Feature distribution alignments for object detection in the thermal domain

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

Noname manuscript No.
(will be inserted by the editor)

Feature distribution alignments for object detection in the thermal domain

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

Received: date / Accepted: date

Abstract Infrared imaging has recently played an important role in a wide range of applications including video surveillance, robotics and night vision. However, the manufacturing cost of high-resolution infrared cameras is more expensive regarding similar quality in visible cameras. This could explain the fact that thermal databases are less available compared to visible ones. In this paper, we mainly emphasis the need for aligning features from visible and thermal domains for object detection in order to ensure effective results in both domains without the need to retrain data and to perform additional annotations. To address that, we incorporate feature distribution alignments into Faster R-CNN architecture at different levels. The resulting proposed adaptive detector has the advantage of covering different aspects of the domain shift in order to improve the overall performance. Using KAIST and FLIR ADAS datasets, the effectiveness of the proposed detector is assessed and better results are obtained compared to the baseline detector and to the obtained results by other existing works.Our code is available at `https://github.com/AmineMarnissi/UDAT`.

Mohamed Amine Marnissi
Université de Sfax, Ecole Nationale d'Ingénieurs de Sfax, Sfax 3038, Tunisie;
Université de Sousse, Ecole Nationale d'Ingénieurs de Sousse, LATIS-Laboratory of Advanced Technology and Intelligent Systems, Sousse 4023, Tunisie;
E-mail: marnissi.mohamedamine@eniso.u-sousse.tn

Hajer Fradi
Université de Sousse, Institut Supérieur des Sciences Appliquées, LATIS-Laboratory of Advanced Technology and Intelligent Systems, Sousse 4023, Tunisie;
E-mail: hajer.fradi@issatso.rnu.tn

Anis Sahbani
Sorbonne Université, CNRS, Ins titule for intelligent Systems and Robotics (ISIR), Paris, France
Enova Robotics S.A.
E-mail: anis.sahbani@enovarobotics.com

Najoua Essoukri Ben Amara
Université de Sousse, Ecole Nationale d'Ingénieurs de Sousse, LATIS-Laboratory of Advanced Technology and Intelligent Systems, Sousse 4023, Tunisie;
E-mail: najoua.benamara@eniso.rnu.tn

Keywords Unsupervised domain adaptation · adversarial loss · object detection · thermal and visible · domain classifier

1 Introduction

An infrared camera is a device that forms an image using infrared radiations, compared to the commonly used visible cameras that form an image using visible light [26,6,49]. This camera forms an image by detecting infrared radiations emitted by an object based on its temperature. The main advantage using infrared cameras is that they could easily distinguish warm objects from other surrounding objects even in bad lighting and adverse weather conditions [15].

For the aforementioned reasons, 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 [9,6,15,32,10,42,28,22]. Despite the usefulness of these cameras mainly at nighttime, there are some limitations that have to be considered, essentially about the compromise between the cost and the image quality. It is important to mention that the manufacturing cost of high-resolution infrared cameras is more expensive regarding similar quality in visible cameras. This could explain the fact that thermal data is less available compared to visible one. Also, for the few available datasets, the visual quality of thermal images is usually poor since low-resolution thermal cameras are more commonly used [49,26,10]. Low-resolution together with bad acquisition conditions in some cases present multiple challenges that impede infrared imaging applications to perform well. It is essentially the case of various video analysis applications such as object detection, object tracking and activity recognition [10,22,21].

Precisely, in this paper, we focus on the problem of object detection and localization from infrared cameras for surveillance applications. This problem has been extensively studied using visible data and good results are usually obtained [20,23]. However, in some situations for instance in nighttime, bad lighting conditions, total darkness, or in adverse weather conditions, the performance of the state-of-the-art detectors dramatically drops [41,7]. Here comes the importance of using thermal cameras for detection since they could better discern warmer target objects than other surrounding ones.

Even though object detection using infrared cameras is more convenient in some situations, it is still subject to errors if we consider the inherent problems of low-resolution and insufficient available data. This topic has been widely studied in the recent years using deep neural networks [20,11,25,40,29,23]. 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 data is available compared to the visible domain; second, the annotation task for object detection is particularly time-consuming process since each object category in every image must be precisely delimited with a bounding box.

To mitigate these problems, we intend in this present work to harness the abundance of annotated visible images by adapting them to the thermal domain at no additional annotation cost. Basically inspired from [4] and [3], we propose to cover the domain shift by means of multiple feature distribution alignments.

The proposed alignment process, where different cues are merged together, is integrated into a two-stage-detector in order to improve its domain adaptability. Precisely, feature alignments are performed at image and instance levels in order to reduce the shift between thermal and visible domains caused by variations in scale, illumination, object appearance and size, etc. To reinforce alignments at these two levels, a consistency regularization is added. The proposed adaptation scheme also includes local and global alignments of low-level and high-level features extracted from the backbone network in order to strictly align the image style across domains.

The resulting proposed adaptive detector aligning feature distributions at multiple levels is considered as the main contribution of this present paper. It has also the advantage of being trained in an unpaired setting with unlabeled thermal images. This unsupervised adaptation of object detectors from source to target domains has been commonly employed in the visible domain to essentially deal with adverse weather conditions [41,3,4,36]. In this paper, it is proposed for the first time in thermal and visible domains, as far as we know. Targeting such dissimilar domains is challenging, since they exhibit different visual characteristics [39]. Unlike existing multispectral detectors from both visible and thermal domains, our proposed adaptation scheme is of significant interest since it enables detections in the thermal domain even though the corresponding annotations are not used in the training step. The idea consists of reducing the domain shift to adapt the same detector from the source to the target domain without the need to retrain data, and more importantly, at no additional annotation cost in the target domain. The effectiveness of the proposed adaptation method is demonstrated on the detection performance by obtaining better results with a significant margin compared to the baseline detector and to recently published works in the field of domain adaptation using two popular datasets.

The remainder of the paper is organized as follows: in Section 2, an overview of the existing transfer learning and domain adaptation methods for object detection is presented. Then, our proposed approach of unsupervised domain adaptation for object detection in thermal and visible domains is detailed in Section 3. The conducted experiments and the obtained results are discussed in Section 4. Finally, we briefly conclude and give an outlook of possible future works in Section 5.

2 Related work

In this section, we give an overview of some existing methods that perform domain adaptation in thermal and visible domains. Precisely, we focus on the studies that perform this adaptation for object detection by means of generative and non generative models. This overview includes as well a general introduction of transfer learning and domain adaptation for computer vision applications.

2.1 Transfer learning and domain adaptation

It is commonly known that deep learning networks mostly require large-scale datasets because of the huge number of parameters that have to be trained [1].

Since collecting and annotating datasets for every new task or domain is time-consuming process, transfer learning can be used as an alternative solution to handle that. Globally, transfer learning can be divided into three main categories: unsupervised, inductive and transductive methods [30,39]. Domain Adaptation (DA) is a special case of transfer learning that makes use of annotated data in a source domain to perform another or the same task in a target domain. According to [39], DA belongs to transductive transfer learning solutions with the assumption that tasks are similar and the difference is only due to the domain divergence. Depending on the domain divergence, domain adaptation can be itself divided into homogeneous and heterogeneous DA. Homogeneous DA refers to the case of data distribution shift, and heterogeneous DA refers to the case of feature space difference due to the use of different sensors.

To cover the domain divergence, discrepancy-based methods can be used if fine-tuning deep network models is effective enough [19,45]. Reconstruction-based methods in which data reconstruction of the source or target samples acts as an auxiliary task to feature invariance can be instead employed [44,13]. As another alternative solution, adversarial-based methods can be applied using domain discriminators to boost domain confusion through an adversarial objective [43,33]. These adversarial-based methods could be performed by means of generative or non-generative models.

2.2 Domain adaptation for object detection in thermal and visible domains

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 [31,27,18,4,50,37]. 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 fall into two categories: homogeneous and heterogeneous DA. As already mentioned, heterogeneous DA is the case of feature space divergence using different sensors (e.g. thermal vs. visible or RGB vs. depth). In this section, we essentially present the related work of DA for detection in thermal and visible domains. We also review some existing methods of homogeneous DA for detection, mainly those requiring only annotations in the source domain (refer to unsupervised DA), which is our primary goal in this work. More details about these methods are given in the following sections by categorizing them into generative and non-generative models.

2.2.1 Generative models

Generative models basically consist of generating synthetic data that is similar to the target data and shares the annotations of the source domain. One typical example is the generation of a visible image from an input thermal image. This transformation aims at enhancing thermal image quality by converting it to perceptually realistic RGB image in order to enable better content interpretation, usually difficult for untrained operators. Commonly known as colorization, this transformation has been the subject of many studies in the literature [17,2,7]. For

instance, in [17] TIC-CGAN which refers to conditional generative adversarial network is proposed to address thermal infrared colorization problem. Compared to [2] that only restored rough luminance and chrominance information, TIC-CGAN uses a coarse-to-fine generator and a composite objective function that combines content, adversarial, perceptual and total variation losses to produce results with realistic colors and fine details. However, it is important to remind that in this study we are merely interested in performing this transformation to enable better detection. It is the case of [7], where a Cycle-GAN for unpaired image-to-image translation of thermal to pseudo-RGB data is proposed to adapt a multimodal Faster-RCNN detector.

The transformation mapping from the source thermal domain to the target visible domain, as proposed in [2,7], could potentially fail in our case since it aims at generating a realistic RGB image, without any focus on particular objects in the image, which is our primary goal in this study. This motivates us to rather refer to other existing works that incorporate this transformation in the detection architecture in order to enable better detection beyond generation. More details are given in the next section dedicated to present non-generative models.

2.2.2 Non-Generative Models

Instead of generating synthetic target data, non-generative models aim at learning a domain-invariant representation, where the distribution of both domains can be similar enough such that the classifier can be directly employed in the target domain, while being trained on samples from the source domain. Non-generative models are basically inspired from GAN architectures, but trained through a domain-confusion loss without generators [8,38]. In the specific case of thermal and visible domains, only very few works that tackled the task of DA for detection by means of non-generative models have been proposed. For instance, in [16], an unified detection network by defining a common feature space, which makes intermediate features from the two domains is proposed. To be trained, the proposed unified detector requires supervision in both domains.

Unlike existing adaptation works for detection that require annotated data for both thermal and visible domains, in this paper we intend to perform adaptation without the need to annotate thermal data. To the best of our knowledge, such unsupervised domain adaptation has not been proposed before in heterogeneous domains such as the case of thermal and visible. This category of domain adaptation has been mainly studied in homogeneous domains, where the difference is only in terms of data distributions as proposed in [4,36,3].

Domain Adaptive Faster R-CNN (DA-Faster) [4] 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 of Faster R-CNN model. The robustness of the proposed adaptive detector is evaluated on different domain shift scenarios. Scale-Aware Domain Adaptive Faster R-CNN [5] is an extension of DA-Faster which is mainly proposed to improve the detection results by dealing with the variation of object scales in cross domain-adaptation. To handle that, the object scale is considered in the alignment

process by adopting the feature pyramid network (FPN) that produces feature maps at different scales.

In [36], another adaptive object detector based on strong local-alignment and weak-global alignment is proposed. The weak alignment model focuses on the adversarial alignment on images that are globally similar and puts less emphasis on aligning images that are globally dissimilar. The strong domain alignment model is designed to only consider local receptive fields of the feature maps. Another approach called Hierarchical Transferability Calibration Network (HTCN) [3] to harmonize transferability and discriminability in the context of adversarial adaptation 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 intend to make use of such techniques of adaptation without additional supervision in the target domain. But differently from the previous works, where homogeneous domains are aligned, our proposed architecture aims at aligning images from heterogeneous domains (thermal and visible) exhibiting different features. In addition, compared to [4, 36, 3] we propose a more complete architecture, where feature distributions are aligned at different levels and at two phases of the network.

3 Proposed Approach

In this paper, we propose a novel approach of unsupervised domain adaptation for object detection dedicated to operate in the target domain at no additional annotation cost. In our specific case, it is about heterogeneous adaptation setting, where visible is the source domain, since it is more commonly used, and thermal is the target domain. The proposed adaptation scheme falls into the category of adversarial-based DA methods by means of non-generative models. To cover different aspects of the domain divergence, we propose to perform feature alignments at multiple levels. The alignment process is incorporated at two phases of a baseline detector. The resulting overall proposed architecture is shown in Figure 1, in which the notations are the same used in the remainder of this section, where we describe each of the architecture components. Precisely, the problem of domain adaptation is formulated in section 3.1. The detection loss function is then presented in section 3.2. Details about how the problem of adaptation is solved is provided in Section 3.3, along with the corresponding loss functions.

3.1 Problem Formulation

Following the common terminology in the field of domain adaptation, a domain $\mathcal{D}$ is characterized by a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, where $X = \{x_1, ..., x_n\} \in \mathcal{X}$. As defined in [39], a task consists of a feature space $\mathcal{Y}$ and a conditional probability distribution $P(Y|X)$. In our case, it is about detection task in heterogeneous DA setting, in which the feature spaces between the source (visible) and target (thermal) domains are different ($\mathcal{X}^s \neq \mathcal{X}^t$). The dimension of feature spaces could be different as well. We assume that the source

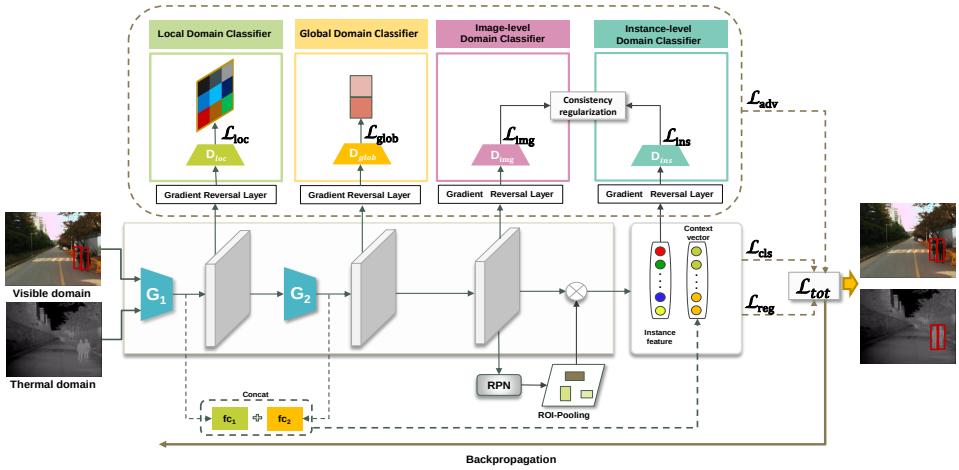

Fig. 1 The proposed architecture for object detector adaptation in visible and thermal domains. This adaptation is performed by means of multiple feature alignments and is integrated in two phases of Faster R-CNN detector. The notations used in this figure are the same used in section 3.

domain $\mathcal{V} = \{X^v, P(X^v)\}$ is labeled, and by means of unsupervised domain adaptation the detection model is adapted to the target domain $\mathcal{T} = \{X^t, P(X^t)\}$. For the task of object detection, $P(Y|X)$ can be formulated as $P(C, B \mid I)$, where $C \in \{1, ...K\}$ is the object class with K is the total number of classes, B refers to the bounding box of an object, and I is the image representation.

Following the notations of [4] and by fitting them to our case, a domain shift between visible and thermal domains can be expressed as: $P_{\mathcal{V}}(C, B, I) \neq P_{\mathcal{T}}(C, B, I)$, where $P(C, B, I)$ refers to the joint probability distribution of training samples for object detection. During the training process, we analyze the problem of domain shift by calculating $P_{\mathcal{T}}(C, B, I)$, even though the corresponding annotations of bounding-box B and class C are unknown for training samples in the thermal domain.

3.2 Detection loss

For detection, we choose Faster R-CNN [34] as a representative two-stage detector since it is one of the most popular object detectors that achieves a good compromise between the speed and the accuracy. In a first stage, Region Proposal Network (RPN) generates region proposals based on the feature map produced by the backbone network. Then, Faster R-CNN feeds the region proposals and feature map into ROI pooling layer in a second stage. As backbone network, we use ResNet-101 which is composed of two feature extractors G_1 and G_2 (see Figure 1). The first one is employed for the extraction of low-level features and the second one for high-level features.

Since in our case thermal images are not annotated, the detection loss is computed in the visible domain. For a given set of annotated source visible images $\{X^v, Y^v\}$, where each image is denoted as x_i^v and y_i^v is its corresponding bounding

boxes, the detection loss $\mathcal{L}_{det}$ that combines the classification and the regression losses is defined as:

$$\mathcal{L}_{det}(X^v, Y^v) = \frac{1}{N_v} \sum_{i=1}^{N_v} \mathcal{L}_{reg}(R(G_2(x_i^v)), y_i^v) + \mathcal{L}_{cls}(R(G_2(x_i^v)), y_i^v) \tag{1}$$

where the output of G_2 is fed to RPN module (denoted as R) and N_v indicates the total number of samples from the source visible domain.

3.3 Domain Adaptation Components

In this section, the main domain adaptation components of our proposed adaptive detector are presented. This adaptation consists of aligning feature distributions at two phases of the network and at different levels. For each 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. Precisely, in the sub-network composed of RPN and ROI, the alignments are performed at image and instance levels, with a consistency regularization. Details about alignments at this phase are given in sections 3.3.1, 3.3.2 and 3.3.3. In addition, global and local alignments are performed at the backbone network (ResNet-101) on the features generated by G_1 and G_2, as explained in section 3.2. Details about alignments at this phase are presented in sections 3.3.4, 3.3.5 and 3.3.6.

3.3.1 Image-Level Alignment

As already formulated in section 3.1, the domain shift problem can be expressed by the joint probability distribution $P(C, B, I)$, which can be decomposed according to Bayes' theorem as:

$$P(C, B, I) = P(C, B \mid I)P(I) \tag{2}$$

As any classification problem, we adopt the covariate shift hypothesis for object detection [4], i.e. the conditional probability $P(C, B|I)$ has to be the same for both domains, and the domain shift is only due to the marginal distribution difference $P(I)$.

Basically, 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 in fact the resulting feature map of backbone network. Consequently, to solve the domain shift problem, the distributions of the image representation from the two domains (i.e. $P^{\mathcal{V}}(I) = P^{\mathcal{T}}(I)$) have to be the same.

Practically, since it is not trivial to reach such alignment at image level, a domain classifier D_{img} is employed to minimize the domain divergence. D_{img} is trained at each activation of feature maps, then, it predicts the domain label

for every image patch. The advantage of this image-level alignment is that it can generally reduce the amount of shift caused by the domain differences in the global image such as style, scale and illumination [14,4]. Using the cross entropy, the adaptation loss at image level denoted as $\mathcal{L}_{img}^{\mathcal{V}}$ and $\mathcal{L}_{img}^{\mathcal{T}}$ in visible and thermal domains is accordingly defined as follows:

$$\mathcal{L}_{img}^{\mathcal{V}} = -\frac{1}{N_v}\sum_{i}^{N_v}\sum_{u,v}\log(1-p_i(u,v)) \tag{3}$$

$$\mathcal{L}_{img}^{\mathcal{T}} = -\frac{1}{N_t}\sum_{i}^{N_t}\sum_{u,v}\log(p_i(u,v)) \tag{4}$$

where $p_i(u,v)$ is the output of the domain classifier D_{img} for given activations of feature maps located at (u,v) position after applying backbone network on the i^{th} input image. N_v and N_t indicate the total number of visible and thermal samples, respectively. The two defined adaptation loss functions are combined in $\mathcal{L}_{img}$ by:

$$\mathcal{L}_{img} = \frac{1}{2}(\mathcal{L}_{img}^{\mathcal{V}} + \mathcal{L}_{img}^{\mathcal{T}}) \tag{5}$$

In addition, we employ a Gradient Reverse Layer (GRL) [8] 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 goal of aligning the domain distributions by applying adversarial learning.

3.3.2 Instance-Level Alignment

Instance-level alignment is also considered in order to reduce the difference of local instance representations between the two domains, such as the appearance and the object size. The joint probability distribution can be alternatively decomposed as:

$$P(C,B,I) = P(C \mid B,I)P(B,I) \tag{6}$$

Following again the hypothesis of covariate shift, the conditional probability $P(C \mid B,I)$ is the same in the two domains and the shift is due to the difference in the marginal distribution $P(B,I)$ [4]. 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. Consequently, the distribution representation of instances is the same in both domains i.e. $P_{\mathcal{V}}(B,I) = P_{\mathcal{T}}(B,I)$.

In our proposed adaptive detector, the instance-level representation (B,I) is employed 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 only known in the visible domain, a domain classifier D_{ins} is trained on the feature vectors in order to perform alignment at instance level. The adaptation loss at instance level denoted as $\mathcal{L}_{ins}^{\mathcal{V}}$ and $\mathcal{L}_{ins}^{\mathcal{T}}$ in visible and thermal domains is defined as follows:

$$\mathcal{L}_{ins}^{\mathcal{V}} = -\frac{1}{N_v}\sum_{i}^{N_v}\sum_{j}\log(1-s_i(j)) \tag{7}$$

$$\mathcal{L}_{ins}^{\mathcal{T}} = -\frac{1}{N_t}\sum_{i}^{N_t}\sum_{j}\log(s_i(j)) \tag{8}$$

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

$$\mathcal{L}_{ins} = \frac{1}{2}(\mathcal{L}_{ins}^{\mathcal{V}} + \mathcal{L}_{ins}^{\mathcal{T}}) \tag{9}$$

Similar to the domain classifier at image level, the gradient reverse layer is added beforehand in order to fit the adversarial training strategy.

3.3.3 Consistency Regularization

To align domains at image and instance levels, the distributions of image $P(I)$ and instance representations $P(B, I)$ have to be the same in both domains. Since it is not trivial to reach such alignments, two domain classifier are trained. For each classifier, the input could be either the image representation I or the instance representation (B, I) and the output is a probability to predict if an input sample belongs to the thermal domain or not.

We denote by $P(D \mid I)$ and $P(D \mid B, I)$ the outputs of the image-level and the instance-level domain classifiers, respectively, with D is the domain label. Following [4], the problem can be formulated according the Bayes' theorem as:

$$P(D \mid B, I)P(B \mid I) = P(B \mid D, I)P(D \mid I) \tag{10}$$

where $P(B \mid D, I)$ and $P(B \mid I)$ are a domain-dependent and a domain-invariant bounding box predictors, respectively. Since we are only able to learn $P_{\mathcal{V}}(B \mid D, I)$ in the visible domain, $P(B|D, I)$ can be calculated by enforcing the consistency (eq. 10) between the two domain classifiers, i.e., $P(D \mid B, I) = P(D \mid I)$. Once $P(B \mid D, I)$ is calculated, $P(B \mid I)$ can be estimated in the thermal domain as well.

The consistency between the domain classifiers at two levels (image and instance) allows to learn the cross-domain robustness of the bounding box predictor [47]. It can be defined for visible and thermal domains as:

$$\mathcal{L}_{cst} = \sum_{i}\sum_{j} \parallel \frac{1}{\mid I \mid}\sum_{u,v} p_i(u, v) - s_i(j) \parallel_2 \tag{11}$$

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

3.3.4 Local Feature Alignment

Given two images from the two domains, since some local regions can 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_{loc}, which is defined to highlight local features

by producing a domain prediction map, that has the same size WxH as the output of the feature extractor G_1 in the backbone network.

D_{loc} is considered as a pixel-wise discriminator based on few convolutional layers with a kernel size of 1 [3]. Following [24], 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:

$$\mathcal{L}_{loc}^{\mathcal{V}} = \frac{1}{HWN_v} \sum_{i}^{N_v} \sum_{w=1}^{W} \sum_{h=1}^{H} D_{loc}(G_1(x_i^v))_{wh}^2 \tag{12}$$

$$\mathcal{L}_{loc}^{\mathcal{T}} = \frac{1}{HWN_t} \sum_{i}^{N_t} \sum_{w=1}^{W} \sum_{h=1}^{H} (1 - D_{loc}(G_1(x_i^t))^2)_{wh} \tag{13}$$

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

$$\mathcal{L}_{loc} = \frac{1}{2}(\mathcal{L}_{loc}^{\mathcal{V}} + \mathcal{L}_{loc}^{\mathcal{T}}) \tag{14}$$

3.3.5 Global Feature Alignment

Following [36], the target samples can be divided into two parts: Easy-to-classify if they are far in the feature space from the source samples, and Hard-to-classify otherwise. Therefore, to align domains, we focus on samples that are Hard-to-classify and we put less emphasis on Easy-to-classify samples. The feature alignment at this stage is performed at global level for fully matching of the distributions between source and target images. This kind of matching is expected to perform well in the case of small domain divergence. In this context, the focal loss is used to train a domain classifier D_{glob} for global features alignment. This loss function is used instead of the cross-entropy function since the latter reinforces the domain classifier D_{glob} to consider both of Easy-to-classify and Hard-to-classify samples [36].

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

$$\mathcal{L}_{glob}^{\mathcal{V}} = -\frac{1}{N_v} \sum_{i}^{N_v} (1 - D_{glob}(G_2(x_i^v)))^{\gamma} \log(D_{glob}(G_2(x_i^v)) \tag{15}$$

$$\mathcal{L}_{glob}^{\mathcal{T}} = -\frac{1}{N_t} \sum_{i}^{N_t} (D_{glob}(G_2(x_i^t)))^{\gamma} \log(1 - D_{glob}(G_2(x_i^t)) \tag{16}$$

$$\mathcal{L}_{glob} = \frac{1}{2}(\mathcal{L}_{glob}^{\mathcal{V}} + \mathcal{L}_{glob}^{\mathcal{T}}) \tag{17}$$

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

3.3.6 Contextual Regularization

In the field of adaptive domain segmentation [37], where the goal is to simultaneously generate the domain label and a 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 in order to enhance its performance and to stabilize the training of the domain classifier using the detection loss computed on visible samples. Practically, this regularization is applied on the two extracted feature vectors fc_1 and fc_2 (outputs of G_1 and G_2, respectively). Each vector includes some contextual information describing the image content. These vectors are then concatenated with the output of ROI-pooling [3]. 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_{loc} and D_{glob}, as illustrated in Figure 1.

3.4 Summary of the proposed adaptive detector

In order to cover different aspects of the domain shift, all alignments detailed in the previous section are combined in a total adversarial loss $\mathcal{L}_{adv}$ defined as:

$$\mathcal{L}_{adv} = \mathcal{L}_{img} + \mathcal{L}_{ins} + \mathcal{L}_{cst} + \mathcal{L}_{glob} + \mathcal{L}_{loc} \tag{18}$$

Afterwards, to train our proposed adaptive detector, the resulting adversarial loss $\mathcal{L}_{adv}$ is added to the detection loss $\mathcal{L}_{det}$ defined in eq. 1 in an overall objective function formulated as:

$$\mathcal{L}_{tot} = \mathcal{L}_{det} + \alpha\mathcal{L}_{adv} \tag{19}$$

where α is used to weight the adversarial loss.

Once the total loss $\mathcal{L}_{tot}$ is calculated by mini-batch size in the training set, its gradient is backpropagated and the weights are updated accordingly following:

$$W \leftarrow W - \mu(\frac{\partial\mathcal{L}_{det}}{\partial W} + \alpha\frac{\partial\mathcal{L}_{adv}}{\partial W}) \tag{20}$$

Also, backpropagation steps are conducted for each domain classifier using the gradient of the corresponding loss functions at the same learning rate μ:

$$\begin{aligned} W_{loc} &\leftarrow W_{loc} - \mu\frac{\partial\mathcal{L}_{loc}}{\partial W_{loc}} \\ W_{glob} &\leftarrow W_{glob} - \mu\frac{\partial\mathcal{L}_{glob}}{\partial W_{glob}} \\ W_{ins} &\leftarrow W_{ins} - \mu\frac{\partial\mathcal{L}_{ins}}{\partial W_{ins}} \\ W_{img} &\leftarrow W_{img} - \mu\frac{\partial\mathcal{L}_{img}}{\partial W_{img}} \end{aligned} \tag{21}$$

4 Experimental results

4.1 Datasets

The proposed approach is evaluated within two challenging datasets widely used for multispectral object detection, namely FLIR ADAS [46] and KAIST [12] datasets. FLIR (Forward Looking InfraRed) ADAS is a recently published dataset for multi-object detection, that approximately contains 10k thermal and visible images collected during daytime and nighttime. Since in this work annotated visible data is required, we made use of an updated version of FLIR published in [46]. Following [46], only "Bicycle", "Car" and "Person" classes are considered. This version of FLIR contains 5142 well-aligned image pairs of resolution 640 $\times$ 512. Precisely 4129 pairs are used for training and 1013 pairs for testing.

KAIST (Korea Advanced Institute of Science & Technology) [12] is another widely used dataset to assess pedesterian detection algorithms. 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 approximately consists of 95k frames of resolution 640 x 480 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, thermal and visible images from this dataset are used, but only labels of visible data are employed to train the proposed architecture.

4.2 Experiments

For object detection, we train the proposed detector following the benchmark protocols and the evaluation metrics that come with the datasets. Precisely, while all images from FLIR ADAS dataset are assessed for training and testing the detector, only every 3 frames from training sets and every 20 frames from testing sets are selected for KAIST dataset. In addition, for the latter, only non occluded, non-truncated and large instances (> 50) are considered. 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, 797 night) for KAIST dataset.

For both datasets, the performance of the detector trained on visible data and tested on thermal images is evaluated in terms of mean Average Precision (mAP) at Intersection Over Union (IOU) equal to 0.5 regarding the ground truth. These results are compared to those obtained by our proposed adaptive architecture. Also, comparisons to the existing unsupervised domain adaptation methods for detection are considered, namely, Domain Adaptive Faster R-CNN (DA-Faster) [4], Scale-Aware Domain Adaptive Faster R-CNN (SA-DA-Faster) [5], Strong-Weak Distribution Alignment (SWDA) [36], and Hierarchical Transferability Calibration Network (HTCN) [3]. These comparisons include as well other existing methods based on generative models from thermal to visible domains, precisely, CycleGAN [48] and TIC-CGAN [17].

4.3 Implementation details

Following the configuration of Faster-RCNN architecture [35], 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 detector is trained on 12 epochs. γ defined in equations 15 and 16 to weight Hard-to-classify samples is set to 1.0 and the weight α of adversarial loss defined in eq. 19 is empirically set to 0.1. For all the experiments, PyTorch framework is used and we learned our model on NVIDIA Titan RTX GPU with 24 GB RAM.

4.4 Results and Comparisons with State-of-the-Arts

At a first stage, we evaluate the performance of the baseline Faster R-CNN detector trained on visible data and tested on thermal images for both datasets. By doing that, only 34.04% in terms of mAP is obtained compared to 51.35% if the tests are performed on visible data using FLIR dataset. The same observation is made on KAIST dataset, in which 21.98% as mAP is obtained compared to 58% if the tests are performed on visible data. As expected, the domain mismatch between training visible samples and testing thermal samples leads to a significant performance drop (of 17.31% and 36.2% in terms of mAP for FLIR and KAIST datasets, respectively). These results comply with our main observation stated at the beginning of the paper, that feature spaces between thermal and visible domains are nonequivalent. These initial low results justify the need for performing domain adaptation by means of feature alignment in order to perform well in the target domain, which is our main proposal in this paper.

Our obtained results of the proposed unsupervised domain adaptive detector and the results of the baseline Faster R-CNN without adaptation (both trained on visible data and tested on thermal data) on FLIR dataset are reported in Table 1. These results are compared in the same table to other unsupervised domain adaptation detectors (DA-Faster, SA-DA-Faster, SWDA, and HTCN). For the different detectors, the results are given for three object classes ("Car", "Bicycle" and "Person") and the average results are shown in the last column.

Table 1 Comparisons of the proposed unsupervised adaptive detector to the baseline Faster R-CNN detector and to other existing domain adaptation detectors on FLIR dataset, all evaluated in terms of mAP. The average mAP is given in the last column.

Detector	Car	Bicycle	Person	average mAP
Baseline	53.19	23.95	24.98	34.04
DA-Faster [4]	59.90	24.30	26.60	36.93
SA-DA-Faster [5]	**70.38**	33.30	**47.27**	50.30
SWDA [36]	58.96	32.02	32.32	41.40
HTCN [3]	56.37	37.95	33.17	42.49
Ours	66.83	**49.34**	43.41	**53.19**

Following the same protocol of evaluation, these comparisons on KAIST dataset, but only on pedestrian class are reported in Table 2. Results are given at daytime, nighttime and for all images.

Table 2 Comparisons of our proposed detector to the baseline method and to other existing domain adaptation methods, all evaluated in terms of mAP on KAIST dataset. Details of the corresponding results at daytime and nighttime are given as well.

Detector	night	day	all
Baseline	22.20	21.86	21.98
DA-Faster [4]	54.19	42.99	45.50
SA-DA-Faster [5]	64.6	41.6	48.7
SWDA [36]	61.04	43.03	48.90
HTCN [3]	70.70	55.40	59.75
Ours	**74.05**	**60.07**	**64.01**

As shown in tables 1 and 2, for both datasets, our proposed adaptive detector outperforms the baseline method and the other existing unsupervised adaptive detectors. Using FLIR dataset, it achieves the best average result on the three classes; 53.19% in terms mAP with a significant margin of 19.15% compared to the baseline Faster R-CNN detector and a margin of 16.26%, 2.89%, 11.79%, and 10.7% compared to DA-Faster, SA-DA-Faster, SWDA and HTCN, respectively. Also, the best result is obtained on KAIST dataset; 64.01% in terms mAP with a margin of 42,03%, 18.52%, 15.31%, 15.11%, and 4,26% compared to the baseline detector, DA-Faster, SA-DA-Faster, SWDA, and HTCN, respectively. On the same dataset KAIST, by comparing day and night results, better results are obtained at nighttime (74.05% as mAP) since thermal data is proven to be more effective at that time.

From these obtained results, the relevance of performing domain adaptation for detection to be tested on another domain is proven. The margin regarding the baseline method is significant for our proposed detector and for the other adaptive detectors as well. Also, by comparing different adaptive detectors, our obtained results are consistent with our expectations, since the proposed detector has the advantage of performing multiple 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 [36,3] and alignments at RPN and ROI are performed in [4, 5]. Combining different alignments in our proposed adaptive detector leads to an overall performance.

The corresponding qualitative results on some sample images from FLIR and KAIST datasets are shown in Figure 2. These results also indicate the performance increase by our adaptive detector compared to others. Precisely, in the sample visual results, it is shown that some false positives and false negatives results are corrected by the proposed detector compared to [4,5,36,3].

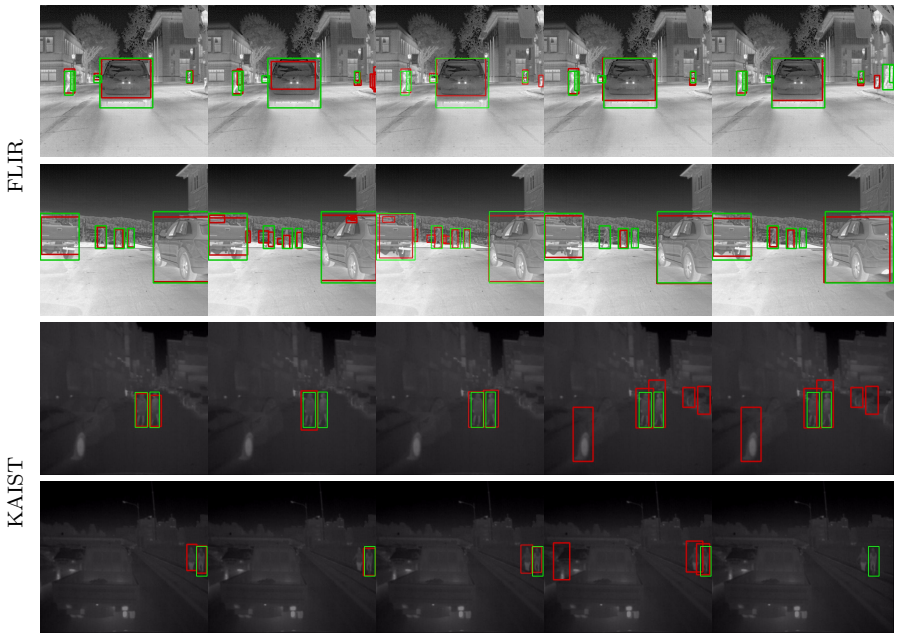

Fig. 2 Qualitative detection results of our proposed detector on FLIR ADAS and KAIST datasets compared to other adaptive detectors. From left to right results of (a) Our proposed detector (b) DA-Faster [4], (c) SA-DA-Faster [5] (d) SWDA [36], and (e) HTCN [3]. From top to bottom: results of 4 sample images in thermal domain, the two first rows showing results on FLIR dataset and the two last rows for KAIST dataset. For each sample image, the detection results are shown in red color and the corresponding annotated bounding boxes in green color.

4.5 Ablation Study

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 shown in Table 3. As depicted in this table and following the same notations in [36], G, L, CTX refer to global alignment, local alignment, and context-vector based regularization, respectively. Compared to [36], since we also consider other alignments (image, instance, and consistency regularization) in the second phase of the network, we add R to refer to them.

Table 3 Results on FLIR ADAS and KAIST datasets of our proposed detector in terms of mAP 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 phase of the network.

Detector	G	L	CTX	R	FLIR	KAIST
Ours	✗	✓	✓	✓	47.87	26.2
	✓	✗	✓	✓	50.11	54.4
	✓	✓	✓	✗	48.59	59.5
	✓	✓	✓	✓	**53.19**	**64.01**

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 phase of the detector, the results are further improved, whether is the dataset. These results justify our choice of combining different alignments at different levels and phases in order to respond to different aspects of the domain divergence.

4.6 Comparisons with generative models

To further justify the choices presented in this paper, mainly about the non-generative vs. generative models, we evaluate in this section some existing generative models for detection. As already discussed, the goal of this particular type of adaptation is to generate a realistic RGB image from every thermal input image in order to improve its visual quality. Particularly, we evaluate and compare the detection performance of the existing generative models mapping from thermal to visible domains presented in section 2.2.1, namely, CycleGAN [48] and TIC-CGAN [17]. The corresponding results are reported in Table 4 and compared to the baseline detector in terms of mAP.

Table 4 Comparisons in terms of mAP of the baseline detector to some existing generative models from visible to thermal domains for detection using Faster R-CNN.

Model	FLIR	KAIST
Baseline	34.04	21.98
CycleGAN [48]	8.60	21.82
TIC-CGAN [17]	16.00	18.31

As shown in the table, the baseline results are not improved by means of generative models from thermal to visible domains. Worse results are instead obtained, mainly on FLIR dataaset since its visual quality in the thermal domain is initially satisfactory. These results are explained by the fact that such techniques aim at improving the visual quality, without focusing on any target object in the image. This observation stated at the beginning of the current paper, is confirmed by the underlying obtained results. This was our main motivation to rather refer to the non-generative models, which are incorporated in the detector architecture in order to align features. To better confirm these observations, some qualitative results on four sample images from KAIST and FLIR datasets are also shown in Figure 3. From these results, we can clearly observe that whether is the used generative model (CycleGAN [48] or TIC-CGAN [2]), target objects mainly persons on KAIST dataset are usually suppressed in the generated images.

5 Conclusion

In this paper, we proposed a novel approach for domain adaptation in the thermal domain at no additional annotation cost. As far as we know, it is the first time

(a) Input image (b) Ground truth (c) CycleGAN (d) TIC-CGAN

Fig. 3 Qualitative results of CycleGAN [48] and TIC-CGAN [2] on two sample thermal images (the input) with their corresponding visible images (the target) for each dataset. From top to bottom: results of 4 sample images, two first rows showing results on FLIR dataset and the two last rows on KAIST dataset. For each sample image, annotations of target objects present in the input thermal images are drawn with green bounding boxes.

that such unsupervised domain adaptation method is employed for detection in heterogeneous thermal and visible domains. The proposed detector incorporating feature distribution alignments into Faster R-CNN architecture has the advantage of combining different domain classifiers in order to achieve an overall performance. Despite its relevance for practical applications, such adaptation for detection in thermal and visible domains is not yet investigated. By means of tests on two widely used datasets for multispectral detection, the effectiveness of the proposed detector is proven by obtaining better results compared to the baseline detector with an outstanding margin. Its performance also exceeds some existing unsupervised domain adaptation methods for detection in homogeneous domains.

There are several possible extensions of this work. For instance, as detection is a basis step to perform other tasks in video analytic, the impact of the obtained improvement could be investigated on other applications such as tracking and activity recognition. Also, the proposed feature alignment approach which is incorporated into Faster R-CNN architecture for illustrative purpose, can readily be replaced by other deep detectors.

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

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Bayoudh, K., Knani, R., Hamdaoui, F., Mtibaa, A.: A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. The Visual Computer pp. 1–32 (2021)
2. Berg, A., Ahlberg, J., Felsberg, M.: Generating visible spectrum images from thermal infrared. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1143–1152 (2018)
3. Chen, C., Zheng, Z., Ding, X., Huang, Y., Dou, Q.: Harmonizing transferability and discriminability for adapting object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8869–8878 (2020)
4. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3339–3348 (2018)
5. Chen, Y., Wang, H., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Scale-aware domain adaptive faster r-cnn. International Journal of Computer Vision **129**(7), 2223–2243 (2021)
6. Dai, X., Yuan, X., Wei, X.: Tirnet: Object detection in thermal infrared images for autonomous driving. Applied Intelligence **51**(3), 1244–1261 (2021)
7. Devaguptapu, C., Akolekar, N., M Sharma, M., N Balasubramanian, V.: 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)
8. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning, pp. 1180–1189. PMLR (2015)
9. Gautam, A., Singh, S.: Neural style transfer combined with efficientdet for thermal surveillance. The Visual Computer pp. 1–17 (2021)
10. Ghose, D., Desai, S.M., Bhattacharya, S., Chakraborty, D., Fiterau, M., Rahman, T.: Pedestrian detection in thermal images using saliency maps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–0 (2019)
11. Huang, X.: Moving object detection in low-luminance images. The Visual Computer pp. 1–13 (2021)
12. Hwang, S., Park, J., Kim, N., Choi, Y., So Kweon, I.: Multispectral pedestrian detection: Benchmark dataset and baseline. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1037–1045 (2015)
13. Jiang, B., Chen, C., Jin, X.: Unsupervised domain adaptation with target reconstruction and label confusion in the common subspace. Neural computing and applications **32**(9), 4743–4756 (2020)
14. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision, pp. 694–711. Springer (2016)
15. Kieu, M., Bagdanov, A.D., Bertini, M., Bimbo, A.D.: Task-conditioned domain adaptation for pedestrian detection in thermal imagery. In: Computer Vision - ECCV (2020)
16. Kim, M., Joung, S., Park, K., Kim, S., Sohn, K.: Unpaired cross-spectral pedestrian detection via adversarial feature learning. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 1650–1654. IEEE (2019)
17. Kuang, X., Zhu, J., Sui, X., Liu, Y., Liu, C., Chen, Q., Gu, G.: Thermal infrared colorization via conditional generative adversarial network. Infrared Physics & Technology p. 103338 (2020)
18. Li, W., Xu, Z., Xu, D., Dai, D., Van Gool, L.: Domain generalization and adaptation using low rank exemplar svms. IEEE transactions on pattern analysis and machine intelligence **40**(5), 1114–1127 (2017)

19. Li, X., Hu, Y., Zheng, J., Li, M., Ma, W.: Central moment discrepancy based domain adaptation for intelligent bearing fault diagnosis. Neurocomputing **429**, 12–24 (2021)
20. Lin, C., Lu, J., Wang, G., Zhou, J.: Graininess-aware deep feature learning for robust pedestrian detection. IEEE transactions on image processing **29**, 3820–3834 (2020)
21. Liu, H., Wang, X., Zhang, W., Zhang, Z., Li, Y.F.: Infrared head pose estimation with multi-scales feature fusion on the irhp database for human attention recognition. Neurocomputing **411**, 510–520 (2020)
22. Liu, Q., Li, X., He, Z., Li, C., Li, J., Zhou, Z., Yuan, D., Li, J., Yang, K., Fan, N., et al.: Lsotb-tir: A large-scale high-diversity thermal infrared object tracking benchmark. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 3847–3856 (2020)
23. Liu, W., Liao, S., Ren, W., Hu, W., Yu, Y.: High-level semantic feature detection: A new perspective for pedestrian detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5187–5196 (2019)
24. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp. 2794–2802 (2017)
25. Mhalla, A., Chateau, T., Gazzah, S., Essoukri Ben Amara, N.: An embedded computer-vision system for multi-object detection in traffic surveillance. IEEE Transactions on Intelligent Transportation Systems **20**(11), 4006–4018 (2018)
26. Mohamed Amine, M., Hajer, F., Anis, S., Najoua, E.B.A.: Thermal image enhancement using generative adversarial network for pedestrian detection. International Conference on Pattern Recognition (2020)
27. Motiian, S., Piccirilli, M., Adjeroh, D.A., Doretto, G.: Unified deep supervised domain adaptation and generalization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5715–5725 (2017)
28. Nasiri, A., Taheri-Garavand, A., Omid, M., Carlomagno, G.M.: Intelligent fault diagnosis of cooling radiator based on deep learning analysis of infrared thermal images. Applied Thermal Engineering **163**, 114410 (2019)
29. Ouyang, W., Zeng, X., Wang, X.: Learning mutual visibility relationship for pedestrian detection with a deep model. International Journal of Computer Vision **120**(1), 14–27 (2016)
30. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering **22**(10), 1345–1359 (2010). DOI 10.1109/TKDE.2009.191
31. Panareda Busto, P., Gall, J.: Open set domain adaptation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 754–763 (2017)
32. Park, S., Hwang, J., Park, J.E., Ahn, Y.C., Kang, H.W.: Application of ultrasound thermal imaging for monitoring laser ablation in ex vivo cardiac tissue. Lasers in surgery and medicine **52**(3), 218–227 (2020)
33. Rahman, M.M., Fookes, C., Baktashmotlagh, M., Sridharan, S.: Correlation-aware adversarial domain adaptation and generalization. Pattern Recognition **100**, 107124 (2020)
34. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp. 91–99 (2015)
35. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp. 91–99 (2015)
36. Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Strong-weak distribution alignment for adaptive object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6956–6965 (2019)
37. Sankaranarayanan, S., Balaji, Y., Jain, A., Nam Lim, S., Chellappa, R.: Learning from synthetic data: Addressing domain shift for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3752–3761 (2018)
38. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7167–7176 (2017)
39. Wang, M., Deng, W.: Deep visual domain adaptation: A survey. Neurocomputing **312**, 135–153 (2018)
40. Wei, L., Cui, W., Hu, Z., Sun, H., Hou, S.: A single-shot multi-level feature reused neural network for object detection. The Visual Computer **37**(1), 133–142 (2021)

41. Xu, C.D., Zhao, X.R., Jin, X., Wei, X.S.: Exploring categorical regularization for domain adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11724–11733 (2020)
42. Xu, D., Ouyang, W., Ricci, E., Wang, X., Sebe, N.: Learning cross-modal deep representations for robust pedestrian detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5363–5371 (2017)
43. Xu, M., Zhang, J., Ni, B., Li, T., Wang, C., Tian, Q., Zhang, W.: Adversarial domain adaptation with domain mixup. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 6502–6509 (2020)
44. Yang, J., An, W., Wang, S., Zhu, X., Yan, C., Huang, J.: Label-driven reconstruction for domain adaptation in semantic segmentation. In: European Conference on Computer Vision, pp. 480–498. Springer (2020)
45. Zellinger, W., Moser, B.A., Saminger-Platz, S.: On generalization in moment-based domain adaptation. Annals of Mathematics and Artificial Intelligence **89**(3), 333–369 (2021)
46. Zhang, H., Fromont, E., Lefèvre, S., Avignon, B.: Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In: 2020 IEEE International Conference on Image Processing (ICIP), pp. 276–280. IEEE (2020)
47. Zhang, H., Zhang, Z., Odena, A., Lee, H.: Consistency regularization for generative adversarial networks. arXiv preprint arXiv:1910.12027 (2019)
48. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp. 2223–2232 (2017)
49. Zoetgnande, Y.W.K., Dillenseger, J.L., Alirezaie, J.: Edge focused super-resolution of thermal images. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2019)
50. Zou, Y., Yu, Z., Vijaya Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Proceedings of the European conference on computer vision (ECCV), pp. 289–305 (2018)