Unsupervised thermal-to-visible domain adaptation method for pedestrian detection

Mohamed Amine Marnissi, Hajer Fradi, Anis Sahbani, Najoua Essoukri Ben Amara

Unsupervised thermal-to-visible domain adaptation method for pedestrian detection

Mohamed Amine Marnissi*†, Hajer Fradi‡, Anis Sahbani§¶ and Najoua Essoukri Ben Amara*
*Université de Sousse, Ecole Nationale d'Ingénieurs de Sousse, LATIS-Laboratory of Advanced Technology and Intelligent Systems, 4023, Sousse, Tunisie;
‡Université de Sousse, Institut Supérieur des Sciences Appliquées, LATIS-Laboratory of Advanced Technology and Intelligent Systems, 4023, Sousse, Tunisie;
†Université de Sfax, Ecole Nationale d'Ingénieurs de Sfax, 3038, Sfax, Tunisie;
§Enova Robotics S.A.
¶Sorbonne Université, CNRS, Ins titule for intelligent Systems and Robotics (ISIR), Paris, France
Emails: medamine.marnissi@eniso.u-sousse.tn, hajer.fradi@issatso.rnu.tn, anis.sahbani@enovarobotics.com and najoua.benamara@eniso.rnu.tn

Abstract—**Pedestrian detection is a common task in the research area of video analysis and its results lay the foundations of a wide range of applications. It is commonly known that under challenging illumination and weather conditions, conventional visible cameras perform poorly and this limitation could be catered using thermal imagery. But, due to the fact that annotated thermal datasets are less available compared to the visible ones, in this paper we emphasis the need for leveraging information from both domains at no additional annotation cost. Precisely, we propose a domain adaptation method by incorporating feature distribution alignments into Faster R-CNN architecture at different levels and at different stages of the network. The resulting proposed thermal-to-visible adaptive detector has the advantage of covering different aspects of the domain shift in order to improve the overall performance. The proposed detector is evaluated on KAIST multispectral dataset and the obtained results demonstrate its effectiveness by improving the adaptability in the thermal domain. Also, by means of comparisons to other existing works, better results are obtained.**

I. Introduction

The past few decades have witnessed a widespread growth in the use of infrared cameras in many fields, including military and civilian ones especially for automotive applications, medical imaging, robotics and video surveillance [1], [2], [3]. These cameras form an image by detecting infrared radiations emitted by an object having a high temperature. The main advantage using infrared cameras is that they could easily distinguish warm objects from other surrounding objects. Consequently, these cameras are proven to be more convenient at nighttime and in adverse weather conditions compared to the conventional visible cameras [4], [1], [5]. Despite the usefulness of these cameras in such situations, there are some limitations that have to be considered, essentially about the expensive cost of high-resolution ones. This could explain the fact that thermal data is usually of bad quality and is less available compared to visible one [6], [7], [8], [4].

In this paper, we essentially focus on the problem of pedestrians detection and localization. This problem has been extensively studied on visible datasets using deep learning networks [9], [10], [11], [12]. It is commonly known that these networks rely on a large labeled training data, which might incur at least two problems in the thermal domain: first, less available data compared to the visible domain; second, the annotation task for object detection which is particularly time-consuming task since each object category in every image must be precisely delimited with a bounding box.

To mitigate these problems, we intend in this present paper to adapt the abundance of annotated visible images to the thermal domain at no additional annotation cost. Precisely, we propose to incorporate feature distribution alignments into a baseline detector. The key idea is basically inspired from [13] and [14], but a more complete architecture is proposed by performing alignments at multiple levels and at two phases of the network for complementary aspect to further improve the overall performance.

The proposed adaptive detector covering different aspects of the domain shift is considered as the main contribution of this current paper. It has also the advantage to be trained in unpaired setting with unlabeled thermal images. Such unsupervised adaptation of object detectors from source to target domains has been previously employed on other domains in the visible spectrum. But it is proposed for the first time in thermal and visible domains. Targeting such dissimilar domains is particularly challenging, since they exhibit different visual characteristics. The proposed thermal-to-visible adaptation is of significant interest since it allows faster execution time and less consumption of resources by reducing the annotation costs associated with detection and applying a single adaptive detector for both domains. Despite its relevance in real applications, such unsupervised adaptation for detection in these domains is not yet investigated in the literature, as far as we know. Consequently, we consider that this paper could potentially open a new path for improving the domain adaptability of existing detectors in thermal and visible domains. In addition, the effectiveness of the proposed adaptation method is demonstrated by improving the detection results with a significant margin compared to the baseline method and to recently published works in the field.

The rest of the paper is organized as follows: in Section II, an overview of the existing methods for object detection and domain adaptation is presented. Then, our proposed approach of thermal-to-visible domain adaptation for detection is detailed in Section III. The conducted experiments and the obtained results are discussed in Section IV. Finally, in Section

V we conclude and give some potential perspectives.

II. Related Work

In this section, we first give an overview of the existing detectors in visible, thermal and in both domains. This overview includes adaptive person detectors in different domains as well.

A. Pedestrian Detection

Object detection consists of precisely identifying and localizing pertinent objects in a given image by classification or by regression. The current popular detectors make use of deep learning networks such as Fast R-CNN [15], Faster R-CNN [16], Single Shot Detector (SSD) [17], You Only Look Once (YOLO) [18], [19], EfficientDet [20], and RetinaNet [21]. Generally, object detection models can be divided into two categories: one-stage and two-stage detectors. The first category is based on one single shot to detect several objects such as SSD and YOLO detectors. The second category requires two stages: the first one consists of generating region proposal networks (RPN) and the second one aims at detecting objects of each proposal as the case of Fast R-CNN and Faster R-CNN detectors.

In this paper, we precisely focus on the problem of pedestrians detection and localization. This problem has attracted research attention because of its usefulness in many applications including video surveillance and driving assistance systems. Thanks to the availability of visible cameras, it has been often studied in the visible spectrum [9], [10], [11], [22], [12]. Despite their widespread applications, visible cameras are not convenient in some situations for instance, in nighttime, bad lighting conditions, adverse weather conditions or in total darkness [23], [24]. In such situations, thermal cameras have instead been proven effective.

In this context, some research studies for detection using thermal imagery have been conducted in the literature [4], [8], [24], [25], [26], [27]. These studies can be categorized into three parts: thermal images only, fusion of thermal and visible images, and thermal images with transfer. For the first category, few works have addressed the problem of detection using only thermal imagery [4], [8], [28], [25]. For instance, in [4] thermal images augmented with their saliency maps to serve as an attention mechanism for the pedestrian detector are employed. From the obtained results, it has been shown that the saliency maps provide complementary information to the pedestrian detector resulting in a significant improvement in performance over the baseline approach. Also, an enhancement architecture based on Generative Adversarial Network, and composed of contrast enhancement and denoising modules is proposed in [8]. The proposed architecture has shown its advantage to enhance the overall thermal image quality and to further improve the detection results.

To deal with vast range of weather and lighting conditions (rain, fog, daytime and nighttime), multispectral detectors that combine information from thermal and visible images have been proposed [29], [30], [23], [31], [32], [33]. Related work in this context includes the different fusion schemes (early, halfway, late and score) introduced in [32] to combine pairs of visible and thermal images. Also, an aligned region CNN to handle the weakly aligned multispectral data in an end-to-end way is proposed in [33]. MSDS-RCNN [23] is another fusion method composed of a multispectral proposal network (MPN) and a multispectral classification network (MCN).

Usually, these multispectral detectors are based on more complex network architectures compared to the detection in a single spectrum visible/thermal. Moreover, these detectors rely in most cases on aligned sensors (thermal and visible) at inference time, which could limit their feasibility in real-time applications. Because of the aforementioned reasons, some recent works instead operate on one single-modality (usually thermal one) and leverage information from the visible spectrum by means of transfer learning or domain adaptation. Related works in this field are reviewed in details in the next section.

B. Domain Adaptation for Detection

Domain adaptation has been widely studied in the field of computer vision for various visual applications including image classification, object detection, fine-grained recognition and semantic segmentation [34], [35], [36], [14], [37], [38]. Conventional methods essentially include domain transfer multiple kernel learning, asymmetric metric learning, feature alignment, subspace interpolation, subspace alignment, and covariance matrix alignment [14].

Different from other domain adaptation methods, adaptation for detection is particularly challenging since both object category and location have to be predicted. Related works for detection operate either in one single spectrum (visible or thermal) or between thermal and visible spectrums. In the thermal spectrum, a task-conditioned domain adaptation between daytime and nighttime is proposed in [1]. Precisely, an auxiliary classification task that distinguishes between daytime and nighttime thermal images is added to the main detection task. This classification task is used to condition a YOLOv3 in order to improve its adaptation to the thermal domain.

Other research studies are conducted for adaptation from thermal to visible domains. Among these studies, some works aim at generating a perceptually realistic RGB image from an input thermal image (commonly known as colorization) usually by means of generative networks [39], [26], [40]. However, in this overview, we are interested in the existing works that perform this transformation to enable better detection. It is the case of [26], where a Cycle-GAN for unpaired image-to-image translation of thermal to pseudo-RGB data is proposed to fine-tune a multimodal Faster-RCNN detector.

For adaptation in the opposite direction from visible to thermal domains, in [27] visible images are transformed to synthetic thermal images. This transformation acts as a data augmentation for training a pedestrian detector to work on thermal imagery. Also, a cross-modality learning framework composed of a Region Reconstruction Network (RRN) and Multi-Scale Detection Network (MDN) is proposed in [5]. RRN is used to learn a non-linear mapping from the RGB channels to the thermal channel in order to improve detection results from visible data. In [41], a domain adaptation method based on style consistency is used to transfer low-level features from the visible to the infrared domains. The cross-domain model with style consistency is used for object detection in the infrared spectrum. Compared to the previous works, in [42] an unified detection network by defining a common feature space, which makes intermediate features from the two domains is proposed.

Unlike the existing adaptation works for detection that require annotated data for both thermal and visible domains, in this paper our primary goal is to perform adaptation without the need for annotating thermal data. In the literature, only few works that tackled the task of unsupervised adaptation for detection have been proposed in other domains but always in the visible spectrum [14], [43], [13]. Domain adaptive Faster R-CNN [14] is one of the most known unsupervised adaptive detectors, where an image-level and an instance-level adaptation components are proposed to alleviate the performance drop caused by the domain shift. These adaptation components are based on adversarial training of H-divergence. A consistency regularizer is also employed to learn a domain-invariant RPN for the Faster R-CNN model. The robustness of the proposed adaptive detector is evaluated on different domain shift scenarios using different datasets such as Cityscapes and KITTI.

In [43], another unsupervised adaptive object detector is proposed. It is based on strong local and weak global alignments for unsupervised adaptation. The weak alignment model focuses on the adversarial alignment loss on images that are globally similar and puts less emphasis on aligning images that are globally dissimilar. And the strong domain alignment model is designed to only consider local receptive fields of the feature maps. Another recent approach called Hierarchical Transferability Calibration Network (HTCN) [13] to harmonize transferability and discriminability for cross-domain object detection is proposed. The idea consists of regularizing the adversarial adaptation by calibrating the representation transferability with improved discriminability.

Following the same strategy, in this current paper, we make use of such techniques of adaptation without additional supervision in the target domain. But differently from the previous works, where images from two domains but always in the same spectrum (visible) are aligned, our proposed architecture aims at aligning images from different domains (thermal and visible) exhibiting different visual features. In addition, we propose a more complete architecture, where feature distributions are aligned at different levels and at different stages of the network.

III. Domain Adaptation Components for Detection

In this section, the main domain adaptation components of our proposed adaptive detector are presented. Precisely, we choose to incorporate this adaptation into Faster R-CNN which is a representative two-stage detector [16]. Practically, an input image is fed to the backbone network in order to produce a feature map. Then, Region Proposal Network (RPN) generates region proposals based on this feature map in a first stage and Faster R-CNN feeds the region proposals and feature map into ROI pooling layer in a second stage.

In our proposed approach, the adaptation consists of aligning feature distributions at both sub-networks of Faster R-CNN. Precisely, global and local alignments are performed at the backbone network. Details about alignments at this first sub-network are presented in section III-A. In the second sub-network that consists of RPN and ROI, the alignments are performed at image and instance levels, with a consistency regularization. Details about alignments at this second sub-network are given in section III-B.

For every alignment step, a domain classifier is defined. It is a neural network that aims at predicting whether the feature distribution is from the source or the target domain. These alignments at multiple levels and at different sub-networks are combined in section III-C in order to cover different aspects of the domain shift such as image style, scale, illumination, object appearance, and size. An algorithm showing in pseudo-code an overview of the proposed detector is given as well. The overall proposed architecture is shown in Fig. 1. The same notations in this figure are used in the remainder of this section, where we describe each of these architecture components.

A. First Sub-network Alignments

In our architecture, we use ResNet-101 as backbone network which is composed of two feature extractors G_1 and G_2 (see Fig. 1). The first one is employed for the extraction of low-level features and the second one for high-level features. In this section, we explain how local and global alignments are performed at low-level and high-level features, respectively.

1) Local Feature Alignment: Given two images from the two domains, since some local regions could be more important than others, we propose an attention module that matches the corresponding regions from both domains in unsupervised way. The semantic coherence between domains is considered by calculating masks of local features. It can be done by a local domain classifier D_1, which is defined to highlight local features by producing a domain prediction map, that has the same size WxH as the output of G_1 in the backbone network.

D_1 is considered as a pixel-wise discriminator based on few convolutional layers with a kernel size 1 [13]. Following [44], the least square loss is employed since it has been proven to be stable in the training of the domain classifier and to be useful for aligning low-level features. The pixel-wise adversarial training loss $\mathcal{L}^{loc}$ of local alignment for each domain is defined as:

$$\mathcal{L}_V^{loc} = \frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H} D_1(G_1(x^v))_{wh}^2 \qquad (1)$$

$$\mathcal{L}_T^{loc} = \frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H}(1 - D_1(G_1(x^t))^2)_{wh} \qquad (2)$$

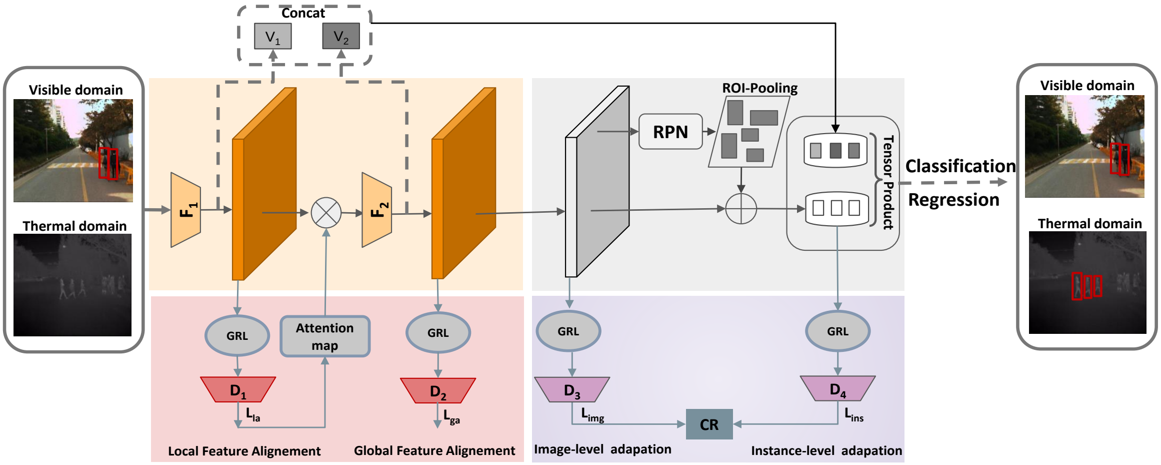


Fig. 1. The architecture of the proposed thermal-to-visible domain adaptation for detection. This adaptation is performed at different levels in the two phases of Faster R-CNN detector. The notations used in this figure are the same used in section III.

where x^v and x^t denote unpaired visible and thermal input images. The two defined pixel-wise adversarial training losses are combined in $\mathcal{L}^{loc}$ by:

$$\mathcal{L}^{loc} = \frac{1}{2}(\mathcal{L}_V^{loc} + \mathcal{L}_T^{loc}) \tag{3}$$

2) Global Feature Alignment: Following [43], in the feature space the target samples can be divided into two parts: Easy-to-classify if they are far from the source samples, and Hard-to-classify if they are close to the source. Therefore, to align domains, we focus on samples that are Hard-to-classify and we put less emphasis on Easy-to-classify samples. At this stage, the focal loss is used to train a domain classifier D_2 for global features alignment. This loss function is used instead of the cross-entropy loss since this latter reinforces the domain classifier D_2 to classify both of the Easy-to-classify and Hard-to-classify samples.

To train a global-level domain classifier D_2, the loss function $\mathcal{L}^{glob}$ based on the focal loss is expressed for visible and thermal domains, respectively as follows:

$$\mathcal{L}_V^{glob} = -(1 - D_2(G_2(x^v)))^\gamma \log(D_2(G_2(x^v))) \tag{4}$$

$$\mathcal{L}_T^{glob} = -(D_2(G_2(x^t)))^\gamma \log(1 - D_2(G_2(x^t)) \tag{5}$$

$$\mathcal{L}^{glob} = \frac{1}{2}(\mathcal{L}_S^{glob} + \mathcal{L}_T^{glob}) \tag{6}$$

where γ is used to weight Hard-to-classify samples during the training process.

3) Contextual Regularization : In the field of adaptive domain segmentation [38], where the goal is to simultaneously generate the domain label and semantic segmentation map, the regularization of the domain classifier together with the segmentation loss have been proven to be efficient to stabilize the adversarial training. Based on that, we choose to integrate a regularization technique in our adaptive model to enhance its performance and to stabilize the training of the domain classifier. Practically, this regularization is applied on the two extracted feature vectors fc_1 and fc_2 (outputs of G_1 and G_2). Each vector includes some contextual information describing the image content. These vectors are then concatenated with the output of ROI-pooling [13]. By doing that, the contextual regularization aims at minimizing the detection loss on visible samples and the domain classification loss while training the domain classifiers D_1 and D_2, as illustrated in Fig. 1.

B. *Second Sub-network Alignments*

At this stage, alignments are applied in the second sub-network of Faster R-CNN composed of RPN and ROI layer at image and instance levels, with a consistency regularization.

1) Image-Level Alignment: To enforce coherency between the two domains, the detection results have to be the same for a given image whether is the domain to which it belongs. In Faster R-CNN model, the image representation I is the resulting feature map of backbone network. Consequently, to solve the domain shift problem, the distributions of the image representation from the two domains have to be the same.

Since, in practice, it is not trivial to reach such alignment at image level, a domain classifier D_3 is employed to minimize the domain distribution difference. D_3 is trained at each activation of feature maps. Then, it predicts the domain label for every image patch, with 0 for the visible domain and 1 for the thermal domain. The advantage of this image-level alignment is that it can generally reduce the amount of shift caused by the differences in the global image such as style, scale and illumination [45], [14]. Using the cross entropy, the adaptation loss at image level denoted as $\mathcal{L}_V^{img}$ and $\mathcal{L}_T^{img}$ in visible and thermal domains is defined as follows:

$$\mathcal{L}_V^{img} = -\sum_{m,n} \log(1 - p_{m,n}) \tag{7}$$

$$\mathcal{L}_T^{img} = -\sum_{m,n} \log(p_{m,n}) \tag{8}$$

where $p_{m,n}$ is the output of the domain classifier D_3 for given activations of feature maps located at (m, n) position after applying backbone network on input image. The two defined adaptation loss functions are combined in $\mathcal{L}^{img}$ by:

$$\mathcal{L}^{img} = \frac{1}{2}(\mathcal{L}_V^{img} + \mathcal{L}_T^{img}) \tag{9}$$

In addition, we employ a gradient reverse layer (GRL) [46] which aims at optimizing the parameters of the domain classifier and the base network, simultaneously. The gradient sign is inversed while passing through the GRL layer to achieve the primary objective of aligning the domain distributions by applying adversarial learning.

2) Instance-Level Alignment: In our method, we also consider instance-level adaptation. It enables reducing the difference of local instance representations between the two domains, such as the appearance and the size of the objects. To reach the semantic consistency, the image region that contains an object and its corresponding category label have to be the same in the visible and thermal domains.

In our proposed adaptive detector, the instance-level representation is used based on the output feature vectors of ROI pooling, that are obtained before being fed to the final category classifier. Since bounding box annotations are known in visible domain but not in thermal domain, a domain classifier D_4 is trained on the feature vectors to align distributions at instance level. The adaptation loss at instance level denoted as $\mathcal{L}_V^{ins}$ and $\mathcal{L}_T^{ins}$ in visible and thermal domains is defined as follows:

$$\mathcal{L}_V^{ins} = -\sum_j \log(1 - s_j) \tag{10}$$

$$\mathcal{L}_T^{ins} = -\sum_j \log(s_j) \tag{11}$$

where s_j is the output of the domain classifier of the j-th region proposal in the input image. The two defined adaptation loss functions are combined in $\mathcal{L}^{ins}$ by:

$$\mathcal{L}^{ins} = \frac{1}{2}(\mathcal{L}_V^{ins} + \mathcal{L}_T^{ins}) \tag{12}$$

Similar to the domain classifier at image level, we add the gradient reverse layer beforehand in order to apply the adversarial training strategy.

3) Consistency Regularization: To align domains at image and instance levels, the distributions of image representation and instance representation have to be the same in the two domains. To solve that, two domain classifiers are trained, where the input could be either the image representation or the instance representation and the output is a probability to predict if an input sample belongs to the thermal domain or not. The consistency between the domain classifiers at different levels (image and instance) allows to learn the cross-domain robustness of the bounding box predictor [47]. It can be defined in visible and thermal domains as:

$$\mathcal{L}^{cst} = \sum_j \parallel \frac{1}{\mid I \mid} \sum_{m,n} p_{m,n} - s_j \parallel_2 \tag{13}$$

where $\mid I \mid$ is the total number of activations in a feature map, and $\parallel . \parallel$ is the l_2 norm.

C. Overall Loss

For a given input visible image x^v with its corresponding bounding boxes y^v, the detection loss of our proposed adaptive Faster R-CNN detector is defined as:

$$\mathcal{L}^{det} = \mathcal{L}^{reg}(R(G_2(x^v)), y^v) + \mathcal{L}^{cls}(R(G_2(x^v)), y^v) \tag{14}$$

where the output of G_2 is fed to RPN module (denoted as R). $\mathcal{L}_{det}$ combines the classification and the regression losses per bounding box.

To train the proposed adaptive detector, the detection loss $\mathcal{L}^{det}$ defined in eq.14 is combined with the adversarial loss $\mathcal{L}^{adv}$:

$$\mathcal{L}^{adv} = \mathcal{L}^{glob} + \mathcal{L}^{loc} + \mathcal{L}^{img} + \mathcal{L}^{ins} + \mathcal{L}^{cst} \tag{15}$$

in an overall objective function $\mathcal{L}$ defined as:

$$\mathcal{L} = \mathcal{L}^{det} + \lambda \mathcal{L}^{adv} \tag{16}$$

where λ is used to weight the adversarial loss.

D. Summary of the Adaptive Detection Algorithm

Algorithm 1 shows in pseudo-code an overview of the training phase of our proposed adaptive detector by integrating different alignments into the baseline detector Faster R-CNN. Once the detector model is obtained, tests can be performed either in the thermal or the visible domains.

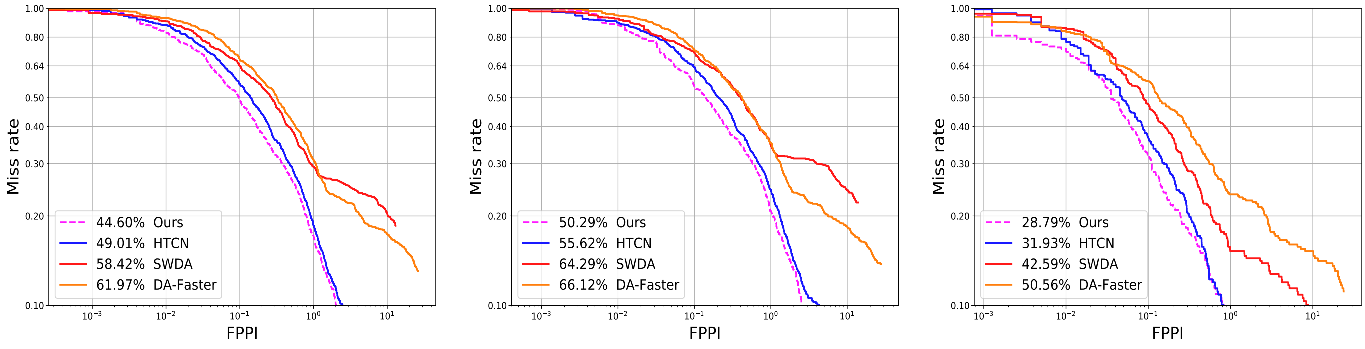
IV. Experimental results

A. Dataset and Experiments

The proposed unsupervised adaptive detector is evaluated on KAIST (Korea Advanced Institute of Science & Technology) dataset [48]. It is one of the largest multi-spectral pedestrian dataset composed of aligned visible and Long-Wave Infrared (LWIR) images under adverse illumination conditions, day and night. It roughly contains 95k frames on urban traffic environment and of dense annotations for 1182 different pedestrians. This dataset is divided into a training set of 50.2k images from Set 00 to Set 05, and a test set of 45.1k images from Set 06 to Set 11. In our work, both thermal and visible images of this dataset are used, but only labels of visible data are employed to train the proposed architecture.

For pedestrian detection, we train our proposed adaptive detector following the benchmark protocol that comes with KAIST dataset and we adopt the evaluation method presented in [4]. Precisely, we select every 3 frames from training sets and every 20 frames from testing sets, and we only consider the non-occluded, non-truncated and large instances (> 50). This results in a training set of 7601 images for both thermal and visible sets, and a testing set of 2252 thermal images (1455 day and 797 night).

The performance of pedestrian detection is evaluated in terms of miss rate as a function of False Positives Per Image (FPPI) and log-average miss rate over the range of $[10^{-2}, 10^{0}]$. Intersection Over Union (IOU) equal to 0.5 compared to the ground truth is used. The obtained results by our proposed adaptive detector are to the baseline detector trained on visible



(a) Reasonable all (b) Reasonable daytime (c) Reasonable nighttime

Fig. 2. Comparisons using miss rate vs. FPPI curves of our proposed adaptive detector to other unsupervised domain adaptation methods on KAIST pedestrian dataset under reasonable setting. Log-average miss rates are given in the legends (lower is better).

Algorithm 1: Proposed adaptive detector

```
Input :
$\mathscr{V} = \{x_i^v, y_i^v\}_{1 \leq i \leq N}$: a set of labeled visible images.
$\mathscr{T} = \{x_i^t\}_{1 \leq i \leq N}$: a set of unlabeled thermal images.
Output :
Training model of the adaptive detector.
Initialize:
Initialize $(\theta_d, \theta_{loc}, \theta_{glob}, \theta_{ins}, \theta_{img})$ parameters of the
  network and learning rate $\mu$
while not converge do
    for $(i \leftarrow 1$ to $N)$ do
        1) Adaptation losses at backbone network
            • Compute local adaptation loss $\mathcal{L}^{loc}$
            • Compute global adaptation loss $\mathcal{L}^{glob}$
        2) Consistency regularization loss
            • Compute image-level loss $\mathcal{L}^{img}$ to train $D_3$
            • Set $p_{m,n}$ to the output of $D_3$ for given activation
              of feature maps
            • Compute instance-level loss $\mathcal{L}^{ins}$ to train $D_4$
            • Set $s_j$ to the output of $D_4$ of j-$th$ region proposal
            • Compute consistency regularization $\mathcal{L}^{cst}(p_{m,n}, s_j)$
        3) Combine in adversarial loss:
            $\mathcal{L}^{adv} = \mathcal{L}^{glob} + \mathcal{L}^{loc} + \mathcal{L}^{img} + \mathcal{L}^{ins} + \mathcal{L}^{cst}$
        4) Compute detection loss $\mathcal{L}^{det}$
        5) Backpropagation step
            $\theta_d \leftarrow \theta_d - \mu \left(\frac{\partial \mathcal{L}^{det}}{\partial \theta_d} - \lambda \frac{\mathcal{L}^{adv}}{\theta_d}\right)$
            $\theta_{loc} \leftarrow \theta_{loc} - \mu \frac{\partial \mathcal{L}^{loc}}{\partial \theta_{loc}}$
            $\theta_{glob} \leftarrow \theta_{glob} - \mu \frac{\partial \mathcal{L}^{glob}}{\partial \theta_{glob}}$
            $\theta_{ins} \leftarrow \theta_{ins} - \mu \frac{\partial \mathcal{L}^{ins}}{\partial \theta_{ins}}$
            $\theta_{img} \leftarrow \theta_{img} - \mu \frac{\partial \mathcal{L}^{img}}{\partial \theta_{img}}$
```

data and tested on thermal images. Also, comparisons to other existing unsupervised domain adaptation methods for detection are considered, namely, Domain Adaptive Faster R-CNN (DA-Faster) [14], Strong-Weak Distribution Alignment (SWDA) [43], and Hierarchical Transferability Calibration Network (HTCN) [13].

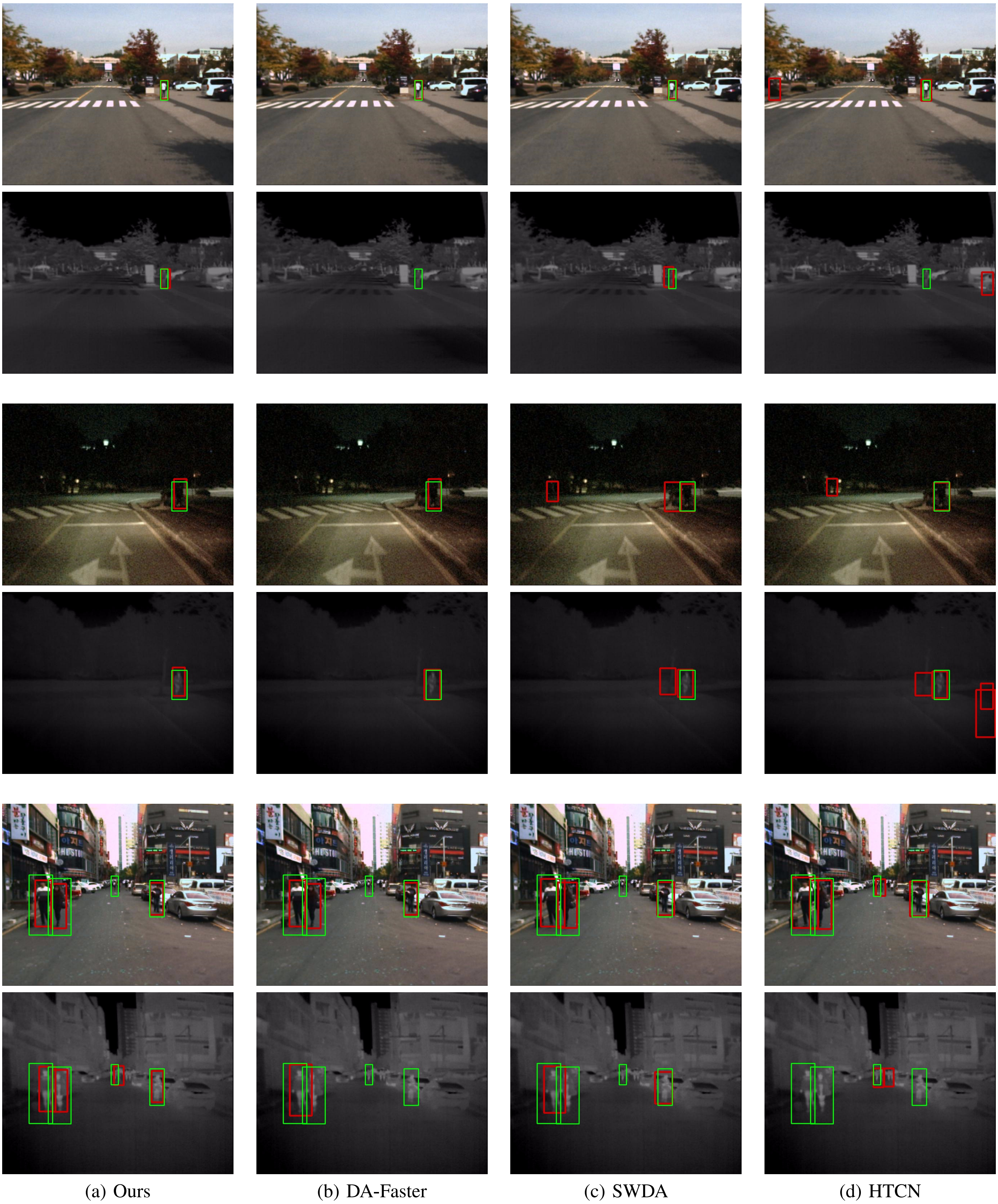
B. Implementation Details

Following Faster-RCNN configuration [16], we used the ResNet-101 [49] architecture as backbone network. The parameters of ResNet-101 are fine-tuned from the pre-trained model on ImageNet and the shorter side of every image is set to 600. For optimization, we use the stochastic gradient descent (SGD) optimizer in the training step, with an initial learning rate set to 0.001 which is brought down to 0.0001 after 50K iterations. We use a mini-batch size of one visible image and one thermal image. Also, our proposed adaptive detector is trained on 6 epochs. γ defined in equations 4 and 5 to weight Hard-to-classify examples is set to 3.0 and the weight λ of adversarial loss defined in eq. 16 is set to 0.1. For all experiments, PyTorch framework is used and the model is learned on NVIDIA Titan RTX GPU with 24 GB RAM.

C. Results and Analysis

At a first stage, the performance of the baseline Faster R-CNN detector trained on visible data and tested on thermal images from KAIST dataset is evaluated. In terms of log-average miss rate 87.1% is obtained compared to 48.59% when tests are performed on visible data. This performance drop is expected because of the distribution mismatch between training (visible) and testing (thermal) data. This result complies with our observation stated at the beginning of the paper, that object detection typically assumes that training and testing data are drawn from similar distribution. That was our motivation to rather refer to adaptation by aligning feature distributions in order to perform well in both domains at no additional annotation cost, which is our main proposal in this paper.

Fig. 2 presents the miss rate vs. FPPI curves and their corresponding log-average miss rates (reported in the figure legend) of our proposed adaptive detector. Not surprisingly, as shown



(a) Ours

(b) DA-Faster

(c) SWDA

(d) HTCN

Fig. 3. Qualitative detection results of our proposed detector compared to other existing adaptive detectors. From left to right: results of (a) our method (b) DA-Faster [14], (c) SWDA [43], and (d) HTCN [13]. From top to bottom: results of 3 sample images in both visible and thermal domains, each sample is shown in two rows (the first for the results in the visible domain and the second for the thermal domain). For each sample image, the detection results are shown in red color and the corresponding annotated bounding boxes in green color.

the figure the detection performance is significantly improved by adaptation compared to the baseline method (original Faster R-CNN trained on visible data and tested on thermal data without adaptation) by 42.5% relative reduction of the error. This significant margin of improvement regarding the baseline detector proves the relevance of applying domain adaptation for detection to be tested on another domain (thermal). Also, by comparing day and night results, better results are obtained in nighttime (only 28.79% as miss rate) since thermal data is proven to be more effective at that time.

In the same figure, our results are compared to other existing adaptation methods, precisely we only consider unsupervised methods (DA-Faster, SWDA, and HTCN) for comparisons. It can be clearly observed that our proposed adaptive detector outperforms the existing unsupervised adaptation methods by a significant margin, on daytime, nighttime and all images. It achieves the best result by 17.37%, 13.82%, and 4.41% relative reduction of the error compared to DA-Faster, SWDA and HTCN, respectively. These obtained results comply with our expectations, since the proposed detector has the advantage of performing alignments at different levels in the two phases of the network. This results in a more complete architecture compared to the other adaptive detectors, where only alignments in the backbone network are performed in [43], [13] and alignments at RPN and ROI stages are performed in [14]. Combining different alignments in our proposed adaptive detector leads to an overall performance.

The corresponding qualitative results on some sample images from KAIST dataset are shown in Fig. 3. These results also indicate the performance increase by our adaptive detector compared to the others. Precisely, in the sample visual results, it is shown that some false positives and false negatives are corrected by our detector compared to [14], [43], [13]. It is also important to mention that our proposed adaptive detector performs well in the visible domain as well. It achieves 40.01% in terms of miss rate, which is a good result compared to the thermal domain (44.6%), and more importantly, it is better than the original result of Faster R-CNN trained and tested on visible data without adaptation (48.59%).

To better highlight the importance and the relevance of each alignment considered in our adaptation method, we evaluate the results by removing one of them each time. As depicted in Table I and following the same notations in [43], [13], G, L, CTX refer to global alignment, local alignment, and context-vector based regularization, respectively. Compared to [43], [13], since we also consider other alignments in the second phase of the network (RPN and ROI), we add R to refer to them. As demonstrated from these results, even though some alignments affect the results more than others (such as the case of the global alignment G), but combining all of them together achieves an overall performance. Mainly by considering other alignments at the second part of the detector, the results are further improved. These results justify our choice of combining different alignments at different levels and phases in order to respond to different aspects of the domain shift.

TABLE I
RESULTS OF OUR PROPOSED ADAPTIVE DETECTOR IN TERMS OF LOG-AVERAGE MISS RATE BY ELIMINATING EACH TIME ONE OF THE ALIGNMENTS. G, L, CTX, AND R REFER TO GLOBAL ALIGNMENT, LOCAL ALIGNMENT, CONTEXT-VECTOR AND ALIGNMENTS IN THE SECOND SUB-NETWORK.

Method	G	L	CTX	R	MR (%)
Ours	✗	✓	✓	✓	66.00
	✓	✓	✗	✓	55.76
	✓	✗	✓	✓	51.00
	✓	✓	✓	✗	49.01
	✓	✓	✓	✓	44.46

V. CONCLUSION

In this paper, we proposed a novel thermal-to-visible adaptive detector leveraging information from both domains in unpaired setting and at no additional annotation cost. This detector incorporating feature distribution alignments into Faster R-CNN has the advantage of combining different domain classifiers in order to achieve an overall performance. Despite its relevance for practical applications since only one model that bridges the gap between the two domains without additional annotations, such unsupervised adaptation for detection in thermal and visible domains is not yet investigated. By means of tests on KAIST dataset, the effectiveness of the proposed detector is proven by obtaining better results compared to the baseline method with an outstanding margin. Its performance also exceeds some recently published works in the field of unsupervised domain adaptation for detection in other domains.

There are several possible extensions of this work. For example, the proposed features distribution alignment which is incorporated into Faster R-CNN architecture for illustrative purpose, can readily be replaced by other deep detectors. Also, since this work specifically addresses the problem of pedestrian detection, natural directions for future research include investigating the detection of other objects.

ACKNOWLEDGMENTS

This work has been supported by the DGVR research fund from the Tunisian Ministry of Higher Education and Scientific Research that is gratefully acknowledged.

REFERENCES

[1] M. Kieu, A. D. Bagdanov, M. Bertini, and A. D. Bimbo, "Task-conditioned domain adaptation for pedestrian detection in thermal imagery," in *Computer Vision - ECCV*, 2020.

[2] S. Park, J. Hwang, J.-E. Park, Y.-C. Ahn, and H. W. Kang, "Application of ultrasound thermal imaging for monitoring laser ablation in ex vivo cardiac tissue," *Lasers in surgery and medicine*, vol. 52, no. 3, pp. 218–227, 2020.

[3] G. Lu, Y. Yan, L. Ren, P. Saponaro, N. Sebe, and C. Kambhamettu, "Where am i in the dark: Exploring active transfer learning on the use of indoor localization based on thermal imaging," *Neurocomputing*, vol. 173, pp. 83–92, 2016.

[4] D. Ghose, S. M. Desai, S. Bhattacharya, D. Chakraborty, M. Fiterau, and T. Rahman, "Pedestrian detection in thermal images using saliency maps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[5] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, "Learning cross-modal deep representations for robust pedestrian detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5363–5371.

[6] F. Lai, J. Kandukuri, B. Yuan, Z. Zhang, and M. Jin, "Thermal image enhancement through the deconvolution methods for low-cost infrared cameras," *Quantitative infrared thermography journal*, vol. 15, no. 2, pp. 223–239, 2018.

[7] Y. W. K. Zoetgnande, J.-L. Dillenseger, and J. Alirezaie, "Edge focused super-resolution of thermal images," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[8] M. Mohamed Amine, F. Hajer, S. Anis, and E. B. A. Najoua, "Thermal image enhancement using generative adversarial network for pedestrian detection," *International Conference on Pattern Recognition*, 2020.

[9] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for robust pedestrian detection," *IEEE transactions on image processing*, vol. 29, pp. 3820–3834, 2020.

[10] A. Mhalla, T. Chateau, S. Gazzah, and N. Essoukri Ben Amara, "An embedded computer-vision system for multi-object detection in traffic surveillance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 11, pp. 4006–4018, 2018.

[11] W. Ouyang, X. Zeng, and X. Wang, "Learning mutual visibility relationship for pedestrian detection with a deep model," *International Journal of Computer Vision*, vol. 120, no. 1, pp. 14–27, 2016.

[12] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5187–5196.

[13] C. Chen, Z. Zheng, X. Ding, Y. Huang, and Q. Dou, "Harmonizing transferability and discriminability for adapting object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8869–8878.

[14] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. 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*, 2018, pp. 3339–3348.

[15] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[19] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[20] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 781–10 790.

[21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[22] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast r-cnn for pedestrian detection," *IEEE transactions on Multimedia*, vol. 20, no. 4, pp. 985–996, 2017.

[23] C. Li, D. Song, R. Tong, and M. Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," in *British Machine Vision Conference, BMVC*, 2018.

[24] M. Kieu, A. D. Bagdanov, M. Bertini, and A. Del Bimbo, "Domain adaptation for privacy-preserving pedestrian detection in thermal imagery," in *International Conference on Image Analysis and Processing*. Springer, 2019, pp. 203–213.

[25] J. Baek, S. Hong, J. Kim, and E. Kim, "Efficient pedestrian detection at nighttime using a thermal camera," *Sensors*, vol. 17, no. 8, p. 1850, 2017.

[26] C. Devaguptapu, N. Akolekar, M M Sharma, and V. N Balasubramanian, "Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[27] T. Guo, C. P. Huynh, and M. Solh, "Domain-adaptive pedestrian detection in thermal images," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1660–1664.

[28] V. John, S. Mita, Z. Liu, and B. Qi, "Pedestrian detection in thermal images using adaptive fuzzy c-means clustering and convolutional neural networks," in *2015 14th IAPR International Conference on Machine Vision Applications (MVA)*. IEEE, 2015, pp. 246–249.

[29] D. Guan, Y. Cao, J. Yang, Y. Cao, and M. Y. Yang, "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection," *Information Fusion*, vol. 50, pp. 148–157, 2019.

[30] D. Konig, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, "Fully convolutional region proposal networks for multispectral person detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 49–56.

[31] S. W. Jingjing Liu, Shaoting Zhang and D. Metaxas, "Multispectral deep neural networks for pedestrian detection," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.

[32] J. Wagner, V. Fischer, M. Herman, and S. Behnke, "Multispectral pedestrian detection using deep fusion convolutional neural networks." in *ESANN*, 2016.

[33] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5127–5137.

[34] P. Panareda Busto and J. Gall, "Open set domain adaptation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 754–763.

[35] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5715–5725.

[36] W. Li, Z. Xu, D. Xu, D. Dai, and L. Van Gool, "Domain generalization and adaptation using low rank exemplar svms," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1114–1127, 2017.

[37] Y. Zou, Z. Yu, B. Vijaya Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 289–305.

[38] S. Sankaranarayanan, Y. Balaji, A. Jain, S. Nam Lim, and R. Chellappa, "Learning from synthetic data: Addressing domain shift for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3752–3761.

[39] A. Berg, J. Ahlberg, and M. Felsberg, "Generating visible spectrum images from thermal infrared," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1143–1152.

[40] X. Kuang, J. Zhu, X. Sui, Y. Liu, C. Liu, Q. Chen, and G. Gu, "Thermal infrared colorization via conditional generative adversarial network," *Infrared Physics & Technology*, p. 103338, 2020.

[41] F. Munir, S. Azam, M. A. Rafique, A. M. Sheri, and M. Jeon, "Thermal object detection using domain adaptation through style consistency," *arXiv preprint arXiv:2006.00821*, 2020.

[42] M. Kim, S. Joung, K. Park, S. Kim, and K. Sohn, "Unpaired cross-spectral pedestrian detection via adversarial feature learning," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1650–1654.

[43] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6956–6965.

[44] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[45] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[46] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[47] H. Zhang, Z. Zhang, A. Odena, and H. Lee, "Consistency regularization for generative adversarial networks," *arXiv preprint arXiv:1910.12027*, 2019.

[48] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, “Multispectral pedestrian detection: Benchmark dataset and baseline,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1037–1045.

[49] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.