

# Borrow from Anywhere: Pseudo Multi-modal Object Detection in Thermal Imagery

Chaitanya Devaguptapu<sup>1</sup>, Ninad Akolekar<sup>1</sup>, Manuj M Sharma<sup>2</sup>, Vineeth N Balasubramanian<sup>1</sup>

<sup>1</sup>Indian Institute of Technology, Hyderabad, India

<sup>2</sup>ANURAG, Defense Research and Development Organization, India

<sup>1</sup>{chaitanyadevaguptapu, cs16btech11024, vineethnb}@iith.ac.in

<sup>2</sup>{m.sharma}@anurag.drdo.in

## Abstract

*Can we improve detection in the thermal domain by borrowing features from rich domains like visual RGB? In this paper, we propose a ‘pseudo-multimodal’ object detector trained on natural image domain data to help improve the performance of object detection in thermal images. We assume access to a large-scale dataset in the visual RGB domain and relatively smaller dataset (in terms of instances) in the thermal domain, as is common today. We propose the use of well-known image-to-image translation frameworks to generate pseudo-RGB equivalents of a given thermal image and then use a multi-modal architecture for object detection in the thermal image. We show that our framework outperforms existing benchmarks without the explicit need for paired training examples from the two domains. We also show that our framework has the ability to learn with less data from thermal domain when using our approach.*

## 1. Introduction

As indicated by the recent fatalities [29], the current sensors in self-driving vehicles with level 2 and level 3 autonomy (lacking thermal imaging) do not adequately detect vehicles and pedestrians. Pedestrians are especially at risk after dark, when 75% of the 5,987 U.S. pedestrian fatalities occurred in 2016 [33]. Thermal sensors perform well in such conditions where autonomy level 2 and level 3 sensor-suite technologies are challenged. As is well-known, thermal IR cameras are relatively more robust to illumination changes, and can thus be useful for deployment both during the day and night. In addition, they are low-cost, non-intrusive and small in size. Consequently, thermal IR cameras have become increasingly popular in applications such as autonomous driving recently, as well as in other mainstream applications such as security and military surveillance operations. Detection and classification of objects in

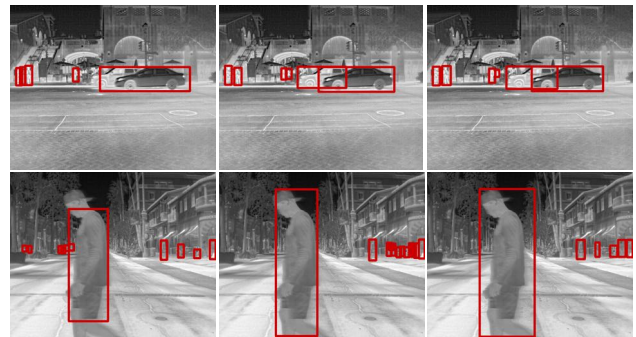


Figure 1. *Left*: Detection with single mode Faster-RCNN; *Middle*: Detection using the proposed method; *Right*: Annotated Ground Truth as provided in FLIR dataset [13].

thermal imagery, is thus an important problem to be addressed and invested in, to achieve successes that can be translated to deployment of such models in real-world environments.

Although object detection has always remained an important problem in computer vision, most of the efforts have focused on detecting humans and objects in standard RGB imagery. With the advent of Deep Convolutional Neural Networks (CNNs) [18], object detection performance in the RGB domain has been significantly improved using region-based methods, such as the R-CNN [12] and Fast R-CNN [11] that use selective search, as well as Faster R-CNN [32] that uses region-proposal networks to identify regions of interest. Object detection methods such as YOLO [31] rephrase the object detection problem into a regression problem, where the coordinates of the bounding boxes and the class probability for each of those boxes are generated simultaneously. This makes YOLO [31] extremely fast, although its performance is lower than R-CNN based counterparts [39]. The aforementioned object detection methods rely, however, on architectures and models that have been trained on large-scale RGB datasets such as ImageNet, PASCAL-VOC, and MS-COCO. A relative dearth of such

publicly available large-scale datasets in the thermal domain restricts the achievement of an equivalent level of success of such frameworks on thermal images. In this work, we propose a ‘pseudo multi-modal’ framework for object detection in the thermal domain, consisting of two branches. One branch is pre-trained on large-scale RGB datasets (such as PASCAL-VOC or MS-COCO) and finetuned using a visual RGB input that is obtained using an image-to-image (I2I) translation framework from a given thermal image (and hence the name ‘pseudo multi-modal’). The second branch follows the standard training process on a relatively smaller thermal dataset. Our multi-modal architecture helps borrow complex high-level features from the RGB domain to improve object detection in the thermal domain. In particular, our multi-modal approach does not need paired examples from two modalities; our framework can borrow from any large-scale RGB dataset available for object detection, and does not need the collection of a synchronized multi-modal dataset. This setting makes this problem challenging too. Our experimental results demonstrate that using our multi-modal framework significantly improves the performance of fully supervised detectors in the thermal domain. The proposed framework also overcomes the problem of inadequacy of training examples in the thermal domain. Furthermore, we also study the relevance of this methodology when there is very limited data in the thermal domain. Our experimental results on the recently released FLIR ADAS[13] thermal imagery dataset show that, using only a quarter of the thermal dataset, the proposed multi-modal framework achieves a higher mAP than a single-mode fully-supervised detector trained on the entire dataset.

The remainder of this paper is organized as follows. Section 2 provides the context for study including a brief overview of the early and recent work on applying deep learning for thermal imagery. Section 3 describes our approach and methodology. Section 4 describes the experiments carried out and their results. Section 5 investigates the impact of size of training set, image resolution and ends with a discussion on some of the cases where our model fails to perform as well.

## 2. Related Work

Detection and classification of objects in the thermal imagery has been an active area of research in computer vision [41][35][27][14], especially in the context of military and surveillance[26]. There has been a significant amount of work on classifying and detecting people and objects in thermal imagery using standard computer vision and machine learning models, even before deep learning became popular. Bertozzi *et al.* [5] proposed a probabilistic template-based approach for pedestrian detection in far infrared (IR) images. They divided their algorithm into three parts: candidate generation, candidate filtering and

validation of candidates. One main weakness of this approach is that it assumes the human is hotter than the background which may not be the case in many real-world scenarios. Davis *et al.* [9] proposed a two-stage template-based method to detect people in widely varying thermal imagery. To locate the potential person locations, a fast screening procedure is used with a generalized template and then an AdaBoost ensemble classifier is used to test the hypothesized person locations. Kai *et al.* [17] proposed a local feature-based pedestrian detector on thermal data. They used a combination of multiple cues to find interest points in the images and used SURF [2] as features to describe these points. A codebook is then constructed to locate the object center. The challenge of this detector is whether a high performance can be achieved when local features are not obvious.

While these efforts have shown good performance for IR image classification and detection tasks over a small number of objects, they have been outperformed in recent years by deep learning models that enable more descriptive features to be learned. With the increase in popularity of deep neural networks, several methods have been proposed for applying deep learning methods to thermal images. Peng *et al.* [30] proposed a Convolutional Neural Network (CNN) for face identification in near IR images. Their CNN is a modification of GoogLeNet but has a more compact structure. Lee *et al.* [20] designed a lightweight CNN consisting of two convolutional layers and two subsampling layers for recognizing unsafe behaviors of pedestrians using thermal images captured from moving vehicles at night. They combined their lightweight CNN with a boosted random forest classifier. Chevalier *et al.* [6] proposed LR-CNN for automatic target recognition which is a deep architecture designed for classification of low-resolution images with strong semantic content. Rodger *et al.* [15] developed a CNN trained on short-to-mid range high resolution IR images containing six object classes (person, land vehicle, helicopter, aeroplane, unmanned aerial vehicle and false alarm) using an LWIR sensor. This network was successful at classifying other short to mid-range objects in unseen images, although it struggled to generalize to long range targets. Abbott *et al.* [1] used a transfer learning approach with the YOLO [31] framework to train a network on high-resolution thermal imagery for classification of pedestrians and vehicles in low-resolution thermal images. Berg *et al.* [4][3] proposed an anomaly-based obstacle detection method using a train-mounted thermal camera. Leykin *et al.* [22] proposed a fusion tracker and pedestrian classifier for multispectral pedestrian detection. Proposals for performing detection are generated using background subtraction and evaluated using periodic gait analysis.

Among efforts that use a multimodal approach, Wagner *et al.* [36] applied Aggregated Channel Features (ACF) and

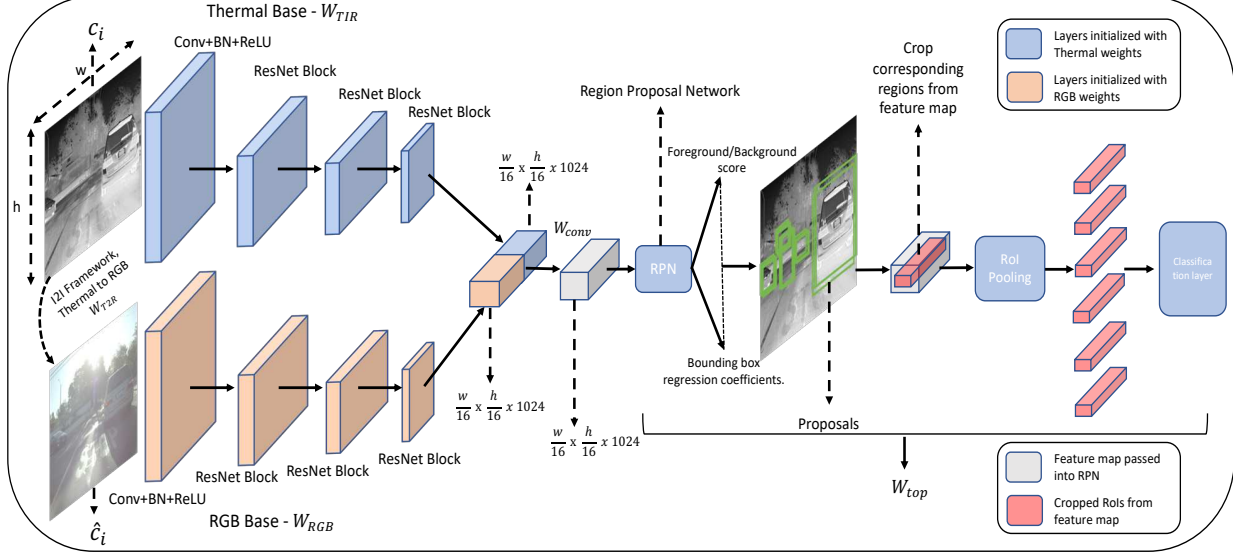


Figure 2. Adaptation of proposed Multi-modal framework for Faster-RCNN

Boosted Decision trees (BDT) for proposal generation and classified these proposals with a CNN, which fuses Visual and IR information. Choi *et al.* [8] uses two separate region proposal networks for both Visual and IR images and evaluates the proposals generated by both the networks with Support Vector Regression on fused deep features. The efforts closest to our work are that of König *et al.* [19] and Liu *et al.* [24], both of which propose a multi-modal framework that combines RGB and thermal information in a Faster-RCNN architecture by posing it as a convolutional network fusion problem. However, all of these multimodal efforts assume the availability of a dataset with paired training examples from the visual and thermal domain. On the other hand, our work assumes only the presence of thermal imagery, and seeks to leverage the use of publicly available RGB datasets (which may not be paired with the thermal dataset) to obtain significant improvement in thermal object detection performance.

### 3. Methodology

Our overall proposed methodology for ‘pseudo multi-modal’ object detection for thermal images is summarized in Figure 2. The key idea of our methodology is to borrow knowledge from data-rich domains such as visual (RGB) without the explicit need for a paired multimodal dataset. We achieve this objective by leveraging the success of recent image-to-image translation methods [40, 25] to automatically generate a pseudo-RGB image from a given thermal image, and then propose a multimodal Faster R-CNN architecture to achieve our objective. Image-to-Image translation models aim to learn the visual mapping between a

source domain and target domain. Learning this mapping becomes challenging when there are no paired images in source and target domains. Recently, there have been noteworthy efforts on addressing this problem using unpaired images [40][38][7][25][34][28][21]. While one could use any unsupervised image-to-image translation framework in our overall methodology, we use CycleGAN[40] and UNIT[25] as I2I frameworks of choice in our work, owing to their wide use and popularity. We begin our discussion with the I2I translation frameworks used in this work.

**Unpaired Image-to-Image Translation:** CycleGAN [40] is a popular unpaired image-to-image translation framework that aims to learn the mapping functions  $F : \mathcal{X} \rightarrow \mathcal{Y}$  and  $G : \mathcal{Y} \rightarrow \mathcal{X}$  where  $\mathcal{X}$  and  $\mathcal{Y}$  are source and target domains respectively. It maps the images onto two separate latent spaces and employs two generators  $\mathcal{G}_{\mathcal{X} \rightarrow \mathcal{Y}}, \mathcal{G}_{\mathcal{Y} \rightarrow \mathcal{X}}$  and two discriminators  $\mathcal{D}_{\mathcal{X}}, \mathcal{D}_{\mathcal{Y}}$ . The generator  $\mathcal{G}_{\mathcal{X} \rightarrow \mathcal{Y}}$  attempts to generate images  $\hat{y}_i$  that look similar to images from domain  $\mathcal{Y}$ , while  $\mathcal{D}_{\mathcal{Y}}$  aims to distinguish between the translated samples  $\hat{y}_i$  and real samples  $y_i$ . This condition is enforced using an adversarial loss. To reduce the space of possible mapping functions, a cycle-consistency constraint is also enforced, such that a source-domain image  $x_i$  when transformed into target domain ( $\hat{y}_i$ ) and re-transformed back to source domain ( $\hat{x}_i$ ) will ensure in  $\hat{x}_i$  and  $x_i$  will belong to the same distribution. For more details, please see [40].

Unlike CycleGAN [40], UNIT [25] tackles the unpaired image-to-image translation problem assuming a shared latent space between both the domains. It learns the joint distribution of images in different domains using the marginal

distribution in individual domains. The framework is based on variational autoencoders  $VAE_1, VAE_2$  and generative adversarial networks  $GAN_1, GAN_2$  with a total of six sub-networks including two image encoders  $\mathcal{E}_1, \mathcal{E}_2$ , two image generators  $\mathcal{G}_1, \mathcal{G}_2$  and two adversarial discriminators  $\mathcal{D}_1, \mathcal{D}_2$ . Since they assume a shared latent space between the two domains, a weight sharing constraint is enforced to relate the two VAEs. Specifically, weight sharing is done between the last few layers of encoders  $\mathcal{E}_1, \mathcal{E}_2$  that are responsible for higher level representations of the input images in the two domains and the first few layers of the image generators  $\mathcal{G}_1, \mathcal{G}_2$  responsible for decoding the high-level representations for reconstructing the input images. The learning problems of  $VAE_1, VAE_2, GAN_1, GAN_2$  for image reconstruction, image translation and cyclic reconstruction are jointly solved. For more information, please see [25].

In case of both CycleGAN and UNIT, the trained model provides two generators which perform the translation between source and target domains. In our case, we use the generator which performs the Thermal-to-RGB translation, which is given by  $G : \mathcal{X} \rightarrow \mathcal{Y}$  in case of a CycleGAN and  $G_1$  in case of UNIT (we used Thermal as the source domain, and RGB as the target domain while training these models). We refer to the parameters of these generators as  $W_{T2R}$  in our methodology.

---

**Algorithm 1:** MMTOD: Multi-modal Thermal Object Detection Methodology

---

**Input:** Thermal image training data:  $\{(c_i, y_i)\}_{i=1}^m$ ;  
 Generator of I2I framework:  $W_{T2R}$ ; Pre-trained RGB base network:  $W_{RGB}$ ; Pre-trained thermal base network:  $W_{TIR}$ , Pre-trained thermal top network  $W_{top}$ ; Randomly initialised 1x1 conv weights:  $W_{conv}$ ; Number of epochs:  $num\_epochs$ ; Loss function:  $\mathcal{L}(\cdot)$

**Output:** Trained MMTOD model,  $\mathcal{F}(\cdot)$

**for**  $num\_epochs$  **do**

**for**  $c_i, i = 1, \dots, m$  **do**

        Generate a pseudo-RGB image  $\hat{c}_i$  using  $W_{T2R}$ .

        Generate feature maps by passing  $c_i$  and  $\hat{c}_i$  to base networks  $W_{TIR}$  and  $W_{RGB}$  respectively

        Stack the feature maps and use  $W_{conv}$  to get  $1 \times 1$  conv output

        Pass the 1x1 conv output to  $W_{top}$

        Update weights:

$W_{RGB}, W_{TIR}, W_{top}, W_{conv}, W_{T2R}$  by minimizing  $\mathcal{L}$  of the object detection framework.

**end**

**end**

---

**Pseudo Multi-modal Object Detection:** As shown in Figure 2, our object detection framework is a multi-modal architecture consisting of two branches, one for the thermal image input and the other for the RGB input. Each branch is initialized with a model pre-trained on images from that domain (specific details of implementation are discussed in Section 4). To avoid the need for paired training examples from two modalities but yet use a multi-modal approach, we use an image-to-image (I2I) translation network in our framework. During the course of training, for every thermal image input, we generate a pseudo-RGB using  $W_{T2R}$  and pass the pseudo-RGB and Thermal to the input branches (parametrized by  $W_{RGB}$  and  $W_{TIR}$  respectively). Outputs from these branches are stacked and passed through a  $1 \times 1$  convolution ( $W_{conv}$ ) to learn to combine these features appropriately for the given task. The output of this  $1 \times 1$  convolution is directly passed into the rest of the Faster-RCNN network (denoted by  $W_{top}$ ). We use the same Region Proposal Network (RPN) loss as used in Faster-RCNN, given as follows:

$$L(\{p_i\}, \{t_i\}) = \frac{1}{N} \sum_i L(p_i, p_i^*) + \lambda \frac{1}{N} \sum_i p_i^* R(t_i, t_i^*)$$

where  $i$  is the index of an anchor,  $p_i$  is the predicted probability of anchor  $i$  being an object,  $p_i^*$  is the ground truth,  $t_i$  represents the coordinates of the predicted bounding box,  $t_i^*$  represents the ground truth bounding box coordinates,  $L$  is log loss,  $R$  is the robust loss function (smooth  $L_1$ ) as defined in [11], and  $\lambda$  is a hyperparameter. We use the same multi-task classification and regression loss as used in Faster-RCNN [11] at the end of the network.

While the use of existing I2I models allow easy adoption of the proposed methodology, the images generated from such I2I frameworks for thermal-to-RGB translation are perceptually far from natural RGB domain images (like MS-COCO[23] and PASCAL-VOC [10]), as shown in Figure 6. Therefore, during the training phase of our multi-modal framework, in order to learn to combine the RGB and thermal features in a way that helps improve detection, we also update the weights of the I2I generator  $W_{T2R}$ . This helps learn a better representation of the pseudo-RGB image for borrowing relevant features from the RGB-domain, which we found to be key in improving detection in the thermal domain. The proposed methodology provides a fairly simple strategy to improve object detection in the thermal domain. We refer to the proposed methodology as MMTOD (Multimodal Thermal Object Detection) hereafter. Our algorithm for training is summarized in Algorithm 1. More details on the implementation of our methodology are provided in Section 4.



## 4. Experiments

### 4.1. Datasets and Experimental Setup

**Datasets:** We use the recently released FLIR ADAS [13] dataset and the KAIST Multispectral Pedestrian dataset [14] for our experimental studies. FLIR ADAS [13] consists of a total of 9,214 images with bounding box annotations, where each image is of  $640 \times 512$  resolution and is captured using a FLIR Tau2 camera. 60% of the images are collected during daytime and the remaining 40% are collected during night. While the dataset provides both RGB and thermal domain images (not paired though), we use only the thermal images from the dataset in our experiments (as required by our method). For all the experiments, we use the training and test splits as provided in the dataset benchmark, which contains the person (22,372 instances), car (41,260 instances), and bicycle (3,986 instances) categories. Some example images from the dataset are shown in Figure 3.

The KAIST Multispectral pedestrian benchmark dataset [14] contains around 95,000 8-bit day and night images (consisting of only the Person class). These images are collected using a FLIR A35 microbolometer LWIR camera with a resolution of  $320 \times 256$  pixels. The images are then upsampled to  $640 \times 512$  in the dataset. Sample images from the dataset are shown in Figure 3. Though the KAIST dataset comes with fully aligned RGB and Thermal, we choose not to use the RGB images as our goal to improve the detection in the absence of paired training data.



Figure 3. Row 1 & Row 2: Example images from FLIR [13] ADAS dataset, Row 3: Example Images from KAIST [14] dataset

Our methodology relies on using publicly available large-scale RGB datasets to improve thermal object detection performance. For this purpose, we use RGB datasets with the same classes as in the aforementioned thermal image datasets. In particular, we perform experiments using two popular RGB datasets namely, PASCAL VOC [10] and MS-COCO [23]. In each experiment, we pre-train an object detector on either of these datasets and use these parameters to initialise the RGB branch of our multimodal framework. We also compare the performance of these two initializations in our experiments. In case of thermal image datasets, an end-to-end object detector is first trained on the dataset and used to initialize the thermal branch of our framework.

We use mean Average Precision (mAP) as the performance metric, as is common for the object detection task.

**Baseline:** A Faster-RCNN trained in a fully supervised manner on the thermal images from the training set is used as the baseline method for the respective experiments in our studies. We followed the original paper [32] for all the hyperparameters, unless specified otherwise. The FLIR ADAS dataset [13] also provides a benchmark test mAP (at IoU of 0.5) of 0.58 using the more recent RefineDetect-512 [39] model. We show that we beat this benchmark using our improved multi-modal Faster-RCNN model.

**Image-to-Image Translation (IR-to-RGB):** For our experiments, we train two CycleGAN models: one for FLIR  $\leftrightarrow$  RGB which uses thermal images from FLIR [13] and RGB images from PASCAL VOC [10], and another for KAIST  $\leftrightarrow$  RGB which uses thermal images from KAIST [14] and RGB images from PASCAL VOC [10]. We use an initial learning rate of  $1e-5$  for the first 20 epochs, which is decayed to zero over the next 20 epochs. The identity mapping is set to zero, i.e., the identity loss and the reconstruction loss are given equal weightage. The other hyperparameters of training are as described in [40]. For training of the UNIT framework, all the hyperparameters are used as stated in the original paper, without any alterations. Since UNIT takes a long time to train (7 to 8 days on an NVIDIA P-100 GPU), we trained it only for FLIR  $\leftrightarrow$  RGB, so the experiments on KAIST are performed using CycleGAN only. Our variants are hence referred to as MMTOD-CG (when I2I is CycleGAN) and MMTOD-UNIT (when I2I is UNIT) in the remainder of the text.

We use the same metrics as mentioned in CycleGAN [40] and UNIT [25] papers for evaluating the quality of translation. In an attempt to improve the quality of generated images in CycleGAN [40], we tried adding feature losses in addition to cycle consistency loss and adversarial loss. However, this did not improve the thermal to visual RGB translation performance. We hence chose to finally use the same loss as mentioned in [40].

**Training our Multi-modal Faster-RCNN:** Our overall architecture (as in Figure 2) is initialized with pre-trained RGB and Thermal detectors as described in Section 3. Since our objective is to improve detection in thermal domain, the region proposal network (RPN) is initialized with weights pre-trained on thermal images. The model is then trained on the same set of images on which the thermal detector was previously pre-trained. The I2I framework generates a pseudo-RGB image corresponding to the input thermal image. The thermal image and the corresponding pseudo-RGB image are passed through the branches of the multi-modal framework to obtain two feature maps of

1024 dimension each, as shown in figure 2. These two feature maps are stacked back-to-back and passed through a  $1 \times 1$  convolution, which is then passed as input to the Region Proposal Network (RPN). RPN produces the promising Regions of Interest (RoIs) that are likely to contain a foreground object. These regions are then cropped out of the feature map and passed into a classification layer which learns to classify the objects in each ROI. Note that as mentioned in Section 3, during the training of the MM-TOD framework, the weights of the I2I framework are also updated which allows it to learn a better representation of the translated image for improved object detection in thermal domain. We adapted the Faster-RCNN code provided at [37] for our purpose. The code for the CycleGAN and UNIT was taken from their respective official code releases[40][16][25]. Our code will be made publicly available for further clarifications.

**Experimental Setup:** To evaluate the performance of the proposed multi-modal framework, the following experiments are carried out:

- MMTOD-CG with RGB branch initialized by PASCAL-VOC pre-trained detector, thermal branch initialized by FLIR ADAS pre-trained detector
- MMTOD-UNIT with RGB branch initialized by PASCAL-VOC pre-trained detector, thermal branch initialized by FLIR ADAS pre-trained detector
- MMTOD-CG with RGB branch initialized by MS-COCO pre-trained detector, thermal branch initialized by FLIR ADAS pre-trained detector
- MMTOD-UNIT with RGB branch initialized by MS-COCO pre-trained detector, thermal branch initialized by FLIR ADAS pre-trained detector
- MMTOD-CG with RGB branch initialized by PASCAL-VOC pre-trained detector, thermal branch initialized by KAIST pre-trained detector
- MMTOD-CG with RGB branch initialized by COCO pre-trained detector, thermal branch initialized by KAIST pre-trained detector

## 4.2. Results

**IR-to-RGB Translation Results:** Figure 6 shows the results of CycleGAN and UNIT trained for Thermal  $\leftrightarrow$  RGB translation. As mentioned in Section 3, the generated pseudo-RGB images are perceptually far from natural domain images. This can be attributed to the fact that the domain shift between RGB and Thermal domains is relatively high compared to other domains. In addition, RGB images have both chrominance and luminance information, while thermal images just have the luminance part which makes estimating the chrominance for RGB images a difficult task. However, we show that using our method, these generated images add value to the detection methodology.



Figure 6. Row 1: Thermal images from FLIR ADAS[13] dataset; Row 2: Translations generated using UNIT[25]; Row 3: Translations generated using CycleGAN[40].

**Thermal Object Detection Results:** Tables 1 and 2 show the comparison of AP for each class and the mAP of our framework against the baseline detector when trained on FLIR ADAS and KAIST datasets respectively. (Note that the KAIST dataset has only one class, the Person.) We observe that in all the experiments, our framework outperforms the baseline network across all the classes.

<i>Method</i>		<i>AP across each class</i>			
		<i>Bicycle</i>	<i>Person</i>	<i>Car</i>	<i>mAP</i>
Baseline		39.66	54.69	67.57	53.97
Framework	RGB Branch				
MMTOD-UNIT	MSCOCO	49.43	<b>64.47</b>	<b>70.72</b>	<b>61.54</b>
	Pascal VOC	45.81	59.45	70.42	58.56
MMTOD-CG	MSCOCO	50.26	63.31	70.63	61.40
	Pascal VOC	43.96	57.51	69.85	57.11

Table 1. Performance comparison of proposed methodology against baseline on FLIR [13]

<i>Method</i>		<i>mAP</i>
Baseline		49.39
Framework	RGB Branch	
MMTOD-CG	MS-COCO	<b>53.56</b>
	Pascal VOC	52.26

Table 2. Performance comparison of proposed methodology against baseline on KAIST [14]

In case of FLIR, we observe that initializing the RGB branch with MS-COCO obtains better results than those with PASCAL-VOC. This can be attributed to the fact that MS-COCO has more instances of car, bicycle, and person as compared to PASCAL VOC. Also, experimental results show that employing UNIT as the I2I framework achieves better performance than CycleGAN. Our framework with MS-COCO initialization and UNIT for I2I translation results in an increase in mAP by at least 7 points. In particular, as mentioned earlier, the FLIR ADAS dataset provides a

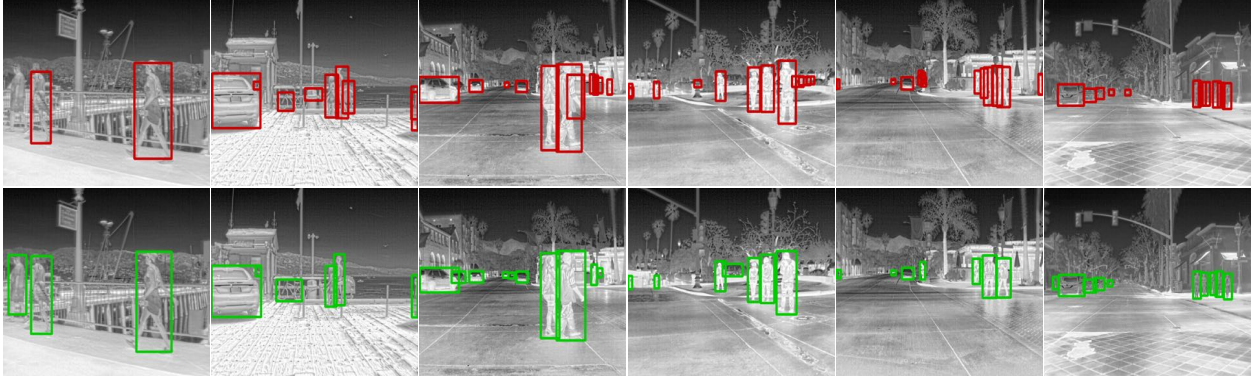


Figure 4. Qualitative results of detections on the FLIR ADAS dataset. *Row 1*: Baseline *Row 2*: MMTOD

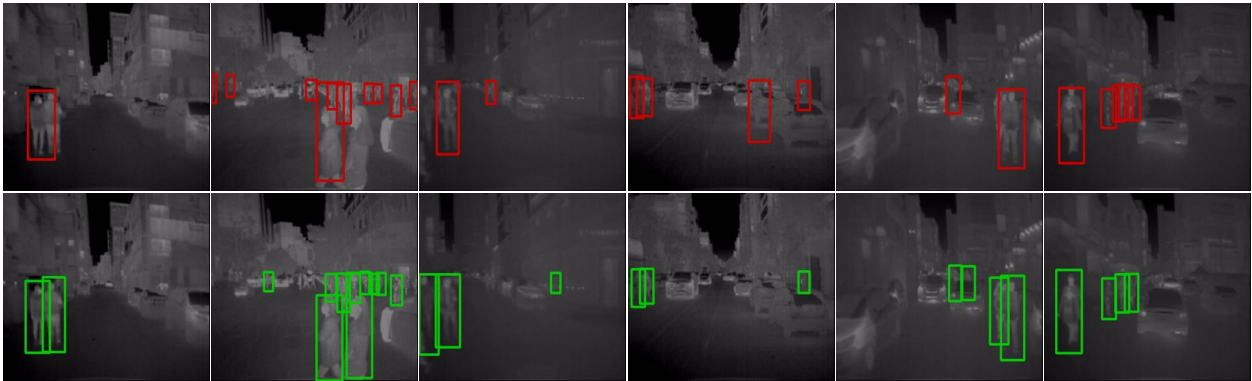


Figure 5. Qualitative results of detections on the KAIST. *Row 1*: Baseline. *Row 2*: MMTOD

benchmark test mAP (at IoU of 0.5) of 0.58 using the more recent RefineDetect-512 [39] model. Our method outperforms the benchmark despite using a relatively older object detection model such as the Faster-RCNN.

As shown in Table 2, our improved performance on the KAIST dataset shows that although this dataset has more examples of the ‘Person’ category than the RGB dataset used such as PASCAL-VOC, our framework still improves upon the performance of the baseline method. This allows us to infer that the proposed framework can be used in tandem with any region-CNN based object detection method to improve the performance of object detection in thermal images. On average our framework takes 0.11s to make detections on a single image, while the baseline framework takes 0.08s. Our future directions of work include improving the efficiency of our framework while extending the methodology to other object detection pipelines such as YOLO and SSD.

## 5. Discussion and Ablation Studies

**Learning with limited examples:** We also conducted studies to understand the capability of our methodology

when there are limited samples in the thermal domain. Our experiments on the FLIR ADAS dataset showed that our framework outperforms the current state-of-the-art detection performance using only half the training examples. Moreover, our experiments show that using only a quarter of the training examples, our framework outperforms the baseline on full training set. Table 3 presents the statistics of the dataset used for this experiment. Note that the test set used in these experiments is still the same as originally provided in the dataset.

Dataset	Number of Instances		
	Car	Person	Bicycle
FLIR	41,260	22,372	3,986
FLIR (1/2)	20,708	11,365	2,709
FLIR (1/4)	10,448	5,863	974

Table 3. Statistics of the datasets we used for our experiments.

We perform the same set of experiments (as discussed in Section 4) on FLIR(1/2) and FLIR(1/4) datasets. Tables 4 and 5 present the results.



Method		AP across each class			
		Bicycle	Person	Car	mAP
Baseline (FLIR)		39.66	54.69	67.57	53.97
Baseline (FLIR-1/2)		34.41	51.88	65.04	50.44
Framework	RGB Branch				
MMTOD-UNIT	MSCOCO	49.84	<b>59.76</b>	<b>70.14</b>	<b>59.91</b>
	Pascal VOC	45.53	57.77	69.86	57.72
MMTOD-CG	MSCOCO	50.19	58.08	69.77	59.35
	Pascal VOC	40.17	54.67	67.62	54.15

Table 4. Performance comparison of proposed methodology against baseline on FLIR (1/2)

Method		AP across each class			
		Bicycle	Person	Car	mAP
Baseline(FLIR)		39.66	54.69	67.57	53.97
Baseline(FLIR-1/4)		33.35	49.18	60.84	47.79
Framework	RGB Branch				
MMTOD-UNIT	MSCOCO	<b>44.24</b>	<b>57.76</b>	<b>69.77</b>	<b>57.26</b>
	Pascal VOC	35.23	54.71	67.83	52.59
MMTOD-CG	MSCOCO	41.29	57.08	69.10	55.82
	Pascal VOC	35.02	51.62	66.09	50.91

Table 5. Performance comparison of proposed methodology against baseline on FLIR (1/4)

Table 4 shows the baselines for training the Faster-RCNN on the complete FLIR training dataset as well as FLIR (1/2). We observe that both MMTOD-UNIT and MMTOD-CG trained on FLIR(1/2) outperform both the baselines, even when Faster-RCNN is trained on the entire training set. Similarly, Table 5 shows the baselines for training the Faster-RCNN on the complete FLIR training dataset as well as FLIR (1/4). Once again, we observe that both MMTOD-UNIT and MMTOD-CG trained on FLIR(1/4) outperform both the baselines, even when Faster-RCNN is trained on the entire training set. In other words, the MMTOD framework requires only a quarter of the thermal training set to surpass the baseline accuracy achieved using the full training set. The results clearly demonstrate the proposed framework’s ability to learn from less examples. This shows that our framework effectively borrows features from the RGB domain that help improve detection in the thermal domain. This is especially useful in the context of thermal and IR images, where there is a dearth of publicly available large-scale datasets.

**Effect of Image Resolution:** To understand the effect of image resolution on object detection performance, we repeated the above experiments were conducted using sub-sampled images of the FLIR ADAS dataset. Table 6 presents these results for  $400 \times 400$  input images. We observe that our multi-modal framework improves the object detection performance significantly even in this case. Our future work will involve extending our work to images of even lower resolutions.

Dataset	Method	AP across each class			
		Bicycle	Person	Car	mAP
FLIR	Baseline	29.25	43.13	58.83	43.74
	P-VOC + CycleGAN	<b>39.42</b>	<b>52.75</b>	<b>62.05</b>	<b>51.41</b>
FLIR (1/2)	Baseline	23.31	40.82	56.25	40.13
	P-VOC + CycleGAN	<b>33.32</b>	<b>48.32</b>	<b>60.87</b>	<b>47.50</b>
FLIR (1/4)	Baseline	18.81	35.42	52.82	35.68
	P-VOC + CycleGAN	<b>30.63</b>	<b>45.45</b>	<b>60.32</b>	<b>45.47</b>

Table 6. Performance comparison of proposed methodology against baseline on FLIR  $400 \times 400$  images

**Missed Detections:** We tried to analyze the failure cases of the proposed methodology, by studying the missed detections. Some examples of these missed detections are shown in figure 7. We infer that MMTOD finds object detection challenging when: (i) the objects are very small and located far from the camera; (ii) two objects are close to each other, and are detected as a single object; and (iii) there is heavy occlusion and crowd. Our future efforts will focus on addressing these challenges.

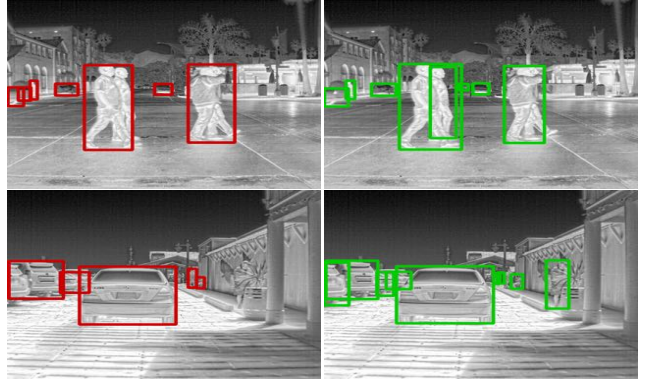


Figure 7. Some examples of missed detections, Red: Predictions using MMTOD, Green: Ground Truth

## 6. Conclusion

We propose a novel multi-modal framework to extend and improve upon any Region-CNN-based object detector in the thermal domain by borrowing features from the RGB domain, without the need of paired training examples. We evaluate the performance of our framework applied to a Faster-RCNN architecture in various settings including the FLIR ADAS and KAIST datasets. We demonstrate that our framework achieves better performance than the baseline, even when trained only on quarter of the thermal dataset. The results suggest that our framework provides a simple and straightforward strategy to improve the performance of object detection in thermal images.

## Acknowledgements

This work was carried out as part of a CARS project supported by ANURAG, Defence Research and Development Organisation (DRDO), Government of India.



## References

- [1] R. Abbott, J. Del Rincon, B. Connor, and N. Robertson. Deep object classification in low resolution lwir imagery via transfer learning. In *Proceedings of the 5th IMA Conference on Mathematics in Defence*, 11 2017. 2
- [2] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. volume 3951, pages 404–417, 07 2006. 2
- [3] A. Berg. *Detection and Tracking in Thermal Infrared Imagery*. Linkping Studies in Science and Technology. Thesis No. 1744, Linkping University, Sweden, 2016. 2
- [4] A. Berg, K. Öfjäll, J. Ahlberg, and M. Felsberg. Detecting rails and obstacles using a train-mounted thermal camera. In R. R. Paulsen and K. S. Pedersen, editors, *Image Analysis*, Cham, 2015. Springer International Publishing. 2
- [5] M. Bertozzi, A. Broggi, C. Hilario, R. Fedriga, G. Vezzoni, and M. Del Rose. Pedestrian detection in far infrared images based on the use of probabilistic templates. pages 327 – 332, 07 2007. 2
- [6] M. CHEVALIER, N. Thome, M. Cord, J. Fournier, G. Henaff, and E. Dusch. LOW RESOLUTION CONVOLUTIONAL NEURAL NETWORK FOR AUTOMATIC TARGET RECOGNITION. In *7th International Symposium on Optronics in Defence and Security*, Paris, France, Feb. 2016. 2
- [7] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *CoRR*, abs/1711.09020, 2017. 3
- [8] H. Choi and S. Kim, , and K. Sohn. Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 621–626, Dec 2016. 3
- [9] J. W. Davis and M. 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*, volume 1, pages 364–369, Jan 2005. 2
- [10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>. 4, 5
- [11] R. B. Girshick. Fast R-CNN. *CoRR*, abs/1504.08083, 2015. 1, 4
- [12] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013. 1
- [13] F. A. Group. Flir thermal dataset for algorithm training. <https://www.flir.in/oem/adas/adas-dataset-form/>, 2018. 1, 2, 5, 6
- [14] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1037–1045, June 2015. 2, 5, 6
- [15] B. C. I. Rodger and N. Robertson. Classifying objects in lwir imagery via cnns. In *In Proc. SPIE: Electro-Optical and Infrared Systems: Technology and Applications XII*, 2016. 2
- [16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017. 6
- [17] K. Jüngling and M. Arens. Feature based person detection beyond the visible spectrum. *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 30–37, 2009. 2
- [18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 1
- [19] D. Knig, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch. Fully convolutional region proposal networks for multispectral person detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 243–250, July 2017. 3
- [20] E. J. Lee, B. C. Ko, and J.-Y. Nam. Recognizing pedestrians unsafe behaviors in far-infrared imagery at night. *Infrared Physics Technology*, 76:261 – 270, 2016. 2
- [21] H. Lee, H. Tseng, J. Huang, M. K. Singh, and M. Yang. Diverse image-to-image translation via disentangled representations. *CoRR*, abs/1808.00948, 2018. 3
- [22] A. Leykin, Y. Ran, and R. Hammoud. Thermal-visible video fusion for moving target tracking and pedestrian classification. 06 2007. 2
- [23] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 4, 5
- [24] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas. Multispectral deep neural networks for pedestrian detection. *CoRR*, abs/1611.02644, 2016. 3
- [25] M. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. *CoRR*, abs/1703.00848, 2017. 3, 4, 5, 6
- [26] S. Liu and Z. Liu. Multi-channel cnn-based object detection for enhanced situation awareness. *CoRR*, abs/1712.00075, 2017. 2
- [27] S. Mangale and M. Khambete. Moving object detection using visible spectrum imaging and thermal imaging. In *2015 International Conference on Industrial Instrumentation and Control (ICIC)*, pages 590–593, May 2015. 2
- [28] S. Mo, M. Cho, and J. Shin. Instagan: Instance-aware image-to-image translation. *CoRR*, abs/1812.10889, 2018. 3
- [29] NTSB. Preliminary report highway hwy18mh010, 2018. 1
- [30] M. Peng, C. Wang, T. Chen, and G. Liu. Nirfacenet: A convolutional neural network for near-infrared face identification. *Information*, 7(4), 2016. 2
- [31] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018. 1, 2
- [32] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015. 1, 5

- [33] R. Retting and S. Schwatz. Governors highway safety association pedestrian traffic fatalities by state (2017 preliminary data). <https://www.ghsa.org/sites/default/files/2018-02/pedestrians18.pdf>, 2017. 1
- [34] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. *CoRR*, abs/1611.02200, 2016. 3
- [35] J. H. M. L. E. S. E. B. Toby P. Breckon, Anna Gaszczak. Multi-modal target detection for autonomous wide area search and surveillance, 2013. 2
- [36] J. Wagner, V. Fischer, M. Herman, and S. Behnke. Multi-spectral pedestrian detection using deep fusion convolutional neural networks. 04 2016. 2
- [37] J. Yang, J. Lu, D. Batra, and D. Parikh. A faster pytorch implementation of faster r-cnn. <https://github.com/jwyang/faster-rcnn.pytorch>, 2017. 6
- [38] Z. Yi, H. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. *CoRR*, abs/1704.02510, 2017. 3
- [39] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li. Single-shot refinement neural network for object detection. In *CVPR*, 2018. 1, 5, 7
- [40] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 3, 5, 6
- [41] T. T. Zin, H. Takahashi, and H. Hama. Robust person detection using far infrared camera for image fusion. In *Second International Conference on Innovative Computing, Information and Control (ICICIC 2007)*, pages 310–310, Sep. 2007. 2