GEM

Glare or Gloom, I Can Still See You - End-to-End Multi-Modal Object Detection

Mazhar, Osama; Babuska, Robert; Kober, Jens

DOI
10.1109/LRA.2021.3093871

Publication date
2021

Document Version
Final published version

Published in
IEEE Robotics and Automation Letters

GEM: Glare or Gloom, I Can Still See You – End-to-End Multi-Modal Object Detection

Osama Mazhar, Robert Babuška, and Jens Kober

***Abstract*—Deep neural networks designed for vision tasks are often prone to failure when they encounter environmental conditions not covered by the training data. Single-modal strategies are insufficient when the sensor fails to acquire information due to malfunction or its design limitations. Multi-sensor configurations are known to provide redundancy, increase reliability, and are crucial in achieving robustness against asymmetric sensor failures. To address the issue of changing lighting conditions and asymmetric sensor degradation in object detection, we develop a multi-modal 2D object detector, and propose deterministic and stochastic sensor-aware feature fusion strategies. The proposed fusion mechanisms are driven by the estimated sensor measurement reliability values/weights. Reliable object detection in harsh lighting conditions is essential for applications such as self-driving vehicles and human-robot interaction. We also propose a new "r-blended" hybrid depth modality for RGB-D sensors. Through extensive experimentation, we show that the proposed strategies outperform the existing state-of-the-art methods on the FLIR-Thermal dataset, and obtain promising results on the SUNRGB-D dataset. We additionally record a new RGB-Infra indoor dataset, namely L515-Indoors, and demonstrate that the proposed object detection methodologies are highly effective for a variety of lighting conditions.**

***Index Terms*—computer vision for automation, deep learning for visual perception, object detection, RGB-D perception, segmentation and categorization, sensor fusion.**

I. Introduction

MODERN intelligent systems such as autonomous vehicles or assistive robots should have the ability to reliably detect objects in challenging real-world scenarios. Object detection is one of the widely studied problems in computer vision. It has been addressed lately by employing deep convolutional neural networks where the state-of-the-art methods have achieved fairly accurate detection performances on the existing datasets [1]–[4]. However, these vision models are fragile and do not generalize across realistic unconstrained scenarios, such as changing lighting conditions or other environmental circumstances which were not covered by the training data [5]. The failure of the detection algorithms in such conditions could lead to potentially catastrophic results, as in the case of self-driving vehicles.

One way of addressing this problem is to employ a data-augmentation strategy [6]. It refers to the technique of perturbing data without altering class labels, and it has been proven to greatly improve robustness and generalization performance [7]. Nevertheless, this is insufficient for the cases where the sensor fails to acquire information due to malfunction or its technical limitations. For example, the output of standard passive cameras degenerates with reduced ambient light, while thermal cameras or LiDARs are less affected by illumination changes.

Multi-sensor configurations are known to provide redundancy and often enhance the performance of the detection algorithms. Moreover, efficient sensor fusion strategies minimize uncertainties, increase reliability, and are crucial in achieving robustness against asymmetric sensor failures [8]. Although, increasing the number of sensors might enhance the performance of detection algorithms, this comes with a considerable computational and energy cost. This is not desirable in mobile robotic systems, which typically have constraints in terms of computational power and battery consumption. In such cases, intelligent choice and combination of sensors are crucial.

Furthermore, multi-modal data fusion often requires an estimate of the sensor signal uncertainty to guarantee efficient fusion and reliable prediction without a priory knowledge of the sensor characteristics [9]. The existing multi-modal object detection methods fuse the sensor data streams without explicitly modeling the measurement reliability. This may have severe consequences when the data from an individual sensor degrades or is missing due to sheer sensor failure.

To address the above problems, we propose sensor-aware multi-modal fusion strategies for object detection in harsh lighting conditions, thus the title "GEM: Glare or gloom, I can still see you - End-to-end Multimodal object detection". The output samples of GEM are shown in Figure 1. Two fusion methods are proposed: deterministic weighted fusion and stochastic feature fusion. In the deterministic weighted fusion, the measurement certainty of each sensor is estimated either by learning scalar weights or masks through separate neural networks. The learned weights are then assigned to the feature maps extracted from the feature extractor backbones for each sensor modality. The weighted feature maps can be fused either by averaging or concatenation. Moreover, we can visualize and interpret the

Manuscript received February 24, 2021; accepted June 8, 2021. Date of publication June 30, 2021; date of current version July 15, 2021. The research presented in this article was carried out as part of the OpenDR project, which has received funding from the European Unions Horizon 2020 Research and Innovation Programme under Grant Agreement 871449. *(Corresponding author: Osama Mazhar.)*

Osama Mazhar and Jens Kober are with the Cognitive Robotics Department, Delft University of Technology, Delft, CD 2628, The Netherlands (e-mail: o.mazhar@tudelft.nl; j.kober@tudelft.nl).

Robert Babuška is with the Cognitive Robotics Department, Delft University of Technology, Delft, CD 2628, The Netherlands, and also with the Czech Institute of Informatics, Robotics, and Cybernetics, Czech Technical University in Prague, Prague 16636, Czech (e-mail: r.babuska@tudelft.nl).

Digital Object Identifier 10.1109/LRA.2021.3093871

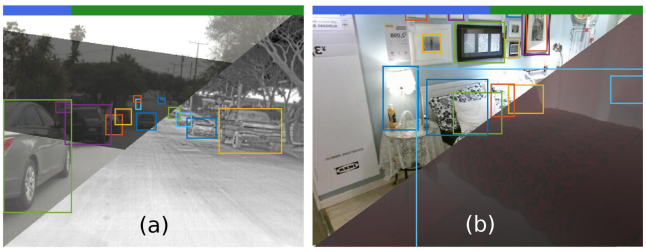

Fig. 1. Output samples of the proposed multi-modal object detector. The blue/green bar at the top illustrates the contribution/reliability of each sensor modality in obtaining the final output. Images from two modalities are merged diagonally only for illustration purposes. (a) Shows the results on the FLIR-Thermal dataset with RGB and thermal sensor modalities, (b) Shows the output on the SUNRGB-D dataset with RGB and our proposed "r-blended" hybrid depth modality.

measurement certainty of each sensor in the execution phase, which provides deeper insights into the relative strengths of each data stream. The stochastic feature fusion creates a one-hot encoding of the feature maps of each sensor, which can be assumed as a discrete switch that allows only the dominant/relevant features to pass. The obtained selected features are then concatenated before they are passed to the object detection and classification head. The proposed sensor-aware multi-modal object detector, referred to simply as GEM in the rest of the paper, is trained in an end-to-end fashion along with the fusion mechanism.

Most modern object detectors, including YOLO, Faster-RCNN and SSD employ many hand-crafted features such as anchor generation, rule-based assignment of classification and regression targets as well as weights to each anchor, and non-maximum suppression postprocessing. The overall performance of these methods often relies on careful tuning of the above-mentioned hyper-parameters. Following their success in sequence/language modeling, transformers have lately emerged in vision applications, outperforming competitive baselines and demonstrating a strong potential in this field. Therefore, we employ transformers in our work as in [10], which thanks to their powerful relational modeling capability eliminates the need of hand-crafted components in object detection. Our main contributions in this paper are:

- Evaluation of feature fusion in two configurations, i.e., deterministic weighted fusion and stochastic feature fusion for multi-modal object detection.
- Estimation of measurement reliability of each sensor as scalar or mask multipliers through separate neural networks for each modality to efficiently drive the deterministic weighted fusion.
- Use of transformers for multi-modal object detection to harness the efficacy of self-attention in sensor fusion.

II. Related Work

In this section, we first review deep learning-based object detection strategies, followed by a discussion on existing methods for multi-modal fusion methods in relevant tasks.

A. Deep Learning-Based Object Detection

Detailed literature surveys for deep learning-based object detectors have been published in [11], [12]. Here we briefly discuss some of the well-known object detection strategies. Typically, object detectors can be classified into two types, namely two-stage and singe-stage object detectors.

1) Two-Stage Object Detection: Two-stage object detectors exploit a region proposal network (RPN) in their first stage. RPN ranks region boxes, alias *anchors*, and proposes the ones that most likely contain objects as candidate boxes. The features are extracted by region-of-interest pooling (RoIPool) operation from each candidate box in the second stage. These features are then utilized for bounding-box regression and classification tasks.

2) Single-Stage Object Detection: Single-stage detectors propose predicted boxes from input images in one forward pass directly, without the region proposal step. Thus, this type of object detectors are time efficient and can be utilized for real-time operations. Lately, an end-to-end object detection strategy has been proposed in [10] that eliminates the need for hand-crafted components like anchor boxes and non-maximum suppression. The authors employ transformers in an encoder-decoder fashion, which have been extremely successful and become a de facto standard for natural language processing tasks. The transformer implicitly performed region proposals instead of using an R-CNN. The multi-head attention module in transformers jointly attended to different semantic regions of an image/feature maps and linearly aggregates the outputs through learnable weights. The learned attention maps can be visualized without requiring dedicated methods, as in the case of convolutional neural networks. The inherent non-sequential architecture of transformers allows parallelization of models. Thus, we opted to build upon the methodology of [10] for our multi-modal object detector for harsh lighting conditions.

B. Sensor Fusion

Sensor fusion strategies can be roughly divided into three types according to the level of abstraction where fusion is performed or in which order transformations are applied compared to feature combinations, namely low-level, mid-level, and high-level fusion [13]. In low-level or early fusion, raw information from each sensor is fused at pixel level, e.g., disparity maps in stereovision cameras [14]. In mid-level fusion, a set of features is extracted for each modality in a pre-processing stage, while multiple approaches [15] are exploited to fuse the extracted features. Late-fusion often employs a combination of two fusing methods, e.g., convolution of stacked feature maps followed by several fully connected layers with dropout regularization [16]. In high-level fusion or ensemble learning methods, predictions are obtained individually for each modality and the learnt scores or hypotheses are subsequently combined via strategies such as weighted majority votes [17]. Deep fusion or cross fusion [18] is another type of fusion strategy which repeatedly combines inputs, then transforms them individually. In each repetition, the transformation learns different features. For example in [8], features from the layers of VGG network are exchanged among

all modalities driven by sensor entropy after each pooling operation.

C. Multi-Sensor Object Detection

Most of the efforts on multi-modal object detection in the literature are focused on pedestrian or vehicle detection in the automotive context. Sensor fusion strategies are typically proposed for camera-LiDAR, camera-radar, and camera-radar-LiDAR setups. Here, we briefly go through the relevant state-of-the-art methods.

The authors in [8] proposed an entropy-steered multi-modal deep fusion architecture for adverse weather conditions. The sensor modalities exploited in their method include RGB camera, gated camera (NIR band), LiDAR, and radar. Instead of employing BeV projection or point cloud representation for LiDAR, the authors encoded depth, height, and pulse intensity on an image plane. Moreover, the radar output was also projected onto an image plane parallel to the image horizontal dimension. Considering the radar output invariant along the vertical image axis, the scan was replicated across the horizontal image axis. They utilized a modified VGG architecture for feature extraction, while features were exchanged among all modalities driven by sensor entropy after each pooling operation. Fused feature maps from the last 6 layers of the feature extractors were passed to the SSD object detection head. In [19], the authors proposed a pseudo multi-modal object detector from thermal IR images in a Faster-RCNN setting. The features from ResNet-50 backbones for the two modalities are concatenated and a 1×1 convolution is applied to the concatenated features before they are passed to the rest of Faster-RCNN network. They exploited I2I translation networks, namely CycleGAN [20] and UNIT [21] to transform thermal images from the FLIR Thermal [22] and KAIST [23] datasets to the RGB domain, thus the names MM-CG and MM-UNIT.

Here, we also discuss some fusion strategies which were originally proposed for applications other than object detection but are relevant to our work. In [24], the authors proposed a sensor fusion methodology for RGB and depth images to steer a self-driving vehicle. The latent semantic vector from an encoder-decoder segmentation network trained on RGB images was fused with the depth features. The fusion architecture proposed by [24] is similar to the gating mechanism driven by the learned scalar weights presented in [25]. The method proposed in [26] is closest to our work. The authors proposed two sensor fusion strategies for Visual-Inertial Odometry (VIO), namely soft fusion and hard fusion. In soft fusion, they learned soft masks which were subsequently assigned to each element in the feature vector. Hard fusion employed a variant of the Gumbel-max trick, which is often used to sample discrete data from categorical distributions. Learning masks equal to the size of feature vectors might introduce computational overhead. Therefore, we learn dynamic scalar weights for each sensor modality, which adapt to the environmental/lighting conditions. These scalar weights represent the reliability or relevance of the sensor signals. Moreover, we also learn single-channel masks with a spatial size equal to that of the feature maps obtained from the feature extractor backbone. Nevertheless, we also implement the Gumbel-Softmax trick for comparison, as a stochastic feature fusion for multi-modal object detectors.

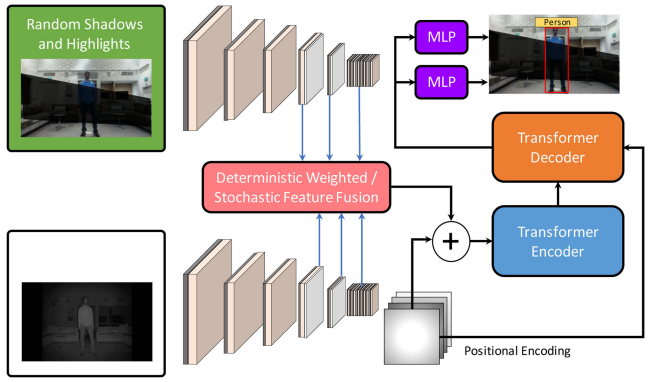

Fig. 2. Our proposed pipeline for a multi-modal object detector with transformers. The features from each backbone are fused and passed to the transformer encoder-decoder network. The decoder output is subsequently exploited by Multilayer Perceptrons (MLPs) for bounding box regression and object classification.

III. Sensor-Aware Multi-Modal Fusion

In this work, we propose a new method for sensor-aware feature selection and multi-modal fusion for object detection. We actually evaluate feature fusion in two configurations, i.e., deterministic weighted fusion with scalar and mask multipliers, and stochastic feature fusion driven by the Gumbel-Softmax trick that enables sampling from a discrete distribution. The overall pipeline of the proposed multi-modal object detector is illustrated in Figure 2. The proposed methodologies are trained and evaluated on datasets with RGB and thermal or depth images. However, it can be extended to include data from other sensors like LiDAR or radar, either by projecting the sensor output onto an image plane as proposed in [8] or by employing sensor-specific feature extractors such as [27].

A. Deterministic Weighted Fusion

The proposed deterministic weighted fusion scheme is conditioned on the measurement certainty of each sensor. These values are obtained either by learning scalar weights or masks through separate neural networks. Subsequently, the weights are assigned to the feature maps extracted from the backbones as (scalar or mask) multipliers for each sensor modality. Given the output of the backbone feature extractors $\mathbf{s}$ for a single modality, the neural network f optimizes parameters $\boldsymbol{\theta}$ to obtain measurement certainty w of the corresponding sensor as described as follows:

$$w = f(\mathbf{s}, \boldsymbol{\theta}) \times \frac{1}{\text{rows} \times \text{cols} \times k} \sum_{l=1}^{\text{rows}} \sum_{m=1}^{\text{cols}} \sum_{n=1}^{k} \mathbf{s}(l, m, n) \quad (1)$$

where k is the selected number of channels. The network f learns the parameters $\boldsymbol{\theta}$ in an end-to-end fashion. In the case of sensor degradation, the output of the neurons in the early layers of the corresponding backbone will remain close to zero. Thus,

we multiply the output of the network f by the mean of first k feature maps, 16 in our case, from $\mathbf{s}$ in a feed-forward setting to obtain w. This allows f to dynamically condition its output to changing lighting/sensor degradation scenarios, which subsequently guides the transformers to focus on the dominant sensor modality for object detection. Furthermore, the multiplication of the raw output of f with the mean of the selected feature maps is performed without gradient calculation to prevent the distortion of the feature maps in the back-propagation phase.

The weighted feature maps are fused by either taking an average of the two feature sets, or by concatenating them. The fused features are then passed to the transformer for object detection and localization. Our scalar fusion functions g_{sa} (averaging) and g_{sc} (concatenating) are represented as:

$$\begin{aligned} g_{\text{sa}}(\mathbf{s}_{\text{RGB}}, \mathbf{s}_{\text{IR}}) &= \phi(w_{\text{RGB}} \odot \mathbf{s}_{\text{RGB}}, w_{\text{IR}} \odot \mathbf{s}_{\text{IR}}) \\ g_{\text{sc}}(\mathbf{s}_{\text{RGB}}, \mathbf{s}_{\text{IR}}) &= [w_{\text{RGB}} \odot \mathbf{s}_{\text{RGB}}; w_{\text{IR}} \odot \mathbf{s}_{\text{IR}}] \end{aligned} \tag{2}$$

where ϕ denotes the mean operation, $\mathbf{s}_{\text{RGB}}$ and $\mathbf{s}_{\text{IR}}$ are feature maps obtained from the backbone feature extractor for RGB and thermal/IR imagers respectively, while w_{RGB} and w_{IR} are the sensor measurement certainty weights obtained through Equation (1). Similar to the scalar fusion method, feature selection is also modelled by learning masks for each modality, in this case $\mathbf{m}_{\text{RGB}}$ and $\mathbf{m}_{\text{IR}}/\mathbf{m}_{\text{depth}}$, with a spatial size equal to that of the features maps. The fusion scheme with mask multipliers is represented as:

$$\begin{aligned} g_{\text{ma}}(\mathbf{s}_{\text{RGB}}, \mathbf{s}_{\text{IR}}) &= \phi(\mathbf{m}_{\text{RGB}} \odot \mathbf{s}_{\text{RGB}}, \mathbf{m}_{\text{IR}} \odot \mathbf{s}_{\text{IR}}) \\ g_{\text{mc}}(\mathbf{s}_{\text{RGB}}, \mathbf{s}_{\text{IR}}) &= [\mathbf{m}_{\text{RGB}} \odot \mathbf{s}_{\text{RGB}}; \mathbf{m}_{\text{IR}} \odot \mathbf{s}_{\text{IR}}] \end{aligned} \tag{3}$$

B. Stochastic Feature Fusion

In addition to the weighted fusion schemes, we exploit a variant of the Gumbel-max trick to learn a one-hot encoding that either propagates or blocks each component of the feature maps for intelligent fusion. The Gumbel-max resampling strategy allows to draw discrete samples from a categorical distribution during the forward pass through a neural network. It exploits the reparametrization trick to separate out the deterministic and stochastic parts of the sampling process. However, it adds Gumbel noise instead of that from a normal distribution, which is actually used to model the distribution of the maximums for samples taken from other distributions. Gumbel-max then employs the $\arg\max$ function to find the class that has the maximum value for each sample.

Considering α be the n-dimensional probability variable conditioned for every row on each channel of the feature volume such that $\alpha = [\pi_1, \ldots, \pi_n]$, representing the probability of each feature at location n, the Gumbel-max trick can be represented by the following equation:

$$Q = \arg\max_i (\log \pi_i + G_i) \tag{4}$$

where, Q is a categorical variable with class probabilities $\pi_1, \pi_2, \ldots, \pi_n$ and $\{G_i\}_{i \leq n}$ is an i.i.d. sequence of standard Gumbel random variables which is given by:

$$G = -\log(-\log(U)), \quad U \sim \text{Uniform}[0, 1] \tag{5}$$

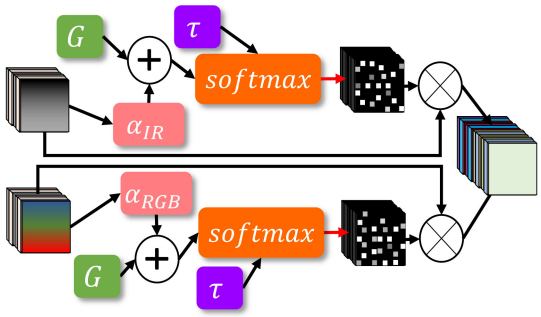

Fig. 3. Illustration of our stochastic feature fusion strategy that employs the Gumbel-Softmax sampling trick.

The use of $\arg\max$ makes the Gumbel-max trick *non*-differentiable. However, it can be replaced by *Softmax* with a temperature factor τ, thus making it a fully-differentiable resampling method [28]. Softmax with temperature parameter τ can be represented as:

$$f_\tau(x)_i = \frac{\exp(x_i/\tau)}{\Sigma_{j=1}^{n} \exp(x_j/\tau)} \tag{6}$$

where τ determines how closely the Gumbel-Softmax distribution matches the categorical distribution. With low temperatures, e.g., $\tau = 0.1$ to $\tau = 0.5$, the expected value of a Gumbel-Softmax random variable approaches the expected value of a categorical random variable [28]. The Gumbel-Softmax resampling function can therefore be written as

$$Q_i^\tau = f_\tau(\log \pi_i + G_i) = \frac{\exp((\log \pi_i + G_i)/\tau)}{\Sigma_{j=1}^{n} \exp((\log \pi_j + G_j)/\tau)} \tag{7}$$

with $i = 1, \ldots, n$.

We set $\tau = 1$ and obtain feature volume approximate one-hot categorical encodings for each modality $\mathbf{e}_{\mathbf{RGB}}$ and $\mathbf{e}_{\mathbf{IR}}$. Then a Hadamard product is taken between the encodings and the feature volumes and the resultants are subsequently concatenated and passed on to the bounding box regressor and classification head. We illustrate our selective fusion process developed for multi-modal object detector in Figure 3, while the selective fusion function g_{sf} is given as follows

$$g_{\text{sf}}(\mathbf{s}_{\text{RGB}}, \mathbf{s}_{\text{IR}}) = [\mathbf{e}_{\text{RGB}} \odot \mathbf{s}_{\text{RGB}}; \mathbf{e}_{\text{IR}} \odot \mathbf{s}_{\text{IR}}]. \tag{8}$$

IV. Experiments

A. Datasets

Three datasets are utilized in the training and evaluation of GEM, including the FLIR Thermal, SUNRGB-D [29] and a new L515-Indoor dataset that we recorded for this research. The FLIR Thermal dataset provides 8862 training and 1366 test samples of thermal and RGB images recorded in the streets and highways in Santa Barbara, California, USA. Only the thermal images in the dataset are annotated with four classes, i.e., *People*, *Bicycle*, *Car* and *Dog*. The given RGB images in the dataset are neither annotated nor aligned with their thermal counterparts, while the camera matrices are also not provided. Thus, to utilize this dataset in a multi-modal setting, the given RGB images

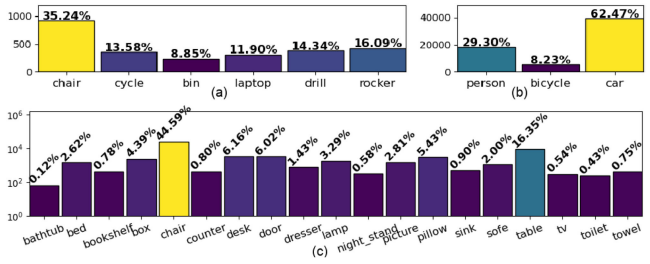

Fig. 4. Class distributions of the datasets (a) L515-Indoor (b) FLIR-Thermal (c) SUNRGB-D. The number of annotations in (c) are presented in the logarithmic scale.

must be annotated or aligned with their corresponding thermal images. One way to address this problem is to create artificial RGB images from input thermal images through GANs or similar neural networks as performed in [19]. However, we opted to employ the concept of homography by manually selecting matching features in multiple RGB and thermal images. The selected feature points are then employed to estimate a transformation matrix between the two camera modalities. RGB images are subsequently transformed with the estimated homography matrix such that they approximately align with their thermal equivalents. The *Dog* class constitutes only 0.29% of all the annotations in the aligned FLIR-Thermal dataset, thus it is not included in our experiments.

The SUNRGB-D dataset contains 10,335 RGB-D images taken by Kinect v1, Kinect v2, Intel RealSense, and Asus Xtion cameras. The annotations provided consist of 146,617 2D polygons and 64,595 3D bounding boxes, while 2D bounding boxes are obtained by projecting the coordinates of 3D bounding boxes onto the image plane. Although the dataset contains labels for approximately 800 objects, we evaluate our method on the selected 19 objects similar to [29]. We first divide the dataset into three subsets such that the training set consists of 4255 images, the validation set has 5050 images, while the test set contains 1059 images.

The L515-Indoor dataset provides 482 training and 207 validation RGB and IR images recorded with Intel RealSense L515 camera with various ambient light conditions in an indoor scene. It contains annotations of 1819 2D bounding boxes of 6 object categories in total. The IR images are aligned with their RGB counterparts through a homography matrix which is computed in a similar fashion as explained for the FLIR-Thermal dataset. The population distributions of the datasets are illustrated in Figure 4.

B. Pre-Processing Sensor Outputs

For the FLIR-Thermal and L515-Indoor datasets, aligned RGB and thermal/IR images are fed into our feature extractor backbones without any pre-processing. However, techniques that exploit datasets with depth images including [29] often apply HHA encoding [30] on the depth sensor modality for early feature extraction prior to being fed into the neural networks. HHA is a geocentric embedding for depth images that encodes horizontal disparity, height above ground, and angle with gravity for each pixel. In a multi-threading setup on a 12-Core Intel Core$^{\text{TM}}$ i7-9750H CPU, HHA encoding of a batch of 32 images takes approximately 119 seconds, which is far from its application in real-time object detection or segmentation tasks.

To address this problem, assuming that we are working with RGB and depth modalities, we create a new hybrid image that introduces scene texture in a depth image. As the red light is scattered the least by air molecules, we blend the depth images and the red image channels through a blend weight α. Thus, we name our hybrid depth image as "r-blended" depth image.

$$\text{img}_{\text{r-blended}} = \alpha\,\text{img}_{\text{depth}} + (1-\alpha)\,\text{img}_{\text{red}} \tag{9}$$

The value of α is set to 0.9 for depth images while the weight value for the red channels becomes 0.1. This is to make sure that when the neural network is trained with "r-blended depth" image, it should focus on learning the depth features while information from the red channel only complements the raw depth map. The idea to blend the red channel is also supported by the fact that CMOS cameras are often more sensitive to green and red light. We first train our multi-modal object detector on RGB and HHA encoded depth images. Later, we fine-tune the trained model by replacing HHA encoded images with r-blended depth images and achieve comparable results in terms of detection accuracy, while the fine-tuned model can indeed be used for real-time multi-modal object detection.

C. Training

GEM is trained with scalar fusion and mask fusion methods, i.e., g_{sa}, g_{sc}, g_{ma} and g_{mc} for deterministic weighted fusion driven by Equations (2) and (3), while it is also trained with g_{sf} for stochastic feature fusion. The backbone feature extractors for both sensor streams and the transformer block are pre-trained on MSCOCO dataset on RGB images as in [10]. For the FLIR thermal dataset, each model is trained on a cluster with 2 GPUs for 100 epochs while the models for the SUNRGB-D dataset are trained with 4 GPUs for 300 epochs. Similarly, the models for L515-Indoor are trained on a cluster with 2 GPUs for 300 epochs with a batch size of 1. The batch size for the FLIR-Thermal and SUNRGBD datasets is set to 2, while the learning rate for the feature extractor backbones, fusion networks, and transformer block is set to 8×10^{-6} for all datasets. We employ ResNet-50 as the feature extractor, while we also train g_{sc} on MobileNet v2 [31] for the FLIR thermal dataset. To guide the fusion process and mimic harsh lighting conditions for the RGB sensor, we also employ Random Shadows and Highlights (RSH) data augmentation as proposed in [7]. RSH develops immunity against lighting perturbations in the convolutional neural networks, which is desirable for real world applications. We additionally implement SSD512 object detector with VGG16 backbones in a multi-modal setting in two configurations, i.e., a simple averaging fusion as the baseline method (SSD-BL) and a weighted averaging fusion scheme (SSD-WA) similar to g_{sa}. The anchor/default boxes are configured for both SSD-BL and SSD-WA in a fashion similar to that for the MS-COCO dataset. These models are trained for 800 epochs in a single GPU setup with a batch size 1 and a learning rate 1×10^{-4} which decays with a decaying factor 0.2 after the first 520,000 iterations.

D. Evaluation

FLIR-Thermal: performance evaluation of the proposed networks on the FLIR-Thermal dataset is shown in Table II. We show Average Precision (AP) values at Intersection over Union (IoU) of 0.5 for each dominant class, while the mean Average Precision (mAP) is also estimated with and without lighting perturbations. These lighting corruptions are introduced by creating Random Shadows and Highlights (RSH) on the test RGB images. The evaluation with lighting perturbation is performed for 10 trials in all experiments, while the average of the obtained mAP is shown in the table. The results are compared with the single modality object detector, the multi-modal baseline fusion networks, and the existing state-of-the-art methods on this dataset. In the baselines, the features from the backbones are fused in two configurations: averaged and concatenated, without any weighing or re-sampling mechanism. Additionally, we compare the performance of SSD-BL and SSD-WA on the FLIR-Thermal dataset. It is clear from the evaluation results, that our proposed methodologies, i.e., g_{sa}, g_{sc}, g_{ma}, g_{mc}, and g_{sf}, outperform the previously reported results on this dataset. Our methods also demonstrate robustness against lighting perturbation, while a significant performance drop of single modality and baseline methods can be seen when tested with RSH. The "avg-baseline" obtained comparable results, but as it is only a blind fusion, hence no sensor contribution or reliability measure can be obtained with this methodology. Additionally, its performance can significantly drop in the case of asymmetric sensor failure. This can partially be observed in Table II where the baselines are tested with RSH perturbations. Concerning the evaluation of multi-modal SSD on the FLIR-Thermal dataset, SSD-WA certainly improves the results compared to SSD-BL, specifically in terms of robustness against lighting perturbations introduced by RSH. The overall performance of SSD-based detectors turned out to be inferior to that of our transformer-based multi-modal object detection methods.

Among our proposed fusion methods, g_{sa} obtained the best overall performance on the FLIR-Thermal dataset. Scalar multiplication amplifies the information in the feature maps by retaining the learned structure. Nevertheless, mask multiplication may amplify a certain spatial portion of the feature maps in some channels, but it could also potentially distort the learned information depth-wise. Concatenation might be useful when the feature spaces of the utilized sensor modalities differ, e.g., image versus point cloud. However, in our case of image modalities, the averaging features g_{sa} performed better than concatenation g_{sc}. Similarly, switching off the features with selective fusion g_{sf} has affected the performance of the model adversely. We plan to explore this method further in our future research, especially in the cases when information from the sensor modalities of dissimilar domains are fused, e.g., camera versus LiDAR/radar.

SUNRGB-D: The evaluation results on SUNRGB-D dataset are shown in Table I. We not only present a comparison of single vs. multi-modal settings on the selected 19 categories of the SUNRGB-D dataset, but also between raw vs. processed depth images. The table only shows the results for eight categories due to limited space. Two single modality networks are trained, one with RGB images and the other with HHA-encoded depth images. We also evaluate the performance of "conc-baseline" and "avg-baseline" with RGB and HHA-encoded depth modalities. Motivated by the performance of g_{sa} and g_{sc} on the FLIR-Thermal dataset, we chose to evaluate their performance on SUNRGB-D dataset exclusively. Since HHA-encoding introduces a significant computational burden inhibiting the possibility of real-time object detection, we first train g_{sa} and g_{sc} with on RGB and HHA-encoded depth images, later we fine-tune these models on raw-depth images as well as on our "r-blended" hybrid depth images. It is evident in Table I that both g_{sa} and g_{sc} obtain promising results on this dataset with RGB and "r-blended" depth images. Further analysing the results of Table II, we observe that the comparative performance of the models on the *Bicycle* class is not stable. Looking at the distribution of the datasets in Figure 4, we realize that the *Bicycle* class only constitutes 8.23% of the dataset. This indicates its comparative inconsistent performance on various models. However, analysing the results in Table I, we realize this performance instability might also be related to the object size. The proposed networks are able to distinguish large sized objects even if their contribution in the dataset is relatively small e.g., *Baththub* and *Bed* classes. This problem can be traced back to [10] which itself struggles to perform equally on detecting small sized objects.

TABLE I
PERFORMANCE EVALUATION ON SUNRGB-D DATASET

Model w/ RSH	w/o Random Shadows and Highlights				w/ RSH
	AP@IoU=0.5			mAP@	mAP@
	Person	Bicycle	Car	IoU=0.5	IoU=0.5
FLIR Baseline	0.794	0.580	0.856	0.743	-
rgb-only	0.383	0.168	0.638	0.395	0.376
thermal-only	0.683	0.499	0.783	0.655	0.316
MM-UNIT [19]	0.644	0.494	0.707	0.615	-
MM-CG [19]	0.633	0.502	0.706	0.614	-
SSD-BL	0.450	0.341	0.719	0.503	0.478
SSD-WA	0.526	0.314	0.718	0.519	0.516
avg-baseline	0.801	0.562	0.879	0.747	0.731
conc-baseline	0.533	0.417	0.675	0.541	0.492
g_{sa}	**0.828**	0.593	**0.891**	**0.770**	**0.769**
g_{sc}	0.809	**0.637**	0.862	0.769	0.764
g_{ma}	0.803	0.575	0.862	0.746	0.744
g_{mc}	0.800	0.611	0.857	0.755	0.755
g_{sf}	0.790	0.584	0.874	0.749	0.756
g_{sc} m-net	0.696	0.472	0.823	0.663	0.664

L515-Indoor: Table III presents the evaluation results of L515-Indoor dataset. We tested the performance RGB-only and IR-only networks, as well as the g_{sa} variant of GEM on this dataset. Evidently, g_{sa} outperformed both single modality detectors providing an additional functionality of switching between the dominant sensors in changing lighting conditions. The performance of g_{sa} with MobileNet v2 backbone is also presented in the table. The qualitative results on all three datasets are shown in Figures 5 and 6.

MobileNet v2: On a mobile platform having a 12-Core Intel Core™ i7-9750H CPU, and Nvidia GeForce RTX 2080 GPU, with ResNet-50 backbones, it takes approx 106.0 ms for a single forward pass on the proposed multi-modal object detector.

TABLE II
PERFORMANCE EVALUATION ON FLIR-THERMAL DATASET

Models	Tested without Random Shadows and Highlights (RSH)											w/ RSH
	AP@IoU=0.5										mAP@ IoU=0.5	mAP@ IoU=0.5
	bathtub	bed	bookshelf	box	chair	...	door	dresser	lamp	night stand		
RGB-only	0.116	0.461	0.038	0.084	0.457	...	0.370	0.085	0.185	0.095	0.224	0.169
HHA-only	0.355	0.409	0.002	0.020	0.413	...	0.113	0.024	0.199	0.057	0.165	0.093
conc-baseline	0.333	0.440	0.002	0.068	0.456	...	0.367	0.056	0.195	0.041	0.211	0.155
avg-baseline	0.174	0.461	0.032	0.062	0.470		0.339	0.030	0.220	0.049	0.207	0.166
g_{sc}(raw-depth)	0.404	0.411	0.008	0.073	0.487	...	0.360	0.051	0.225	0.044	0.226	0.209
g_{sc}(r-blended)	0.350	0.457	0.040	0.085	**0.490**	...	**0.381**	**0.107**	**0.226**	**0.108**	**0.242**	0.230
g_{sa}(raw-depth)	0.204	0.399	0.033	**0.087**	0.478		0.344	0.102	0.220	0.025	0.221	0.214
g_{sa}(r-blended)	0.253	0.423	**0.106**	0.080	0.474		0.379	0.035	0.219	0.079	0.239	**0.236**

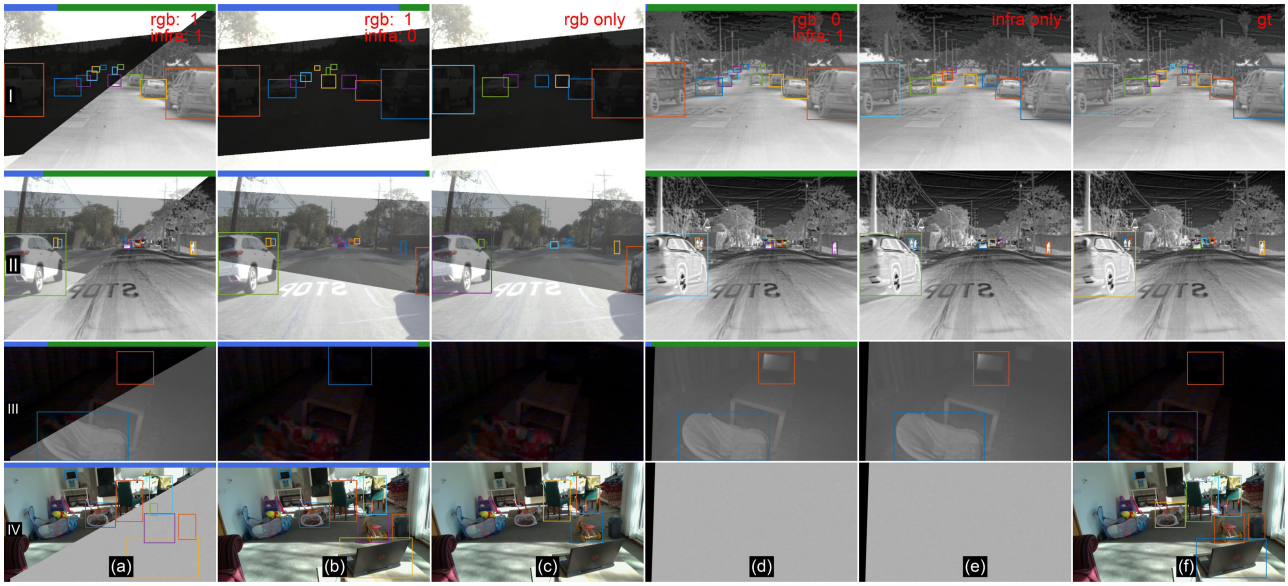

Fig. 5. Qualitative analysis of our multi-modal object detector, g_{sa} in this case. Columns (a), (b) and (d) are the outputs of g_{sa} in various asymmetric sensor failure conditions imitated artificially, which are mentioned on the upper-right corner of each image in row I. The top blue/green bar represents the contribution of each sensor modality in obtaining the final results (RGB: blue and Thermal/Infra: green). (c) and (e) are the outputs from single modal baselines. (f) is the ground-truth. Rows I and II are from FLIR-Thermal dataset while III and IV are from L515-Indoor dataset. Row IV represents a true sensor failure case when IR camera gets saturated due to sun-light even in indoors.

TABLE III
PERFORMANCE EVALUATION ON L515-INDOOR DATASET

Model w/ RSH	w/o Random Shadows and Highlights					w/ RSH
	AP@IoU=0.5				mAP@ IoU=0.5	mAP@ IoU=0.5
	Chair	Cycle	Bin	Laptop		
rgb-only	0.909	0.912	0.920	0.911	0.912	0.769
ir-only	0.141	0.557	0.012	0.690	0.386	0.311
g_{sa} m-net	0.851	0.895	0.705	0.740	0.811	0.685
g_{sa}	0.968	0.998	0.997	0.979	0.982	0.945

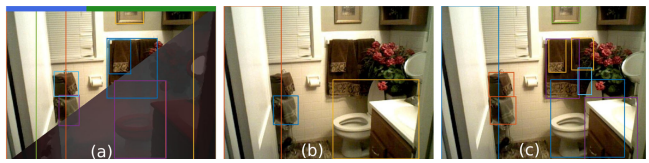

Fig. 6. (a) Sample output of GEM (g_{sc}) on the SUNRGB-D dataset with RGB images and "r-blended" depth modality. In (b), the output of single modal object detector trained only on RGB images is shown, while (c) is the ground truth.

However, with MobileNet v2 backbone feature extractors, the time for a single forward pass reduces to 49.7 ms obtaining approximately 20.1 fps. The drop in prediction accuracy of the deep models with the decrease in the number of network parameters for faster detection speed, is a well-known dilemma (e.g., in our case 23 million parameters for ResNet-50 to 3.4 million parameters for MobileNet v2). A compromise on prediction accuracy should only be made in non-critical cases where human safety is not at stake. Otherwise, the use of lightweight backbones should be avoided

V. CONCLUSION

In this paper, we propose GEM, a novel sensor-aware multi-modal object detector, with immunity against adverse lighting scenarios. Among the proposed sensor fusion configurations, the scalar averaging variant of the deterministic weighted fusion outscored the state-of-the-art and other fusion methods. The mask multipliers may amplify a certain spatial portion of the feature maps, but could also potentially distort the learned features depth-wise. Concatenation might be useful in cases where the

feature spaces of the utilized sensor modalities differ. Regarding RGB-D data, the proposed "r-blended" hybrid depth modality has proven to be a promising and lightweight alternative to the commonly employed HHA-encoded depth images. However, instead of employing a fixed blend weight α, dynamic adaptation driven by ambient light intensity could demonstrate a more realistic use of the proposed hybrid image. GEM brings along the shortcomings of [10] in multi-modal object detection setting as well, e.g., it struggles to detect small objects and suffers from the computational complexity of the attention layers. These issues will be addressed in the future work.

REFERENCES

[1] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Intl. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[2] W. Liu *et al.*, "Ssd: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[3] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018, *arXiv:1804.02767*.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Adv. Neural Inf. Process. Syst.*, vol. 28, pp. 91–99, 2015.

[5] D. Yin, R. G. Lopes, J. Shlens, E. D. Cubuk, and J. Gilmer, "A fourier perspective on model robustness in computer vision," in *Proc. of Intl. Conf. on Neural Inf. Process. Syst.*, 2015, pp. 13276–13286, *arXiv:1906.08988*.

[6] D. Hendrycks *et al.*, "The many faces of robustness: A critical analysis of out-of-distribution generalization," 2020, Online]. Available: *arXiv:2006.16241*.

[7] O. Mazhar and J. Kober, "Random shadows and highlights: A new data augmentation method for extreme lighting conditions," 2021, *arXiv:2101.05361*.

[8] M. Bijelic *et al.*, "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11 682–11692.

[9] C. M. Martinez, F. Zhang, D. Clarke, G. Hinz, and D. Cao, "Feature uncertainty estimation in sensor fusion applied to autonomous vehicle location," in *Proc. Intl. Conf. Inf. Fusion*, 2017, pp. 1–7.

[10] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conf. on Comput. Vision*, pp. 213–229, 2020, *arXiv:2005.12872*.

[11] L. Jiao *et al.*, "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128 837–128868, 2019.

[12] L. Liu *et al.*, "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, 2020.

[13] F. Garcia, D. Martin, A. De La Escalera, and J. M. Armingol, "Sensor fusion methodology for vehicle detection," *IEEE Intell. Transp. Syst. Mag.*, vol. 9, no. 1, pp. 123–133, Mar–Jun. 2017.

[14] J. Weichselbaum, C. Zinner, O. Gebauer, and W. Pree, "Accurate 3d-vision-based obstacle detection for an autonomous train," *Comput. Ind.*, vol. 64, no. 9, pp. 1209–1220, 2013.

[15] E. Park, X. Han, T. L. Berg, and A. C. Berg, "Combining multiple sources of knowledge in deep cnns for action recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–8.

[16] G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara, "Poseidon: Face-from-depth for driver pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4661–4670.

[17] T. T. Nguyen, J. Spehr, S. Zug, and R. Kruse, "Multisource fusion for robust road detection using online estimated reliabilities," *IEEE Trans. Ind. Informat.*, vol. 14, no. 11, pp. 4927–4939, Nov. 2018.

[18] L. Caltagirone, M. Bellone, L. Svensson, and M. Wahde, "Lidar-camera fusion for road detection using fully convolutional neural networks," *Robot. Auton. Syst.*, vol. 111, pp. 125–131, 2019.

[19] C. Devaguptapu, N. Akolekar, M. M Sharma, and V. N Balasubramanian, "Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 1029–1038.

[20] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1125–1134.

[21] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," *Adv. Neural Inf. Process. Syst.*, 2017, pp. 700–708, *arXiv:1703.00848*.

[22] "FLIR Thermal Dataset," 2018. https://www.flir.eu/oem/adas/adas-dataset-form/

[23] "Multispectral pedestrian detection: Benchmark dataset and saseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1037–1045.

[24] Q. Khan, T. Schön, and P. Wenzel, "Towards self-supervised high level sensor fusion," 2019, *arXiv:1902.04272*.

[25] N. Patel, A. Choromanska, P. Krishnamurthy, and F. Khorrami, "Sensor modality fusion with cnns for ugv autonomous driving in indoor environments," in *Proc. IEEE/RSJ Intl. Conf. Intell. Robots Syst.*, 2017, pp. 1531–1536.

[26] C. Chen *et al.*, "Selective sensor fusion for neural visual-inertial odometry," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10 542–10551.

[27] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 652–660.

[28] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *Proc. Intl. Conf. on Learning*, 2017, pp. 1–13, *arXiv:1611.01144*.

[29] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 567–576.

[30] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *Proc. Eur. Conf. Comput. Vis.*. Springer, 2014, pp. 345–360.

[31] A. G. Howard *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017, Online]. Available: *arXiv:1704.04861*.