Enhanced Thermal-RGB Fusion for Robust Object Detection

Wassim El Ahmar
University of Ottawa
Ottawa, Ontario, Canada
welahmar@uottawa.ca

Yahya Massoud
University of Ottawa
Ottawa, Ontario, Canada
ymass049@uottawa.ca

Dhanvin Kolhatkar
Sensor Cortek Inc
Ottawa, Ontario, Canada
dhanvin@sensorcortek.ai

Hamzah AlGhamdi
University of Ottawa
Ottawa, Ontario, Canada
halgh091@uottawa.ca

Mohammad Alja'afreh
University of Ottawa
Ottawa, Ontario, Canada
jaafreh@uottawa.ca

Riad Hammoud
Plus AI
Santa Clara, California, USA
riad.hammoud@plus.ai

Robert Laganiere
University of Ottawa
Ottawa, Ontario, Canada
laganier@eecs.uottawa.ca

Abstract

Thermal imaging has seen rapid development in the last few years due to its robustness in different weather and lighting conditions and its reduced production cost. In this paper, we study the performance of different RGB-Thermal fusion methods in the task of object detection, and introduce a new RGB-Thermal fusion approach that enhances the performance by up to 9% using a sigmoid-activated gating mechanism for early fusion. We conduct our experiments on an enhanced version of the City Scene RGB-Thermal MOT Dataset where we register the RGB and corresponding thermal images in order to conduct fusion experiments. Finally, we benchmark the speed of our proposed fusion method and show that it adds negligible overhead to the model processing time. Our work would be useful for autonomous systems and any multi-model machine vision system. The improved version of the dataset, our trained models, and source code are available at https://github.com/wassimea/rgb-thermal-fusion.

1. Introduction

The rapid evolution of machine learning models in computer vision has led to increased focus and effort in deploying such systems in real-world environments. Such systems require reliable and robust performance under different lighting and weather conditions. For this reason, machine vision systems often utilize different types of sensors to mitigate the limitations of individual sensors. With the advent of deep convolutional neural networks (CNNs), computer vision research has seen significant improvements in many different applications such as image classification [17, 26, 38, 47], object detection [9, 21, 40], semantic segmentation [4, 5, 35, 50], instance segmentation [1, 16], multiple object tracking (MOT) [45, 46, 48, 49] and odometry estimation [22, 29, 42]. Solutions to these tasks tend to rely on a single type of input data for predictions, but can be improved by using multi-modal sensor fusion to combine information from multiple, ideally complementary, types of data [6, 20, 23, 30, 31, 43].

1.1. Multi-modal Sensor Fusion

Multi-modal sensor fusion can be used in computer vision tasks to alleviate the disadvantages of each unique type of sensor. For example, sensors operating in the visible spectrum suffer in poor lighting and weather conditions but perform quite well when those conditions are decent. Thermal sensors, on the other hand, operate well in poor lighting and weather conditions but usually have a lower resolution [10, 39]. Fusing information from both spectra can be expected to mitigate the limitations of each sensor. Data collected from the different sensors is fused in one of many ways to enable the system to predict more accurate results. These fusion techniques are classified into three main groups depending on where in the processing pipeline fusion is implemented: early-, mid-, and late-fusion (before, during, and after feature extraction respectively). These techniques also vary from computationally simple but limited operations (e.g. element-wise arithmetic [23]) to more complex ones with the potential for better fusion results [27, 30].

1.2. 2D Object Detection

Deep CNNs have achieved impressive results in 2D object detection [3,9,16,21,24,40,41] - the task of classifying and predicting the location of objects in a scene - which has been used for applications such as face detection [18] and autonomous driving [2, 6, 15, 44]. In addition, object detection is often the first step in other computer vision tasks like MOT [46,48]. Deep learning-based object detection models are most often trained to detect objects on frames from RGB cameras rather than other types of sensors. Thermal sensors prove to be more robust than RGB cameras in some scenarios [10, 39], particularly in low-light settings such as nighttime assisted driving. The performance of systems relying on thermal and RGB sensors can be improved through sensor fusion, which should increase a network's performance by improving the quality of the input information, leveraging the important features from both sensors.

1.3. Contributions

This paper introduces the following contributions to the field of RGB-Thermal sensor fusion, applicable for autonomous systems and multi-modal machine vision systems:

- A novel fusion approach for RGB and thermal images that improves performance by up to 9% compared to existing fusion approaches for the task of 2D object detection.

- An augmented public benchmark of $10,000$ registered RGB-Thermal frames on which we report comparative results of our proposed approach and with state-of-the-art techniques. This enhanced dataset will be a valuable resource for RGB-Thermal fusion future research.

- Comparison of the performance of object detectors using different backbones, with and without the utilization of existing fusion operations. We benchmark all experiments on 3 different computing platforms and show that our proposed fusion method does not add any noticeable overhead.

The remainder of this paper is organized as follows. We provide a review of standard sensor fusion operations, object detectors, and the use of thermal sensors in the literature in Section 2. We elaborate on our work to improve the City Scene RGB-Thermal MOT Dataset in Section 3. We discuss the sensor fusion methodology and architecture in Section 4. We elaborate on our experiments and the results achieved in Section 5, and conclude this paper in Section 6.

2. Literature Review

Sensor fusion, object detection, and the utilization of thermal sensors in computer vision tasks have been significantly explored in the literature. In this section, we provide an overview of some of this work that we have built upon.

2.1. Object Detection

Object detection methods are divided into two categories: two-stage and one-stage. Methods in the former category use a first stage to generate object proposals that are refined and classified by a second stage [3, 9, 13, 14, 16, 34], while those in the one-stage category bypass the region proposal stage and generate predictions directly [21,24,25,32]. One-stage detectors have become the more popular and standard apporach for 2D object detection as they are generally faster, and the accuracy of these models have become comparable to that of two-stage approaches.

The Task-aligned Head (T-Head) and Task Alignment Learning (TAL) concepts were devised by Feng et al. [11] for their Task-aligned One-stage Object Detection (TOOD) network. The T-head's purpose is to increase feature sharing between the localization and classification tasks, as compared to the traditional independent design (one set of features for localization and one for classification). These shared features are then fed through one Task-aligned Predictor (TAP) for each task to generate aligned final predictions. On the other hand, the TAL is designed to improve anchor selection to ensure task alignment. Combining ResNeXt-101 with TOOD, Feng et al. achieved an mAP of 51.1 on the MS-COCO *test-dev* set.

2.1.1 ResNet

He et al. [17] introduced the residual block to enable the training of significantly deeper models than previously possible. The architecture of the residual block uses identity mappings as shortcuts around consecutive convolutional layers. The resulting residual network, or ResNet, is presented in multiple versions depending on its number of layers: ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152. Each version is composed of 5 stages, each composed of residual blocks of two or three convolutional layers.

2.1.2 ConvNeXt

Liu et al. [26] proposed ConvNeXt as an update of the ResNet architecture. They change training parameters (using the AdamW optimizer and new data augmentation techniques), the number of residual blocks in each stage (making the 3rd stage heavier and the other stages lighter), the structure of the first layer (using a smaller non-overlapping kernel), and how downsampling layers are implemented (between stages instead of at the end of each stage). Additionally, the authors use depth-wise convolutions, inverted bottleneck design, and larger kernel sizes, replace ReLU activation functions with GELU and Batch Normalization with Layer Normalization, and reduce the number of each. All these design changes result in a significant increase in accuracy for multiple tasks.

2.2. Multi-Modal Fusion

Sensor fusion can be used with a variety of combinations of sensors but is expected to achieve the most noticeable performance by fusing data from complimentary sensors. Fusion methods can be classified in one of three categories depending on when fusion occurs in the network's structure: early-fusion, mid-fusion, or late-fusion.

Massoud [27] implements learnable fusion from a LiDAR scanner and an RGB camera. The approach uses fusion methods with the LiDAR bird's eye view (BEV) representation, the LiDAR's frontal view (FV) representation, and RGB frames for the tasks of 2D, 3D, and BEV object detection. The method proposed for learnable fusion, Multi-modal Factorized Bilinear Pooling (MFB), is tested as an early and mid-fusion approach and is compared with the more straight-forward solutions of element-wise addition [23] and element-wise multiplication.

Pasandi et al. [30] expand upon this work by comparing the same set of fusion operations, but add in feature concatenation and their proposed Bilaterally-Guided Fusion (BGF). The authors compare these techniques as either early-fusion or mid-fusion in the context of the 2D object detection task on the KITTI dataset. They present a detailed analysis of performance and computational complexity.

As the work in [27, 30] shows that early fusion methods out-perform mid and late fusion approaches, we focus on early fusion in our work.

2.3. Thermal Sensors in Computer Vision

Thermal sensors are increasingly used in recent work [7, 10, 19, 28], especially due to their ability to operate well regardless of lighting conditions [10, 39].

Nowruzi et al. [28] proposed a computationally cheap CNN approach to detect the number of passengers in a vehicle. Using an in-vehicle thermal sensor and a small neural network enables deployment in embedded systems while preserving competitive performance and protecting the privacy of the passengers. Kristo et al. [19] experiment with and compare the performance of various object detection approaches [3, 25, 33, 34] with thermal sensors. The dataset used for the experiments is composed of images captured at night in various weather conditions. Dai et al. [7] propose the TIRNet object detection approach based on VGG [37] and collect the China Thermal Infrared (CTIR) dataset.

El Ahmar et al. [10] collected the City Scene RGB-Thermal MOT Dataset and compare tracking performance on its RGB and thermal images, achieving superior results on the latter. The dataset is made up of 15 sequences for a total of 1997 annotated frames for each sensor.

3. Improvements to City Scene Dataset

We conduct our experiments on the City Scene RGB-Thermal MOT Dataset [10]. The dataset is composed of thermal and RGB images collected through two different sensors mounted on a fixed support. During data collection, the sensors were static and aimed at a city intersection. Cars and pedestrians were annotated up to a distance of 300m and 100m respectively.

The collected data contains sequences taken during different times of the day (morning, afternoon, and night) making it ideal for experimenting the efficacy of different sensor fusion approaches under different conditions. However, even though the sensors are fixed close to each other, the captured images were not registered and had different resolutions. In order to conduct fusion experiments, the collected RGB and corresponding thermal images needed to be aligned to be in the same coordinate system.

To achieve this, we utilize an ORB feature extractor [36] to extract 5000 features from the thermal image and the corresponding RGB image (after being converted to grayscale). We then take the top 90% matched features and find the homography between the two images using the RANSAC algorithm [12] running for 10000 iterations. Calculating the homography matrix is not guaranteed to be accurate from one pair of thermal-RGB images. However, since the cameras and support are fixed, it is enough for the homography to be accurate for only one pair of thermal-RGB images, and then this calculated homography can be applied to all images in the dataset.

As a result of this transformation, both thermal and RGB images in the enhanced dataset have the same resolution of 500×425, and share the same annotation file. This contribution allows for thermal-RGB sensor fusion experiments to be conducted and simplifies multi-modal training of object detection and MOT algorithms.

4. RGB-T Fusion Experiment

4.1. Architectural Overview

The architecture of our RGB-T fusion framework consists of (1) a data preprocessing module, (2) a backbone network, and (3) an object detector. The data preprocessing module loads the registered frames from the City Scene dataset. Each frame contains a pair of images: an RGB image and a thermal image. We compare the performance of fusion experiments to two different baseline models. The first baseline model uses RGB-only input. The second baseline uses thermal-only input. Other experiments perform a fusion operation on both modalities before proceeding to the next step. This type of fusion is denoted as early fusion and is applied before feature extraction. Moreover, the data preprocessing module standardizes the input images before feeding them to the backbone network.

The second component is the backbone network, which performs feature extraction on either its uni-modal or multi-modal input. We experiment using ResNet-50 [17] and ConvNeXt [26]. Both backbone networks are pre-trained on the ImageNet [8] dataset. The final component is the detector. We use TOOD [11] as our single-stage 2D object detector.

4.2. Fusion Operators

We aim to experiment with our multi-modal detector with two categories of fusion operators: (1) arithmetic-based and (2) learnable fusion. The arithmetic-based operations are advantageous in terms of processing speed since they do not include any learnable weights. Meanwhile, learnable fusion operators have processing overhead due to the layers of trainable parameters. That being said, learnable fusion operators provide an edge over arithmetic-based fusion when it comes to their learning capacity. This is due to their ability to capture and learn sophisticated interactions between features from different input modalities.

Following recent work [27, 30], we use two arithmetic-based operators: (1) element-wise addition and (2) element-wise multiplication, presented in Equation 1 and Equation 2 respectively. To further enrich our set of experiments, we use two learnable fusion operators: (1) multi-modal factorized bilinear pooling and (2) bilaterally-guided fusion, illustrated in Figure 1. We implement the aforementioned fusion operators using convolutional layers in order to keep our architecture fully convolutional. We use ReLU as the main activation function, a padding value of 1, a kernel size of 3 in all of our convolution operations, and batch normalization.

$$\mathcal{F}_{add} = \mathcal{I}_{rgb} + \mathcal{I}_{thermal} \tag{1}$$

$$\mathcal{F}_{mul} = \mathcal{I}_{rgb} \cdot \mathcal{I}_{thermal} \tag{2}$$

4.3. Enhanced Sigmoid Gating

In order to enhance the robustness of fusion operators in multi-modal training, we introduce a novel fusion gating mechanism that relies on the sigmoid function. The gating mechanism is shown in Equation 3. The sigmoid-based gating mechanism, which is simple yet effective, is designed to learn the significance of each modality on a per-pixel basis before performing the multi-modal fusion operation.

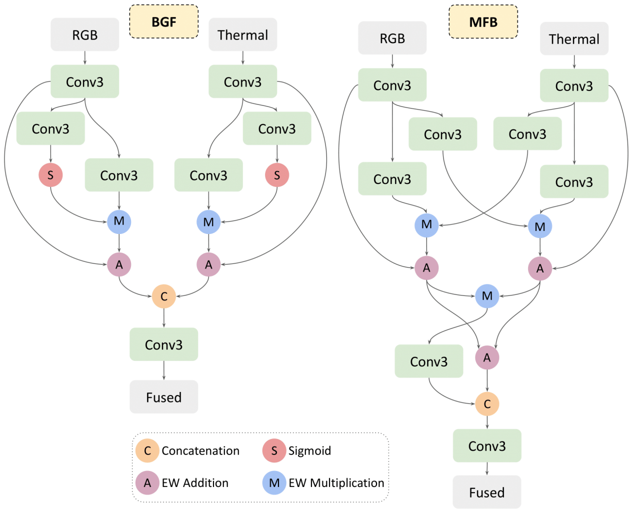

Figure 1. Left: A diagram illustrating the bilaterally-guided fusion "BGF" operator. Right: A diagram illustrating the multi-modal factorized bilinear pooling "MFB" fusion operator.

$$\begin{aligned} \mathcal{G}_{\sigma} &= \text{Conv}_{c \to c}(\sigma(\mathcal{I})) \\ \mathcal{I} &= \mathcal{I} \cdot \mathcal{G}_{\sigma} \end{aligned} \tag{3}$$

The gating mechanism is performed by feeding an input modality $\mathcal{I}$, either an RGB or thermal input, into a sigmoid-activated point-wise convolution operation while preserving the number of channels of the input modality. The resulting sigmoid gate produces a probability value for each pixel, indicating its importance. The gate $\mathcal{G}_{\sigma}$ is comprised of probability values ranging from 0 to 1, where a value of 1 indicates a significant pixel and a value of 0 indicates insignificance. The higher the pixel probability, the higher its significance. Finally, the probability gate $\mathcal{G}_{\sigma}$ is multiplied by the original input via point-wise operation. The multiplication will result in keeping the values of significant pixels, but more importantly, diminishing the values of insignificant pixels.

For RGB-thermal fusion, this gating mechanism proves quite effective as thermal and RGB images are complimentary to each other (Thermal sensors perceive well at night, while RGB sensors do not. RGB sensors perceive well when there is little variance between foreground and background, while thermal sensors do not. etc.). In addition, since there is a $1:1$ mapping between RGB pixels and corresponding thermal pixels, the gating mechanism can reliably learn the significance of each pixel in each modality.

5. Experiments

In this section, we present the results of applying both fixed and learnable fusion operators as an early fusion technique for RGB-T data to perform the task of 2D object detection. We first provide our detailed experimental setting in 5.1, then we provide a quantitative analysis of early sensor fusion in 5.2, followed by reporting results of an enhanced fusion method in 5.3. Moreover, we qualitatively analyze in 5.4 the performance of our trained detectors on a set of chosen scenarios from the City Scene dataset. Lastly, we provide an analysis of inference time in 5.5 by benchmarking our detectors on three different computing platforms.

5.1. Experimental Setting

To assess the performance of employing early sensor fusion mechanisms, we train and evaluate two different models on the City Scene RGB-T dataset on the task of 2D object detection with a special focus on the 'Car' class. The City Scene dataset contains a total of 1968 samples: each sample corresponds to a registered pair of RGB and thermal images along with their corresponding 2D annotations. We use a 75/25 split on the dataset, where 1562 samples are used for training and 406 are used for validation.

We perform our experiments using two variants of the TOOD [11] object detector, each variant contains a different backbone network, namely, ResNet-50 [17] and ConvNeXt [26]. Backbone networks are pre-trained on ImageNet [8]. We freeze the first stage of ResNet-50 while keeping the remaining three stages trainable. We do not freeze any stage in ConvNeXt. Both variants of TOOD are trained using the SGD optimizer, with an initial learning rate of 0.01, a momentum of 0.9, and a weight decay of 0.0001. Models are trained for 12 epochs, and we report the best-performing checkpoint out of the 12. We use average precision (AP%) with 50% intersection-over-union as a quantitative metric to assess detection accuracy, along with providing precision-recall curves. Moreover, we provide a qualitative analysis of scenes chosen from the City Scene dataset.

5.2. Evaluation of Early Sensor Fusion

In Table 1, we report the results of training two variants of the TOOD 2D object detector with four different fusion operators. These fusion operators are grouped into two subcategories: (1) arithmetic-based fusion operators, and (2) learnable fusion operators. The arithmetic-based operators are (1) element-wise addition (ADD) and (2) element-wise multiplication (MUL). The learnable operators are (1) multi-model factorized bilinear pooling (MFB) and (2) bilaterally-guided fusion (BGF). We specify two baselines in our comparison: (1) RGB-only and (2) thermal-only. In each of these baseline models, only one modality is fed to the detector. The choice of our baseline models helps us find the delta of performance between uni- and multi-modal settings.

We first assess our baselines: RGB-only and thermal-only detectors, which represent the uni-modal setting. As observed in Table 1, thermal-only consistently outperforms RGB-only with a considerable margin when using both ResNet-50 and ConvNeXt with a delta performance of $(+28.6\%)$ and $(+38.5\%)$, respectively. This finding renders thermal-only a more suitable modality in the 2D object detection task as thermal data is independent of light and is prone to extreme light conditions as opposed to RGB sensors. This result is expected as one-third of the validation set of the City Scene dataset is taken in poor lighting conditions.

We aim to enhance the performance of the detector by integrating the data from both modalities. In the first set of fusion experiments, we train and evaluate a TOOD [11] 2D object detector with a ResNet-50 [17] backbone network. Applying early-fusion results in a large performance boost compared to the RGB-only baseline, with a maximum of $(+31.3\%)$ increase for element-wise multiplication, and a minimum of $(+22.1\%)$ increase for multi-modal factorized bilinear pooling. Furthermore, both arithmetic-based fusion operators outperform the thermal-only baseline, with a $(+1.9\%)$ increase for element-wise addition and $(+2.7\%)$ increase for element-wise multiplication. Fusing RGB with thermal with the multi-modal factorized bilinear pooling results in a (-6.5%) decrease in the average-precision metric, while bilateral-guided fusion results in a $(+2.1\%)$ increase. With the exception of multi-modal factorized bilinear pooling, RGB data can contribute to the increase of the overall performance of an object detector. The best-performing detector fused both modalities by applying element-wise multiplication to reach an average precision of 75.4%.

In our second set of fusion experiments, we train and evaluate a TOOD [11] 2D object detector with a ConvNeXt [26] backbone network. Compared to the RGB-only baseline, all fusion operators are showing large performance boosts following the same trend of training with ResNet-50. Both arithmetic-based fusion operators resulted in decreased performance when compared to the thermal-only baseline. However, multi-modal bilinear fusion and bilateral guided fusion both showed a performance increase of $(+3.3\%)$ and $(+2.2\%)$, respectively, showing the superiority of learning-based fusion operators. Multi-modal bilinear fusion helped the best-performing detector to achieve 71.8% average precision.

In Figure 2, we plot the precision-recall curves for all fusion operators, grouped by the corresponding backbone network. Consistent trends are observed with the superiority of fusion operators to both baselines in the majority of experiments.

5.3. Evaluation of Enhanced Sigmoid Gating

In Table 2, we demonstrate the performance of the enhanced sigmoid gates that are incorporated with the already

Object Detector	Backbone Network	Fusion Operation	Average Precision Car (AP_{50})	Δ from Thermal-only	Δ from RGB-only
TOOD	ResNet50	RGB	44.1%	-28.6%	-
		Thermal	72.7%	-	+28.6%
		ADD	74.1%	+1.9%	+30.0%
		MUL	**75.4%**	**+2.7%**	**+31.3%**
		MFB	66.2%	-6.5%	+22.1%
		BGF	74.8%	+2.1%	+30.7%
TOOD	ConvNeXt	RGB	30.1%	-38.5%	-
		Thermal	68.6%	-	+38.5%
		ADD	64.0%	-4.6%	+33.9%
		MUL	67.7%	-0.9%	+37.6%
		MFB	**71.9%**	**+3.3%**	**+41.8%**
		BGF	70.8%	+2.2%	+40.7%

Table 1. Contrasting 2D detection accuracy, i.e. average precision, of four different early fusion operators.

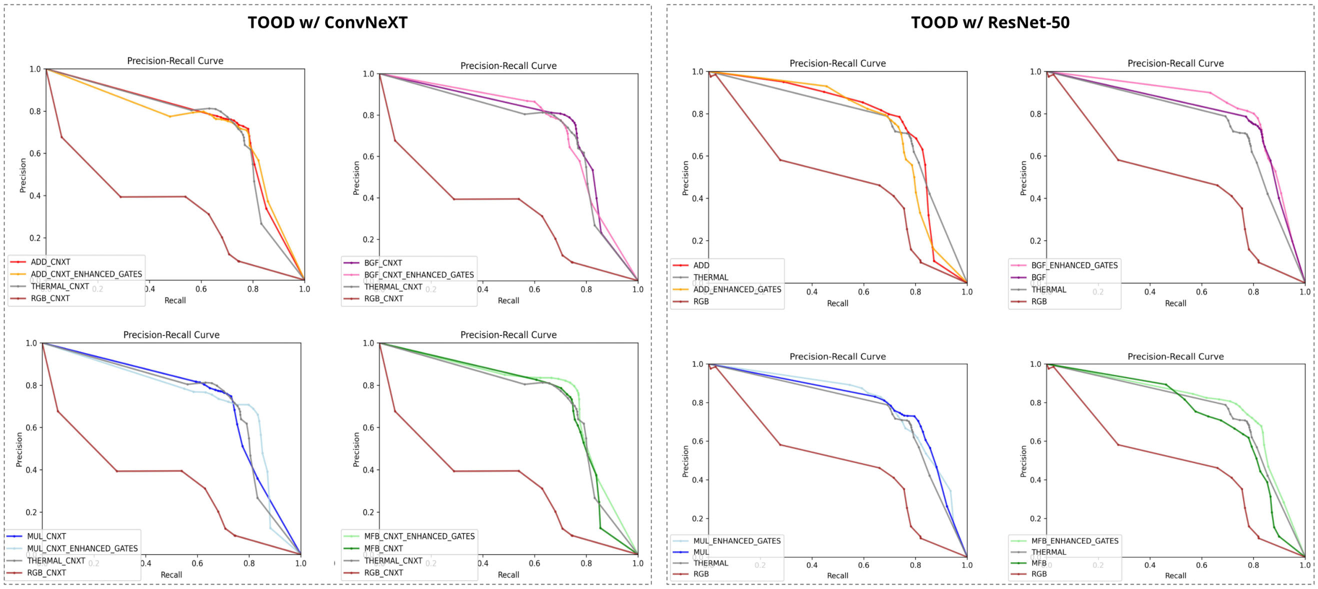

Figure 2. Right: Precision-Recall curves for TOOD with ResNet-50 as a backbone network. Left: Precision-Recall curves for TOOD with ConvNeXt as a backbone network. All plots aim to contrast the detection accuracy of fusion operators versus baseline models.

proposed fusion operators. Sigmoid gates show significant improvements when added to the TOOD ResNet-50 experiments, as both learnable fusion operators, namely multimodal factorized bilinear pooling and bilateral guided fusion, with staggering (+9%) and (+5%) increases in detection accuracy, respectively. Meanwhile, the sigmoid gates have shown less impact on arithmetic-based fusion operators, with a (-2%) decrease for element-wise addition, and a (+0.5%) increase for element-wise multiplication. After adding the sigmoid gates to TOOD with the ResNet-50 backbone network, the best-performing model has changed from element-wise multiplication which scored 75.4% AP_{50} to bilateral-guided fusion, scoring 79.8%, achieving a (+4.4%) increase.

Furthermore, we add the sigmoid gates to the TOOD detector with ConvNeXt backbone network. From Table 2, we observe an increase of average precision for both arithmetic-based fusion operators, addition and multiplication, of (+7.6%) and (+1.8%) respectively. When incorporated with learnable fusion operators, sigmoid gates results in a slight decrease in average precision for both multimodal factorized pooling (-0.2%) and bilateral-guided fusion (-0.7%). To conclude this analysis, we note the importance of carefully selecting the appropriate fusion operators when using sigmoid gates in multi-modal sensor fusion.

Backbone Network	Fusion Op.	Sig. Gates ✗	Sig. Gates ✓	Δ
ResNet50	ADD	**74.1%**	72.1%	-2.0%
	MUL	75.4%	**75.9%**	+0.5%
	MFB	66.2%	**75.2%**	+9.0%
	BGF	74.8%	**79.8%**	+5.0%
ConvNeXt	ADD	64.0%	**71.6%**	+7.6%
	MUL	67.7%	**69.5%**	+1.8%
	MFB	**71.9%**	71.7%	-0.2%
	BGF	**70.8%**	70.1%	-0.7%

Table 2. Demonstrating the performance boost that results from adding the sigmoid gated mechanism to the fusion operators. The experiments are conducted on the TOOD object detector with two different backbone networks. Experiments are compared based on their average precision (AP_{50}).

5.4. Qualitative Analysis

In this subsection, we aim to showcase and qualitatively analyze the performance of the fusion operators. We put RGB and thermal images side-by-side and visualize the bounding boxes on top of them. We visualize groundtruth bounding boxes with a green color and the predictions with a red color. We filter all bounding boxes with a minimum area of 150 pixels to simplify the visualization. Also, we apply a minimum detection confidence score of 40% for all detectors.

In Figure 3, we choose a day frame from City Scene [10] dataset. Our first baseline, RGB-only, does not perform well on this frame, predicting only two true positives out of five cars, as well as predicting one false positive. This performance reflects the detection accuracy in Table 1 and Figure 2. Our second baseline, thermal-only, performs slightly better, by predicting three out of five cars. Fusing both modalities using the bilinear-guided fusion results in detecting all five cars in this frame. Moreover, when observing the confidence scores of all three detectors, we find that the fusion-based detector produces the most confident bounding boxes, as four out of the five bounding boxes have a confidence score over 90%.

We also choose a night frame from the City Scene dataset and visualize the results in Figure 4. Our first baseline, RGB-only, does not perform well by predicting three cars out of six in addition to one false positive. Our second baseline, thermal-only, detects all six, but with relatively low confidence scores. The use of bilinear-guided fusion results in the detection of all six cars in this frame, with all detections having confidence scores above 90%. This highlights the effectiveness of multi-modal training to leverage the strengths of multiple sensors which produces more accurate and confident detections.

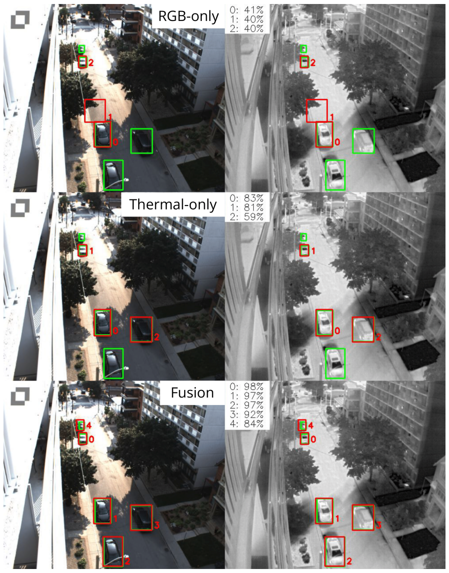

Figure 3. A day frame from the City Scene [10] dataset. The frame is fed into both baselines (RGB-only and thermal-only) as well as a fusion method (BGF). Green boxes correspond to the groundtruth. Red boxes correspond to predictions. Each bounding box has an index from 0 to N. Confidence scores (%) of bounding boxes are positioned to the left of each box's index.

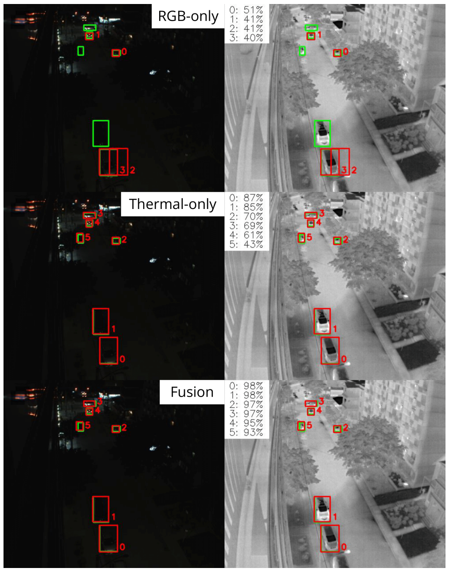

Figure 4. A night frame from City Scene [10] dataset. The frame is fed into both baselines (RGB-only and thermal-only) as well as a fusion method (BGF). Green boxes correspond to the groundtruth. Red boxes correspond to predictions. Each bounding box has an index from 0 to N. Confidence scores (%) of bounding boxes are positioned to the left of each box's index.

5.5. Inference Time and Benchamrking

In order to study the computational cost of the different fusion operations and the proposed enhanced gating

Object Detector	Input Modality	Goose		Xavier		TX2	
		sec	*fps*	*sec*	*fps*	*sec*	*fps*
TOOD w/ ResNet50	RGB-only	0.06	17.46	0.77	1.29	1.34	0.74
	Thermal-only	0.06	17.46	0.77	1.29	1.34	0.74
	ADD	0.06	17.57	0.78	1.28	1.34	0.74
	ADD w/ Gating	0.06	17.34	0.78	1.28	1.35	0.73
	MUL	0.06	17.65	0.77	1.29	1.34	0.74
	MUL w/ Gating	0.06	17.34	0.78	1.28	1.35	0.73
	MFB	0.06	16.54	0.82	1.22	1.44	0.69
	MFB w/ Gating	0.09	16.95	0.84	1.20	1.45	0.70
	BGF	0.06	16.20	0.95	1.05	1.78	0.56
	BGF w/ Gating	0.06	16.05	0.96	1.14	1.81	0.55
TOOD w/ ConvNeXt	RGB-only	0.06	16.18	0.89	1.12	1.62	0.62
	Thermal-only	0.06	16.18	0.89	1.12	1.62	0.62
	ADD	0.06	16.29	0.89	1.12	1.62	0.62
	ADD w/ Gating	0.06	15.53	0.90	1.11	1.65	0.60
	MUL	0.06	15.88	0.89	1.12	1.63	0.61
	MUL w/ Gating	0.06	15.53	0.90	1.11	1.63	0.61
	MFB	0.06	15.51	0.94	1.06	1.73	0.58
	MFB w/ Gating	0.07	14.99	0.95	1.05	1.75	0.57
	BGF	0.07	14.94	1.07	0.94	2.06	0.49
	BGF w/ Gating	0.07	14.43	1.08	0.93	2.09	0.48

Table 3. Inference time and benchmarking results on NVIDIA GeForce RTX 3090, Jetson Xavier, and Jetson TX2. The reported results include inference time in seconds (*sec*) and the total number of frames-per-second (*fps*).

method, we benchmark the inference speed of TOOD with ResNet50 and ConvNext backbones when applying the different fusion operations. The benchmarking experiments were conducted on three different computing platforms:

- Goose: Powerful computing machine equipped with an NVIDIA RTX 3090 GPU (10496 cores).
- Xavier: NVIDIA Jetson Xavier edge computing platform (512-Core Volta GPU).
- TX2: NVIDIA Jetson TX2 edge computing platform (256-Core Pascal GPU).

The benchmarking results are given in Table 3. The results confirm the statement that learnable fusion operators (MFB and BGF) have a higher computational requirement and thus higher inference latency than arithmetic fusion operators. However, the results also show that the utilization of our proposed enhanced sigmoid gating adds negligible overhead to the fusion operators.

6. Conclusion

In this paper, we contribute an improvement to the City Scene RGB-Thermal MOT Dataset by registering RGB and thermal frames, making the dataset an important resource for RGB-thermal fusion research. In addition, we conduct experiments to study the efficacy of different existing arithmetic and learnable fusion methods for the task of object detection using thermal and RGB images. We propose a novel enhanced sigmoid gating method to enhance the fusion performance of thermal and RGB images, and report the results of applying this mechanism to both arithmetic and learnable fusion methods. Finally, we benchmark the inference speed of the proposed fusion operators on three different computing platforms with different specifications and show that the proposed enhanced gating adds negligible overhead.

References

[1] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. YOLACT: Real-time instance segmentation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Apr. 2019.

[2] Garrick Brazil, Xi Yin, and Xiaoming Liu. Illuminating pedestrians via simultaneous detection and segmentation. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, June 2017.

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2017.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, June 2016.

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, June 2017.

[6] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Nov. 2016.

[7] Xuerui Dai, Xue Yuan, and Xueye Wei. TIRNet: Object detection in thermal infrared images for autonomous driving. *Applied Intelligence*, 51(3):1244–1261, sep 2020.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[9] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, oct 2019.

[10] Wassim A El Ahmar, Dhanvin Kolhatkar, Farzan Erlik Nowruzi, Hamzah AlGhamdi, Jonathan Hou, and Robert Laganiere. Multiple object detection and tracking in the thermal spectrum. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 277–285, 2022.

[11] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3490–3499. IEEE Computer Society, 2021.

[12] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[13] Ross Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, dec 2015.

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Nov. 2014.

[15] Irtiza Hasan, Shengcai Liao, Jinpeng Li, Saad Ullah Akram, and Ling Shao. Pedestrian detection: Domain generalization, cnns, transformers and beyond. In *ECCV2022*, Jan. 2022.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2017.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Dec. 2015.

[18] Peiyun Hu and Deva Ramanan. Finding tiny faces. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Dec. 2016.

[19] Mate Kristo, Marina Ivasic-Kos, and Miran Pobar. Thermal object detection in difficult weather conditions using YOLO. *IEEE Access*, 8:125459–125476, 2020.

[20] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven Waslander. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Dec. 2018.

[21] Hei Law and Jia Deng. CornerNet: Detecting objects as paired keypoints. *International Journal of Computer Vision*, 128(3):642–656, aug 2018.

[22] Qing Li, Shaoyang Chen, Cheng Wang, Xin Li, Chenglu Wen, Ming Cheng, and Jonathan Li. LO-net: Deep real-time lidar odometry. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Apr. 2019.

[23] Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In *ECCV 2018*, page 663–678. Cham: Springer International Publishing, 2018.

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2017.

[25] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot MultiBox detector. *Computer Vision – ECCV 2016*, pages 21–37, Dec. 2015.

[26] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jan. 2022.

[27] Yahya Massoud. Sensor fusion for 3d object detection for autonomous vehicles. Master's thesis, Université d'Ottawa/University of Ottawa, 2021.

[28] Farzan Erlik Nowruzi, Wassim A. El Ahmar, Robert Laganiere, and Amir H. Ghods. In-vehicle occupancy detection with convolutional networks on thermal images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, jun 2019.

[29] Farzan Erlik Nowruzi, Dhanvin Kolhatkar, Prince Kapoor, and Robert Laganiere. Point cloud based hierarchical deep odometry estimation. In *Proceedings of the 7th International*

Conference on Vehicle Technology and Intelligent Transport Systems, Mar. 2021.

[30] Morteza Mousa Pasandi, Tianran Liu, Yahya Massoud, and Robert Laganière. Sensor fusion operators for multimodal 2d object detection. In *ISVC 2022*, volume 13598 of *LNCS*, page 184–195. Springer, 2022.

[31] Charles R. Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J. Guibas. Frustum PointNets for 3d object detection from RGB-d data. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Nov. 2017.

[32] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2015.

[33] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement, Apr. 2018.

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, June 2015.

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science*, pages 234–241, May 2015.

[36] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011.

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, Sept. 2014.

[38] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning, 2019*, May 2019.

[39] Michael Teutsch, Angel D Sappa, and Riad I Hammoud. Computer vision in the infrared spectrum: challenges and approaches. *Synthesis Lectures on Computer Vision*, 10(2):1–138, 2021.

[40] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Scaled-YOLOv4: Scaling cross stage partial network. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2021.

[41] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *arXiv preprint arXiv:2207.02696*, 2022.

[42] Guangming Wang, Xinrui Wu, Zhe Liu, and Hesheng Wang. PWCLO-net: Deep LiDAR odometry in 3d point clouds using hierarchical embedding mask optimization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Dec. 2020.

[43] Shenlong Wang, Simon Suo, Wei-Chiu Ma, Andrei Pokrovsky, and Raquel Urtasun. Deep parametric continuous convolutional neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2018.

[44] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Nov. 2017.

[45] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *Computer Vision – ECCV 2020*, pages 107–122. Springer International Publishing, Sept. 2019.

[46] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, Mar. 2017.

[47] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Nov. 2016.

[48] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 1–21. Springer, 2022.

[49] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. FairMOT: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129(11):3069–3087, Apr. 2020.

[50] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Dec. 2016.