Pedestrian detection in low-light conditions

A comprehensive survey

Ghari, Bahareh; Tourani, Ali; Shahbahrami, Asadollah; Gaydadjiev, Georgi

DOI
10.1016/j.imavis.2024.105106

Publication date
2024

Document Version
Final published version

Published in
Image and Vision Computing

Image and Vision Computing 148 (2024) 105106

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Review article

Pedestrian detection in low-light conditions: A comprehensive survey

Bahareh Ghari [a], Ali Tourani [b,*], Asadollah Shahbahrami [a], Georgi Gaydadjiev [c]

[a] Department of Computer Engineering, University of Guilan, Rasht, Iran
[b] Interdisciplinary Centre for Security, Reliability, and Trust (SnT), University of Luxembourg, Luxembourg City, Luxembourg
[c] Department of Quantum and Computer Engineering, Delft University of Technology, Delft, The Netherlands

ARTICLE INFO

Keywords:
Pedestrian detection
Object detection
Computer vision
Autonomous vehicles

ABSTRACT

Pedestrian detection remains a critical problem in various domains, such as computer vision, surveillance, and autonomous driving. In particular, accurate and instant detection of pedestrians in low-light conditions and reduced visibility is of utmost importance for autonomous vehicles to prevent accidents and save lives. This paper aims to comprehensively survey various pedestrian detection approaches, baselines, and datasets that specifically target low-light conditions. The survey discusses the challenges faced in detecting pedestrians at night and explores state-of-the-art methodologies proposed in recent years to address this issue. These methodologies encompass a diverse range, including deep learning-based, feature-based, and hybrid approaches, which have shown promising results in enhancing pedestrian detection performance under challenging lighting conditions. Furthermore, the paper highlights current research directions in the field and identifies potential solutions that merit further investigation by researchers. By thoroughly examining pedestrian detection techniques in low-light conditions, this survey seeks to contribute to the advancement of safer and more reliable autonomous driving systems and other applications related to pedestrian safety. Accordingly, most of the current approaches in the field use deep learning-based image fusion methodologies (*i.e.,* early, halfway, and late fusion) for accurate and reliable pedestrian detection. Moreover, the majority of the works in the field (approximately 48%) have been evaluated on the *KAIST* dataset, while the real-world video feeds recorded by authors have been used in less than 6 % of the works.

1. Introduction

Automatic identification and localization of pedestrians in images or video frames captured by visual sensors have become increasingly vital in computer vision. Pedestrian detection has use cases in various fields, such as autonomous vehicles [1,2], surveillance systems [3–5], and robotics [6–8]. This task can be challenging to resolve in real-world scenarios, as there are different factors to consider for accurate performance. Accordingly, various illumination conditions, dissimilar pedestrian appearances and poses, occlusion, camouflage, and cluttered backgrounds can bring about issues for pedestrian detection systems [9]. Regarding the introduced challenges, low illumination is the leading problem compared to others, as it has a natural or environment-related cause and cannot be prevented or naturally handled. It can be due to the time of the day, geographical location of where the scene is captured, weather conditions, *etc.* For instance, in some Scandinavian countries, especially during winter, the sunrise to sunset time can be less than ten hours, and the pedestrian detection systems need to be adapted to the challenging scenarios.

Although some approaches have used Light Detection And Ranging (LiDAR) sensors, which are accurate remote sensing technologies that use laser light for distance measurement and obstacle avoidance, such sensors do not provide rich information from the surroundings [10]. There are many recently introduced pedestrian detection approaches such as [11,12,13–16] that cover the task in a wide range of scenarios. However, few recent works focus on detecting pedestrians at night and in low visibility conditions. In low-illumination scenarios, it is much more difficult for autonomous vehicles equipped only with vision sensors to detect moving objects on the road and prevent incidents. Fig. 1 depicts how challenging the pedestrian detection task is compared to the same ordinary task during the daytime. The mentioned fact has resulted in an increase in demand for developing computer vision algorithms that can work under various illumination conditions.

Accordingly, this survey gives a deep review of 130 state-of-the-art low-light condition pedestrian detection algorithms. The research questions that the present work has aimed to answer are:

* Corresponding author.
E-mail address: ali.tourani@uni.lu (A. Tourani).

https://doi.org/10.1016/j.imavis.2024.105106
Received 23 January 2024; Received in revised form 23 April 2024; Accepted 30 May 2024
Available online 4 June 2024

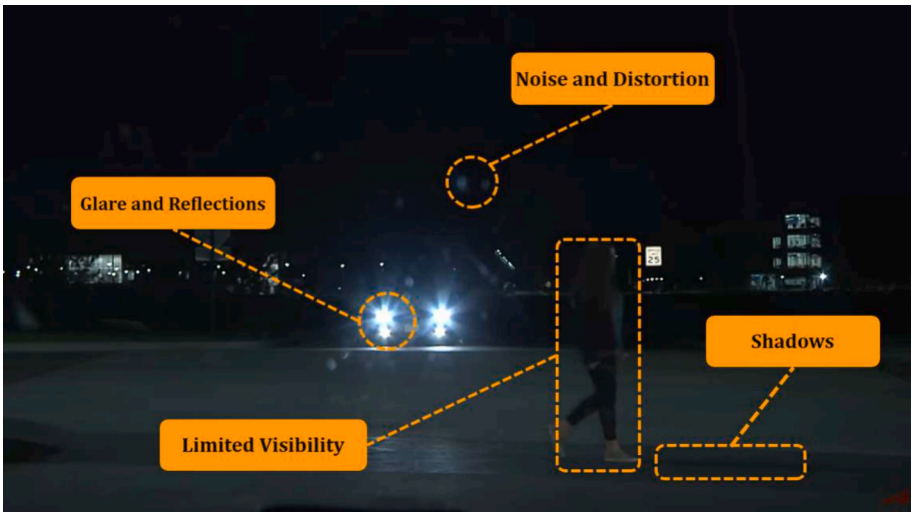

Fig. 1. Challenges of detecting pedestrians at night (image taken from 2019 Traffic Safety conference nighttime visibility report by Texas A&M Transportation Institute).

- **RQ1** Which datasets and baselines are mainly employed in low-light pedestrian detection tasks?
- **RQ2** What are the current deep learning-based algorithmic trends in low-illumination pedestrian detection?
- **RQ3** Regarding the state-of-the-art solutions, what are the currently existing and unresolved challenges in practical large-scale applications, such as fully autonomous vehicles?

To answer these questions, the paper in hand contributes to the body of knowledge in the field by providing contributions listed below:

- A survey of more than a hundred papers in the field of nighttime pedestrian detection,
- Review and classification of well-known baselines and datasets used for this purpose,
- Categorization of the state-of-the-art approaches in nighttime pedestrian detection regarding their architectural variations,
- Identification of the current trends and future methodologies in the field,

The rest of the paper is organized as follows: Section 2 reviews the currently available surveys in pedestrian detection in low-light conditions. Section 3 introduces the existing baselines and datasets employed by available approaches (**RQ1**). In Section 4, a set of recently introduced nighttime pedestrian detection approaches are introduced. Some discussions on the revealed trends and future insights in the field are presented in Section 5 (**RQ2**, **RQ3**). Finally, the paper concludes in Section 6.

2. Related surveys

Given the significant importance of the pedestrian detection chore in cutting-edge domains such as autonomous vehicles, an extensive collection of surveys has been publicly available. These surveys delve into various aspects, including methodological approaches, target environmental contexts, evaluation procedures, and pre-defined presumptions. This section offers a concise study of these existing reviews and identifies the unexplored factors within them. This identification of gaps underscores the unique contribution that the manuscript in hand aims to provide in this field.

Chen et al. [17] analyzed various object detection methodologies along with robust feature extractors employed in the fields of vehicle and pedestrian detection. They also employed extensive experiments on the *KITTI* vision benchmark [18] as a well-known street dataset to assess the performance of the studied algorithms in terms of accuracy, inference time, memory consumption, model size, and the number of Floating-Point Operations per Second (FLOPS). It is important to note that their research primarily focuses on the algorithmic perspective within practical frameworks, addressing the algorithms' efficiency across diverse scenarios. Hou et al. [19] studied pixel-level image fusion strategies for vision-based pedestrian detection that works in all daytime/nighttime conditions and discussed efficient strategies of combining such methods with Convolutional Neural Network (CNN)-based fusion architectures. The primary aim of their research is to discuss a pixel-level fusion of strategies adopted from various approaches that result in better performance for multi-spectral pedestrian detection tasks. Accordingly, their approach does not cover the future guidelines and possible strategies for such tasks. Authors in [20] studied deep learning-based methodologies employed in pedestrian detection tasks and provided informative discussions on how effective they are compared to other traditional algorithms. Although the mentioned survey also covers nighttime pedestrian detection, the comparison among various methodologies was mainly established on different datasets with low-quality and multi-spectral instances. Other works such as [21,22] surveyed approaches that targeted occlusion and scale variance challenges in pedestrian detection. They discussed solutions introduced in various papers that show acceptable performance in diverse conditions with occlusion, deformation, clutter, and scale difficulties.

Considering the introduced survey works, the present survey aims to provide specificities to set it apart from other available works, particularly regarding target scenarios and practical application use cases. In contrast with the previous works, the survey in hand is exclusively dedicated to nocturnal pedestrian detection, shedding light on the distinctive challenges tackled under low-light conditions introduced by state-of-the-art works. To the best of our knowledge, this work is the first study to focus entirely on the detection of pedestrians in low-illumination conditions (nighttime, in particular). To identify pedestrian detection approaches at nighttime that produce substantial results and feature novel architectures, the authors began by gathering and screening highly read and cited works from prominent venues over recent years. The sources included Google Scholar,[1] as well as well-

[1] https://scholar.google.com/, accessed on 10 April 2024

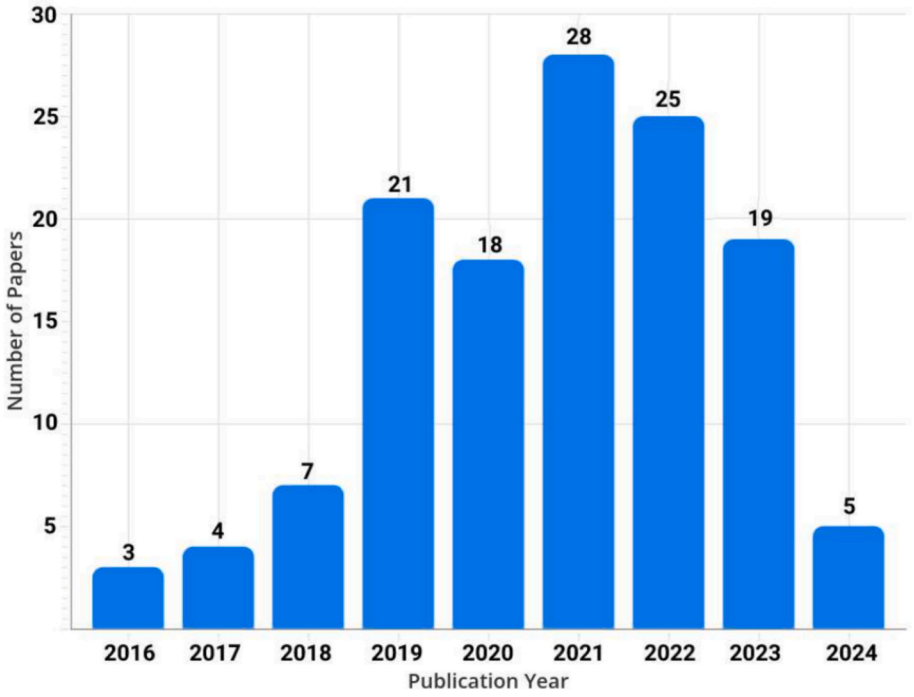

Fig. 2. Distributions of the papers surveyed in the current research work that only focus on pedestrian detection at low-light scenarios from 2016 to 2024 (Total: 130 papers).

established Computer Science bibliography databases, namely Scopus[2] and DBLP.[3] From the publications referenced in these sources, particular attention was given to those directly related to the ones targeted for challenging low-illumination conditions, and further checks were performed to ensure their alignment with the domain. Following an in-depth exploration of the papers, they were systematically categorized based on their primary methodological solutions in addressing nighttime pedestrian detection challenges. Fig. 2 depicts the distribution of papers collected, studied, and analyzed in the current survey in the range of 2016 to 2024.

3. Benchmarking datasets

Evaluation and development of pedestrian detection algorithms highly depend on providing proper data with annotated images/videos containing pedestrian instances. Such datasets should be well-annotated and cover diverse samples of pedestrian shots captured in real-world scenarios with various poses, occlusion levels, appearances, *etc.* to be considered appropriate for accurate training, testing, and validation stages. In this regard, this section collects various standard datasets for pedestrian detection at night that can be used for training and evaluation along with facilitating benchmark creation.

3.1. Ohio State University (OSU) Dataset[4] [23]

As one of the first pedestrian datasets, this thermal database provides a total number of 1.9k thermal frames with a resolution of 360×240, which were captured on campus and street. The *OSU* comprises three different classes of objects, including persons, cars, and poles. A total of 984 people were annotated in this dataset.

3.2. Laboratoire d'Interprétation et de Traitement d'Images et Vidéo (LITIV)[5] [24]

The dataset has nine video sequences, each containing people in an indoor hall with various zoom settings. The main challenges in these video sequences are the strong occlusions of objects and cluttered backgrounds.

3.3. CVC-09 Dataset[6] [25]

As another well-known dataset, CVC-09 is acquired during the day and night with 11k frames. The dataset contains training and testing sets, where the day and night sequences contain $5,990$ and $5,081$ frames, respectively.

3.4. Laboratorio de Sistemas Inteligentes Far-Infrared (LSI-FIR) Dataset[7] [26]

This dataset is composed of classification and detection portions and contains grayscale images collected in different temperatures with varying illumination. The classification part has $16,152$ positive samples (*i.e.,* pedestrian) and $65,440$ negative samples (*i.e.,* background), while the detection part includes $15,224$ images, categorized into $6,159$ train and $9,065$ test instances.

3.5. Thermal Infrared Video (TIV)[8] [27]

The dataset contains video sequences with $63,782$ annotated frames for visual processing tasks, such as detection, counting, group motion estimation, and single-view and multiple-view tracking. Three out of

[2] https://www.dblp.org/, accessed on 10 April 2024
[3] https://www.scopus.com, accessed on 10 April 2024
[4] https://vcipl-okstate.org/pbvs/bench/Data/01/download.html
[5] https://www.polymtl.ca/litiv/en/codes-and-datasets
[6] http://adas.cvc.uab.es/elektra/enigma-portfolio/item-1/
[7] https://www.kaggle.com/datasets/muhammeddalkran/lsi-far-infrared-pedestrian-dataset/code
[8] http://csr.bu.edu/BU-TIV/

sixteen sequences are mainly used for pedestrian detection, while other classes like car, runner, bicycle, and motorcycle are marked in this dataset.

3.6. Korea Advanced Institute of Science and Technology (KAIST)[9] [28]

It is one of the first multi-spectral pedestrian datasets with 95k aligned color-thermal image pairs and 103k dense annotation of samples. Data was captured from various traffic scenarios in the daytime and nighttime for autonomous driving applications. Annotations were manually added, resulting in three primary categories (person, people, and cyclist), and three occlusion levels (no-occlusion, partial-occlusion, and heavy occlusion).

3.7. Night-Time Pedestrian Dataset (NTPD) Dataset [29]

It contains a set of pedestrian images recorded by an active night vision system. The dataset contains $1,998$ positive and $8,730$ negative in the training set and $2,370$ positive and $9,000$ negative samples in the testing set.

3.8. CVC-14 Dataset[10] [30]

An extended version of the CVC-09 dataset, titled CVC-14, was introduced later to facilitate the challenges of automated driving. It contains video sequences of grayscale visible and thermal pairs corresponding to daytime and nighttime, where the daytime and the nighttime shares are $4,401$ and $4,117$ instances, respectively.

3.9. Keimyung University (KMU) Dataset[11] [31]

As a dataset captured using a FIR camera mounted on a vehicle driving in the summer nights for pedestrian detection, it contains three types of videos regarding the driving speed (20-30 km/h). It also covers pedestrians with different activities and poses, such as walking, running, and crossing the road. KMU has $4,474$ positive and $3,405$ negative frames in the training set and $5,045$ frames in the testing set.

3.10. UTokyo[12] [32]

This multi-spectral dataset contains RGB, NIR, MIR, and FIR images collected in a university for object detection in automated driving, including person, car, and bike. It contains $7,512$ images, where $3,740$ was taken during the daytime and the rest at nighttime.

3.11. CAMEL[13] [33]

The dataset provides visible-infrared video sequences for multiple object detection and tracking, where 43k visible-infrared image pairs are annotated with four different object classes, including person, bike, vehicle, and motorcycle. CAMEL covers various real-world scenes, occluded targets, and different illumination conditions.

3.12. NightOwls[14] [34]

This dataset targets the research on pedestrian detection at night and contains videos recorded in seven cities across Germany, the Netherlands, and the United Kingdom. It contains 279*k* frames with 42*k* pedestrians that have been manually labeled. Three primary labels (*i.e.,* far, medium, and near) have been assigned to the pedestrians to categorize them based on the distance they had from the vehicle during data acquisition. Additionally, frame brightness levels (low, medium, and high) and pedestrian pose (frontal and sideways) are other classification metrics employed in NightOwls.

3.13. South China University of Technology (SCUT) Dataset[15] [35]

A large-scale nighttime pedestrian dataset proposed by Xu et al. to motivate more attempts toward the task of on-road FIR pedestrian detection. The dataset contains approximately 11 hours-long image sequences with 211k annotated frames and a total of 477k bounding boxes for 7k unique pedestrians. *SCUT* groups pedestrians into three subsets, including near-scale (*i.e.,* ~ 80 pixels), medium-scale (*i.e.,* ~ 30 to ~ 80 pixels), and far-scale (*i.e.,* less than 30 pixels) subset based on the range of imaging distances.

3.14. YU FIR [36]

This seasonal temperature-based pedestrian detection dataset is captured on campus and urban traffic roads. The temperature was calibrated from -40°C to 150°C and used as the thermal infrared data for pedestrian detection. YU FIR contains a total of $2,802$ frames with $1,803$ and 575 positive images in the training set and test set, respectively.

3.15. Forward Looking InfraRed (FLIR) Dataset[16] [37]

This multi-spectral dataset was collected for Advanced Driver Assistance Systems (ADAS) during daytime and nighttime. It contains visible-thermal image pairs, some of which are not aligned, and the rest contain 5k multi-spectral pairs for training and testing. This version contains three frequent object categories, including persons, bicycles, and cars.

3.16. Zachodniopomorski Uniwersytet Technologiczny (ZUT) Dataset[17] [38]

It is a thermal dataset recorded in four European countries during diverse weather conditions, including sunny, foggy, heavy rain, light rain, and cloudy. The dataset contains 110k frames with 80k pedestrian annotations and provides synchronized Controller Area Network (CAN bus) data, including brake pedal status, driving speed, and outside temperature for ADAS.

3.17. Low Light Visible Image Person (LLVIP) Dataset[18] [39]

The dataset is recorded by a binocular camera containing visible light and Infrared (IR) sensors. Targeting low-illumination surveillance tasks, the dataset contains 15*k* pairs of visible-infrared images. The annotations of IR and visible-light images are the same due to the similar resolution and Field-of-View (FoV) of the cameras.

[9] https://github.com/SoonminHwang/rgbt-ped-detection
[10] http://adas.cvc.uab.es/elektra/enigma-portfolio/cvc-14-visible-fir-day-night-pedestrian-sequence-dataset/
[11] https://cvpr.kmu.ac.kr/KMU-SPC.html
[12] http://www.mi.t.u-tokyo.ac.jp/projects/mil_multispectral/
[13] https://camel.ece.gatech.edu/
[14] https://www.nightowls-dataset.org/
[15] https://github.com/SCUT-CV/SCUT_FIR_Pedestrian_Dataset
[16] https://www.flir.com/oem/adas/adas-dataset-form/
[17] https://ieee-dataport.org/open-access/zut-fir-adas
[18] https://github.com/bupt-ai-cz/LLVIP/

(a) NightOwls [172] (b) LLVIP [173] (c) KAIST [174] (d) SCUT [175]

(e) FLIR [176] (f) CVC-09 [177] (g) CVC-14 [178] (h) ZUT [179]

(i) OSU [180] (j) UTokyo [181] (k) KMU [182] (l) LSI [183]

(m) TIV [184] (n) CAMEL [185] (o) C3I [186] (p) NTPD [187]

Fig. 3. Instances of some datasets introduced for nighttime pedestrian detection. It should be noted that the collected image or video sequences were captured using various sensors.

3.18. C3I Thermal Automotive Dataset[19] *[40]*

The dataset was acquired in various environmental (*i.e.,* roadside, industrial town, alley, and downtown) and weather (*i.e.,* cloudy, foggy, windy, and sunny weather) conditions during daytime, evening, and nighttime. It comprises video sets with 39,770 frames, of which 17,740 frames are recorded in daytime, 12,640 in evening time, and 9,390 frames at nighttime. The frames are annotated in six object classes: person, car, bike, bicycle, bus, and pole.

To provide a comprehensive introduction to the datasets at hand for nighttime pedestrian detection, Fig. 3 depicts some of their instances. Additionally, Table 1 provides an in-depth description, facilitating a deeper understanding of their attributes and characteristics. It should be noted that these datasets have been curated in a way that encompasses a wide range of scenarios, including various lighting conditions, diverse pedestrian poses, occlusions, and complicated backgrounds. Such diversities ensure that provided data can serve as valuable resources for evaluating algorithms under real-world conditions.

[19] https://ieee-dataport.org/documents/c3i-thermal-automotive-dataset/

4. State of the art

When considering the pedestrian detection methodologies for nighttime and low-illumination conditions, one of the primary architectures that come to mind for designing such frameworks is to include an *offline training* procedure that utilizes a dedicated pedestrian images dataset to train a classification model. In this regard, the model learns hidden patterns and characteristics specific to pedestrians in darker environments, enabling it to distinguish them from other objects or background elements. Although the mentioned architecture can be very beneficial, it should be noted that learning-based methodologies are not always used for this task. Fig. 4 shows these typical stages for pedestrian detection at night and how they are connected to each other. We can see the typical stages that form the foundation and serve as critical components in identifying pedestrians, irrespective of the specific methodology employed. Whether the framework employs handcrafted features, machine learning models, or a mixture of both, it typically encompasses standard stages, including Region of Interest (ROI) selection (*i.e.,* identifying potential pedestrian regions within an image), visual feature extraction (*i.e.,* capturing relevant information from the selected ROIs), pedestrian classification (*i.e.,* using the features to classify the detected regions as pedestrians or non-pedestrians), and position calculation (*i.e.,* determining the precise location of the detected pedestrians).

In this survey, nighttime pedestrian detection approaches based on the underlying techniques and methodologies employed have been

Table 1
Various datasets collected for pedestrian detection at night, sorted based on their publication year. Accordingly, dataset instances are collected using various sensors in different spectral ranges, *i.e.,* Near-Infrared (NIR), Middle-Infrared (MIR), and Far-Infrared (FIR).

	Metadata	Data						Sensor			
Dataset	*Published*	*#Videos*	*#Frames*	*#Pedestrians*	*Resolution*	*Frame-rate*[*]	*Bit depth*	*RGB*	*NIR*	*MIR*	*FIR*
C3I [40]	2022	6	~39k	–	640 × 480	30	8		✓	✓	✓
LLVIP [39]	2021	26	~33.6 k	–	1080 × 720	1	24	✓	✓	✓	✓
ZUT [38]	2020	–	~110k	~80k	640 × 480	30	16		✓	✓	✓
FLIR [37]	2020	–	~10.2 k	~28.1 k	640 × 512	24	16				✓
YU FIR [36]	2018	–	~2.8 k	~9.3 k	640 × 480	30	14				✓
SCUT [35]	2018	21	~211k	~477k	720 × 576	25	8				✓
NightOwls [34]	2018	40	~279k	~42k	1024 × 640	15	–	✓			
CAMEL [33]	2018	26	~43k	~80k	336 × 256	30	24	✓	✓	✓	
UTokyo [32]	2017	–	~7.5 k	~2k	640 × 480	1	–	✓	✓	✓	✓
KMU [31]	2016	23	~12.9 k	–	640 × 480	30	24				✓
CVC-14 [178]	2016	4	~8.5 k	~9.3 k	640 × 512	10	–	✓			✓
NTPD [29]	2015	–	~22k	–	64 × 128	–	–		✓	✓	✓
KAIST [28]	2015	12	~95k	~103k	640 × 480	20	8	✓	✓	✓	✓
TIV [27]	2014	16	~63.7 k	–	512 × 512	30	16		✓	✓	✓
LSI-FIR [26]	2013	13	~15.2 k	~16.1 k	164 × 129	–	14				✓
CVC-09 [177]	2013	2	~11k	~14k	640 × 480	–	–				✓
LITIV [24]	2012	9	~6.3 k	–	320 × 240	30	8	✓	✓	✓	✓
OSU [23]	2005	10	~1.9 k	984	360 × 240	30	8		✓	✓	✓

[*] presented in *frames per second (fps)*.

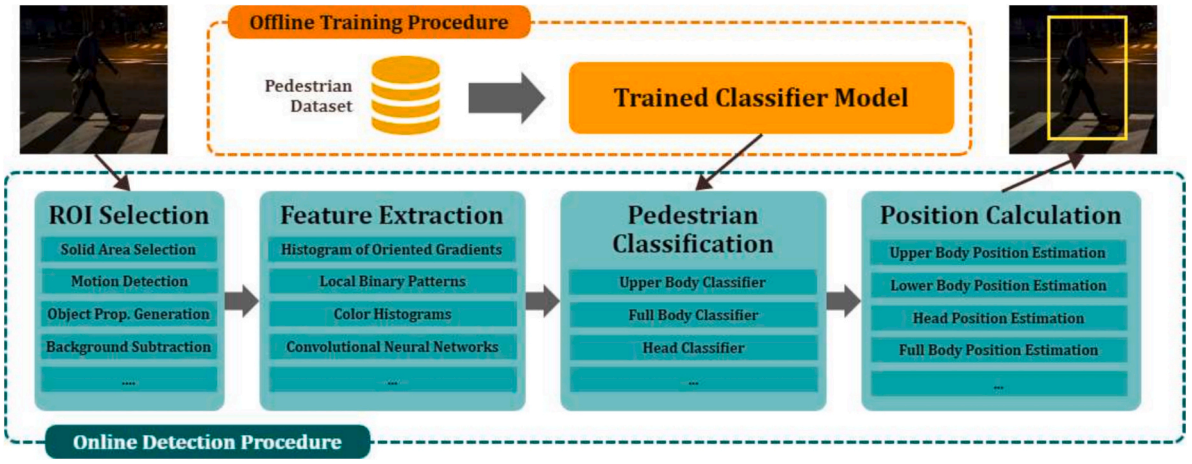

Fig. 4. The overall diagram of nighttime pedestrian detection methodologies.

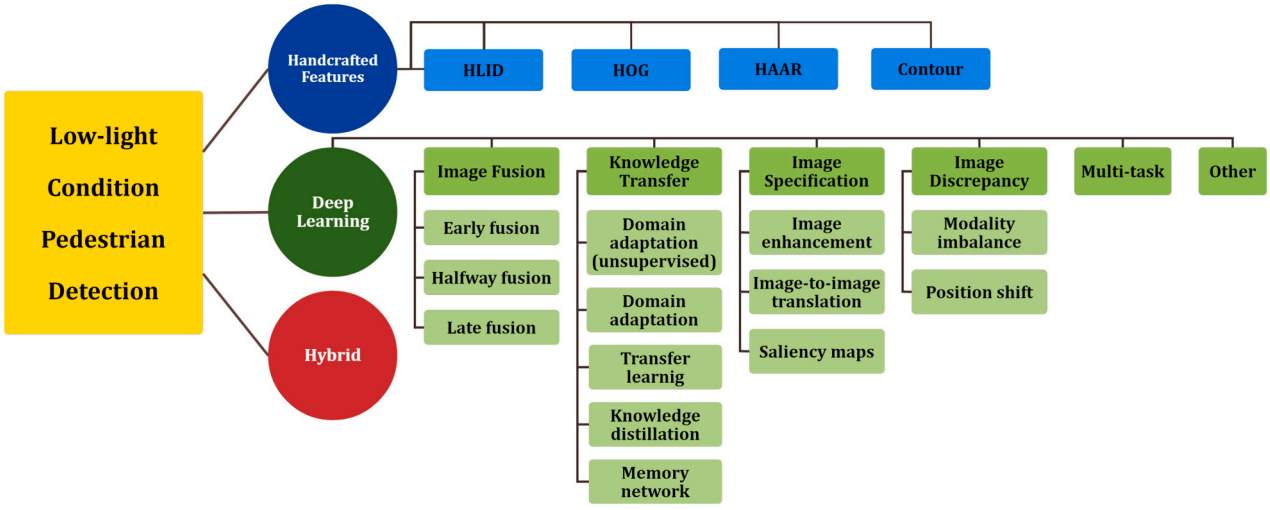

Fig. 5. The primary classification of different nighttime pedestrian detection methodologies considered in this survey.

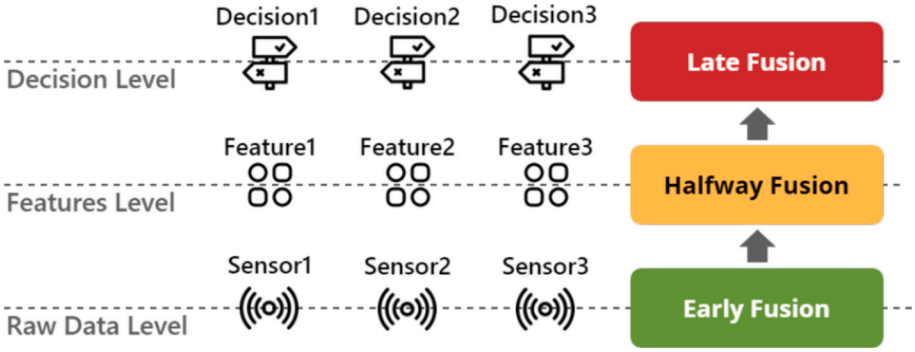

Fig. 6. Various image fusion strategies used in nighttime pedestrian detection approaches.

categorized into three distinct groups, including handcrafted features, deep learning, and hybrid methods. To understand the diverse strategies, along with their set of advantages and limitations, this section provides an in-depth study of the state-of-the-art works classifiable in the mentioned three categories. Fig. 5 depicts the classification strategy considered for different nighttime pedestrian detection methodologies in the current survey.

4.1. Handcrafted features approaches

Handcrafted features contain manual design and selection of particular visual features from the input image/frame. Extracting information using such features has the advantage of simplicity, transparency, explainability, and the ability to provide consistent results across similar scenarios. They generally require lower computational costs and can still work well when there is no access to annotated data. However, they can have adaptability problems regarding their domain expertise and may reach a *performance ceiling*, making them difficult to be improved over a certain point. Considering these trade-offs, handcrafted features can still be found in many pedestrian detection frameworks under challenging illumination scenarios.

Regarding the intrinsic characteristics of handcrafted features, the majority of approaches in this category require employing thermal data. As one of the first works in this category, Davis et al. [23] used a combination of generalized person template derived from Contour Saliency Map (CSM) and background subtraction to identify pedestrians' locations in thermal frames. Then, an *AdaBoost* classifier could validate the candidate regions, which adaptively adjusts the filters from the gradient information of training instances. While the template-based method brings about a quick screening procedure, it is considered a challenging methodology to detect groups of people in the scene. Similarly, Nowosielski et al. [41] presented a HAAR and Adaboost-based night-vision framework to identify humans in thermal images. The proposed algorithm processed all frames independently and without the aggregation mechanism, which increases the false positive rate due to incorrect recognition of the region as a person. An approach titled Thermal Infrared Radiometric Cumulative Channel Feature (TIR-ACF) introduced in [36] employs a thermal normalization methodology to factor in the maximum human body temperature for pedestrian detection. However, the experimental environment of this normalization strategy only includes a specific temperature range of small distant targets. As a more complicated methodology, Jeong et al. [31] presented an approach based on a Cascade Random Forest (CaRF) classifier, low-dimensional Haar-like features, and Oriented Center-Symmetric Local Binary Patterns (OCS-LBPs) for detecting sudden pedestrian crossing in thermal images. As the thermal temperature of the road is similar to or slightly higher than the pedestrians during summer nights, the concentration of this approach is on pedestrian samples in the summer season which leads to high prediction accuracy. Kim et al. [42] designed a pedestrian detector using a multi-level cascade learning algorithm and Histogram of Oriented Gradients (HOG) features. They used a smartphone-based thermal camera to capture human images of indoor environments to validate their work. Additionally, the 2D thermal image is mapped into a 3D space through an inverse perspective transformation method [43] to estimate the distance of the pedestrian detected from the camera.

Infrared images are another source of valuable information for pedestrian detection tasks at night. Zhou et al. [44] designed a pedestrian extraction algorithm for IR images. They build a global model using the weighted HLID and texture weighted Histogram of Local Intensity Difference (HLID) and texture weighted HOG algorithms to locate potential pedestrian regions. Then, using a head template based on the HAAR-like features and incorporating it into a local model for pedestrian head search, the global and head templates are combined to identify pedestrians. As another approach, Khalifa et al. [45] introduced a foreground detection framework that models the background's global motion between consecutive frames by applying the block-matching algorithms to the ROI to compensate for the camera motion. They use a Support Vector Machine (SVM) classifier to differentiate between the image's foreground and background. The evaluation results on the *CVC-14* show that the proposed algorithm can capture the dynamic aspect between frames in a video stream. Shahzad et al. [46] suggested a new procedure for pedestrian detection, tracking, and head detection in IR systems using template matching, Kalman filter, and HAAR cascade classifiers, respectively. The authors confirmed that the template matching method performs better than the contour-based method for pedestrian detection, and pedestrian tracking using the Kalman filter has the highest error rate. Likewise, and based on visual saliency in IR images, Cai et al. [47] proposed a model to focus on ROI generation along with a HLID feature and an SVM classifier to make a final detection. Considering that the visual saliency-based method includes small processing regions for candidate verification, the proposed algorithm demonstrates a fast execution time.

To conclude, the approaches with handcrafted features can provide acceptable results in many cases. However, they generally suffer from their incapability to handle complex scenarios due to their low discriminative nature and seem to act less flexibly while adapting to new scenarios.

4.2. Deep learning approaches

Solutions based on deep learning leverage the potential of neural networks to learn and extract features from raw image data automatically. In this regard, adaptability, versatility, generalization w.r.t. diverse scenarios, and high-performance results are among the expected outcomes of employing Deep Neural Networks (DNNs). They also have the capability to automatically learn features and reduce the requirement for manual feature engineering, along with integrating feature extraction and detection steps to have an end-to-end learning procedure. However, approaches in this category typically require large amounts of

labeled data for training and powerful hardware due to their computationally intensive nature. Additionally, they lack straightforward explainability, leading to challenging interpretations of their decision-making process and, thus, fine-tuning to improve their performance.

Many recent works for low-light pedestrian detection employ DNNs as an inevitable part of their algorithms. These methodologies have been divided into categories below in this survey regarding their use case:

4.2.1. Image fusion methodologies

Image fusion refers to extracting and fusing the most significant characteristics of raw images captured by multiple sensors to generate a single image with complementary information, a compelling description of the scene, *etc.* [48]. Considering the *fusion* stage, CNN fusion architectures can be divided into three primary strategies, namely, *early fusion*, *halfway fusion*, and *late fusion*. Fig. 6 depicts a brief overview of different fusion architectures used in the research works.

4.2.1.1. Early fusion-based methods. In the context of nighttime pedestrian detection, it indicates integrating visual and thermal feature maps right after the first *convolutional* layer of a CNN. However, fusing IR and RGB images to generate a four-channel input before feeding the network is another approach that can execute low-level feature fusion at an early stage. As a recent work introduced in [49], a You Only Look Once (YOLO) v.3-based [50] multi-spectral pedestrian detector is introduced that can use RGB, thermal, and multi-spectral images. The mentioned approach merges the three channels of RGB and the single channel of thermal images to prepare a 4-channels input. The authors also evaluated a YOLO-4 L [51] version to improve the detection accuracy of small-scale pedestrians on the scene. Evaluation results on various datasets demonstrated that the methodology outperforms other image types under all lighting conditions.

4.2.1.2. Halfway fusion-based methods. As a widely explored fusion strategy in recent years, the fusion operation of input modalities happens at the middle stages of a network, after the fourth convolutional layer. As one of the halfway fusion works, Yang et al. [52] designed a Cascaded Information Enhancement Module (CIEM) and a Cross-modal Attention Feature Fusion Module (CAFFM) to enrich the pedestrian information and suppress the interference caused by background noise in the color and thermal modalities. While CIEM uses a *spatial attention mechanism* to weigh the features combined by the cascaded feature fusion block, CAFFM employs *complementary features* to construct global features. In [53], Channel-wise Attention Module (CAM) and Spatial-wise Attention Module (SAM) were integrated into a multi-layer fusion CNN, aiming to re-weigh cross-spectral features at channel-dimension and pixel-level, respectively. Although the SAM methodology results in reduced detection speed, the performance of the approach is substantially improved. Zhang et al. [54] suggested a new halfway fusion strategy that applies cyclical fusion and refinement operations to achieve the consistency and complementary balance of multi-spectral features by controlling the number of loops. Based on the fact that the fused features are more discriminative than the mono-spectral ones, their main idea is to consecutively refine the spectral features with the fused ones and increase the overall feature quality. Hence, according to the analysis on the KAIST and FLIR dataset, the authors suggested that the number of loops should be tuned for any dataset. A cross-modal framework based on YOLO v5 detector introduced in [55] for multi-spectral pedestrian detection. In their study, the information complementarity of RGB-thermal streams was acquired by a Cross-modality Feature Complementary Module (CFCM) to reduce the target loss. They use an Attentionbased Feature Enhancement Fusion Module (AFEFM) to fuse different modalities' essential features and suppress the background noise while strengthening the semantic information. In another approach, and based on YOLO v5 lightweight network, Fu et al. [56] proposed an adaptive spatial and pixel-level feature fusion module, called *ASPFF Net*, to obtain fusion weights of spatial positions and pixel dimensions in two feature maps. The fusion weights are employed to re-calibrate the original feature maps of visible and IR images to acquire multi-scale fusion feature layers. The spatial and pixel attention mechanisms enable the *ASPFF Net* to focus on learning useful information and suppress redundant information to achieve a fast prediction speed of 35 frames per second (fps) and lower Miss Rate (MR) on the night subset of the KAIST dataset. A Multi-Layer Fusion network based on Faster R-CNN (MLF-FRCNN) was proposed by [57], which employs Feature Pyramid Network (FPN) and Region Proposal Network (RPN) as two parallel feature extractors to deal with pedestrian samples with different scales. As a two-stage multi-spectral pedestrian detector, the MLF-FRCNN achieves a running time of 0.14 s per frame and the highest Average Precision (AP) in detecting various pedestrian scales. In [58], four variants of fusion models have been designed at different stages, titled low-level (*i.e.,* early fusion), middle-level (halfway fusion), high-level (late fusion), and confidence-level (score fusion). The first three approaches implement convolutional feature fusions, while the last corresponds to the combination of confidence scores from RGB and thermal CNN branches at the decision stage. The study reveals that the halfway fusion model achieves the lowest overall MR.

In halfway fusion spatial attention-based mechanisms, the importance of each location in the feature map is calculated to highlight the areas with valuable information. Accordingly, Cao et al. [59] used Channel Switching and Spatial Attention (CSSA) in a lightweight fusion module to effectively fuse multi-modal inputs while ensuring low computational cost. During channel switching, the channel of each modality with insufficient features is replaced by the corresponding channel from another modality. Likewise, a bi-directional fusion strategy called *BAANet* is introduced in [60] to ensemble the RGB-thermal features for multi-spectral pedestrian detectors. The strategy distills the high-quality features of two modalities and re-calibrates the representations gradually. It contains intra- and inter-modality attention modules to improve spectra-specific features and adaptive selection of information from the most reliable modalities, respectively. In another similar work, Zhang et al. [61] introduced a two-stream CNN, titled Guided Attentive Feature Fusion (GAFF), to dynamically re-weigh and integrate multi-spectral pedestrian features under the guidance of the intra- and inter-modality attention mechanisms. The intra-modality attention module aims to enhance the visible or thermal features in pedestrian areas, while the inter-modality attention module selects the most reliable modality according to the feature quality, which requires costly annotation information. The authors' solution to this issue is to assign labels based on the prediction of pedestrian masks from the intra-modality attention module and then select the most relevant modality where the prediction mask is closer to the ground truth. Qingyun et al. [62] proposed a Crossmodality Fusion Transformer (CFT) module and embedded it to the YOLO v5 framework. The CFT learns long-range dependencies and focuses on global contextual information. In particular, by leveraging the self-attention mechanism, the network can simultaneously carry out intra-modal and inter-modal fusion and capture the latent interactions between visible and Thermal-Infrared (TIR) spectrums.

Some works discuss the most common feature fusion strategies in CNNs: *concatenation* (*i.e.,* stacking two feature maps at the exact spatial locations), *summation* (*i.e.,* calculating the sum of two feature maps at the exact spatial locations), *maximum* (*i.e.,* obtaining the maximum response of two feature maps at the exact spatial locations), and *mean* (*i.e.,* calculating the mean value of two feature maps at the exact spatial location) [63]. Accordingly, Pei et al. [63] discussed the influence of these strategies in various CNN fusion architectures, including merged Visual-Optical (VIS) and IR images based on *RetinaNet* detector [64]. The results proved that the summation fusion strategy performs better than other methodologies. Ding et al. [65] employed a Network-In-Network (NIN) in Region- based Fully Convolutional Network (R-FCN) framework [66] to merge the image information of two sub-networks to

deal with large-scale and small-scale pedestrian instances. After the concatenation of Conv-VIS and Conv-IR, the small- and large-scale pedestrian candidates generated by RPN are merged with convolutional layers in the middle of the architecture. Yun et al. [67] proposed inter- and intra-weighted cross-fusion networks (Infusion-Net), which use a High-Frequency Assistant (HFA) to integrate color and thermal features regarding the feature level gradually. In this procedure, the HFA block exchanges, purifies, and reinforces the object detection-relevant features based on Discrete Cosine Transform (DCT) and Residual Channel Attention Block (RCAB). Additionally, learnable inter- and intra-weight parameters provide optimal information utilization and feature reinforcement for each stream considering each fusion stage. Bao et al. [68] proposed a dual-YOLO method based on YOLO v7 [69] for integration of IR and visible images. They also designed attention fusion and fusion shuffle modules to alleviate the false detection rate caused by redundant feature information during the fusion process.

Numerous anchor-free pipelines have recently been proposed for multi-spectral pedestrian detection, which speeds up model detection while avoiding the complex hyper-parameter settings of anchor boxes. Two feature fusion schemes based on a dual-branch *CenterNet* [70] anchor-free detector proposed by [71] for multi-spectral and multi-scale pedestrian detection. The first one is Scale-aware Permutated Attention (SPA) module, which combines local attention and global attention sub-modules, enhancing the quality of the feature fusion at different scales. The second is Adjacent Feature Aggregation (AFA), which aggregates features across different scales, considering spatial resolution and semantic context. Likewise, Cao et al. [72] attempted to train a multi-spectral pedestrian detector without anchor boxes *via* a box-level segmentation supervised learning framework and compute heat maps. Consequently, the network can be able to localize the pedestrians on small-size input images.

In an innovative approach, Tang et al. [73] took illumination into account and designed a progressive image fusion network referred to as *PIAFusion*, which can adaptively maintain the intensity distribution of salient targets as well as retain texture information in the background. It uses an illumination-aware sub-network to estimate the illumination situations and exploits the illumination probability to construct illumination-aware loss. Afterward, the Cross-Modality Differential Aware Fusion (CMDAF) module and halfway fusion strategy merge meaningful information of IR and visible images under the guidance of illumination-aware loss. Likewise, Roszyk et al. [74] applied YOLO v4 framework for fast and low-latency multi-spectral pedestrian detection in autonomous driving. Different fusion schemes, as well as different types of models, were investigated, among which feature-level fusion, namely YOLO v4-Middle, demonstrates the best trade-off between accuracy and speed. Peng et al. [75] introduced Hierarchical Attentive Fusion Network (HAFNet) embedded with a Hierarchical Content-dependent Attentive Fusion (HCAF) module and a Multi-modality Feature Alignment (MFA) block to overcome the background noise and modality misalignment issue. The MFA exploits the correlation between the TIR and visible domains to fine-tune the pixel alignment of multi-spectral image pairs. Then, the HCAF utilizes top-level features to guide pixel-wise fusion across two streams, resulting in high-quality feature representation. Yadav et al. [76] built two uni-modal encoded-decoder feature networks for color and thermal individually using Faster Region-based CNN (Faster R-CNN) [77]. Further, they constructed middle-level CNN fusion architecture, which fused the extracted features in the last convolution layer before feeding it to the decoder for providing the final predictions. Zhang et al. [78] presented a Cross-Modality Interactive Attention Network (CIAN) to encode the correlations between two color and thermal spectrums and predict the positions and sizes of pedestrians on a contextual enhanced feature hierarchy. Regarding the halfway fusion strategy, CIAN has investigated three types of operations (*i.e., Elementwise Sum*, *Elementwise Maximation*, and *Concatenation and Channel Reduction*) for how to fuse feature maps. The fusion operation of the *Concatenation and Channel Reduction* shows better performance, which first concatenates the two feature maps, then applies a NIN to reduce the number of channels. To improve the detection accuracy in cases such as occluded objects, light changes, and cluttered backgrounds, Hu et al. [79] proposed a Dual-modal Multi-scale Feature Fusion Network (DMFFNet). In their work, MobileNetv3 [80] extracts multi-scale features of dual-modal images as input for MFA module, which processes the spatial information of input feature maps with different scales and establishes longer-distance channel dependencies, thereby reducing background noise interference. Eventually, the Double Deep Feature Fusion (DDFF) module deeply combines the multi-scale features to maximize the correlation between the multi-scale features, which significantly enhances the representation of semantic information and geometric detail.

Cao et al. [81] modeled a Multi-spectral Channel Feature Fusion (MCFF) module based on YOLO v4 to fuse the multi-spectral features according to the different illumination conditions. The MCFF module first concatenates the features from visible and thermal modalities in the channel dimension, then uses learning weights to adapt aggregate features.

Gated Fusion Units (GFU) [82] adjusts the contribution of the feature maps generated by each modality *via* the gating weighting mechanism. Instead of stacking selected features from each channel and adjusting their weights, and motivated by [82], Gated Fusion Double SSD (GFD-SSD) [83] developed two variations of GFU (*i.e., gated fusion* and *mixed fusion*) to fuse the feature maps generated by the two Single Shot MultiBox Detector (SSD) [84] middle layers for multi-spectral pedestrian detection. Using GFUs on the feature pyramid structure, the authors also designed four mixed architectures of both stack fusion and gated fusion (*i.e., Mixed Even*, *Mixed Odd*, *Mixed Early*, and *Mixed Late*), depending on which layers are selected to use the GFUs. By comparing the experimental results on the KAIST dataset, both the GFD-SSD and *Mixed Early* models are superior to the stack fusion. Redundant Information Suppression Network (RISNet) [85] designed a mutual information minimization module to alleviate the influence of cross-modality redundant information on the fusion of RGB-Infrared complementary information. Besides, the RISNet introduced a classification method of illumination conditions based on histogram statistics. Xie et al. [86] introduced a hallucination branch in order to map from the thermal to the visible spectrum by a three-branch feature extraction module and then fused feature maps from visible, thermal, and hallucination branches. The proposed method shows boosting in overall detection performance. Conventional multi-modal feature fusion methods rely only on local feature correlations, which degrades performance. According to the problem, Lee et al. [87] proposed an attention-based fusion model, named INSANet (INtra-INter Spectral Attention Network) to capture global intra- and inter-information and learn mutual spectral relationships by intra- and inter-spectral attention blocks.

4.2.1.3. Late fusion-based methods. Also known as decision-level fusion, is the high-level fusion technique in which the concatenation is conducted after the last convolutional layer and before fully connected layers or merged the outputs of the two sub-networks such as the location and category prediction. As a work in this category, MultiSpectral Pedestrian DEtection TRansformer (MS-DETR) is introduced by [88], which extracts multi-scale feature maps through two parallel modality-specific CNN backbones, aggregates them within the corresponding modality-specific transformer encoders, and fuses the features using a multi-modal transformer decoder. It also adopts a modality-balanced optimization strategy to measure further and balance the contribution of each modality at the instance level. Khalid et al. [89] proposed two fusion methods to detect people: In the first one, an encoder-decoder architecture was used for image-level fusion, which independently encodes visible and thermal frames and fed the combined frames into a decoder to produce a single fused image, inputting a Residual Network-152 (ResNet-152) architecture. The second one takes ResNet-152 for

feature-level fusion, which extracts features of visible and thermal images separately and concatenates them into a single feature vector as the input of the dense layer. Montenegro et al. [90] customized YOLO v5's architecture for low-light pedestrian detection, and also they have conducted experiments on multiple multi-spectral pedestrian datasets *e.g.,* CVC-09, LSI-FIR, FLIR, CVC-14, NightOwls, and KAIST. By extensive evaluations made on different datasets, the best mean Average Precision (mAP) was obtained on LSI-FIR, followed by CVC-09 and CVC-14. Song et al. [91] designed a MultiSpectral Feature Fusion Network (MSFFN) that uses the extracted features of visible-channel and infrared-channel to obtain integrated features. The MSFFN strikes a favorable trade-off between accuracy and speed, especially on small-size input images. They extract multi-scale semantic features using two sub-networks, including Multiscale Feature Extraction of Visible images (MFEV) and Multiscale Feature Extraction of Infrared images (MFEI) and integrate them with an improved YOLO v3 framework. Selective Kernel Network (SKNet) [92] suggested a dynamic selection scheme to adaptively adjust receptive field size using selective kernel units with different kernel sizes. It uses a NIN-based fusion strategy to fuse RGB-IR image pairs. Park et al. [93] considered all detection probabilities from RGB, IR, and RGB-IR fusion channels in a unified three-branch model and designed a Channel Weighting Fusion (CWF) and an Accumulated Probability Fusion (APF) layers to fuse probabilities from different information streams at a proposal-level. A combination of an adaptive weight adjustment method with the YOLO v4 [94] is introduced by [95] to enrich the multi-spectral complementarity information for score fusion. Authors in [96] introduced Probabilistic Ensembling (ProbEn), a simple non-learning technique for late-fusion of multiple modalities derived from Bayes' first principle, *i.e.,* conditional independence assumptions. Shaikh et al. [97] introduced a probabilistic decision-level fusion approach based on Naïve Bayes to address lighting and temperature changes in color and thermal images by fusion and modeling the detection results of available pedestrian detectors without requiring retraining. In particular, the use of Naïve Bayes for the late fusion strategy enables the network to work with non-registered image pairs as well as poorly registered image pairs.

Zhuang et al. [98] examined the impacts of environmental variables on the efficiency of the pedestrian detector and proposed a lightweight Illumination and Temperature-aware Multispectral Network (IT-MN). The method is built on the SSD architecture with designing a late-fusion strategy and Fusion Weight Network (FWN) to compute the fusion weights. In addition, the default box generation is optimized by reducing the number of bounding boxes and choosing specific box aspect ratios to minimize the inference time. Inspired by YOLO v4, Double-Stream Multispectral Network (DSMN) [99] was designed to carry out pedestrian detection in challenging situations such as insufficient and confusing lighting. Their method extracts multi-spectral information provided by RGB and thermal images *via* two YOLO-based sub-networks. Also, it has an improved Illumination–Aware Network (i-IAN) module to estimate the lighting intensity of varied scenarios and allocate fusion weights to RGB-thermal sub-networks. Li et al. [100] explored various fusion schemes and pointed out their key adaptations. They also designed an Illumination-Aware Faster R-CNN (IAF R-CNN) framework to estimate the illumination value of the input image and incorporate color and thermal sub-networks *via* a gate function defined over the illumination value. In another work, Li et al. [101] introduced an Adaptive Soft-Gated Light Perception Fusion (ASG-LPF) to improve detection performance in varying lighting conditions, which uses a light perception module to distinguish the illumination levels in diverse driving scenarios. Takumi et al. [32] proposed a multi-spectral ensemble method based on YOLO v1 [102], which integrates detection results of the four single-spectral detection models into a single space as the final detection. As another approach, LG-FAPF [103] performed a cross-modal feature aggregation process guided by locality information to learn human-related multi-spectral features and used the obtained spatial locality maps of pedestrians as pixel-wise prediction confidence scores for the adaptive fusion of detection results under complex illumination conditions. Considering Center and Scale Prediction Network (CSPNet) model, Wolpert et al. [104] proposed an anchor-free multi-spectral framework to investigate various fusion strategies. They also introduced a new data augmentation technique for multi-spectral images called *Random Masking*. Kim et al. [105] designed a Multispectral Chain-of-Thought Detection (MSCoTDet) framework, which integrates Large Language Models (LLMs) to understand the complementary information between IR-RGB modalities and facilitate cross-modal reasoning at the semantic level. The proposed framework can generate text descriptions of the pedestrian sample in each modality. Moreover, a Language-driven Multi-modal Fusion (LMF) method was introduced to fuse the vision-driven and language-driven detection.

4.2.2. Knowledge transfer-based methodologies

This section's approaches leverage insights from various domains based on knowledge acquired from diverse sources to facilitate nighttime pedestrian detection capabilities. It contains various categories with different methodologies, including transfer learning, supervised and unsupervised domain adaptation, knowledge distillation, and memory-network methods.

4.2.2.1. Transfer learning methods. Transfer learning is reusing the knowledge obtained from pre-trained models for dissimilar but related tasks. In the context of nighttime pedestrian detection, and to fill in the gap of large-scale TIR dataset, Hu et al. [106] applied CycleGAN [107] to generate synthetic IR images from visible ones to expand the *CVC-09* dataset. They performed experiments using the YOLO v3 and Faster R-CNN models on the *CVC-09* dataset, in which the Faster R-CNN has shown better performance in the transfer learning task. In another work by Vandersteegen et al. [108], a pre-trained YOLO v2 [109] was used to perform real-time visible-thermal pedestrian detection. Their method takes three image channels composed of a combination of four image channels (*i.e., RGB* and *T*) information as input and can work on 80 fps. They discussed the possibility of creating a number of channel combinations as input channels of the YOLO v2 model and designed three models named YOLO-TGB, YOLO-RTB and YOLO-RGT. The YOLO-TGB, which only uses the combination of thermal, green, and blue image channels as input, performs better on the KAIST dataset than other proposed models. Geng et al. [110] replaced the loss function of YOLO v3 model with *DIOU Loss* [111] to accelerate the convergence speed of the network in IR image-based pedestrian detection. Although the loss function curve is more stable, the AP of the Diou-YOLO v3 is not satisfactory.

4.2.2.2. Domain adaptation methods. The critical idea of employing the domain adaptation mechanism in multi-spectral pedestrian detection is to exploit learned knowledge acquired from the color domain in the thermal images. In this regard, Guo et al. [112] focused on image-level domain adaptation by using an image-to-image transformer as a data augmentation tool to convert color images to the thermal spectrum. To aid the joint training process of the domain adapter and the detector, the authors defined a detection loss that back-propagates its gradients to the image transformer to progressively refine synthetic thermal images. The proposed method provides promising results compared to the baseline on the KAIST benchmark. Kieu et al. [113] introduced a task-conditioned training method to help domain adaptation of YOLO v3 to the thermal spectrum. The primary detection network was augmented by adding an auxiliary classification task of day and nighttime thermal images. Additionally, learned representations of this auxiliary task were used to condition YOLO to perform better in the thermal imagery. Authors in [114] addressed three top-down and one bottom-up domain adaptation techniques for pedestrian detection in the nighttime thermal images. They showed that bottom-up domain adaptation achieves better results in challenging illumination conditions. As another work by the

same authors [115], a new bottom-up domain adaptation strategy, known as *layer-wise domain adaptation*, is introduced. The main idea for this method is to adjust the RGB-trained detector to adapt to the thermal spectra gradually. Kristo et al. [116] attempted to improve the typical object detector performance for person detection at night in challenging weather conditions such as heavy rain, clear weather, and fog. The authors retrained YOLO v3, SSD, Faster R-CNN, and Cascade R-CNN detectors on a dataset of thermal images. They found that YOLO v3 is significantly faster than the others with a processing speed of 27, 5 fps. The generalization ability of RPN has been analyzed by [117] for multi-spectral person detection by performing cross-dataset evaluations on several benchmark datasets such as Caltech [118], CityPersons [119], CVC-09, KAIST, OSU, and Tokyo semantic segmentation [120]. They showed that KAIST achieves better results in generalization tasks in both daytime and nighttime conditions.

4.2.2.3. Unsupervised domain-adaptation methods. The objective of unsupervised domain adaptation is to adapt the well-trained detectors on annotated visible images to the thermal target without any manual annotation effort. As a work in this category, Meta-UDA [121] performed Unsupervised Domain Adaptation (UDA) thermal target detection using an online meta-learning strategy, resulting in a short and tractable computational graph. To mitigate the domain shift between the source and target domain, the Meta-UDA uses the adversarial feature alignment at both the image and instance levels, leading to slight improvement. In another work, Lyu et al. [122] used an iterative process to automatically generate the pseudo-training labels from visible and thermal modalities using two single-modality auxiliary detectors. They used the illumination knowledge of daytime and nighttime to assign the fusion priorities of labels for *label fusion*. Without using any manual training labels on the target dataset, the proposed method shows reasonable results on the night scenes of the KAIST dataset. Authors in [123] used transformers to tackle unlabeled data challenges in TIR images. They designed a Self-Supervised Thermal Network (SSTN) to learn feature representation and maximize the mutual data between visible and IR domains by contrastive learning to compensate for the shortage of labeled data. Later, a multi-scale encoder-decoder transformer system was employed for thermal object detection based on the learned feature representations. Inspired by pseudo-training labels, Lyu et al. [124] proposed an unsupervised transfer learning framework in multi-spectral pedestrian detection. Their overall framework is based on a two-step domain adaptation solution, in which the first stage generates intermediate representations of color and thermal images to reduce the domain gap across the source and target domains. The pseudo labels of the target objects are fused *via* an illumination-aware label fusion mechanism. In the second stage, an iterative fine-tuning process is conducted to progressively converge the detector on the target domain. In another work, Cao et al. [125] introduced an auto-annotation framework to iteratively label pedestrian instances in visible and thermal image channels by leveraging the complementary information of multi-modal data. They aim to automatically adapt a pedestrian detector pre-trained on the visible domain to a new multi-spectral domain without manual annotation. The predicted pedestrian labels on both image channels are merged *via* a label fusion scheme to generate the final multi-spectral pedestrian annotations. Then, the automatically generated labels are fed to a Two Stream Region Proposal Network (TS-RPN) detector to achieve unsupervised learning of complementary semantic features. An unsupervised multi-spectral domain adaptation framework was proposed by Guan et al. [126] to generate pseudo-annotations in the source domain, which can be utilized to update the parameters of the model in the target domain according to the complementary information in aligned visible-IR image pairs. Transfer knowledge from thermal to visible domain in unpaired settings and without requiring additional annotations has been performed in [127] by applying image-level and instance-level alignments based on the Faster R-CNN network using adversarial training.

4.2.2.4. Knowledge distillation methods. The concept of the Knowledge Distillation (KD) is based on inheriting the knowledge learned from a large and complex pre-trained teacher model to a smaller and simpler student model through a supervised learning process [128,129,130]. Generally, the main objective of this method is to transfer the applicable and meaningful representations of data to speed up the inference time of the student model without a significant drop in accuracy [128]. According to the teacher-student scheme, Liu et al. [131] developed a knowledge distillation framework as a student network that only takes color images as input and generates distinguishing multi-spectral representations, guided by a two-modalities teacher network. Moreover, Cross-modal Feature Learning (CFL) module based on a split-and-aggregation approach was incorporated into the teacher network to learn the standard and modality-specific characteristics between color and thermal image pairs. Hnewa et al. [128] employed Cross Modality Knowledge Distillation (CMKD) to enhance the performance of RGB-based pedestrian detection under adverse weather and low-light conditions. Two different CMKD methods were developed to transfer the multi-modal information of a teacher detector to a student RGB-only detector. The former uses KD loss, while the latter integrates adversarial training with knowledge distillation. Zhang et al. [130] proposed a Modality Distillation (MD) framework to transfer the knowledge from a high thermal resolution two-stream network with feature-level fusion to a low thermal resolution single-stream network with early fusion strategy. In particular, two specific knowledge distillation modules are used in the MD framework. An attention transfer generates attention masks by GAFF from a two-stream teacher model, which is transferred to a single-stream student model through performing an early fusion. Finally, a semantic transfer resolves the problem of modality imbalance in feature distillation using a new Focal Mean Square Error (F-MSE) cost function.

4.2.2.5. Memory-network methods. Memory Augmented Neural Network (MANN) can memorize and recall the prior information, such as visual appearance in the memory module, so the relevant data can be accessed by calculating the similarity [132]. In [133], a pedestrian detection process is introduced to improve the detector's performance in any modality. In the first stage, a multisensory-matching contrastive loss guides the pedestrian visual representation of two visible and thermal modalities to be similar. In the second, a Multi-Spectral Recalling (MSR) memory improves the visual representation of the single modality features by recalling the visual appearance of multi-spectral modalities and memorizes the multi-spectral contexts through a multi-spectral recalling loss, which encoded more discriminative information from a single input modality. The Large-scale Pedestrian Recalling (LPR) based on key-value memory was proposed by [132], which memorizes visual information of large-scale pedestrians to recall the relevant characteristics to cover inadequate small-scale pedestrian appearances.

4.2.3. Image specification-based methodologies

Another category focuses on methodologies in which image specifications play a crucial role. These methods can be divided into three primary strategies: *image enhancement*, *image-to-image translation*, and *saliency maps* methods.

4.2.3.1. Image enhancement methods. TIR images are characterized by noisy details, blurred edges, low contrast, and low resolution, resulting in a performance drop caused by low discrimination. In this regard, low-light image enhancement techniques are considered to improve the visual quality of thermal images and simplify their challenges. In a work by Marnissi et al. [134], an enhancement method based on images' architecture, title Thermal-Enhancement GAN (TE-GAN) is designed, which constituted of contrast augmentation, noise elimination, and edge

restoration. To enhance the clarity of the IR pedestrian targets with blurred edges, Sun et al. [135] adopted a super-resolution algorithm called Wide Activation Deep Super-Resolution (WDSR)-B [136]. They add the four-time down-sampling layer output to YOLO v3 trained by the enhanced IR images to acquire richer context information for small pedestrian targets. In another work, Marnissi et al. [137] combined Generative Adversarial Network (GAN) and Vision Transformer (ViT) for thermal image enhancement and introduced Thermal Enhancement Vision Generative Adversarial Network (TE-VGAN). TE-VGAN employs the U-Net architecture as an input image generator and two ViT models as global and local discriminators. The thermal loss feature is also introduced in their work to generate high-quality images. They investigated the effect of the thermal image enhancement method on the detection performance of different YOLO versions, resulting in a balance between contrast enhancement and noise reduction.

DIVFusion [138] incorporates a low-light image enhancement task and a dual-modal fusion task in a unified framework to investigate the effect of lighting conditions on image fusion. In their method, firstly, a Scene-Illumination Disentangled Network (SIDNet) is devised to eliminate the illumination degradation in nighttime visible images while maintaining informative features of source images. Then, a Texture-Contrast Enhancement Fusion Network (TCEFNet) is employed to aggregate complementary information and boost fused features' contrast and texture details. Finally, a color consistency loss is used to alleviate color distortion in enhancement and fusion processes. Li et al. [139] built Feature Attention Module (FAM) and Feature Transformation Module (FTM) to improve the efficiency of a pedestrian detector in darkness. FAM is designed to suppress the noisy representations, while FTM allows pedestrian examples under a low-light environment to generate more discriminate feature representations. An attention-based feature fusion module was designed in [140] to enhance pedestrian detection in low-illumination images. They used the brightness channel (*i.e.,* V-channel) from the *HSV* image of the thermal image as an attention map to activate the unsupervised auto-encoder for obtaining more details about the pedestrian. In order to address the challenge of light compensation in low-light conditions, a Brightness Correction Processing (BCP) algorithm is considered to guide self-attention map learning. Eventually, the image enhancement method was integrated into YOLO v4 detection model. They evaluated the proposed architecture on the LLVIP dataset.

To highlight pedestrians in low-resolution and noisy IR images, an Attention-guided Encoder-Decoder Convolutional Neural Network (AED-CNN) [141] is devised. In AED-CNN, the encoder-decoder module generates multi-scale features, and a skip connection block is integrated into the decoder to fuse the feature maps from the encoder and decoder structure. By adding an attention module, the network effectively emphasizes informative features and suppresses background interference while re-weighting the multi-scale features generated by the encoder-decoder module. Patel et al. [142] introduced a computationally compact algorithm based on Depthwise Convolution (DC) with the aim of network parameters reduction. The proposed algorithm enhances the details of the thermal images using Adaptive Histogram Equalization (AHE) and extracts the salient features in these images by a new Convolutional Backbone Network (CBN), where depthwise convolution minimizes the computational complexity. YOLO-FIRI [143] is another method developed for pedestrian detection in IR images, which achieved outstanding results by making improvements on YOLO v5 structure. Firstly, by extending shallow CSPNet in the backbone network and incorporating an improved Select Kernel (SK) attention module in the residual block, it forces the model to focus on shallow and detailed information and learn the distinguishable features. Secondly, the detection accuracy of small and blurry pedestrians in IR images is increased by adding four-scale feature maps to the detection head. Finally, Densefuse [144] is adopted as a data enhancement to fuse visible and infrared images to boost the features of IR images.

4.2.3.2. Image-to-image translation methods. The goal of Image-to-Image (I2I) translation models is to learn the visual mapping between a source and target domain while preserving the essential features. Specifically, I2I has been widely used in image colorization, denoising, and synthesis [145]. In these approaches, thermal image colorization aims to translate from the temperature-channel domain into the RGB channel. PearlGAN presented in [146] to facilitate the translation of nighttime Thermal-Infrared (TIR) image into a daytime color one. By taking advantage of a top-down guided attention module and a corresponding attention loss, *PearlGAN* can produce hierarchical attention distribution and reduce local semantic ambiguity in IR images through context information. In addition, a structured gradient alignment loss was designed to enhance edge consistency during the translation. The colorization of thermal-IR images in pedestrian detection application is accomplished by [147], organized into three main modules: thermal image colorization, improvement of colorized images, and pedestrian detection. The colorized and improved images are fed to the detection head using a pre-trained YOLO v5 framework.

To mitigate color distortion and edge blurring caused by translation from temperature spectrum to color spectrum, [148] considered a one-to-one mapping relationship and introduced an improved CycleGAN [107], called Gray Mask Attention-CycleGAN (GMA-CycleGAN). It first translates the TIR images to Grayscale Visible (GV) and then uses the original CycleGAN to obtain the translation from GV to Color Visible (CV). A mask attention module based on the thermal temperature mask and the color semantic mask has been designed without increasing training parameters to better differentiate between pedestrians and the background. Meanwhile, to make the texture and color of the translated image more realistic in the feature space, a perceptual loss was added to the original CycleGAN loss function. Devaguptapu et al. [149] proposed to borrow knowledge from the large-scale RGB dataset without the need for paired multi-modal training examples and used CycleGAN to implement an unpaired image-to-image translation framework. It can generate pseudo-RGB equivalents of a given thermal image and employs a multi-modal Faster R-CNN detector for pedestrian detection in thermal imagery. To transform the visible domain into the thermal domain, authors in [150] implemented a generative data augmentation method based on the Least-Squares GAN (LS-GAN) [151]. They also used the perceptual loss function to measure the similarity between authentic and synthesized images in pixel space.

4.2.3.3. Saliency maps methods. The purpose of salient object detection is to highlight the most noticeable areas in the given image and distinct the prominent objects from their surroundings using the intensity of each pixel. Accordingly, in TIR images, the saliency maps can be used to detect temperature. Altay et al. [152] presented a two-branch architecture that can incorporate features of thermal images with their correlated saliency maps to acquire better representations of pedestrian regions. Instead of using color-thermal image pairs in the fusion network, Ghose et al. [153] suggested augmenting thermal images with their corresponding saliency maps, which produced by static methods and two deep saliency networks, Pixel-wise Contextual Attention Network (PiCANet) [154] and Recurrent Residual Refinement Network (RRRNet) [155]. Marnissi et al. [156] proposed a bi-spectral image fusion scheme, which was augmented with a corresponding saliency map using Visual Salient Transformer (VST) and also incorporated this fusion process into the YOLO v3 as base architecture for real-time applications. The proposed approach has shown its advantage in low computational cost, which allows faster inference time. Zhao et al. [157] put more emphasis on the temperature information in infrared images by constructing an IR-temperature transformation formula which can convert the IR images into corresponding temperature maps. It finally uses a trained temperature network for pedestrian detection. On the OSU and FLIR datasets, the transformed temperature maps boost the overall performance regardless of external influences.

4.2.4. Images discrepancy-based methodologies

These methods target enhancing the accuracy and reliability of nighttime pedestrian detection by exploiting discrepancies within images and characteristics of different imaging sensors and analyzing variations in image quality and content. The modality discrepancy is alleviated by focusing on *modality imbalance problem* and *position-shift problem*.

4.2.4.1. Modality imbalance problem. Scenes in which one sensor performs considerably better than the others can lead to a *bias* in training toward one dominant input modality. For instance, uneven distribution of training data in multi-modal learning causes less contribution of the non-dominant input modality during network training and, therefore, limits the generalizability of the model. In this regard, Oksuz et al. [158] provided a comprehensive taxonomy of the imbalance problems in object detection. They categorize these problems into four significant categories: *class imbalance* (*i.e.,* inequality distribution of training data among different classes), *scale imbalance* (*i.e.,* various scales of objects), *spatial imbalance* (*i.e.,* spatial properties of the bounding boxes), and *objective imbalance* (*i.e.,* minimization of multiple loss functions). Additionally, in multi-spectral pedestrian detection, the modality imbalance issue substantially impacts the algorithm performance, which can occur in two different ways, including the *illumination modality imbalance problem* and the *feature modality imbalance problem* [159]. Das et al. [160] proposed a training process with a regularization term *i.e., Logarithmic Sobolev Inequalities* [161] to consider the features of both modalities equally during fusion. The proposed regularizer reduces the modality imbalance in the network by equally distributing the training data among the modalities. Li [162] trained YOLO v3 framework to detect pedestrians under insufficient illumination conditions. In their method, focal loss [64] was added to the loss function to overcome the imbalance issue of IR images. Zhou et al. [159] resolved the modality imbalance issue in multi-spectral images through the implementation of a single-stage Modality Balance Network (MB-Net), which included a Differential Modality Aware Fusion (DMAF) and an Illumination Aware Feature Alignment (IAFA) module to extract complementary information and align the two modality features according to the lighting conditions. Dasgupta et al. [163] developed Multi-modal Feature Embedding (MuFEm) module using Graph-Attention Network (GAT) [164] to deal with the imbalance issue between color image branch and thermal image branch. Also, the channel-wise attention block and four-directional IRNN (4Dir-IRNN) block [165] are incorporated in Spatio-Contextual Feature Aggregation (SCoFA) to improve fusion using spatial and contextual information of the pedestrian. The *4Dir-IRNN* block consists of four Recurrent Neural Networks (RNNs), which compute context features in four directions. Kim et al. [166] addressed the problem of dataset bias in multi-spectral pedestrian detection and designed a Causal Mode Multiplexer (CMM) framework, which causalities between multi-spectral inputs and predictions. They also made a new dataset to evaluate bias, which successfully removed the bias. The results show that the CMM approach is generalizable to the existing dataset.

4.2.4.2. Position-shift problem. The physical properties of different cameras (*e.g.,* Field-of-View (FoV), resolutions, wavelengths, *etc.*) can cause weakly aligned image pairs in multi-spectral data, where the positions of the objects are out of synchronization on different modalities. Some works tried to address the mentioned problem in multi-modal sensors using geometrical calibration and image alignment methods. The study by Zhang et al. [167] is the first work providing insights into the position shift problem between color and thermal images. They introduced an Aligned Region CNN (AR-CNN) detection framework to solve the weakly aligned image pairs. The AR-CNN firstly predicts the position shift and adaptively aligns the region feature maps of the two modalities through a Region Feature Alignment (RFA) module. Based on the aligned features, a confidence-aware fusion method is proposed to accomplish feature re-weighting, which selects the highly informative features while suppressing the useless ones. Moreover, a ROI jitter strategy is adopted to enhance the robustness of position shift patterns. Kim et al. [168] used adversarial learning to make each spectrum share its complementary information in a common feature space to compensate for the lack of aligned multi-spectral pedestrian datasets. Kim et al. [169] have constructed uncertainty-aware multi-spectral pedestrian detection architecture to handle miscalibration (*i.e.,* different FoV in color and thermal cameras) and modality discrepancy challenges. For the miscalibration issue, the Uncertainty-aware Feature Fusion (UFF) module was formulated to mitigate the impact of ambiguous Region of Interest (ROI). The modality discrepancy is alleviated through the Uncertainty-aware Cross-modal Guiding (UCG) module, which can encode more discriminative visual representations. Wanchaitanawong et al. [170] introduced a multi-modal Faster R-CNN robustly against significant misalignment between the two modalities. The key points are modal-wise regression for bounding-box regression of each modality to deal with the significant misalignment and multi-modal Intersection over Union (IoU) for mini-batch sampling that combines the IoU for both modalities.

4.2.5. Generative methods

4.2.5.1. Diffusion model-based methods. Yue et al. [171] proposed a model to explore objects with high color fidelity which is used by diffusion. They constructed multi-channel distribution by denoising the network with forward and reversed diffusion processes in a latent space. They also extracted multi-channel diffusion features by using a denoising network and fed the extracted features to the multi-channel fusion module to generate three-channel fused images. Furthermore, intensity loss and multi-channel gradient loss were proposed to retain the intensity and texture information. In order to multi-modal image fusion and leverage strong interpretable data of the GAN-based method, Zhao et al. [172] suggested a model of fusion algorithm based on the denoising diffusion probabilistic model (DDPM), which consists of the unconditional generation and a maximum likelihood sub-problems, and modeled in a hierarchical Bayesian manner with latent variables by the Expectation–Maximization (EM) approach. The proposed model demonstrated improvement in generating high-quality fused images with natural image generative priors and cross-modality information from source images.

4.2.5.2. Structural re-parameterization-based methods. Residual Spatial Fusion Network (RSFNet) [173] adopts an asymmetric dual-encoder to learn the compensating features of RGB-Thermal modalities and uses the saliency map to yield the pseudo-labels to supervise the feature learning. Moreover, to capture more promising features, the Residual Spatial Fusion (RSF) module was developed with structural re-parameterization, which applies the spatial weights with the residual connection to control the cross-modality feature fusion and improves the performance of multi-branch structure fusion without additional inference cost.

Wang et al. [174] investigated automatic driving perception under fog conditions and its accuracy and speed. Firstly, a new dataset was conducted. Then, a detection network for driving in fog based on improved YOLOv5 was presented. Structural re-parameterization modified ResNeXt model serves as the model's backbone and a new Feature Enhancement Module (FEM) was built. The results show improvement in fog multi-target detection in accuracy and speed. In order to slow detection speeds and high-cost challenges, Song et al. [175] introduced the YOLOv5-MS model, a YOLOv5-based solution for target detection. Leveraging re-parameterization, original backbone convolution replacement with RepvggBlock, and reducing convolutional layer channels to enhance speed was applied. Also, the

incorporation of a bioinspired "squeeze and excitation" module was used. In order to achieve experimental results, lower costs and higher speed of detection demonstrated an effective pedestrian detection method. A robust, lightweight network (RSTDet-Lite) [176] was proposed based on an improved version of YOLOv5-x for detecting pedestrians on rainy days, which is challenging. CBP-GNet approach, which incorporates a compact bilinear pooling algorithm, was designed. To improve the performance of the network, the CBAM attention mechanism and the idea of structural re-parameterization were introduced. In this study a new dataset was created named RainDet3000 for experimental evaluations, which result demonstrates performance improvement in rainy days, compared with YOLOv5-x. In [177], a method was proposed in order to detect pedestrians under low-light conditions problems of small and dense objects problems in You Only Look Once X (YOLOX). The model structure was re-parameterized using the Re-parameterization Visual Geometry Group (RepVGG) approach. Also, larger-scale daylight and the smaller-scale nighttime dataset after low-illumination degrading were combined. The result shows effective technical support for the safety of automated driving at night.

4.2.5.3. Variational autoencoder-based methods. To decrease dataset gathering and annotation costs, Nikolov et al. [178] proposed to augment the existing dataset to generate synthetic pedestrian data by training a variational autoencoder on a small subset of annotated pedestrians. In addition, the latent space of the autoencoder is interpolated to generate pedestrian variations and combine them to create new images.

4.2.6. Multi-task methods

Multi-task learning is a training paradigm that aims to learn multiple related tasks simultaneously, using shared feature representations [179]. A cross-task feature alignment method was proposed by [180] to tackle the misalignment of scale and channel of features from image relighting and pedestrian detection tasks by placing four feature alignment layers before the feature fusing and sharing step in cross-task learning. Meanwhile, a multi-scale feature-enhanced detection network expands the receptive field of the multi-scale feature extractor and thereby provides richer semantic information of fused features for the detection head. An illumination-aware weighting mechanism is presented by Guan et al. [181] to adaptively re-weight the detection results of day- and night-illumination sub-networks to learn multi-spectral human-related characteristics to perform pedestrian detection and semantic segmentation under various illumination conditions, simultaneously. Dai et al. [182] developed the Faster R-CNN detector using the ResNet-50 as a feature extractor for pedestrian detection and distance estimation using the NIR-based camera. An Automatic Region Proposal Network (ARPN) was designed by [183] to get bounding boxes. A pedestrian segmentation task is also added based on a Feature Pyramid Network (FPN) [184] to obtain the confidence scores. To distinguish pedestrian examples from complex negative samples, Li et al. [185] added two sub-networks for jointly semantic segmentation and pedestrian detection tasks to the unified fusion network, which is denoted as *MSDS-RCNN*. The paper also studied the effects of training annotation noise by creating a sanitized version of KAIST ground-truth annotations so that the sanitized training annotations significantly reduce the inference error. Evaluations showed that the segmentation supervision benefits multi-spectral pedestrian detection.

4.2.7. Other methods

As for the final category, this subsection introduces works that cannot fit into the previous ones. Accordingly, the authors in [186] employed a region decomposition branch in Faster R-CNN architecture, which exploits the multi-region features, including head, body trunk, and legs, to solve the pedestrian occlusion problem in thermal images. The proposed architecture learns the high-level semantic features by combining the global and partial appearance features step by step. The Center and Scale Prediction Network (CSPNet) [187] has been applied in [188] to obtain three IR pedestrian detection models, namely daytime, nighttime, and full-time. The full-time model has a lower detection loss rate, while the nighttime model and the daytime model perform poorly in detecting small objects in the evening, respectively. Xu et al. [189] aggregated ground-area context information into the Faster R-CNN for pedestrian detection and shared the predicted ground horizon area to a Ground-Region Proposal Network (GRPN), which can only process the pixels on the proposed horizon region to minimize False Positive (FP) rate. Since the output of the *FC layer* is the position vector of pixels in the horizon region, the size of the GRPN model is largely increased and has a high computational cost. Dai et al. [190] compared and analyzed visible and IR images acquired by using visible-spectrum, Near-Infrared (NIR), Short-wave Infrared (SWIR), and Long-wave Infrared (LWIR) cameras. For the first time, they used a nine-layer CNN model with a self-learning SoftMax [191] to detect nighttime pedestrian samples in NIR images. In order to enhance the detection accuracy of multi-scale pedestrians in IR images, in [192], two regional proposal networks based on the Faster R-CNN architecture were designed to focus on near and far away pedestrians. Although the proposed multi-scale RPN has shown improvements in far-away pedestrian detection, it is not optimized to work in real time. Kalita et al. [193] have presented a real-time human detection system using YOLO v3, which achieved a speed of 17 millisecond per image on the KAIST thermal dataset. The brightness aware Faster Region-based CNN (Faster R-CNN) model [194] was proposed to perform the pedestrian prediction under low-light and day-light scenarios. In the first step, the model calculates the brightness of the input image based on the pixel intensity to predict the day or night scenario. In the second step, two separate thermal or color models are employed for pedestrian detection based on the first step output. It should be noted that the authors trained the FLIR dataset for the thermal model and the PASCAL VOC dataset for the color model.

4.3. Hybrid approaches

Hybrid methods combine elements of handcrafted features and deep learning, aiming to harness the strengths of each approach for improved nighttime pedestrian detection performance. Accordingly, they require significantly less computational resources than deep learning methods and overcome the poor generalization of handcrafted methods. However, the performance of such approaches is not properly optimized in terms of prediction accuracy or model running time.

As a hybrid methodology for nighttime pedestrian detection, the study of Kim et al. [195] presented a method to detect pedestrians at night using a visible-light camera and Faster R-CNN model, which can handle the changes of the pedestrians' spatial position by fusing deep convolutional features in successive frames. To make the model robust against noise and illumination, the authors used Additive Random White Gaussian Noise (AWGN) and applied two pre-processing methods, *i.e.,* pixel normalization and Histogram Equalization (HE) mean subtraction, to normalize the illumination and contrast levels of successive frames. Besides, a weighted summation of successive frame features was added to exploit temporal information about the pedestrian, which enhanced the accuracy of the pedestrian detector at nighttime. To find the optimal fusion stage in CNN, authors in [196] used RPN to merge the features of visual and IR images. After halfway feature fusion in RPN, they employed Boosted Decision Tree (BDT) classifier to improve pedestrian detection results and reduce the false positive rate. Tumas et al. [197] eliminated the sliding window technique and applied background subtraction to extract thermally active points as Region of Interest (ROI) for pedestrian detection in Far-Infrared (FIR) domain. The proposed technique accelerates the Histogram of Oriented Gradients (HOG) based pedestrian detector to run at 6 fps using only CPU performance. Narayanan et al. [198] developed a model for low-light pedestrian prediction using HOG and YOLO v3 algorithm. They also experimented the

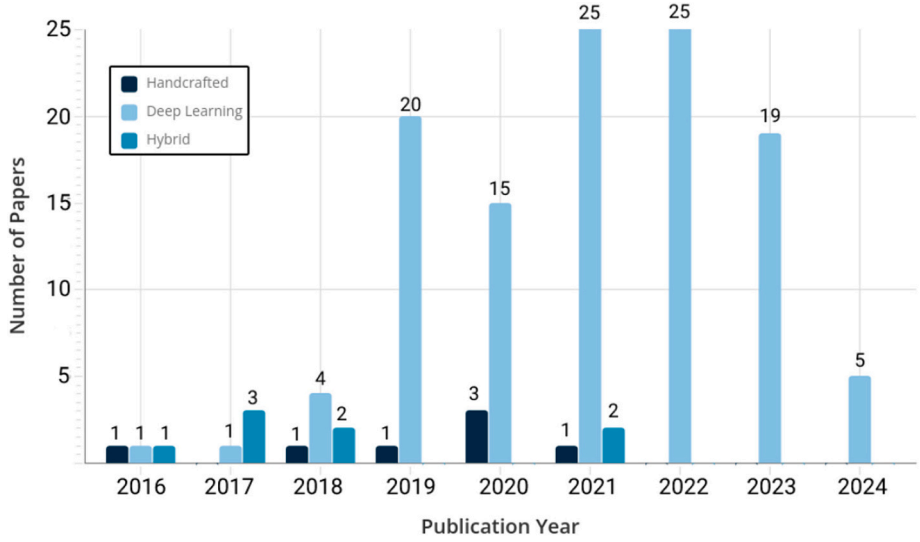

Fig. 7. The trends of employing various approaches explained in this survey by state-of-the-art research works published in different years (2016–2024).

detection accuracy of HOG detector and SVM classifier in thermal images. Xu et al. [199] designed a framework for learning and transferring cross-domain feature representations for pedestrian detection that works based on two different networks. The first one, titled Region Reconstruction Network (RRN), is employed to learn a non-linear feature mapping and model the relations among the color and IR image pairs. Afterward, the cross-modality feature representations learned from RRN are transferred to a second network titled Multi-Scale Detection Network (MSDN), which operates only on RGB inputs and outputs the recognition results. Both RRN and MSDN networks have employed ACF [200] to generate pedestrian proposals. In this way, only color images are considered at the test phase, and no thermal data are needed, which significantly reduces the cost of thermal data annotation. In [201], Support Vector Regression (SVR) was adopted to learn the pedestrians probabilities, which performs well on small-scale pedestrians. Chen et al. [202] utilized a Total Variation (TV) minimization [203] method based on structure transfer to integrate TIR-RGB image pairs, preserving the infrared intensity distribution and the local appearance features. However, when the thermal radiation of the pedestrian and the background are the same, the performance of the detector is affected.

5. Discussion

This section discusses the state-of-the-art methodologies introduced in previous sections, as well as the current trends and future expectations of works targeting nighttime pedestrian detection.

5.1. Employed methodologies trends

Regarding categorizing the works introduced in Section 4, nighttime pedestrian detection approaches can be divided into handcrafted features, deep learning, and hybrid methodologies. In this regard, Fig. 7 shows the distribution of the surveyed paper regarding the primary categories they belong to. According to the figure, it can be seen that the majority of the works published in the last two years consider deep learning-based techniques the most reliable methodology to detect pedestrians in low-light conditions. In other words, recent approaches only focus on employing DNNs instead of handcrafted and hybrid approaches. The main reason may be attributed to *automatic learning of*

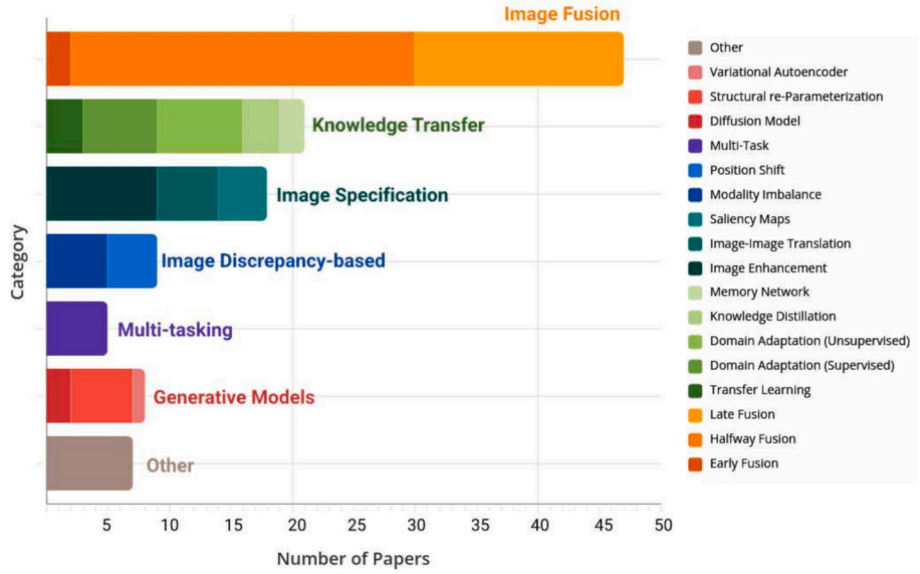

Fig. 8. Distribution of the reviewed papers considering the sub-categories introduced in Section 4.

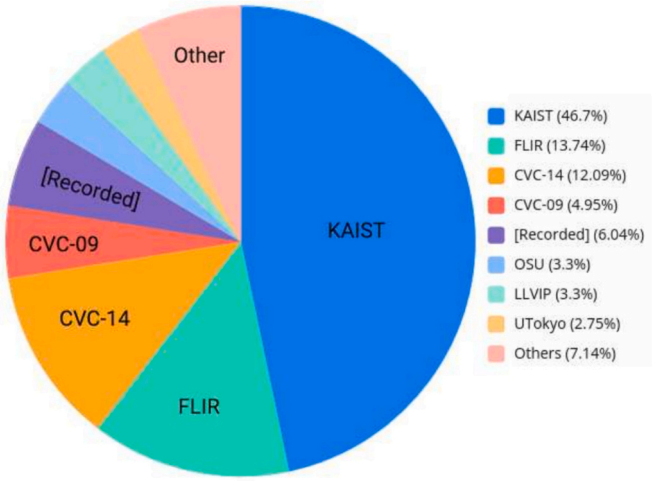

Fig. 9. Distribution of the datasets available for evaluation of nighttime pedestrian detection in different works.

features in DNNs, which cover many possible conditions in which pedestrians are challenging to detect. Additionally, the works are becoming more practical, providing the possibility of being used in real applications, and making domain-specific applications based on handcrafted or hybrid methods is not a practical solution.

Moreover, in a more detailed chart, Fig. 8 shows the distribution of papers surveyed regarding the introduced sub-categories. It can be seen that most of the papers have targeted *image fusion* techniques for nighttime pedestrian detection applications. Since thermal imaging (*i.e.,* long wavelength IR) can capture the infrared radiation from objects and are sensitive to temperature changes, thermal images provide clearer contours information of pedestrians under insufficient lighting conditions. However, the thermal IR modality lacks visual details such as texture, color, and precise edges of the objects, which can be captured by RGB sensors. In addition, the quality of visible images is significantly degraded under severe weather conditions, low resolution, and unfavorable lighting. Considering the characteristics of both visible and thermal sensors, cross-spectral fusion has become a promising alternative solution for overcoming the limitations of an unimodal approach to adapt to the all-weather and all-day situations. By fusing complementary visual features from multiple modalities, pedestrian detectors can be enhanced in stability, reliability, and perceptibility. Despite the great progress made in multi-spectral pedestrian detection, a large gap still exists between the current artificial vision systems and human vision ability. Among them, *halfway fusion* covers most of the works, and *late fusion* is the second preferred approach in *image fusion* subcategory. *Knowledge transfer* and *image specification* methodologies are the following trendy solutions according to the stats in the figure. It can also be seen that *generative models* methodologies have not absorbed massive attention among the papers published in recent years in the domain due to their ability to generate and employ realistic scenes and enrich the training process in challenging scenarios. They can simulate how visual features in low-light conditions are aimed and essential synthetic data. Moreover, *multi-task* methodologies are also gaining attention due to their potential for simultaneous module execution, joint optimization, and advancements in hardware architectures.

It should also be added that three types of deep learning-based architectures have been dedicated to achieving multi-spectral pedestrian detection, which can be categorized into the conventional CNN-based, Auto-Encoder (AE)-based, and GAN-based architectures. End-to-end CNN-based methods, which cover ninety of the studied works (*i.e.,* 69.23%), employ feature extraction, feature fusion, and image reconstruction processes through well-designed loss functions and network architectures. Conversely, the AE-based methods (17.69% of the papers) first train the encoder and decoder as the feature extractor and the image reconstructor, respectively. Then, the multi-image fusion process is accomplished according to the fusion rules. Finally, and in the GAN-based methods (*i.e.,* 9.23% of the studied works), the architecture is suitable for unsupervised pedestrian detection, relying on the adversarial mechanism between the generator and discriminator. The discriminator forces the generator to make the target distribution in the fused images as close as possible to the source images.

Table 2
Performance evaluation of state-of-the-art multi-spectral pedestrian detectors on KAIST test set, sorted based on their publication year. The superscripts *X*, *V*, *K*, *P*, *1*, *2*, and *3* represent NVIDIA GPU models used for evaluation, including *TitanX*, *Tesla V100*, *Tesla K40*, *Tesla P40*, *1080Ti*, *2080Ti*, and *3090Ti*, respectively.

Method	Published	Family	Backbone	Speed (*s/f*)
HAFNet [75]	2023	DL	ResNet-50	0.017^{1}
Dual-YOLO [68]	2023	”	ELAN	0.016^{3}
DSMN [99]	2023	”	CSPDarknet-53	0.76^{3}
Chan et al. [95]	2023	”	CSPDarknet-53	0.76^{3}
YOLO-CMN [55]	2022	”	CSPDarknet-53	0.02^{2}
Marnissi et al. [156]	2022	”	Darknet-53	0.019^{X}
RISNet [85]	2022	”	Custom CNN	0.1^{V}
BAANet [60]	2022	”	ResNet-50	0.07^{1}
ProbEn [96]	2022	”	ResNet-50 + FPN	0.025^{2}
Zou et al. [71]	2022	”	ResNet-50	0.04^{1}
MD [130]	2022	”	ResNet-18	0.007^{1}
DMFFNet [79]	2022	”	MobileNet v3	0.021^{2}
MSR [133]	2022	”	VGG-16	0.04^{1}
LG-FAPF [103]	2022	”	VGG-16	0.14^{X}
GAFF [61]	2021	”	VGG-16	0.009^{1}
Kim et al. [169]	2021	”	VGG-16	0.11^{1}
Ding et al. [92]	2021	”	VGG-16	0.071^{X}
IT-MN [98]	2021	”	MobileNet v2	0.03^{X}
MCFF [81]	2021	”	CSPDarknet-53	0.031^{P}
ASPFF Net [56]	2021	”	CSPDarknet-53	0.028^{1}
TC Det [113]	2020	”	Darknet-53	0.033^{1}
ResNet + FPN [63]	2020	”	ResNet-101	0.129^{X}
MB-Net [159]	2020	”	ResNet-50	0.07^{1}
Ding et al. [65]	2020	”	VGG-16	0.222^{X}
CIAN [78]	2019	”	VGG-16	0.066^{1}
HMFFN [72]	2019	”	VGG-16	0.026^{X}
AR-CNN [167]	2019	”	VGG-16	0.12^{1}
GFD-SSD [83]	2019	”	VGG-16	0.0512^{1}
IAF R-CNN [100]	2019	”	VGG-16	0.21^{X}
IATDNN+IASS [181]	2019	”	VGG-16	0.25^{X}
YOLO-TGB [108]	2018	”	Darknet-19	0.012^{1}
MSDS-RCNN [185]	2018	”	VGG-16	0.22^{X}
Park et al. [93]	2018	”	VGG-16	0.58^{X}
Halfway Fusion [58]	2016	”	VGG-16	0.43^{X}
CMT-CNN [199]	2017	Hybrid	VGG-16+ACF	0.59^{K}

5.2. Dataset trends

Regarding the datasets introduced in Section 3, the surveyed research works have been evaluated on various datasets. Thus, Fig. 9 depicts the distribution of the utilized datasets by the reviewed papers. It can be seen that most of the papers (*i.e.,* around half of them) have utilized KAIST dataset. The second and third datasets other research works use are FLIR and CVC-14, respectively. It should be mentioned that some of the research works (*i.e.,* $\sim$ 5.3 percent) prefer to evaluate their in-house collected, mainly collected from real-world scenarios. As introduced in Section 4, the differences in the physical characteristics of sensors lead to the misalignment of image pairs and have limited applicability in real-life situations. Despite the increasing number of visible-IR datasets in recent years, accessing instances with strictly aligned multi-spectral images is still a challenging problem. The benchmark datasets reported in the scientific literature can only provide information under certain scenes, most of which are recorded by a stationary camera. Therefore, there is a lack of datasets that contain a

Table 3
Miss Rate (MR) comparison of state-of-the-art multi-spectral pedestrian detectors in three subsets of the KAIST test set, *i.e.,* all-day, day-time, and night-time, sorted based on their publication year. The best and second-best results are boldfaced and underlined, respectively. Note that the lower MR is better.

Method	Published	Family	Category	Backbone	All-Day	Day-Time	Night-Time
INSANet [87]	2024	DL	Halfway-Fusion	VGG-16	5.50	6.29	4.20
Beyond Fusion [86]	2024	”	”	ELAN	5.01	**5.89**	3.27
HAFNet [75]	2023	”	”	ResNet-50	6.93	7.68	5.66
Yang et al. [52]	2023	”	”	ResNet-101	10.71	13.09	8.45
RISNet [85]	2022	”	”	Custom CNN	7.89	7.61	7.08
DMFFNet [79]	2022	”	”	MobileNet v3	9.26	12.79	5.17
Zou et al. [71]	2022	”	”	ResNet-50	7.77	9.41	**2.00**
BAANet [60]	2022	”	”	ResNet-50	7.92	8.37	6.98
YOLO-CMN [55]	2022	”	”	CSPDarknet-53	7.85	8.03	7.82
ASPFF Net [56]	2021	”	”	CSPDarknet-53	11.64	14.14	6.73
MCFF [81]	2021	”	”	CSPDarknet-53	**4.91**	6.23	2.90
GAFF [61]	2021	”	”	ResNet-18	7.93	9.79	4.33
ResNet-101 + FPN + Sum [63]	2020	”	”	ResNet-101	27.60	27.92	25.77
CS-RCNN [53]	2020	”	”	ResNet-50	11.43	11.86	8.82
CFR [54]	2020	”	”	VGG-16	6.13	7.68	3.19
Yadav et al. [76]	2020	”	”	VGG-16	29.00	26.00	32.00
Ding et al. [65]	2020	”	”	VGG-16	34	36	35
GFD-SSD [83]	2019	”	”	VGG-16	28.00	25.80	30.03
CIAN [78]	2019	”	”	VGG-16	27.71	30.74	21.07
Halfway Fusion [58]	2016	”	”	VGG-16	36.99	36.84	35.49
DSMN [99]	2023	DL	Late-Fusion	CSPDarknet-53	14.33	13.34	22.36
MS-DETR [88]	2023	”	”	ResNet-50+ResNet-18	6.13	7.78	**3.18**
ProbEn [96]	2022	”	”	ResNet-50+FPN	7.66	9.07	4.89
LG-FAPF [103]	2022	”	”	VGG-16	**5.12**	**5.83**	3.69
Ding et al. [92]	2021	”	”	VGG-16	32	34	34
IT-MN [98]	2021	”	”	MobileNet v2	14.19	14.30	13.98
IAF R-CNN [100]	2019	”	”	VGG-16	15.73	14.55	18.26
Park et al. [93]	2018	”	”	VGG-16	31.36	31.79	30.82
CMM [166]	2024	DL	Modality-Imbalance	VGG-16	8.54	9.60	**5.93**
Das et al. [160]	2023	”	”	PVT	**7.41**	**7.69**	7.03
Dasgupta et al. [163]	2022	”	”	ResNeXt-50	9.23	9.33	8.97
MB-Net [159]	2020	”	”	ResNet-50	8.13	8.28	7.86
Wanchaitanawong et al. [170]	2021	DL	Position-Shift	VGG-16	9.67	10.69	9.24
Kim et al. [169]	2021	”	”	VGG-16	**8.45**	**9.39**	**7.39**
AR-CNN [167]	2019	”	”	VGG-16	9.34	9.94	8.38
Kim et al. [168]	2019	”	”	ResNet-50	42.89	42.42	43.65
BU (VLT, T) [115]	2021	DL	Domain-Adaptation	Darknet-53	**25.61**	**32.69**	10.87
TC-Det [113]	2020	”	”	Darknet-53	27.11	34.81	**10.31**
VGG-16-two-stage [112]	2019	”	”	VGG-16	46.30	53.37	31.63
Marnissi et al. [127]	2022	DL	Unsupervised Domain-Adaptation	ResNet-101	44.60	50.29	28.79
UTL [124]	2022	”	”	VGG-16	**19.98**	**22.17**	**15.78**
Feature-Map Fusion [122]	2021	”	”	VGG-16	23.09	24.55	17.74
U-TS-RPN [125]	2019	”	”	VGG-16	36.42	37.15	33.00
MSR [133]	2022	DL	Memory-Network	ResNet-101	**10.32**	**13.28**	**6.23**
Kim et al. [132]	2021	”	”	VGG-16	19.16	24.70	8.26
IATDNN+IASS [181]	2019	DL	Multi-Task	VGG-16	26.37	27.29	24.41
MSDS-RCNN [185]	2018	”	”	VGG-16	**11.63**	**10.60**	**13.73**
MD [130]	2022	DL	Knowledge-Distillation	ResNet-18	**8.03**	**9.85**	**4.84**
DCRL-PDN [131]	2021	”	”	VGG-16	25.89	27.01	23.82
LS-GAN [150]	2021	DL	I2I-Translation	Darknet-53	25.62	31.86	12.92
YOLO-TGB [108]	2018	DL	Transfer-Learning	Darknet-19	31.2	34.7	23.1
Ghose et al. [153]	2019	DL	Saliency-Maps	VGG-16	–	30.4	21.0
Song et al. [188]	2020	DL	Other	ResNet-50	–	12.23	4.56
Chen et al. [202]	2021	Hybrid	–	Darknet-53	**43.25**	46.99	**35.84**
Kim et al. [195]	2018	”	–	VGG-16	45.36	**41.30**	55.82
CMT-CNN [199]	2017	”	–	VGG-16+ACF	49.55	47.30	54.78
Choi et al. [201]	2016	”	–	VGG-16+ACF	47.31	49.31	43.75

sufficient variety of fine-grained annotated samples taken from a moving camera, as environments can change dynamically.

5.3. Performance evaluations

Considering the performance of the surveyed works, Table 2 analyzes the computational efficiency of the state-of-the-art methods on KAIST test set. It should be noted that they have been evaluated using dissimilar hardware mentioned in the caption of the table. According to the table, the MD [130] method takes only 0.007 seconds to process a single image. The main reason for such performance is due to the theory of knowledge distillation, which accelerates inference by transferring the knowledge learned from a high thermal-resolution model to a low one. Another interesting result is the GAFF [61] model, which requires only 0.009 seconds of inference time. The primary indication for such performance is that the GAFF only includes three convolution layers, so the total number of learnable parameters and the computational cost is low. There are some approaches with performances negligibly less than these approaches, including YOLO-TGB [108] with 0.012, Dual-YOLO [68] with 0.016, HAFNet [75] with 0.017, and Marnissi et al. [156] with 0.019 seconds to process a single image. On the other hand, CMT-CNN [199] as a hybrid approach is the most computationally intensive methodology, with 0.59 seconds to process a single image. The main reason that can be attributed to such low performance is the time-

consuming process of using ACF proposals during the test. Comparing the elapsed times in a frame-based approach using similar hardware in the table implies the significance of design choices and optimization methodologies while employing/merging appropriate algorithms to achieve proper performance. Accordingly, minimizing the computational cost while keeping the hardware and speed trade-off in a reasonable range is directly related to the practicality of the designed system in real-world applications. A clear example can be found in GAFF [61] and Kim et al. [169], where using the same hardware and backbone on the same dataset samples lead to huge difference. Keeping the architecture of GAFF as simple as possible to prioritize the performance over a generalized solution made it much faster than Kim et al.'s method that tackles miscalibration and modality discrepancy challenges, while the latter is primarily designed for sophisticated conditions.

Moreover, Table 3 shows the detection accuracy in evaluating different approaches. The results are reported in terms of MR under *Reasonable* settings, and the approaches are classified according to the categories presented in Section 4. As shown in the Table, the MCFF [81] ranks first as a halfway-fusion strategy in overall performance on the KAIST by a large margin. The main reason for such performance is due to the MCFF transferring the fusion information from the bottom to the top at different stages. It can be observed that in the *Reasonable* nighttime criteria, the MCFF [81] obtains superior results than its daytime experiment. The reason is that the MCFF uses the illumination information to learn the fusion weights. Similarly, LG-FAPF [103] as a late-fusion strategy performs remarkably better compared to the other detectors. The primary reason for such performance is a locality-guided pixel-level fusion scheme that aggregates the human-related features in complementary modalities to integrate the prediction confidence scores in color and thermal channels. Among these methods, only four methodologies (*i.e.,* CMT-CNN [199], Kim et al. [195], Choi et al. [201], and Chen et al. [202]) are hybrid approaches, which witnessed a significant drop in MR. It can be concluded that the hybrid approaches are not highly applicable to around-the-clock applications, and specifications are required.

As the final discussion, it is essential to note that by expanding the use of fully autonomous vehicles and robots, the challenges of correct and real-time detecting pedestrians under various scenarios are becoming inevitable. Accordingly, explainable and interpretable mechanisms to exploit why a system failed/succeeded in a scenario can bring about more public reliability and confidence among people, including pedestrians, for interacting with autonomous systems. Thus, tailoring the current methodologies with the field of Explainable AI (xAI) is another direction to be investigated by researchers.

6. Conclusions

The paper in hand provided a comprehensive survey of pedestrian detection approaches tailored to low-light conditions, addressing a crucial challenge in computer vision, surveillance, and autonomous driving. The accurate and reliable recognition and tracking of pedestrians under reduced visibility is of paramount importance for enhancing the safety of autonomous vehicles and preventing accidents. The survey has examined a wide array of methodologies, including deep learning-based, feature-based, and hybrid approaches, which have demonstrated promising results in improving pedestrian detection performance in challenging lighting scenarios. By delving into the current landscape of low-light pedestrian detection, this work contributes to advancing more secure and dependable autonomous driving systems and other applications related to pedestrian safety. It has also identified ongoing research directions in the field and highlighted potential zones that warrant further research and investigation. The insights provided in this paper aim to inform and inspire future work, ultimately driving innovation and progress in the domain of pedestrian detection under adverse conditions.

CRediT authorship contribution statement

Bahareh Ghari: Writing – original draft, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Ali Tourani:** Writing – review & editing, Writing – original draft, Validation, Supervision. **Asadollah Shahbahrami:** Writing – review & editing, Validation. **Georgi Gaydadjiev:** Writing – review & editing.

Declaration of competing interest

None.

Data availability

No data was used for the research described in the article.

References

[1] Azzedine Boukerche, Mingzhi Sha, Design guidelines on deep learning–based pedestrian detection methods for supporting autonomous vehicles, ACM Comp. Surveys (CSUR) 54 (6) (2021) 1–36. ACM New York, NY, USA.
[2] Mingzhi Sha, Azzedine Boukerche, Performance evaluation of CNN-based pedestrian detectors for autonomous vehicles, Ad Hoc Netw. 128 (2022) 102784. Elsevier.
[3] Sangjun Kim, Sooyeong Kwak, Byoung Chul Ko, Fast pedestrian detection in surveillance video based on soft target training of shallow random forest, IEEE Access 7 (2019) 12415–12426. IEEE.
[4] Oluwakorede Monica Oluyide, Jules-Raymond Tapamo, Tom Mmbasu Walingo, Automatic dynamic range adjustment for pedestrian detection in thermal (Infrared) surveillance videos, Sensors 22 (5) (2022) 1728. MDPI.
[5] Gabriel Oltean, Laura Ivanciu, Horea Balea, Pedestrian detection and behaviour characterization for video surveillance systems, in: 2019 IEEE 25th International Symposium for Design and Technology in Electronic Packaging (SIITME), IEEE, 2019, pp. 256–259.
[6] Meiyuan Zou, Yu Jiajie, Lu Bo, Wenzheng Chi, Lining Sun, Active pedestrian detection for excavator robots based on multi-sensor fusion, in: 2022 IEEE International Conference on Real-Time Computing and Robotics (RCAR), IEEE, 2022, pp. 255–260.
[7] Zhe Zhao, Xianyu Qi, Yufei Zhao, Jiadong Zhang, Wei Wang, Xiangdong Yang, Pedestrian Detection and Tracking Based on 2D Lidar and RGB-D Camera, in: Proceedings of the 2022 3rd International Conference on Control, Robotics and Intelligent System 7–14, 2022.
[8] Lei Pang, Zhiqiang Cao, Yu Junzhi, Shuang Liang, Xuechao Chen, Weimin Zhang, An efficient 3D pedestrian detector with calibrated RGB camera and 3D LiDAR, in: 2019 IEEE International Conference on Robotics and Biomimetics (ROBIO), IEEE, 2019, pp. 2902–2907.
[9] Ujwalla Gawande, Kamal Hajari, Yogesh Golhar, Pedestrian Detection and Tracking in Video Surveillance System: Issues, Comprehensive Review, and Challenges, Recent Trends in Computational Intelligence, IntechOpen, 2020, pp. 1–24.
[10] Wenguang Wang, Xiyuan Chang, Jihuang Yang, Gaofei Xu, LiDAR-based dense pedestrian detection and tracking, Appl. Sci. 12 (4) (2022) 1799. MDPI.
[11] Bahareh Ghari, Ali Tourani, Shahbahrami Asadollah, A robust pedestrian detection approach for autonomous vehicles, in: Proceedings of the 2022 8th Iranian Conference on Signal Processing and Intelligent Systems, IEEE, 2022, pp. 1–5.
[12] Jiang Yu, Guoxiang Tong, Henan Yin, Naixue Xiong, A pedestrian detection method based on genetic algorithm for optimize XGBoost training parameters, IEEE Access 7 (2019) 118310–118321. IEEE.
[13] Guofa Li, Yifan Yang, Qu Xingda, Deep learning approaches on pedestrian detection in hazy weather, IEEE Trans. Ind. Electron. 67 (10) (2019) 8889–8899. IEEE.
[14] Luis Barba-Guaman, Jose Eugenio Naranjo, Anthony Ortiz, Deep learning framework for vehicle and pedestrian detection in rural roads on an embedded GPU, Electronics 9 (4) (2020) 589. MDPI.
[15] Goon Li Hung, Mohamad Safwan Bin Sahimi, Hussein Samma, Tarik Adnan Almohamad, Badr Lahasan, Faster R-CNN deep learning model for pedestrian detection from drone images, SN Comp. Sci. 1 (2020) 1–9. Springer.
[16] Zahid Ahmed, R. Iniyavan, et al., Enhanced vulnerable pedestrian detection using deep learning, in: 2019 International Conference on Communication and Signal Processing (ICCSP), IEEE, 2019, pp. 0971–0974.
[17] Long Chen, Shaobo Lin, Xiankai Lu, Dongpu Cao, Hangbin Wu, Chi Guo, Chun Liu, Fei-Yue Wang, Deep neural network based vehicle and pedestrian detection for autonomous driving: a survey, IEEE Trans. Intell. Transp. Syst. 22 (6) (2021) 3234–3246. IEEE.
[18] Andreas Geiger, Philip Lenz, Raquel Urtasun, Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2012.
[19] Ya-Li Hou, Yaoyao Song, Xiaoli Hao, Yan Shen, Manyi Qian, Houjin Chen, Multispectral pedestrian detection based on deep convolutional neural networks, Infrared Phys. Technol. 94 (2018) 69–77. Elsevier.

[20] Sundas Iftikhar, Zuping Zhang, Muhammad Asim, Ammar Muthanna, Andrey Koucheryavy, Ahmed A. Abd El-Latif, Deep learning-based pedestrian detection in autonomous vehicles: substantial issues and challenges, Electronics 11 (21) (2022) 3551. MDPI.

[21] Wei Chen, Yuxuan Zhu, Zijian Tian, Fan Zhang, Minda Yao, Occlusion and multi-scale pedestrian detection: a review, Array (2023) 100318. Elsevier.

[22] Fang Li, Xueyuan Li, Qi Liu, Zirui Li, Occlusion handling and multi-scale pedestrian detection based on deep learning: a review, IEEE Access 10 (2022) 19937–19957. IEEE.

[23] James W. Davis, Mark A. Keck, A two-stage template approach to person detection in thermal imagery, in: 2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05)-Volume 1, 1, IEEE, 2005, pp. 364–369.

[24] Atousa Torabi, Guillaume Massé, Guillaume-Alexandre Bilodeau, An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications, Comput. Vis. Image Underst. 116 (**2**) (2012) 210–221. Elsevier.

[25] Yainuvis Socarrás, Sebastian Ramos, David Vázquez, Antonio M. López, Theo Gevers, adapting pedestrian detection from synthetic to far infrared images, in: ICCV Workshops 3, 2013.

[26] Daniel Olmeda, Cristiano Premebida, Urbano Nunes, Jose Maria Armingol, Arturo de la Escalera, Pedestrian detection in far infrared images, Integrat. Comp. Aided Eng. 20 (**4**) (2013) 347–360. IOS Press.

[27] Zheng Wu, Nathan Fuller, Diane Theriault, Margrit Betke, A thermal infrared video benchmark for visual analysis, Proc. IEEE Conf. Comp. Vision Pattern Recog. Workshops 201–208 (2014).

[28] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, In So Kweon, Multispectral pedestrian detection: Benchmark dataset and baseline, in: Proceedings of the IEEE conference on computer vision and pattern recognition 1037–1045, 2015.

[29] Piniarski Karol, Pawlowski Pawel, Dabrowski Adam, Video processing algorithms for detection of pedestrians, CMST **21** (**3**) (2015) 141–150. PSNC, Poznan Supercomputing and Networking Center.

[30] Alejandro González, Zhijie Fang, Yainuvis Socarras, Joan Serrat, David Vázquez, Jiaolong Xu, Antonio M. López, Pedestrian detection at day/night time with visible and FIR cameras: a comparison, Sensors 16 (6) (2016) 820. MDPI.

[31] Mira Jeong, Byoung Chul Ko, Jae-Yeal Nam, Early detection of sudden pedestrian crossing for safe driving during summer nights, IEEE Trans. Circuits Syst. Video Technol. 27 (**6**) (2016) 1368–1380. IEEE.

[32] Karasawa Takumi, Kohei Watanabe, Qishen Ha, Antonio Tejero-De-Pablos, Yoshitaka Ushiku, Tatsuya Harada, Multispectral object detection for autonomous vehicles, Proc. Thematic Workshops ACM Multimedia 35–43 (2017) 2017.

[33] Evan Gebhardt, Marilyn Wolf, Camel dataset for visual and thermal infrared multiple object detection and tracking, in: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, 2018, pp. 1–6.

[34] Lukáš Neumann, Michelle Karg, Shanshan Zhang, Christian Scharfenberger, Eric Piegert, Sarah Mistr, Olga Prokofyeva, Robert Thiel, Andrea Vedaldi, Andrew Zisserman, et al., Nightowls: A pedestrians at night dataset, in: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part I 14, Springer, 2019, pp. 691–705.

[35] Zhewei Xu, Jiajun Zhuang, Qiong Liu, Jingkai Zhou, Shaowu Peng, Benchmarking a large-scale FIR dataset for on-road pedestrian detection, Infrared Phys. Technol. 96 (2019) 199–208. Elsevier.

[36] Taehwan Kim, Sungho Kim, Pedestrian detection at night time in FIR domain: comprehensive study about temperature and brightness and new benchmark, Pattern Recogn. **79** (2018) 44–54. Elsevier.

[37] FLIR Thermal Dataset for Algorithm Training, `https://www.flir.in/oem/adas/adas-dataset-form/`, ***TELEDYNE FLIR***, TELEDYNE FLIR, 2021.

[38] Paulius Tumas, Adam Nowosielski, Arturas Serackis, Pedestrian detection in severe weather conditions, IEEE Access 8 (2020) 62775–62784. IEEE.

[39] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, Wenli Zhou, LLVIP: A visible-infrared paired dataset for low-light vision, in: Proceedings of the IEEE/CVF International Conference on Computer Vision 3496–3504, 2021.

[40] Muhammad Ali Farooq, Waseem Shariff, Peter Corcoran, Evaluation of Thermal Imaging on Embedded GPU Platforms for Application in Vehicular Assistance Systems, *arXiv preprint*. *arXiv:2201.01661*, 2022.

[41] Adam Nowosielski, Krzysztof Małecki, Paweł Forczmański, Anton Smoliński, Kazimierz Krzywicki, Embedded night-vision system for pedestrian detection, IEEE Sensors J. 20 (16) (2020) 9293–9304. IEEE.

[42] JongBae Kim, Pedestrian detection and distance estimation using thermal camera in night time, in: 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), IEEE, 2019, pp. 463–466.

[43] Jong Bae Kim, Detection of direction indicators on road surfaces using inverse perspective mapping and NN, J. Inf. Proc. Korean 4 (2015) 201–208.

[44] Dongmei Zhou, Shi Qiu, Song Yang, Kaijian Xia, A pedestrian extraction algorithm based on a single infrared image, Infrared Phys. Technol. 105 (2020) 103236. Elsevier.

[45] Anouar Ben Khalifa, Ihsen Alouani, Mohamed Ali Mahjoub, Najoua Essoukri Ben Amara, Pedestrian detection using a moving camera: A novel framework for foreground detection, Cogn. Syst. Res. **60** (2020) 77–96. Elsevier.

[46] Ali Raza Shahzad, Ahmad Jalal, A smart surveillance system for pedestrian tracking and counting using template matching, in: 2021 International Conference on Robotics and Automation in Industry (ICRAI), IEEE, 2021, pp. 1–6.

[47] Yingfeng Cai, Ze Liu, Hai Wang, Xiaoqiang Sun, Saliency-based pedestrian detection in far infrared images, IEEE Access 5 (2017) 5013–5019. IEEE.

[48] Hao Zhang, Han Xu, Xin Tian, Junjun Jiang, Jiayi Ma, Image fusion meets deep learning: a survey and perspective, Inform. Fusion 76 (2021) 323–336. Elsevier.

[49] Jason Nataprawira, Gu Yanlei, Igor Goncharenko, Shunsuke Kamijo, Pedestrian detection on multispectral images in different lighting conditions, in: 2021 IEEE International Conference on Consumer Electronics (ICCE), IEEE, 2021, pp. 1–5.

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

[51] Jason Nataprawira, Gu Yanlei, Igor Goncharenko, Shunsuke Kamijo, Pedestrian detection using multispectral images and a deep neural network, Sensors **21** (**7**) (2021) 2536. Multidisciplinary Digital Publishing Institute.

[52] Yang Yang, Kaixiong Xu, Kaizheng Wang, Cascaded information enhancement and cross-modal attention feature fusion for multispectral pedestrian detection, Front. Phys. **11** (2023) 1121311. Frontiers.

[53] Yongtao Zhang, Zhishuai Yin, Linzhen Nie, Song Huang, Attention based multi-layer fusion of multispectral images for pedestrian detection, IEEE Access 8 (2020) 165071–165084. IEEE.

[54] Heng Zhang, Elisa Fromont, Sébastien Lefevre, Bruno Avignon, Multispectral fusion for object detection with cyclic fuse-and-refine blocks, in: 2020 IEEE International Conference on Image Processing (ICIP), IEEE, 2020, pp. 276–280.

[55] Qunyan Jiang, Juying Dai, Ting Rui, Faming Shao, Jinkang Wang, Guanlin Lu, Attention-based cross-modality feature complementation for multispectral pedestrian detection, IEEE Access 10 (2022) 53797–53809. IEEE.

[56] Fu Lei, Gu Wen-bin, Yong-bao Ai, Wei Li, Dong Wang, Adaptive spatial pixel-level feature fusion network for multispectral pedestrian detection, Infrared Phys. Technol. 116 (2021) 103770. Elsevier.

[57] Qing Deng, Wei Tian, Lu Yuyao Huang, Xin Bi Xiong, Pedestrian detection by Fusion of RGB and infrared images in low-light environment, in: 2021 IEEE 24th International Conference on Information Fusion (FUSION), IEEE, 2021, pp. 1–8.

[58] Jingjing Liu, Shaoting Zhang, Shu Wang, Dimitris N. Metaxas, Multispectral deep neural networks for pedestrian detection, arXiv preprint. *arXiv:1611.02644*, 2016.

[59] Yue Cao, Junchi Bin, Jozsef Hamari, Erik Blasch, Liu Zheng, Multimodal object detection by channel switching and spatial attention, Proc. IEEE/CVF Conf. Comp. Vision Pattern Recog. 403–411 (2023).

[60] Xiaoxiao Yang, Yeqiang Qian, Huijie Zhu, Chunxiang Wang, Ming Yang, BAANet: Learning bi-directional adaptive attention gates for multispectral pedestrian detection, in: 2022 International Conference on Robotics and Automation (ICRA), IEEE, 2022, pp. 2920–2926.

[61] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, Bruno Avignon, Guided attentive feature fusion for multispectral pedestrian detection, Proc. IEEE/CVF Winter Conf. Appl. Comp. Vision 72–80 (2021).

[62] Fang Qingyun, Han Dapeng, Wang Zhaokui, *Cross-modality fusion transformer for multispectral object detection*, *arXiv preprint*. *arXiv:2111.00273*, 2021.

[63] Dashun Pei, Mingxuan Jing, Huaping Liu, Fuchun Sun, Linhua Jiang, A fast RetinaNet fusion framework for multi-spectral pedestrian detection, Infrared Phys. Technol. 105 (2020) 103178. Elsevier.

[64] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, Piotr Dollár, Focal loss for dense object detection, Proc. IEEE Int. Conf. Comp. Vision 2980–2988 (2017).

[65] Ding Lu, Yong Wang, Robert Laganiere, Dan Huang, Shan Fu, Convolutional neural networks for multispectral pedestrian detection, Signal Process. Image Commun. 82 (2020) 115764. Elsevier.

[66] Jifeng Dai, Yi Li, Kaiming He, Jian Sun, R-FCN: object detection via region-based fully convolutional networks, Adv. Neural Inf. Proces. Syst. 29 (2016).

[67] Jun-Seok Yun, Seon-Hoo Park, Seok Bong Yoo, Infusion-Net: inter-and intra-weighted cross-fusion network for multispectral object detection, Mathematics **10** (**21**) (2022) 3966. MDPI.

[68] Chun Bao, Jie Cao, Qun Hao, Cheng Yang, Yaqian Ning, Tianhua Zhao, Dual-YOLO architecture from infrared and visible images for object detection, Sensors 23 (6) (2023) 2934. MDPI.

[69] Chien-Yao Wang, Alexey Bochkovskiy, Hong-Yuan Mark Liao, YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 7464–7475, 2023.

[70] Xingyi Zhou, Dequan Wang, Philipp Krähenbühl, Objects as points. arXiv 2019, arXiv preprint. *arXiv:1904.07850* **448**, 1904.

[71] Xin Zuo, Zhi Wang, Jifeng Shen, Wankou Yang, Improving multispectral pedestrian detection with scale-aware permutation attention and adjacent feature aggregation, IET Comput. 17 (2022) 726–738. Wiley Online Library.

[72] Yanpeng Cao, Dayan Guan, Yulun Wu, Jiangxin Yang, Yanlong Cao, Michael Ying Yang, Box-level segmentation supervised deep neural networks for accurate and real-time multispectral pedestrian detection, ISPRS J. Photogramm. Remote Sens. 150 (2019) 70–79. Elsevier.

[73] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, Jiayi Ma, PIAFusion: a progressive infrared and visible image fusion network based on illumination aware, Inform. Fusion 83 (2022) 79–92. Elsevier.

[74] Kamil Roszyk, Michał, R. Nowicki, Piotr Skrzypczyński, Adopting the YOLOv4 architecture for low-latency multispectral pedestrian detection in autonomous driving, Sensors 22 (3) (2022) 1082. MDPI.

[75] Peiran Peng, Tingfa Xu, Bo Huang, Jianan Li, HAFNet: hierarchical attentive Fusion network for multispectral pedestrian detection, Remote Sens. 15 (8) (2023) 2041. MDPI.

[76] Yadav Ravi, Ahmed Samir, Hazem Rashed, Senthil Yogamani, Rozenn Dahyot, CNN based color and thermal image fusion for object detection in automated driving, Irish Machine Vision Image Proc. (2) (2020).

[77] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, Faster R-CNN: towards real-time object detection with region proposal networks, Adv. Neural Inf. Proces. Syst. 28 (2015).

[78] Zhang Lu, Zhiyong Liu, Shifeng Zhang, Xu Yang, Hong Qiao, Kaizhu Huang, Amir Hussain, Cross-modality interactive attention network for multispectral pedestrian detection, Inform. Fusion 50 (2019) 20–29. Elsevier.

[79] Ruizhe Hu, Ting Rui, Yan Ouyang, Jinkang Wang, Qunyan Jiang, Yinan Du, DMFFNet: dual-mode multi-scale feature fusion-based pedestrian detection method, IEEE Access (2022), 1–1 IEEE.

[80] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al., Searching for MobileNetV3, in: Proceedings of the IEEE/CVF International Conference on Computer Vision 1314–1324, 2019.

[81] Zhiwei Cao, Huihua Yang, Juan Zhao, Shuhong Guo, Lingqiao Li, Attention fusion for one-stage multispectral pedestrian detection, Sensors 21 (**12**) (2021) 4184. MDPI.

[82] Jaekyum Kim, Jaehyung Choi, Yechol Kim, Junho Koh, Chung Choo Chung, Jun Won Choi, Robust camera lidar sensor fusion via deep gated information fusion network, in: 2018 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2018, pp. 1620–1625.

[83] Zheng Yang, Izzat H. Izzat, Shahrzad Ziaee, *GFD-SSD: gated fusion double SSD for multispectral pedestrian detection, arXiv preprint*. *arXiv:1903.06999*, 2019.

[84] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Fu Cheng-Yang, Alexander C. Berg, SSD: Single Shot Multibox Detector, Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, 21–37, Springer, 2016.

[85] Qingwang Wang, Yongke Chi, Tao Shen, Jian Song, Zifeng Zhang, Yan Zhu, Improving RGB-infrared object detection by reducing cross-modality redundancy, Remote Sens. **14** (**9**) (2020) 2022. MDPI.

[86] Qian Xie, Ta-Ying Cheng, Jia-Xing Zhong, Kaichen Zhou, Andrew Markham, Niki Trigoni, Beyond Fusion, Modality hallucination-based multispectral fusion for pedestrian detection, Proc. IEEE/CVF Winter Conf. Appl. Comp. Vision 655–664 (2024).

[87] Sangin Lee, Taejoo Kim, Jeongmin Shin, Namil Kim, Yukyung Choi, INSANet: INtra-INter spectral attention network for effective feature fusion of multispectral pedestrian detection, Sensors 24 (4) (2024) 1168. MDPI.

[88] Yinghui Xing, Song Wang, Guoqiang Liang, Qingyi Li, Xiuwei Zhang, Shizhou Zhang, Yanning Zhang, Multispectral Pedestrian Detection via Reference Box Constrained Cross Attention and Modality Balanced Optimization, arXiv preprint. *arXiv:2302.00290*, 2023.

[89] Bushra Khalid, Asad Mansoor Khan, Muhammad Usman Akram, Sherin Batool, Person detection by fusion of visible and thermal images using convolutional neural network, in: 2019 2nd International Conference on Communication, Computing and Digital Systems (C-CODE), IEEE, 2019, pp. 143–148.

[90] Bryan Montenegro, Marco Flores-Calero, Pedestrian detection at daytime and nighttime conditions based on YOLO-v5, Ingenius. Rev. Ciencia Tecnol. 27 (2022) 85–95.

[91] Xiaoru Song, Song Gao, Chaobo Chen, A multispectral feature fusion network for robust pedestrian detection, Alexandria Eng. J. 60 (**1**) (2021) 73–85. Elsevier.

[92] Ding Lu, Yong Wang, Robert Laganiere, Dan Huang, Xinbin Luo, Huanlong Zhang, A robust and fast multispectral pedestrian detection deep network, Knowl.-Based Syst. 227 (2021) 106990. Elsevier.

[93] Kihong Park, Seungryong Kim, Kwanghoon Sohn, Unified multi-spectral pedestrian detection based on probabilistic fusion networks, Pattern Recogn. 80 (2018) 143–155. Elsevier.

[94] Alexey Bochkovskiy, Chien-Yao Wang, Hong-Yuan Mark Liao, *Yolov4: Optimal speed and accuracy of object detection, arXiv preprint*. *arXiv:2004.10934*, 2020.

[95] Hung-Tse Chan, Po-Ting Tsai, Chih-Hsien Hsia, Multispectral pedestrian detection via two-stream YOLO with complementarity Fusion for autonomous driving, in: 2023 IEEE 3rd International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB), IEEE, 2023, pp. 313–316.

[96] Yi-Ting Chen, Jinghao Shi, Zelin Ye, Christoph Mertz, Deva Ramanan, Shu Kong, Multimodal object detection via probabilistic ensembling, in: European Conference on Computer Vision, Springer, 2022, pp. 139–158.

[97] Zuhaib Ahmed Shaikh, David Van Hamme, Peter Veelaert, Wilfried Philips, Probabilistic fusion for pedestrian detection from thermal and colour images, Sensors **22** (**22**) (2022) 8637. MDPI.

[98] Yifan Zhuang, Pu Ziyuan, Jia Hu, Yinhai Wang, Illumination and temperature-aware multispectral networks for edge-computing-enabled pedestrian detection, IEEE Trans. Network Sci. Eng. 9 (**3**) (2021) 1282–1295. IEEE.

[99] Chih-Hsien Hsia, Hsiao-Chu Peng, Hung-Tse Chan, All-weather pedestrian detection based on double-stream multispectral network, Electronics 12 (**10**) (2023) 2312. MDPI.

[100] Chengyang Li, Dan Song, Ruofeng Tong, Min Tang, Illumination-aware faster R-CNN for robust multispectral pedestrian detection, Pattern Recogn. 85 (2019) 161–171. Elsevier.

[101] Guofa Li, Weijian Lai, Qu Xingda, Pedestrian detection based on light perception fusion of visible and thermal images, Opt. Laser Technol. 156 (2022) 108466. Elsevier.

[102] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, You Only Look Once, Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 779–788, 2016.

[103] Yanpeng Cao, Xing Luo, Jiangxin Yang, Yanlong Cao, Michael Ying Yang, Locality guided cross-modal feature aggregation and pixel-level fusion for multispectral pedestrian detection, Inform. Fusion **88** (2022) 1–11. Elsevier.

[104] Alexander Wolpert, Michael Teutsch, M. Saquib Sarfraz, Rainer Stiefelhagen, Anchor-free small-scale multispectral pedestrian detection, arXiv preprint. *arXiv:2008.08418*, 2020.

[105] Taeheon Kim, Sangyun Chung, Damin Yeom, Yu Youngjoon, Hak Gu Kim, Yong Man Ro, *MSCoTDet: Language-driven Multi-modal Fusion for Improved Multispectral Pedestrian Detection, arXiv preprint*. *arXiv:2403.15209*, 2024.

[106] Hu Jinda, Yanshun Zhao, Xindong Zhang, Application of transfer learning in infrared pedestrian detection, in: 2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC), IEEE, 2020, pp. 1–4.

[107] Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE international conference on computer vision 2223–2232, 2017.

[108] Maarten Vandersteegen, Kristof Van Beeck, Toon Goedemé, Real-time multispectral pedestrian detection with a single-pass deep neural network, in: Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings 15, Springer, 2018, pp. 419–426.

[109] Joseph Redmon, Ali Farhadi, YOLO9000: Better, faster, stronger, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 7263–7271 (2017).

[110] Shengqi Geng, Infrared image pedestrian target detection based on yolov3 and migration learning, arXiv preprint. *arXiv:2012.11185*, 2020.

[111] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, Dongwei Ren, Distance-IoU loss: faster and better learning for bounding box regression, Proc. AAAI Conf. Artific. Intel. 34 (07) (2020) 12993–13000.

[112] Tiantong Guo, Cong Phuoc Huynh, Mashhour Solh, Domain-adaptive pedestrian detection in thermal images, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 1660–1664.

[113] My Kieu, Andrew D. Bagdanov, Marco Bertini, Alberto Del Bimbo, Task-conditioned domain adaptation for pedestrian detection in thermal imagery, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16, 546–562, Springer, 2020.

[114] My Kieu, Andrew D. Bagdanov, Marco Bertini, Alberto Del Bimbo, Domain adaptation for privacy-preserving pedestrian detection in thermal imagery, in: Image Analysis and Processing–ICIAP 2019: 20th International Conference, Trento, Italy, September 9–13, 2019, Proceedings, Part II 20, Springer, 2019, pp. 203–213.

[115] My Kieu, Andrew D. Bagdanov, Marco Bertini, Bottom-up and layerwise domain adaptation for pedestrian detection in thermal images, ACM Trans. Multimedia Comp. Commun. Appl. (TOMM) 17 (1) (2021) 1–19. ACM New York, NY.

[116] Mate Krišto, Marina Ivasic-Kos, Miran Pobar, Thermal object detection in difficult weather conditions using YOLO, IEEE Access **8** (2020) 125459–125476. IEEE.

[117] Kevin Fritz, Daniel König, Ulrich Klauck, Michael Teutsch, Generalization ability of region proposal networks for multispectral person detection, Automat. Target Recog. XXIX 10988 (2019) 222–235. SPIE.

[118] Piotr Dollar, Christian Wojek, Bernt Schiele, Pietro Perona, Pedestrian detection: an evaluation of the state of the art, IEEE Trans. Pattern Anal. Mach. Intell. 34 (4) (2011) 743–761. IEEE.

[119] Shanshan Zhang, Rodrigo Benenson, Bernt Schiele, Citypersons: a diverse dataset for pedestrian detection, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 3213–3221 (2017).

[120] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, Tatsuya Harada, MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes, in: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2017, pp. 5108–5115.

[121] Vibashan Vs, Domenick Poster, Suya You, Shuowen Hu, Vishal M. Patel, Meta-uda: Unsupervised domain adaptive thermal object detection using meta-learning, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 1412–1423, 2022.

[122] Chengjin Lyu, Patrick Heyer, Asad Munir, Ljiljana Platisa, Christian Micheloni, Bart Goossens, Wilfried Philips, Visible-Thermal pedestrian detection via unsupervised transfer learning, in: 2021 the 5th International Conference on Innovation in Artificial Intelligence, 2021, pp. 158–163.

[123] Farzeen Munir, Shoaib Azam, Moongu Jeon, Sstn: self-supervised domain adaptation thermal object detection for autonomous driving, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2021, pp. 206–213.

[124] Chengjin Lyu, Patrick Heyer, Bart Goossens, Wilfried Philips, An unsupervised transfer learning framework for visible-thermal pedestrian detection, Sensors 22 (**12**) (2022) 4416. Mdpi.

[125] Yanpeng Cao, Dayan Guan, Weilin Huang, Jiangxin Yang, Yanlong Cao, Qiao Yu, Pedestrian detection with unsupervised multispectral feature learning using deep neural networks, Inform. Fusion 46 (2019) 206–217. Elsevier.

[126] Dayan Guan, Xing Luo, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, George Vosselman, Michael Ying Yang, Unsupervised domain adaptation for multispectral pedestrian detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 0–0, 2019.

[127] Mohamed Amine Marnissi, Hajer Fradi, Anis Sahbani, Najoua Essoukri Ben Amara, Unsupervised thermal-to-visible domain adaptation method for pedestrian detection, Pattern Recogn. Lett. 153 (2022) 222–231. Elsevier.

[128] Mazin Hnewa, Alireza Rahimpour, Justin Miller, Devesh Upadhyay, Hayder Radha, cross modality knowledge distillation for robust pedestrian detection in low light and adverse weather conditions, in: ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.

[129] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, Distilling the knowledge in a neural network, arXiv preprint. *arXiv:1503.02531*, 2015.

[130] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, Bruno Avignon, Low-cost multispectral scene analysis with modality distillation, Proc. IEEE/CVF Winter Conf. Appl. Comp. Vision 803–812 (2022).

[131] Tianshan Liu, Kin-Man Lam, Rui Zhao, Guoping Qiu, Deep cross-modal representation learning and distillation for illumination-invariant pedestrian detection, IEEE Trans. Circuits Syst. Video Technol. 32 (1) (2021) 315–329. IEEE.

[132] Jung Uk Kim, Sungjune Park, Yong Man Ro, Robust small-scale pedestrian detection with cued recall via memory learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3050–3059.

[133] Jung Uk Kim, Sungjune Park, Yong Man Ro, Towards versatile pedestrian detector with multisensory-matching and multispectral recalling memory, Proc. AAAI Conf. Artif. Intell. 36 (1) (2022) 1157–1165.

[134] Mohamed Amine Marnissi, Hajer Fradi, Anis Sahbani, Najoua Essoukri Ben Amara, Thermal image enhancement using generative adversarial network for pedestrian detection, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 6509–6516.

[135] Yue Sun, Yifeng Shao, Guanglin Yang, Haiyan Xie, A method of infrared image pedestrian detection with improved YOLOv3 algorithm, Am. J. Optics Photon. 9 (**3**) (2021) 32–38. Science Publishing Group.

[136] Jiahui Yu, Yuchen Fan, Jianchao Yang, Ning Xu, Zhaowen Wang, Xinchao Wang, Thomas Huang, Wide activation for efficient and accurate image super-resolution, arXiv preprint. *arXiv:1808.08718*, 2018.

[137] Mohamed Amine Marnissi, Abir Fathallah, GAN-based vision transformer for high-quality thermal image enhancement, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 817–825.

[138] Linfeng Tang, Xinyu Xiang, Hao Zhang, Meiqi Gong, Jiayi Ma, DIVFusion: darkness-free infrared and visible image fusion, Inform. Fusion 91 (2023) 477–493. Elsevier.

[139] Gang Li, Shanshan Zhang, Jian Yang, Nighttime pedestrian detection based on feature attention and transformation, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 9180–9187.

[140] Chenhang Cui, Jinyu Xie, Yechenhao Yang, Bright Channel Prior Attention for Multispectral Pedestrian Detection, arXiv preprint. *arXiv:2305.12845*, 2023.

[141] Yunfan Chen, Hyunchul Shin, Pedestrian detection at night in infrared images using an attention-guided encoder-decoder convolutional neural network, Appl. Sci. 10 (**3**) (2020) 809. MDPI.

[142] Heena Patel, Kalpesh Prajapati, Anjali Sarvaiya, Kishor Upla, Kiran Raja, Raghavendra Ramachandra, Christoph Busch, Depthwise convolution for compact object detector in nighttime images, Proc. IEEE/CVF Conf. Comp. Vision Pattern Recog. 379–389 (2022).

[143] Shasha Li, Yongjun Li, Yao Li, Mengjun Li, Xiaorong Xu, Yolo-firi: improved yolov5 for infrared image object detection, IEEE Access 9 (2021) 141861–141875. IEEE.

[144] Hui Li, Xiao-Jun Wu, DenseFuse: a fusion approach to infrared and visible images, IEEE Trans. Image Process. 28 (5) (2018) 2614–2623. IEEE.

[145] Yingxue Pang, Jianxin Lin, Tao Qin, Zhibo Chen, Image-to-image translation: methods and applications, IEEE Trans. Multimedia **24** (2021) 3859–3881. IEEE.

[146] Fuya Luo, Yunhan Li, Guang Zeng, Peng Peng, Gang Wang, Yongjie Li, Thermal infrared image colorization for nighttime driving scenes with top-down guided attention, IEEE Trans. Intell. Transp. Syst. 23 (9) (2022) 15808–15823. IEEE.

[147] Anagha Dangle, Radhika Mundada, Sonal Gore, Jeenisha Shrungare, Harish Dalal, Enhanced colorization of thermal images for pedestrian detection using deep convolutional neural networks, Proc. Comp. Sci. 218 (2023) 2091–2101. Elsevier.

[148] Shihao Yang, Min Sun, Xiayin Lou, Hanjun Yang, Hang Zhou, An unpaired thermal infrared image translation method using GMA-CycleGAN, Remote Sens. 15 (**3**) (2023) 663. MDPI.

[149] Chaitanya Devaguptapu, Ninad Akolekar, Manuj M. Sharma, Vineeth N. Balasubramanian, Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops 0–0, 2019.

[150] My Kieu, Lorenzo Berlincioni, Leonardo Galteri, Marco Bertini, Andrew D. Bagdanov, Alberto Del Bimbo, Robust pedestrian detection in thermal imagery using synthesized images, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 8804–8811.

[151] Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, Multi-class generative adversarial networks with the L2 loss function, arXiv preprint. *arXiv:1611.04076*.

[152] Fatih Altay, Senem Velipasalar, The use of thermal cameras for pedestrian detection, IEEE Sensors J. 22 (12) (2022) 11489–11498. IEEE.

[153] Debasmita Ghose, Shasvat M. Desai, Sneha Bhattacharya, Deep Chakraborty, Madalina Fiterau, Tauhidur Rahman, Pedestrian detection in thermal images using saliency maps, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops 0–0, 2019.

[154] Nian Liu, Junwei Han, Ming-Hsuan Yang, Picanet: learning pixel-wise contextual attention for saliency detection, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 3089–3098 (2018).

[155] Zijun Deng, Xiaowei Hu, Lei Zhu, Xuemiao Xu, Jing Qin, Guoqiang Han, Pheng-Ann Heng, R3net: Recurrent residual refinement network for saliency detection, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence 684–690, 2018. AAAI Press Menlo Park, CA, USA.

[156] Mohamed Amine Marnissi, Ikram Hattab, Hajer Fradi, Anis Sahbani, Najoua Essoukri Ben Amara, Bispectral Pedestrian Detection Augmented with Saliency Maps using Transformer, *VISIGRAPP (5: VISAPP)*, 2022, pp. 275–284.

[157] Yifan Zhao, Jingchun Cheng, Wei Zhou, Chunxi Zhang, Xiong Pan, Infrared pedestrian detection with converted temperature map, in: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2019, pp. 2025–2031.

[158] Kemal Oksuz, Baris Can Cam, Sinan Kalkan, Emre Akbas, Imbalance problems in object detection: a review, IEEE Trans. Pattern Anal. Mach. Intell. **43** (**10**) (2020) 3388–3415. IEEE.

[159] Kailai Zhou, Linsen Chen, Xun Cao, Improving multispectral pedestrian detection by addressing modality imbalance problems, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16, Springer, 2020, pp. 787–803.

[160] Arindam Das, Sudip Das, Ganesh Sistu, Jonathan Horgan, Ujjwal Bhattacharya, Edward Jones, Martin Glavin, Revisiting Modality Imbalance In Multimodal Pedestrian Detection, arXiv preprint arXiv:2302.12589, 2023.

[161] Leonard Gross, Logarithmic sobolev inequalities, Am. J. Math. 97 (**4**) (1975) 1061–1083. JSTOR.

[162] Wei Li, Infrared image pedestrian detection via YOLO-V3, in: 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC) 5, IEEE, 2021, pp. 1052–1055.

[163] Kinjal Dasgupta, Arindam Das, Sudip Das, Ujjwal Bhattacharya, Senthil Yogamani, Spatio-contextual deep network-based multimodal pedestrian detection for autonomous driving, IEEE Trans. Intell. Transp. Syst. 23 (9) (2022) 15940–15950. IEEE.

[164] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al., Graph attention networks, stat 1050 (20) (2017) 10–48550.

[165] Sean Bell, C. Lawrence Zitnick, Kavita Bala, Ross Girshick, Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2874–2883, 2016.

[166] Taeheon Kim, Sebin Shin, Youngjoon Yu, Hak Gu Kim, Yong Man Ro, Causal Mode Multiplexer: A Novel Framework for Unbiased Multispectral Pedestrian Detection, arXiv preprint. *arXiv:2403.01300*, 2024.

[167] Zhang Lu, Xiangyu Zhu, Xiangyu Chen, Yang Xu, Zhen Lei, Zhiyong Liu, Weakly aligned cross-modal learning for multispectral pedestrian detection, Proc. IEEE/CVF Int. Conf. Comp. Vision 5127–5137 (2019).

[168] Minsu Kim, Sunghun Joung, Kihong Park, Seungryong Kim, Kwanghoon Sohn, Unpaired cross-spectral pedestrian detection via adversarial feature learning, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 1650–1654.

[169] Jung Uk Kim, Sungjune Park, Yong Man Ro, Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection, IEEE Trans. Circuits Syst. Video Technol. 32 (3) (2021) 1510–1523. IEEE.

[170] Napat Wanchaitanawong, Masayuki Tanaka, Takashi Shibata, Masatoshi Okutomi, Multi-modal pedestrian detection with large misalignment based on modal-wise regression and multi-modal IoU, in: 2021 17th International Conference on Machine Vision and Applications (MVA), IEEE, 2021, pp. 1–6.

[171] Jun Yue, Leyuan Fang, Xia Shaobo, Deng Yue, Jiayi Ma, Dif-fusion: towards high color fidelity in infrared and visible image fusion with diffusion models, IEEE Trans. Image Process. 32 (2023) 5705–5720. IEEE.

[172] Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Xu Shuang, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, Luc Van Gool, DDFM: Denoising diffusion model for multi-modality image fusion, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 8082–8093.

[173] Ping Li, Junjie Chen, Binbin Lin, Xianghua Xu, Residual spatial fusion network for rgb-thermal semantic segmentation, arXiv preprint. *arXiv:2306.10364*, 2023.

[174] Hai Wang, Yansong Xu, Youguo He, Yingfeng Cai, Long Chen, Yicheng Li, Miguel Angel Sotelo, Zhixiong Li, YOLOv5-Fog: A multiobjective visual detection algorithm for fog driving scenes based on improved YOLOv5, in: IEEE Transactions on Instrumentation and Measurement 71, IEEE, 2022, pp. 1–12.

[175] Fangzheng Song, Peng Li, YOLOv5-MS: Real-time multi-surveillance pedestrian target detection model for smart cities, Biomimetics 8 (6) (2023) 480. MDPI.

[176] Yuanjie Chen, Chunyuan Wang, Chi Zhang, A robust lightweight network for pedestrian detection based on YOLOv5-x, Appl. Sci. **13** (18) (2023) 10225. MDPI.

[177] Kefu Yi, Kai Luo, Tuo Chen, Rongdong Hu, An improved YOLOX model and domain transfer strategy for nighttime pedestrian and vehicle detection, Appl. Sci. 12 (23) (2022) 12476. MDPI.

[178] Ivan Adriyanov Nikolov, Variational autoencoders for pedestrian synthetic data augmentation of existing datasets-a preliminary investigation, in: 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, SCITEPRESS Digital Library, 2024, pp. 829–836.

[179] Michael Crawshaw, Multi-task learning with deep neural networks: A survey, arXiv preprint. *arXiv:2009.09796*, 2020.

[180] Yuanzhi Wang, Tao Lu, Yanduo Zhang, Wenhua Fang, Yuntao Wu, Zhongyuan Wang, Cross-task feature alignment for seeing pedestrians in the dark, Neurocomputing **462** (2021) 282–293. Elsevier.

[181] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, Michael Ying Yang, Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection, Inform. Fusion 50 (2019) 148–157. Elsevier.

[182] Xiaobiao Dai, Junping Hu, Hongmei Zhang, Abubakar Shitu, Chunlei Luo, Ahmad Osman, Stefano Sfarra, Multi-task Faster R-CNN for nighttime pedestrian detection and distance estimation, Infrared Phys. Technol. 115 (2021) 103694. Elsevier.

[183] Zhiwei Cao, Huihua Yang, Juan Zhao, Xipeng Pan, Longhao Zhang, Zhenbing Liu, A new region proposal network for far-infrared pedestrian detection, IEEE Access **7** (2019) 135023–135030. IEEE.
[184] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, Serge Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition 2117–2125, 2017.
[185] Chengyang Li, Dan Song, Ruofeng Tong, Min Tang, *Multispectral pedestrian detection via simultaneous detection and segmentation*, *arXiv preprint. arXiv:1808.04818*, 2018.
[186] Yung-Yao Chen, Sin-Ye Jhong, Guan-Yi Li, Ping-Han Chen, Thermal-based pedestrian detection using faster r-cnn and region decomposition branch, in: 2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), IEEE, 2019, pp. 1–2.
[187] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, Yinan Yu, High-level semantic feature detection: a new perspective for pedestrian detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 5187–5196, 2019.
[188] Song Yu, Min Li, Xiaohua Qiu, Weidong Du, Jin Feng, Full-time infrared feature pedestrian detection based on CSP network, in: 2020 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), IEEE, 2020, pp. 516–518.
[189] Zhewei Xu, Chi-Man Vong, Chi-Chong Wong, Qiong Liu, Ground plane context aggregation network for day-and-night on vehicular pedestrian detection, IEEE Trans. Intell. Transp. Syst. 22 (10) (2020) 6395–6406. IEEE.
[190] Xiaobiao Dai, Yuxia Duan, Junping Hu, Shicai Liu, Caiqi Hu, Yunze He, Dapeng Chen, Chunlei Luo, Jianqiao Meng, Near infrared nighttime road pedestrians recognition based on convolutional neural network, Infrared Phys. Technol. **97** (2019) 25–32. Elsevier.
[191] Christopher M. Bishop, Nasser M. Nasrabadi, Pattern Recognition and Machine Learning 4, Springer, 2006, p. 4.
[192] Michelle A. Galarza-Bravo, Marco J. Flores-Calero, Pedestrian detection at night based on Faster R-CNN and far infrared images, in: Intelligent Robotics and Applications: 11th International Conference, ICIRA 2018, Newcastle, NSW, Australia, August 9–11, 2018, Proceedings, Part II, Springer, 2018, pp. 335–345.
[193] Rumi Kalita, Anjan Kumar Talukdar, Kandarpa Kumar Sarma, Real-time human detection with thermal camera feed using YOLOv3, in: 2020 IEEE 17th India Council International Conference (INDICON), IEEE, 2020, pp. 1–5.
[194] Koti Naga Renu Chebrolu, P.N. Kumar, Deep learning based pedestrian detection at all light conditions, in: 2019 International Conference on Communication and Signal Processing (ICCSP), IEEE, 2019, pp. 0838–0842.
[195] Jong Hyun Kim, Ganbayar Batchuluun, Kang Ryoung Park, Pedestrian detection based on faster R-CNN in nighttime by fusing deep convolutional features of successive images, in: Expert Systems with Applications **114**, 2018, pp. 15–33. Elsevier.
[196] Daniel Konig, Michael Adam, Christian Jarvers, Georg Layher, Heiko Neumann, Michael Teutsch, Fully convolutional region proposal networks for multispectral person detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops 49–56, 2017.
[197] Paulius Tumas, Artūras Jonkus, Artūras Serackis, Acceleration of HOG based pedestrian detection in FIR camera video stream, in: 2018 Open Conference of Electrical, Electronic and Information Sciences (eStream), IEEE, 2018, pp. 1–4.
[198] A. Narayanan, R. Darshan Kumar, R. RoselinKiruba, T. Sree Sharmila, Study and analysis of pedestrian detection in thermal images using YOLO and SVM, in: 2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), IEEE, 2021, pp. 431–434.
[199] Dan Xu, Wanli Ouyang, Elisa Ricci, Xiaogang Wang, Nicu Sebe, Learning cross-modal deep representations for robust pedestrian detection, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 5363–5371 (2017).
[200] Piotr Dollár, Ron Appel, Serge Belongie, Pietro Perona, Fast feature pyramids for object detection, IEEE Trans. Pattern Anal. Mach. Intell. 36 (**8**) (2014) 1532–1545. IEEE.
[201] Hangil Choi, Seungryong Kim, Kihong Park, Kwanghoon Sohn, Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks, in: 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE, 2016, pp. 621–626.
[202] Chen Xu, Lei Liu, Xin Tan, Robust pedestrian detection based on multi-spectral image fusion and convolutional neural networks, Electronics 11 (**1**) (2021) 1. MDPI.
[203] Yong Ma, Jun Chen, Chen Chen, Fan Fan, Jiayi Ma, Infrared and visible image fusion using total variation model, Neurocomputing 202 (2016) 12–19. Elsevier.