D3T: Distinctive Dual-Domain Teacher Zigzagging Across RGB-Thermal Gap for Domain-Adaptive Object Detection

Dinh Phat Do[1], Taehoon Kim[1], Jaemin Na[1,2], Jiwon Kim[3], Keonho Lee[3], Kyunghwan Cho[3],
and Wonjun Hwang[1]
[1]Ajou University, Korea, [2]Tech. Innovation Group, KT, [3]Robotics Lab, Hyundai Motor Company
{phatai,th951113,wjhwang}@ajou.ac.kr jaemin.na@kt.com
{jiwon1115,keonho.lee,kyunghwan.cho}@hyundai.com

Abstract

Domain adaptation for object detection typically entails transferring knowledge from one visible domain to another visible domain. However, there are limited studies on adapting from the visible to the thermal domain, because the domain gap between the visible and thermal domains is much larger than expected, and traditional domain adaptation can not successfully facilitate learning in this situation. To overcome this challenge, we propose a Distinctive Dual-Domain Teacher (D3T) framework that employs distinct training paradigms for each domain. Specifically, we segregate the source and target training sets for building dual-teachers and successively deploy exponential moving average to the student model to individual teachers of each domain. The framework further incorporates a zigzag learning method between dual teachers, facilitating a gradual transition from the visible to thermal domains during training. We validate the superiority of our method through newly designed experimental protocols with well-known thermal datasets, i.e., FLIR and KAIST. Source code is available at https://github.com/EdwardDo69/D3T.

1. Introduction

Beyond the significant success of the Convolutional Neural Network (CNN) [17, 24], it has naturally led to recent advancements in CNN-based object detection [28, 33, 34, 38]. These advances hold promise for wide real-world applications such as autonomous driving, surveillance, and human activity recognition. Reflecting on the key contributors to this success, two crucial factors emerge: the development of efficient network architectures [33, 34] and the availability of a sufficient number of trainable RGB images [8, 11, 27] with corresponding supervision signals for supervised learning. It is noteworthy that RGB cameras struggle to provide reliable imaging in scenarios where visible light sensors prove inadequate, particularly during nighttime. In sharp contrast, thermal cameras [12] hold a significant advantage, detecting the heat emitted by objects and facilitating effective operation in complete darkness, through smoke, and in visually obstructive environments. This capability makes them indispensable for various applications, including nighttime surveillance, search and rescue operations, wildlife monitoring, and all-weather autonomous driving systems [1, 16, 21, 22, 32].

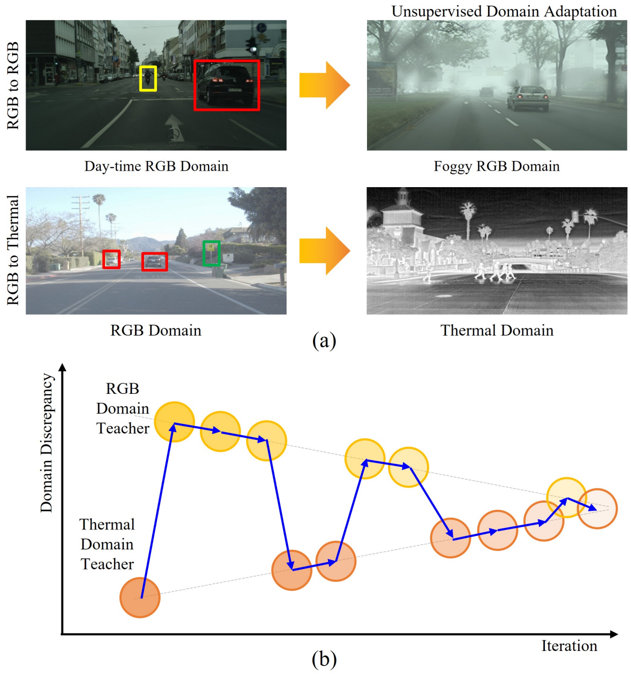


Figure 1. (a) Sample images showing the difference between unsupervised domain adaptation from RGB to RGB domains and unsupervised domain adaptation from RGB to thermal domains. (b) Conceptual illustration of the proposed unsupervised domain adaptation using distinctive dual-domain teachers, demonstrating the zigzag approach across the large RGB-thermal gap.

As we delve into thermal image-based object detection [1, 16], a distinct set of challenges emerges. Foremost among them is the scarcity of annotated thermal datasets essential for training sophisticated detection models. Contrary to the wealth of annotations accessible for RGB object images [27], thermal datasets are notably limited, posing a challenge to the advancement of high-performance thermal detection models using the sufficient training images. The visual features in thermal images diverge significantly from those in RGB images, giving rise to a domain shift problem. This discrepancy leads to performance degradation when models trained on visible datasets are applied to thermal images. Consequently, addressing these challenges necessitates the employment of specialized training and adaptation techniques to construct effective object detection systems capable of harnessing the unique properties inherent in thermal cameras.

In this paper, we leverage Unsupervised Domain Adaptation (UDA) methods [13, 39, 41] to alleviate the domain shift problems from the source domain (e.g., RGB images) to the target domain (e.g., thermal images). We have focused on UDA for object detection [5, 35]. This aims to minimize the discrepancy between source and target domains and enhance model performance without requiring labor-intensive labeling of target data. We have focused to *one-stage object detection* method, e.g., FCOS [38] in this paper, because it is generally faster than two-stage object detection for real-time applications. This is particularly crucial in applications such as autonomous driving, where acquiring labeled thermal images can be both time-consuming and expensive. While the aforementioned methods primarily utilize conventional UDA methods based on only RGB images, they fall short in addressing the fundamental challenge of UDA from RGB to thermal images. As shown in Fig. 1 (a), it stems from the significant disparity between the RGB and thermal domains compared to that between two RGB domains.

To solve this issue, we propose a novel Mean Teacher (MT) framework using Distinctive Dual-Domain Teacher (D3T) for domain adaptive object detection between RGB and thermal domains. Unlike prior MT-based object detections (e.g., single teacher and single student) [9, 26], we employ two distinct teacher models, each specializing in either RGB or thermal domain. This facilitates more effective learning of domain-specific information, particularly in the presence of the substantial discrepancy. This D3T framework, paired with a zigzag learning method (as shown in Fig. 1 (b)) between domains, updates selected domain-specific weights to the single student, enabling a gradual transition from RGB to thermal domains. By zigzagging the teacher network selection, we leverage the observation that, during initial training, the RGB teacher pre-trained from source labels is more likely to predict relatively accurate pseudo-labels on the target, while the thermal teacher performs better as training progresses. To achieve this, we adjust the selection frequency, favoring the RGB teacher more in the early stages of training and gradually increasing the emphasis on the thermal teacher as training progresses. Finally, we verify performances of our method using new established evaluation protocols with well-known thermal datasets such as FLIR [44] and KAIST [19].

We summarize our contributions as follows:

- We introduce the D3T framework, leveraging two distinctive domain teachers for effective domain adaptive object detection between RGB and thermal domains.
- Our zigzag learning method facilitates a gradual shift from RGB to thermal domains, updating domain-specific weights dynamically. This optimizes adaptation, leveraging each teacher's strengths during training.
- We have made our experimental protocols using well-known thermal datasets: FLIR and KAIST, and prove the superiority of our method compared with other methods.

2. Related Work

2.1. Thermal Object Detection

Thermal object detection [1] is pivotal for applications in surveillance, military operations, and autonomous driving. Recent advancements underscore their adaptability and efficiency [16, 21, 22, 43]. A notable trend is the fusion of visible and thermal features, enhancing detection accuracy by capturing more comprehensive environmental information [2, 6, 45, 48]. However, these studies typically assume the simultaneous capture of visible and thermal images and they should be aligned well.

2.2. UDA for Object Detection

UDA for object detection is focusing on adapting detectors from a labeled source domain to an unlabeled target domain. The primary methods in UDA are categorized into domain alignment and self-training. Domain alignment techniques including style transfer [4, 20, 23], adversarial training [5, 18, 35], and graph matching [25, 42, 47] aim to minimize the domain discrepancy by aligning features or visual styles between the source and target domains. However, these methods face challenges in maintaining a balance between feature transferability and discriminability. In contrast, self-training methods leverage inherent information from the target domain. The UMT [9] generates pseudo labels using similar images to the source domain, while the HT [10] emphasizes consistency in classification and localization, using a new sample reweighting scheme. The unified CMT [3] framework employs self-training with contrastive learning in domain-adaptive object detection. This enhances target domain performance by optimizing object-level features using pseudo-labels without requiring target domain labels. UDA is crucial for enabling object detection models to perform accurately across diverse environments, especially where labeling data in the target domain is impractical, such as in autonomous driving at night.

2.3. Domain Adaptive Thermal Object Detection

Domain adaptive thermal object detection aims to enhance object detection in thermal images, especially in suboptimal lighting conditions. This field addresses the limitations inherent in object detectors designed for visible light datasets, which typically underperform in environments with poor or variable lighting. Utilizing UDA, these methods leverage labeled data from the visible spectrum to improve detection in the thermal spectrum with the limited availability of labeled thermal data. Despite its considerable potential for practical applications, this field currently attracts relatively modest research investment. Meta-UDA approach [40] stands out as a significant advancement by leveraging an algorithm-agnostic meta-learning framework for better domain adaptation using labeled data from visible domains. Nakamura et al. [31] introduces a unique data fusion strategy using CutMix. This approach integrates elements of target images into source images, coupled with adversarial learning, resulting in enhanced object detection efficacy.

The previous methods only take advantage of UDA for RGB images and it is not easy to bridge the large gap between the RGB and thermal domains. To overcome this, we propose the D3T framework collaborated with a zigzag learning method specifically designed from RGB to thermal domain adaptation, which is highly efficient and easy to implement for domain adaptive object detection.

3. Proposed Method

In the quest for advancing object detection capabilities across diverse imaging domains, we delve into the Mean Teacher (MT) framework [37] and extend it with a dual teacher-based framework.

3.1. MT Framework with A Single Teacher

The MT framework represents a paradigm in domain adaptation, particularly within the context of object detection tasks [9, 10, 26]. This approach learns knowledge from labeled data in the source domain and adapts it to the unlabeled target domain. Furthermore, it employs the teacher-student mutual learning method, as introduced in [29], to enhance detection accuracy.

Overview: The core idea of the MT framework is a model architecture consisting of a teacher model and a student model, two detectors with identical architectures. The teacher model, pre-trained on labeled data from the source domain, generates pseudo-labels for the target domain data, which lacks labels. The student model is optimized by using these pseudo-labels, and its weights are updated to the single teacher model. The teacher model can be regarded as the ensemble of student models at various time steps, resulting in higher accuracy and the production of better quality pseudo labels.

Training method: The MT framework uses both source and target domains for training at the same time. The source domain data is applied with both strong and weak data augmentation before being directly used for the supervised training of the student model with ground-truth labels. Target domain data employs two types of data augmentation: weak augmentation for the teacher model's input images to ensure reliable pseudo-labels, and strong augmentation for the student model's input images to enhance the model's diversity. This enhances the teacher model since it is updated with the weights from the student model at various time steps.

The overall loss function for the MT framework is defined as follows:

$$\mathcal{L} = \mathcal{L}_{src} + \mathcal{L}_{tgt}, \tag{1}$$

where $\mathcal{L}_{src}$ is the loss in the source domain, including classification and localization loss, and $\mathcal{L}_{tgt}$ is the loss in the target domain which is similarly calculated using pseudo-labels.

Update teacher parameter: The MT framework updates the weights of the teacher model with the weights of the student model via Exponential Moving Average (EMA). This gradual updating process results in the teacher model becoming an ensemble of student models across different time steps and it is derived by

$$\theta^{\mathcal{T}} \leftarrow \alpha\theta^{\mathcal{T}} + (1-\alpha)\theta^{\mathcal{S}}. \tag{2}$$

where $\theta^{\mathcal{T}}$ represents the weights of the teacher model, $\theta^{\mathcal{S}}$ represents the weights of the student model, and α is the EMA coefficient. For simplicity, we set α as 0.9996 in all experiments.

3.2. Distinctive Dual-Domain Teacher (D3T)

UDA for object detection typically employs an MT framework with a single teacher model to adapt across RGB image domains, such as from the Cityscapes [7] to the Foggy Cityscapes dataset [36]. However, the domain gap between the RGB and thermal domains is significantly larger. Therefore, using a single teacher model for both domains can lead to negative effects and diminish the model's effectiveness. To address this issue, we introduce a new initiative called D3T, which is directly inspired by [30] and includes two individual-teacher models for the RGB domain and the thermal domain, respectively. The two teacher models leverage the specialized knowledge of their respective domains and transfer this knowledge to the student model. The overview of D3T is summarized in Fig. 2.

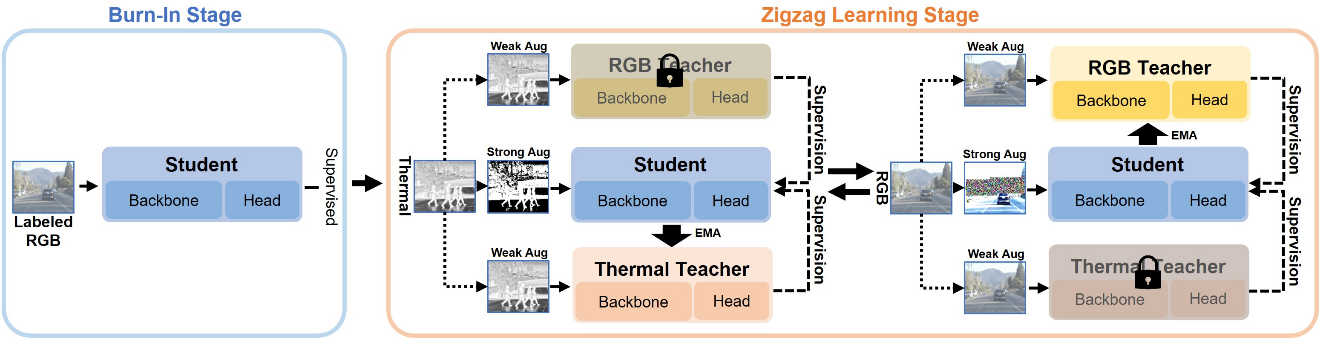


Figure 2. Overview of **D3T**: Our D3T model consists of two stages. **Burn-in Stage**: We initiate the training of the object detector using labeled data from the RGB domain. **Zigzag Learning Stage**: Comprises two distinct and interleaved training components for the Thermal domain and the RGB domain, respectively. During each step of training, the student model utilizes images from a single domain for training but leverages knowledge from two teachers for enhanced learning effectiveness. In each step, only one teacher model is updated corresponding to the trained domain.

Separate teachers: The core idea of our method is to use two separate teachers, an RGB teacher and a thermal teacher, to integrate knowledge from their respective domains. Each teacher's model is updated with the student model's weights only when it is trained with the corresponding domain. As a result, that teacher acquires the specialized knowledge of that domain without being negatively impacted by other domains. The D3T model is trained using thermal images and updates the weights for the corresponding thermal teacher. Similar to the right side, our model is trained with the RGB domain and updates the weights for the RGB teacher model.

Learning knowledge from Dual-Teachers: During each training step of the D3T model, images from only one domain, either RGB or thermal domain, are used. However, to leverage the combined knowledge of both teachers and minimize the domain shift between the two domains, both thermal and RGB teachers are employed to generate pseudo-labels. The dual teaching method not only utilizes the knowledge from the two teachers but also increases the reliability of the pseudo-labels, which leads to more effective training of the student model. The loss functions are defined as follows:

$$\mathcal{L}_{rgb\_sup} = \mathcal{L}_{sup}(f^{\mathcal{S}}(\mathcal{I}_{rgb}), \mathcal{Y}), \tag{3}$$

$$\mathcal{L}_{thr} = \mathcal{L}_{un}(f^{\mathcal{S}}(\mathcal{I}_{thr}), f^{\mathcal{T}}_{thr}(\mathcal{I}_{thr})) + \mathcal{L}_{un}(f^{\mathcal{S}}(\mathcal{I}_{thr}), f^{\mathcal{T}}_{rgb}(\mathcal{I}_{thr})), \tag{4}$$

where, $\mathcal{L}_{thr}$ is the loss for the thermal domain, and $\mathcal{L}_{rgb\_sup}$ represents the supervised loss for the RGB domain. Similarly, $\mathcal{I}_{thr}$ and $\mathcal{I}_{rgb}$ denote the images from thermal and RGB domains, respectively. $f^{\mathcal{S}}$ corresponds to the student model, which generates predictions for the input images. Whereas $f^{\mathcal{T}}_{thr}$ and $f^{\mathcal{T}}_{rgb}$, representing the teacher models for the thermal and RGB domains, are responsible for generating pseudo-labels to train the student model. $\mathcal{Y}$ denotes the ground truth labels for the images in the RGB source domain. The losses consist of the unsupervised loss $\mathcal{L}_{un}$ and the supervised loss $\mathcal{L}_{sup}$, which are used like [10].

3.3. Zigzag Learning Across RGB-Thermal Domains

In traditional UDA methods for object detection, the source and target domains were commonly trained simultaneously. However, due to the substantial domain gap between the RGB and thermal domains, simultaneously training is ineffective. We propose a training approach for domain adaptation from RGB to thermal, which is called zigzag learning.

Distinctive training: The zigzag learning involves separate and alternate training for the RGB and thermal domains to learn the distinct knowledge of each domain effectively. Each time we train a specific domain, we update the weights to the teacher model of the corresponding domain using EMA. This domain specific training and weight updating strategy ensures that the significant domain gap between the RGB and thermal domains does not result in negative cross domain influence.

Progressive training transition: The concept of the zigzag learning method is a progressive training transfer process that starts with a focus on learning knowledge from the labeled RGB domain. Next, the training progressively transitions to the thermal domain by steadily increasing the training frequency for thermal images and simultaneously reducing the training frequency for RGB images. This gradual shift facilitates a smooth domain adaptation from the RGB to the thermal domain, resulting in improved performance within the thermal domain. As illustrated in Fig. 1 (b), for example, the unlabeled thermal domain is trained a single time at first, while the labeled RGB domain is trained

Algorithm 1 Zigzag learning method

```
Require:
  $I$: Total number of iterations, $\alpha$: EMA coefficient, $Z_{thr}$, $Z_{rgb}$: Training iterations of thermal and RGB at each step, $\theta^{\mathcal{S}}$,$\theta^{\mathcal{T}}_{thr}$,$\theta^{\mathcal{T}}_{rgb}$: Weights of student, thermal and RGB teachers, $\mathcal{I}_{thr}$, $\mathcal{I}_{rgb}$: Input of thermal and RGB images.
Ensure:
  $switch \leftarrow Z_{thr}$
  for iteration $i \in \{0, 1, 2, \ldots, I\}$ do
      if $i < switch$ then        ▷ Update only thermal teacher
          Calculate $\mathcal{L}_{thr}$ by $\mathcal{I}_{thr}$
          $\theta^{\mathcal{S}} \leftarrow \mathcal{L}_{thr}$
          $\theta^{\mathcal{T}}_{thr} \leftarrow \alpha\theta^{\mathcal{T}}_{thr} + (1-\alpha)\theta^{\mathcal{S}}$
      else                        ▷ Update only RGB teacher
          Calculate $\mathcal{L}_{rgb}$ by $\mathcal{I}_{rgb}$
          $\theta^{\mathcal{S}} \leftarrow \mathcal{L}_{rgb}$
          $\theta^{\mathcal{T}}_{rgb} \leftarrow \alpha\theta^{\mathcal{T}}_{rgb} + (1-\alpha)\theta^{\mathcal{S}}$
      end if
      if $i > 0$ and $i\%(Z_{thr} + Z_{rgb}) == 0$ then
          $switch \leftarrow switch + Z_{thr} + Z_{rgb}$
      end if
  end for
```

three times to focus on acquiring knowledge from the RGB domain. Subsequently, the frequency of training sessions in the RGB domain is decreased, while it is increased for the thermal domain, facilitating domain adaptation from RGB to thermal between the two domains. The training iterations for the RGB and thermal domains at each step are defined as follows:

$$\begin{aligned} Z^{t}_{thr} &= Z^{t-1}_{thr} + \beta, \\ Z^{t}_{rgb} &= Z^{t-1}_{rgb} - \beta, \end{aligned} \tag{5}$$

where Z^t_{thr} and Z^t_{rgb} are the number of training iterations for the thermal domain and RGB domain at t^{th} step, Z^{t-1}_{thr} and Z^{t-1}_{rgb} are the number of training iterations at $(t-1)^{th}$ step. β indicates the number of iterations that are adjusted after each step. This equation guarantees that the sum of Z^t_{thr} and Z^t_{rgb} remains constant, while the ratio $Z^t_{\text{thr}} : Z^t_{\text{rgb}}$ increases incrementally at each step. We have presented a pseudocode of the zigzag learning algorithm in Algorithm 1.

3.4. Incorporating Knowledge from Teacher Models

Our experiments on the effectiveness of domain adaptation techniques indicate some limitations when training within the RGB domain using only ground truth labels. In this section, we describe the limitations and propose an improved strategy that integrates pseudo-labels to enhance knowledge transfer.

Limitations of training with only ground truth labels: We found that training the student model using only ground truth labels poses challenges because the complexity of the ground truth labels makes it difficult for the student model to learn effectively from strongly augmented input images. This leads us to our first observation: a combination of ground truth and pseudo-labels is more effective for knowledge transfer from the teacher model to the student model than training with only ground truth labels. This combination makes the process of transferring knowledge from the teacher model to the student model more effective.

Secondly, we find that training solely with ground truth labels from the RGB domain does not utilize the knowledge synthesized by the thermal teacher model, thereby reducing the effectiveness of domain adaptation from the RGB to the thermal domain. To address these issues, we strategically integrate pseudo-labels generated by both the RGB and thermal teacher models, as well as ground truth labels, into the training for the RGB domain.

Pseudo label integration: However, the direct use of pseudo-labels leads to poor results. The experiments detailed in Table 6 indicate that using pseudo-labels in the same manner as ground truth labels (with λ equals 1) results in a substantial decline in model performance. As in Section 3.3, our method initially focuses on training with ground truth labels, and then we gradually integrate the pseudo-labels from both teachers, alongside the ground truth labels, into the training process. This approach is defined by the following set of equations:

$$\mathcal{L}_{rgb\_unsup} = \mathcal{L}_{un}(f^{\mathcal{S}}(\mathcal{I}_{rgb}), f^{\mathcal{T}}_{rgb}(\mathcal{I}_{rgb})) + \mathcal{L}_{un}(f^{\mathcal{S}}(\mathcal{I}_{rgb}), f^{\mathcal{T}}_{thr}(\mathcal{I}_{rgb})), \tag{6}$$

$$\mathcal{L}_{rgb} = \mathcal{L}_{rgb\_sup} + \lambda\mathcal{L}_{rgb\_unsup}. \tag{7}$$

In this equation, λ is a hyperparameter that controls the degree to which pseudo-labels are used during training in the RGB domain. This hyperparameter is employed to balance the influence of pseudo-labels and ensure that the student model benefits from the knowledge provided by the teacher models without negative effects. The unsupervised loss $\mathcal{L}_{un}$ is utilized in a similar manner as in Section 3.2.

The total loss for the D3T model is formulated as follows:

$$\mathcal{L}_{all} = \begin{cases} \mathcal{L}_{thr} & \text{training with thermal domain,} \\ \mathcal{L}_{rgb} & \text{training with RGB domain.} \end{cases} \tag{8}$$

4. Experimental Results and Discussions

4.1. Dataset and Evaluation Protocol

We evaluate our proposed method using the following datasets and new designed domain adaptation evaluation protocols from RGB to thermal domains;

FLIR [44]: In our research, we chose the updated FLIR dataset over the older one [15] because it has many labeling errors. The dataset includes 5,142 precisely aligned pairs of color and infrared images, with 4,129 used to train our method and 1,013 used for testing it. These images are from the view of a car driver and include both daytime and nighttime scenes. We are only looking at objects like "people," "cars," and "bicycle" that have complete labels to make sure our evaluation is accurate.

KAIST [19]: The renowned KAIST dataset comprises 95,328 pairs of color and thermal images. We employ an updated version with more precise labeling as provided by [46]. This version contains 8,892 accurately adjusted pairs of RGB-Thermal images for training and 2,252 pairs for evaluation purposes.

RGB→Thermal FLIR evaluation: The FLIR dataset, known for its precisely aligned image pairs, can cause models to overfit and may not accurately reflect the true performance of domain adaptation algorithms. To address this, we introduce a disjointed image training approach. We use the first 2,064 RGB images as the source domain and a separate set of 2,064 thermal images as the target domain for training. Note that RGB source and thermal target images are exclusively selected. This method guarantees that the training does not use any matching RGB-Thermal image pairs, preventing overfitting and providing a more reliable assessment of the domain adaptation algorithm's effectiveness.

RGB→Thermal KAIST evaluation: Like with the FLIR dataset, we apply a disjointed image training approach for the KAIST dataset. We select the initial 4,446 RGB images as the source domain and the subsequent 4,446 thermal images as the target domain, ensuring that training does not involve any matched image pairs. RGB source and thermal target images are exclusively selected. Furthermore, we have removed any images without labels, resulting in a total of 1,216 images to validate the algorithm's performances.

4.2. Implemental Details

Following the baseline [10], we deploy the FCOS detector [38] equipped with a VGG-16 backbone for the FLIR dataset and a ResNet-50 backbone for the KAIST dataset in our experiments. Our experiments run on a batch size of 8 using 4 NVIDIA RTX A5000 GPUs. In accordance with [10], we initiate the learning rate at 0.005 and not apply any decay. For data augmentation, we adopt the same strategy as in [10, 29], resizing the shortest edge of images to a maximum of 800 pixels. In Section 3.3, for the FLIR dataset, Z^0_{thr} and Z^0_{rgb} are initialized to 50 and 150, respectively. They will be adjusted every 10k iterations by a β value of 50 as specified in the equation (5). For the KAIST dataset, Z^0_{thr}, Z^0_{rgb}, and β are initially established at 25, 75, and 25 respectively, and each adjustment step comprises 10k iterations.

Method	Person	Bicycles	Car	mAP
Source only	28.54	28.28	47.22	34.68
DANN [14]	32.02	30.52	48.88	37.14
SWDA [35]	30.91	36.03	47.94	38.29
EPM [18]	40.97	38.95	53.83	44.60
HT [10]	**70.87**	48.11	78.45	65.81
D3T(Ours)	70.77	**57.44**	**79.68**	**69.30**

Table 1. Adaptation results of FLIR dataset from RGB images to thermal images with VGG16 backbone. The best accuracy is indicated in bold, and the second-best accuracy is underlined.

4.3. Performance Comparison Table

We compare our proposed method with the well-known domain adaptation methods.

RGB→Thermal FLIR evaluation: The adaptation results for RGB to thermal image conversion on the FLIR dataset, as presented in Table 1, indicate that our D3T method has achieved remarkable performance, surpassing other advanced technologies in domain adaptation. Specifically, the D3T method outperforms the HT [10] algorithm, which is a significant player in the field utilizing a student-teacher framework, by 3.49% in mean Average Precision (mAP). Notably, HT [10] itself had previously set a high benchmark by outperforming the EPM [18] method, which does not use the student-teacher approach, by 21.21% in mAP.

The advancements observed in our study highlight a critical insight: previous algorithms have not adequately tackled the considerable domain gap between RGB and thermal domains. This gap poses a more formidable challenge than those encountered in typical adaptation scenarios, such as transitioning from Cityscapes to Foggy Cityscapes. Our experimental results unequivocally showcase the efficacy of the proposed D3T method in effectively addressing this issue, marking a substantial leap forward in domain adaptation.

RGB→Thermal KAIST evaluation: The domain adaptation results for converting RGB images to thermal images on the KAIST dataset, as depicted in Table 2, demonstrate the superior performance of our D3T algorithm. D3T outperforms the HT [10] algorithm, which is one of the most advanced algorithms in this domain, by a significant margin of 5.51% in mAP. Furthermore, when compared to the EPM [18] algorithm, which does not utilize a student-teacher framework, the D3T method shows an even greater improvement of 9.41% mAP. This impressive advancement illustrates the effectiveness of the D3T algorithm in addressing the challenges of domain adaptation from RGB to thermal domain.

4.4. Ablation Experiments

We make ablations and detail discussions in this section.

Method	Person
RGB Source only	9.09
DANN [14]	9.17
SWDA [35]	31.30
EPM [18]	39.55
HT [10]	43.45
D3T(Ours)	**48.96**

Table 2. Adaptation results on KAIST dataset from RGB images to thermal images with Resnet-50 backbone. The best accuracy is indicated in bold, and the second-best accuracy is underlined.

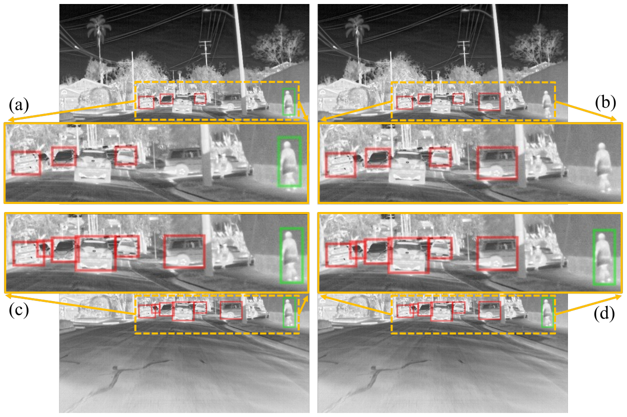


Figure 3. Dual-teachers' pseudo-labels at different training stages. (a) and (b) are pseudo-labels from the RGB and thermal teacher models in early training stages, respectively, while (c) and (d) are pseudo-labels from the same models in later training stages.

Visualization: Fig. 3 illustrates the effectiveness of our D3T model. At the early training steps, each teacher holds specific knowledge relevant to their respective domain. Therefore, the teachers created different pseudo labels as illustrated in Fig. 3 (a) and (b). In the final training steps, the two teachers provided pseudo labels of high quality that were similar. This indicates that our D3T algorithm improves model efficiency and bridges the domain gap. This corresponds to our concept presented in Fig. 1 (b). We also show object detection results on two datasets, FLIR and KAIST, in Fig. 4 and Fig. 5 to provide a visual comparison of the effectiveness of our D3T method.

Dual-Teachers	Zigzag-Learn	Incor-Know	mAP
			65.81
✓			66.93
✓	✓		68.46
✓	✓	✓	**69.30**

Table 3. Ablation studies of D3T on FLIR dataset from RGB images to thermal images. Dual-Teachers, Zigzag-Learn and Incor-Know refer to Distinctive Dual-Domain teachers, zigzag learning Across RGB-Thermal domains and Incorporating Knowledge from teacher Models.

Oracle		Ours
RGB only	Thermal only	
34.68	65.04	69.30

Table 4. Comparison of our D3T on FLIR dataset from RGB images to thermal images with the oracle single FCOS models.

Method Iteration	Fix 50	Fix 100	Fix 1,000	Zigzag Dynamic
mAP	65.57	68.28	65.36	**69.30**

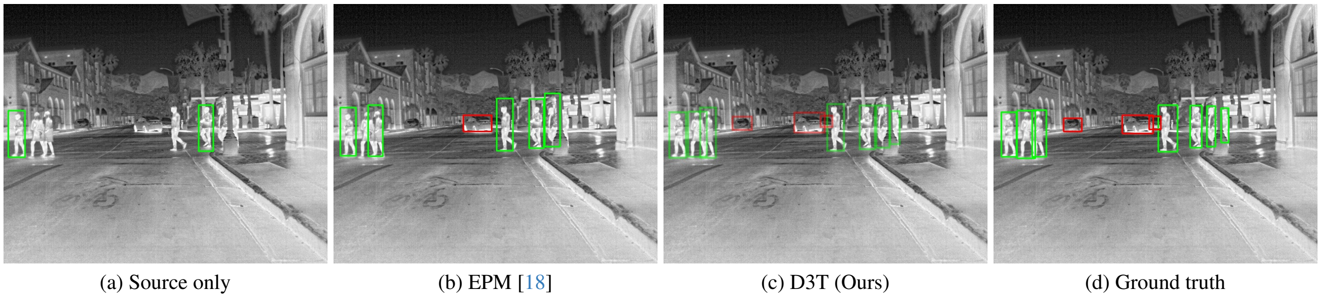
Table 5. Comparison of zigzag learning across RGB-Thermal domains on FLIR dataset from RGB images to thermal images with different iteration settings.

Ablation study: The ablation study for the D3T model, as detailed in Table 3, assesses the adaptation from RGB to thermal images on the FLIR dataset and sheds light on the contributions of various model components. Starting from a baseline mAP of 65.81%, the addition of Dual-Teachers provides a notable improvement, bringing the mAP to 66.93%. Further incorporation of Zigzag-Learn with Dual-Teachers enhances the mAP marginally to 68.46%, suggesting the effectiveness of alternating training strategies in domain adaptation. The full model, integrating Dual-Teachers, Zigzag-Learn, and Incor-Know, achieves the most significant performance leap, culminating in an mAP of 69.30%. This comprehensive approach highlights the synergistic impact of combining domain-specific knowledge acquisition, specialized training methods, and robust cross-domain knowledge integration to effectively adapt RGB image detection models for thermal image applications.

Gap between RGB and Thermal: We provide both qualitative and quantitative evidences to support our motivation on the large domain gap between RGB and thermal domains. Fig. 6, featuring KAIST thermal images, distinctly highlights the unique characteristics of different sensors. Table 4 shows a substantial performance gap, with results showing 34.68% for RGB and 65.04% for thermal oracles, providing quantitative evidence for our motivation.

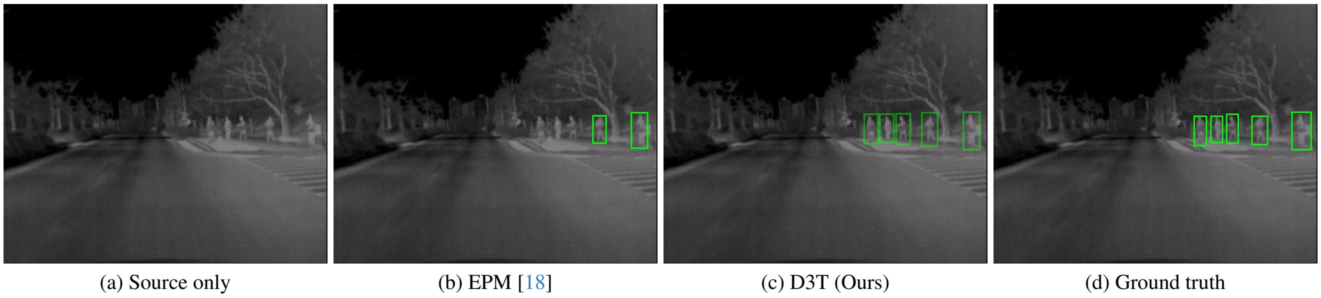
Effect of zigzag learning across RGB-thermal gap: Table 5 presents the outcomes of employing zigzag learning across RGB-Thermal domains on the FLIR dataset for adapting from RGB to thermal images, utilizing dynamic iteration settings. The 'Fix' setting refers to a consistent training regime where each domain is trained for an equal number of iterations, e.g., 100. In contrast, the 'zigzag' setting, as detailed in Section 3.3, begins with a focus on the RGB domain before progressively shifting emphasis towards the thermal domain. Note that the frequency of teacher selection is dynamically changed as learning advances. The results indicate that the 'zigzag' approach yields a superior mAP by 1.02%, demonstrating its effectiveness over the 'Fix' setting method.

Effect of incorporating knowledge from teacher models: Table 6 illustrates the impact of employing pseudo labels to enhance learning capabilities and bridge the domain gap between RGB and thermal domains on the FLIR dataset. The table compares the performance of models



(a) Source only (b) EPM [18] (c) D3T (Ours) (d) Ground truth

Figure 4. Visualization of UDA results for object detection models: Source only, EPM [18], our D3T, and ground truth labels in the FLIR dataset RGB → thermal domain. The green and red boxes represent the classes of person and car.



(a) Source only (b) EPM [18] (c) D3T (Ours) (d) Ground truth

Figure 5. Visualization of UDA results for object detection models: Source only, EPM [18], our D3T, and ground truth labels in the KAIST dataset RGB → thermal domain. The green boxes represent the classes of person.

Figure 6. KAIST RGB and thermal images illustrating disparities between the two domains. Evaluation is conducted without the use of this paired information.

Method λ	Fixed 0	Fixed 1	Fixed 0.1	Dynamic 0→1
mAP	68.46	55.12	68.57	**69.30**

Table 6. Comparison of incorporating knowledge from teacher models on FLIR dataset from RGB images to thermal images with different λ values.

trained without pseudo labels, with a fixed λ hyperparameter and with a dynamically λ hyperparameter for equation (7). The results reveal that not using pseudo labels results in an mAP of 68.46%, whereas using pseudo labels that closely resemble real labels with a λ of 1 leads to a notable decrease in performance, dropping it to 55.12%. A fixed λ of 0.1 improves the mAP to 68.57%, and a dynamically changing λ of 0→1 achieves the best mAP at 69.30%. This suggests that dynamically adjusting the level of pseudo labels usage during training is an effective strategy to alleviate RGB-Thermal gap, as teachers adapt to the target domain in the later stages of training, instilling trust in their accuracy.

5. Conclusion

Our research has effectively navigated the challenges of domain adaptation for object detection from RGB to the thermal domain, a task typically constrained by a lack of extensive thermal datasets and significant domain disparities stemming from RGB data. We have put forth the D3T framework, a novel approach that leverages a dual-teacher model coupled with a zigzag learning regimen, meticulously tailored for adapting from RGB to thermal image. This method markedly enhances model performance, enabling smooth transitions and a focused application of domain-specific knowledge. The results highlight the efficacy of our approach. Our method establishes a solid foundation for subsequent innovations in UDA and sets a groundbreaking benchmark for thermal object detection, bolstering applications dependent on trustworthy vision systems across diverse conditions.

Acknowledgement: This work was partially supported by Hyundai Motor Company and Kia, and supported by Korea IITP grants (IITP-2024-No.RS-2023-00255968, AI Convergence Innovation Human Resources Dev.; RS-2023-00236245, Dev. of Perception/Planning AI SW for Seamless Auton. Driving in Adverse Weather/Unstructured Env.; No.2021-0-02068, AI Innovation Hub) and by NRF grant (NRF-2022R1A2C1091402). W. Hwang is the corresponding author.

References

[1] KR Akshatha, A Kotegar Karunakar, Satish B Shenoy, Abhilash K Pai, Nikhil Hunjanal Nagaraj, and Sambhav Singh Rohatgi. Human detection in aerial thermal images using faster r-cnn and ssd algorithms. *Electronics*, 11(7):1151, 2022. 1, 2

[2] Bing Cao, Yiming Sun, Pengfei Zhu, and Qinghua Hu. Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23555–23564, 2023. 2

[3] Shengcao Cao, Dhiraj Joshi, Liang-Yan Gui, and Yu-Xiong Wang. Contrastive mean teacher for domain adaptive object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23839–23848, 2023. 2

[4] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8869–8878, 2020. 2

[5] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 2

[6] Yi-Ting Chen, Jinghao Shi, Zelin Ye, Christoph Mertz, Deva Ramanan, and Shu Kong. Multimodal object detection via probabilistic ensembling. In *European Conference on Computer Vision*, pages 139–158. Springer, 2022. 2

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 3

[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. Ieee, 2009. 1

[9] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4101, 2021. 2, 3

[10] Jinhong Deng, Dongli Xu, Wen Li, and Lixin Duan. Harmonious teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23829–23838, 2023. 2, 3, 4, 6, 7

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 1

[12] Rikke Gade and Thomas B Moeslund. Thermal cameras and applications: a survey. *Machine vision and applications*, 25: 245–262, 2014. 1

[13] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 2

[14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 6, 7

[15] FA Group et al. Flir thermal dataset for algorithm training, 2018. 6

[16] Meryem Mine Gündoğan, Tolga Aksoy, Alptekin Temizel, and Ugur Halici. Ir reasoner: Real-time infrared object detection by visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 422–430, 2023. 1, 2

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1

[18] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 733–748. Springer, 2020. 2, 6, 7, 8

[19] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015. 2, 6

[20] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5001–5009, 2018. 2

[21] Rohan Ippalapally, Sri Harsha Mudumba, Meghana Adkay, and Nandi Vardhan HR. Object detection using thermal imaging. In *2020 IEEE 17th India Council International Conference (INDICON)*, pages 1–6. IEEE, 2020. 1, 2

[22] Chenchen Jiang, Huazhong Ren, Xin Ye, Jinshun Zhu, Hui Zeng, Yang Nan, Min Sun, Xiang Ren, and Hongtao Huo. Object detection from uav thermal infrared images and videos using yolo models. *International Journal of Applied Earth Observation and Geoinformation*, 112:102912, 2022. 1, 2

[23] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12456–12465, 2019. 2

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012. 1

[25] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Com-*

puter Vision and Pattern Recognition, pages 5291–5300, 2022. 2

[26] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7581–7590, 2022. 2, 3

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 2

[28] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016. 1

[29] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*, 2021. 3, 6

[30] Jaemin Na, Jung woo Ha, Hyung Jin Chang, Dongyoon Han, and Wonjun Hwang. Switching temporary teachers for semi-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 36, 2023. 3

[31] Yuzuru Nakamura, Yasunori Ishii, Yuki Maruyama, and Takayoshi Yamashita. Few-shot adaptive object detection with cross-domain cutmix. In *Proceedings of the Asian Conference on Computer Vision*, pages 1350–1367, 2022. 3

[32] Heena Patel and Kishor P Upla. Night vision surveillance: Object detection using thermal and visible images. In *2020 International Conference for Emerging Technology (INCET)*, pages 1–6. IEEE, 2020. 1

[33] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015. 1

[35] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6956–6965, 2019. 2, 6, 7

[36] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018. 3

[37] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30, 2017. 3

[38] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 1, 2, 6

[39] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 2

[40] Vibashan Vs, Domenick Poster, Suya You, Shuowen Hu, and Vishal M Patel. Meta-uda: Unsupervised domain adaptive thermal object detection using meta-learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1412–1423, 2022. 3

[41] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *International conference on machine learning*, pages 5423–5432. PMLR, 2018. 2

[42] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12355–12364, 2020. 2

[43] Shengbo Yao, Qiuyu Zhu, Tao Zhang, Wennan Cui, and Peimin Yan. Infrared image small-target detection based on improved fcos and spatio-temporal features. *Electronics*, 11 (6):933, 2022. 2

[44] Heng Zhang, Elisa Fromont, Sébastien Lefevre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *2020 IEEE International conference on image processing (ICIP)*, pages 276–280. IEEE, 2020. 2, 6

[45] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon. Guided attentive feature fusion for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 72–80, 2021. 2

[46] Lu Zhang, Xiangyu Zhu, Xiangyu Chen, Xu Yang, Zhen Lei, and Zhiyong Liu. Weakly aligned cross-modal learning for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5127–5137, 2019. 6

[47] Yixin Zhang, Zilei Wang, and Yushi Mao. Rpn prototype alignment for domain adaptive object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12425–12434, 2021. 2

[48] Kailai Zhou, Linsen Chen, and Xun Cao. Improving multispectral pedestrian detection by addressing modality imbalance problems. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 787–803. Springer, 2020. 2