial Intelligence Research*, 71:999–1024, 2022.

[208] Chao Jia et al. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916, 2021.

[209] Xiujun Li et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. *European Conference on Computer Vision*, pages 121–137, 2020.

[210] Yen-Chun Chen et al. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120, 2019.

[211] Junnan Li et al. Align before fuse: Vision and language representation learning with momentum distillation. *arXiv preprint arXiv:2107.07651*, 2021.

[212] Feng Yang et al. Token contrastive learning of image-text joint embeddings. *arXiv preprint arXiv:2111.01279*, 2021.

[213] Amanpreet Singh et al. Flava: A foundational language and vision alignment model. *arXiv preprint arXiv:2112.04482*, 2021.

[214] Lu Yuan et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[215] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2803–2812, 2021.

[216] Michael Schlichtkrull et al. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607, 2018.

[217] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. Heterogeneous graph attention network. *The World Wide Web Conference*, pages 2022–2032, 2019.

[218] Luowei Zhou et al. A unified framework for cross-modal information retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 125–134, 2020.

[219] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[220] Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR 2019*. ICLR, April 2019.

[221] Yen-Chun Li, Yao-Hung Hubert Tsai, Jianwei Wu, et al. Uniter: Universal image-text representation learning. In *ECCV*, 2020.

[222] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[223] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

[224] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011.

[225] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.

[226] Ning Xie, Farley Lai, Hal Daume III, et al. Visual entailment task for visually-grounded language learning. In *NAACL*, 2019.

[227] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. Evaluating visual reasoning through grounded language understanding. *AI Magazine*, 39(2):45–52, 2018.

[228] Pengchuan Zhang, Xiujun Li, Romer Thoudam, Lijuan Wu, Jianwei Zhang, Lei Qi, Ying Chen, Yizhe Yang, Jianfeng Tang, Mohan Zhou, et al. Vinvl: Making visual representations matter in vision-language models. *arXiv preprint arXiv:2101.00529*, 2021.

[229] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

[230] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, 2016.

[231] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[232] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021.

[233] Wei Li, Can Gao, Guocheng Niu, et al. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *ACL*, 2020.

[234] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, 2020.

[235] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[236] Shahin Sharifi Noorian, Jie Yang, and Alessandro Bozzon. Graphfusion: Unified vision-language representation learning using heterogeneous graph neural networks. In *The 2025 International Conference on Learning Representations, ICLR25 (under review)*, 2024.

[237] Shahin Sharifi Noorian and Christian E Murphy. Balanced allocation of multi-criteria geographic areas by a genetic algorithm. In *Advances in Cartography and GIScience: Selections from the International Cartographic Conference 2017 28*, pages 417–433. Springer International Publishing, 2017.

Curriculum Vitæ

Shahin Sharifi

16-07-1988 Born in Karaj, Iran

Professional Experience

2024–present	Data Science Practice Lead, Infosys, The Netherlands
2024–2025	Postdoctoral Researcher, University of British Columbia, Canada
2017–2024	PhD Candidate, Delft University of Technology, The Netherlands
2015–2024	AI Solution Architect, WIGeoGIS GmbH, Munich, Germany
2013–2015	Data Science Researcher, WIGeoGIS GmbH, Munich, Germany
2012–2015	Software Engineer, Technical University of Munich, Germany

Education

2017–2024	Doctor of Philosophy (PhD), Computer Science Delft University of Technology, The Netherlands
2012–2015	Master of Science (MS), Geo-Informatics Technical University of Munich, Germany
2006–2010	Bachelor of Science (BS), Computer Science University of Science and Culture, Tehran, Iran

Project Highlights

2022–2024	Spearheaded the creation and deployment of a Multimodal Neural Search and Recommender System that leverages Large Language Models and Graph Neural Networks to efficiently search for points of interest across a comprehensive European dataset.
2021–2022	Led the design and development of The Scalpel-HS framework, an innovative approach for detecting and characterizing "unknown unknowns" in image recognition, significantly improving their identification and characterization compared to current methods.
2018–2021	Led the successful implementation of a real-time Computer Vision system that automates the extraction and geolocation of commercial points of interest from street-level imagery, improving the accuracy and speed of geospatial data processing.
2017–2019	Pioneered the design and development of multi-objective optimization algorithms currently employed by several enterprise organizations to tackle challenges like Facility Siting, Sales Territory Planning, and Delivery Time Optimization through complex algorithms for practical, large-scale problem-solving.

List of Publications

List of Publications

8. Shahin Sharifi Noorian, Jie Yang, and Alessandro Bozzon. Graphfusion: Unified vision-language representation learning using heterogeneous graph neural networks. In *The 2025 International Conference on Learning Representations, ICLR25 (under review)*, 2024.

7. Shahin Sharifi Noorian, Sihang Qiu, Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. What should you know? a human-in-the-loop approach to unknown unknowns characterization in image recognition. In *Proceedings of the ACM Web Conference 2022*, pages 882–892, 2022.

6. Shahin Sharifi Noorian, Sihang Qiu, Burcu Sayin, Agathe Balayn, Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. Perspective: Leveraging human understanding for identifying and characterizing image atypicality. In *IUI '23: Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages Pages–650, 2023.

5. Shahin Sharifi Noorian, Sihang Qiu, Achilleas Psyllidis, Alessandro Bozzon, and Geert-Jan Houben. Detecting, classifying, and mapping retail storefronts using street-level imagery. In *Proceedings of the 2020 international conference on multimedia retrieval*, pages 495–501, 2020.

4. Shahin Sharifi Noorian, Achilleas Psyllidis, and Alessandro Bozzon. St-sem: A multimodal method for points-of-interest classification using street-level imagery. In *International Conference on Web Engineering*, pages 32–46. Springer International Publishing Cham, 2019.

3. Alessandro Bozzon Shahin Sharifi Noorian, Achilleas Psyllidis. A time-varying p-median model for location-allocation analysis. In *21st Conference on Geo-information Science (AGILE 2018)*, 2018.

2. Vasileios Milias, Shahin Sharifi Noorian, Alessandro Bozzon, and Achilleas Psyllidis. Is it safe to be attractive? disentangling the influence of streetscape features on the perceived safety and attractiveness of city streets. volume 4, page 8. Copernicus Publications Göttingen, Germany, 2023.

1. Shahin Sharifi Noorian and Christian E Murphy. Balanced allocation of multi-criteria geographic areas by a genetic algorithm. In *Advances in Cartography and GIScience: Selections from the International Cartographic Conference 2017 28*, pages 417–433. Springer International Publishing, 2017.

SIKS Dissertation Series

Since 1998, all dissertations written by PhD. students who have conducted their research under auspices of a senior research fellow of the SIKS research school are published in the SIKS Dissertation Series (following are all the dissertations since 2016).

2016
01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
04 Laurens Rietveld (VUA), Publishing and Consuming Linked Data
05 Evgeny Sherkhonov (UvA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
07 Jeroen de Man (VUA), Measuring and modeling negative emotions for virtual training
08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
09 Archana Nottamkandath (VUA), Trusting Crowdsourced Information on Cultural Artefacts
10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
11 Anne Schuth (UvA), Search Engines that Learn from Their Users
12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
13 Nana Baah Gyan (VUA), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
16 Guangliang Li (UvA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17 Berend Weel (VUA), Towards Embodied Evolution of Robot Organisms
18 Albert Meroño Peñuela (VUA), Refining Statistical Data on the Web
19 Julia Efremova (TU/e), Mining Social Structures from Genealogical Data
20 Daan Odijk (UvA), Context & Semantics in News & Web Search
21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22 Grace Lewis (VUA), Software Architecture Strategies for Cyber-Foraging Systems
23 Fei Cai (UvA), Query Auto Completion in Information Retrieval

24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26 Dilhan Thilakarathne (VUA), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30 Ruud Mattheij (TiU), The Eyes Have It
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UvA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TU/e), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UvA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40 Christian Detweiler (TUD), Accounting for Values in Design
41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
42 Spyros Martzoukos (UvA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44 Thibault Sellam (UvA), Automatic Assistants for Database Exploration
45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46 Jorge Gallego Perez (UT), Robots to Make you Happy
47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48 Tanja Buttler (TUD), Collecting Lessons Learned
49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis

	50	Yan Wang (TiU), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains
2017	01	Jan-Jaap Oerlemans (UL), Investigating Cybercrime
	02	Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
	03	Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
	04	Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
	05	Mahdieh Shadi (UvA), Collaboration Behavior
	06	Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
	07	Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
	08	Rob Konijn (VUA), Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
	09	Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
	10	Robby van Delden (UT), (Steering) Interactive Play Behavior
	11	Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
	12	Sander Leemans (TU/e), Robust Process Mining with Guarantees
	13	Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
	14	Shoshannah Tekofsky (TiU), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
	15	Peter Berck (RUN), Memory-Based Text Correction
	16	Aleksandr Chuklin (UvA), Understanding and Modeling Users of Modern Search Engines
	17	Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
	18	Ridho Reinanda (UvA), Entity Associations for Search
	19	Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
	20	Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
	21	Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
	22	Sara Magliacane (VUA), Logics for causal inference under uncertainty
	23	David Graus (UvA), Entities of Interest — Discovery in Digital Traces
	24	Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
	25	Veruska Zamborlini (VUA), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
	26	Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
	27	Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
	28	John Klein (VUA), Architecture Practices for Complex Contexts

	29	Adel Alhuraibi (TiU), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"
	30	Wilma Latuny (TiU), The Power of Facial Expressions
	31	Ben Ruijl (UL), Advances in computational methods for QFT calculations
	32	Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
	33	Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
	34	Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
	35	Martine de Vos (VUA), Interpreting natural science spreadsheets
	36	Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
	37	Alejandro Montes Garcia (TU/e), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
	38	Alex Kayal (TUD), Normative Social Applications
	39	Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
	40	Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
	41	Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
	42	Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
	43	Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
	44	Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
	45	Bas Testerink (UU), Decentralized Runtime Norm Enforcement
	46	Jan Schneider (OU), Sensor-based Learning Support
	47	Jie Yang (TUD), Crowd Knowledge Creation Acceleration
	48	Angel Suarez (OU), Collaborative inquiry-based learning
2018	01	Han van der Aa (VUA), Comparing and Aligning Process Representations
	02	Felix Mannhardt (TU/e), Multi-perspective Process Mining
	03	Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
	04	Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
	05	Hugo Huurdeman (UvA), Supporting the Complex Dynamics of the Information Seeking Process
	06	Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
	07	Jieting Luo (UU), A formal account of opportunism in multi-agent systems
	08	Rick Smetsers (RUN), Advances in Model Learning for Software Systems
	09	Xu Xie (TUD), Data Assimilation in Discrete Event Simulations

	10	Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
	11	Mahdi Sargolzaei (UvA), Enabling Framework for Service-oriented Collaborative Networks
	12	Xixi Lu (TU/e), Using behavioral context in process mining
	13	Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
	14	Bart Joosten (TiU), Detecting Social Signals with Spatiotemporal Gabor Filters
	15	Naser Davarzani (UM), Biomarker discovery in heart failure
	16	Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
	17	Jianpeng Zhang (TU/e), On Graph Sample Clustering
	18	Henriette Nakad (UL), De Notaris en Private Rechtspraak
	19	Minh Duc Pham (VUA), Emergent relational schemas for RDF
	20	Manxia Liu (RUN), Time and Bayesian Networks
	21	Aad Slootmaker (OU), EMERGO: a generic platform for authoring and playing scenario-based serious games
	22	Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
	23	Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
	24	Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
	25	Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
	26	Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
	27	Maikel Leemans (TU/e), Hierarchical Process Mining for Scalable Software Analysis
	28	Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
	29	Yu Gu (TiU), Emotion Recognition from Mandarin Speech
	30	Wouter Beek (VUA), The "K" in "semantic web" stands for "knowledge": scaling semantics to the web
2019	01	Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
	02	Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
	03	Eduardo Gonzalez Lopez de Murillas (TU/e), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
	04	Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
	05	Sebastiaan van Zelst (TU/e), Process Mining with Streaming Data
	06	Chris Dijkshoorn (VUA), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
	07	Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
	08	Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes

09 Fahimeh Alizadeh Moghaddam (UvA), Self-adaptation for energy efficiency in software systems
10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
12 Jacqueline Heinerman (VUA), Better Together
13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
16 Guangming Li (TU/e), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts
18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication
19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents
20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
21 Cong Liu (TU/e), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
22 Martin van den Berg (VUA),Improving IT Decisions with Enterprise Architecture
23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
24 Anca Dumitrache (VUA), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
25 Emiel van Miltenburg (VUA), Pragmatic factors in (automatic) image description
26 Prince Singh (UT), An Integration Platform for Synchromodal Transport
27 Alessandra Antonaci (OU), The Gamification Design Process applied to (Massive) Open Online Courses
28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
29 Daniel Formolo (VUA), Using virtual agents for simulation and training of social skills in safety-critical circumstances
30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
31 Milan Jelisavcic (VUA), Alive and Kicking: Baby Steps in Robotics
32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
33 Anil Yaman (TU/e), Evolution of Biologically Inspired Learning in Artificial Neural Networks
34 Negar Ahmadi (TU/e), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES
35 Lisa Facey-Shaw (OU), Gamification with digital badges in learning programming
36 Kevin Ackermans (OU), Designing Video-Enhanced Rubrics to Master Complex Skills

	37	Jian Fang (TUD), Database Acceleration on FPGAs
	38	Akos Kadar (OU), Learning visually grounded and multilingual representations
2020	01	Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
	02	Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
	03	Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
	04	Maarten van Gompel (RUN), Context as Linguistic Bridges
	05	Yulong Pei (TU/e), On local and global structure mining
	06	Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
	07	Wim van der Vegt (OU), Towards a software architecture for reusable game components
	08	Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search
	09	Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
	10	Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining
	11	Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models
	12	Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
	13	Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
	14	Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
	15	Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
	16	Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
	17	Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
	18	Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
	19	Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
	20	Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations
	21	Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be
	22	Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
	23	Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
	24	Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
	25	Xin Du (TU/e), The Uncertainty in Exceptional Model Mining
	26	Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization

	27	Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
	28	Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
	29	Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
	30	Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst
	31	Gongjin Lan (VUA), Learning better – From Baby to Better
	32	Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising
	33	Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation
	34	Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development
	35	Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production
2021	01	Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space
	02	Rijk Mercuur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models
	03	Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices
	04	Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
	05	Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
	06	Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
	07	Armel Lefebvre (UU), Research data management for open science
	08	Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
	09	Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play
	10	Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
	11	Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
	12	Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
	13	Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning
	14	Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support
	15	Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm
	16	Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues
	17	Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks
	18	Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks
	19	Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management

	20	Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems
	21	Pedro Thiago Timbó Holanda (CWI), Progressive Indexes
	22	Sihang Qiu (TUD), Conversational Crowdsourcing
	23	Hugo Manuel Proença (UL), Robust rules for prediction and description
	24	Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing
	25	Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
	26	Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables
	27	Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
	28	Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs
2022	01	Judith van Stegeren (UT), Flavor text generation for role-playing video games
	02	Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
	03	Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
	04	Ünal Aksu (UU), A Cross-Organizational Process Mining Framework
	05	Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
	06	Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
	07	Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
	08	Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
	09	Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
	10	Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines
	11	Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring
	12	Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases
	13	Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge
	14	Michiel Overeem (UU), Evolution of Low-Code Platforms
	15	Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining
	16	Pieter Gijsbers (TU/e), Systems for AutoML Research
	17	Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification

	18	Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation
	19	Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation
	20	Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media
	21	Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments
	22	Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations
	23	Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents
	24	Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values
	25	Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty
	26	Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences
	27	Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences
	28	Onuralp Ulusoy (UU), Privacy in Collaborative Systems
	29	Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality
	30	Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays
	31	Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management
	32	Cezara Pastrav (UU), Social simulation for socio-ecological systems
	33	Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management
	34	Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change
	35	Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction
2023	01	Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions
	02	Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts
	03	Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations
	04	Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval
	05	Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications
	06	António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment
	07	Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning

08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning
09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques
10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing
11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications
12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries
13 Injy Sarhan (UU), Open Information Extraction for Knowledge Representation
14 Selma Čaušević (TUD), Energy resilience through self-organization
15 Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models
16 Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters
17 Meike Nauta (UT), Explainable AI and Interpretable Computer Vision – From Oversight to Insight
18 Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation
19 George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals
20 Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning
21 Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain
22 Alireza Shojaifar (UU), Volitional Cybersecurity
23 Theo Theunissen (UU), Documentation in Continuous Software Development
24 Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning
25 Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs
26 Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour
27 Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions
28 Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts
29 Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results

---

2024 01 Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education
02 Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems
03 Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis
04 Mike Huisman (UL), Understanding Deep Meta-Learning
05 Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair

06 Azqa Nadeem (TUD), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence
07 Parisa Shayan (TiU), Modeling User Behavior in Learning Management Systems
08 Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation
09 Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks
10 Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science
11 withdrawn
12 Peide Zhu (TUD), Towards Robust Automatic Question Generation For Learning
13 Enrico Liscio (TUD), Context-Specific Value Inference via Hybrid Intelligence
14 Larissa Capobianco Shimomura (TU/e), On Graph Generating Dependencies and their Applications in Data Profiling
15 Ting Liu (VUA), A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health
16 Arthur Barbosa Câmara (TUD), Designing Search-as-Learning Systems
17 Razieh Alidoosti (VUA), Ethics-aware Software Architecture Design
18 Laurens Stoop (UU), Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations
19 Azadeh Mozafari Mehr (TU/e), Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior
20 Ritsart Anne Plantenga (UL), Omgang met Regels
21 Federica Vinella (UU), Crowdsourcing User-Centered Teams
22 Zeynep Ozturk Yurt (TU/e), Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts
23 Jie Luo (VUA), Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution
24 Nirmal Roy (TUD), Exploring the effects of interactive interfaces on user search behaviour
25 Alisa Rieger (TUD), Striving for Responsible Opinion Formation in Web Search on Debated Topics
26 Tim Gubner (CWI), Adaptively Generating Heterogeneous Execution Strategies using the VOILA Framework
27 Lincen Yang (UL), Information-theoretic Partition-based Models for Interpretable Machine Learning
28 Leon Helwerda (UL), Grip on Software: Understanding development progress of Scrum sprints and backlogs
29 David Wilson Romero Guzman (VUA), The Good, the Efficient and the Inductive Biases: Exploring Efficiency in Deep Learning Through the Use of Inductive Biases
30 Vijanti Ramautar (UU), Model-Driven Sustainability Accounting
31 Ziyu Li (TUD), On the Utility of Metadata to Optimize Machine Learning Workflows
32 Vinicius Stein Dani (UU), The Alpha and Omega of Process Mining

	33	Siddharth Mehrotra (TUD), Designing for Appropriate Trust in Human-AI interaction
	34	Robert Deckers (VUA), From Smallest Software Particle to System Specification - MuDForM: Multi-Domain Formalization Method
	35	Sicui Zhang (TU/e), Methods of Detecting Clinical Deviations with Process Mining: a fuzzy set approach
	36	Thomas Mulder (TU/e), Optimization of Recursive Queries on Graphs
	37	James Graham Nevin (UvA), The Ramifications of Data Handling for Computational Models
	38	Christos Koutras (TUD), Tabular Schema Matching for Modern Settings
	39	Paola Lara Machado (TU/e), The Nexus between Business Models and Operating Models: From Conceptual Understanding to Actionable Guidance
	40	Montijn van de Ven (TU/e), Guiding the Definition of Key Performance Indicators for Business Models
	41	Georgios Siachamis (TUD), Adaptivity for Streaming Dataflow Engines
	42	Emmeke Veltmeijer (VUA), Small Groups, Big Insights: Understanding the Crowd through Expressive Subgroup Analysis
	43	Cedric Waterschoot (KNAW Meertens Instituut), The Constructive Conundrum: Computational Approaches to Facilitate Constructive Commenting on Online News Platforms
	44	Marcel Schmitz (OU), Towards learning analytics-supported learning design
	45	Sara Salimzadeh (TUD), Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making
	46	Georgios Stathis (Leiden University), Preventing Disputes: Preventive Logic, Law and Technology
	47	Daniel Daza (VUA), Exploiting Subgraphs and Attributes for Representation Learning on Knowledge Graphs
	48	Ioannis Petros Samiotis (TUD), Crowd-Assisted Annotation of Classical Music Compositions
2025	01	Max van Haastrecht (UL), Transdisciplinary Perspectives on Validity: Bridging the Gap Between Design and Implementation for Technology-Enhanced Learning Systems
	02	Jurgen van den Hoogen (JADS), Time Series Analysis Using Convolutional Neural Networks
	03	Andra-Denis Ionescu (TUD), Feature Discovery for Data-Centric AI