Bottom-up and Layer-wise Domain Adaptation for Pedestrian Detection in Thermal Images

Bottom-up and Layer-wise Domain Adaptation for Pedestrian Detection in Thermal Images

MY KIEU, Media Integration and Communication Center (MICC), University of Florence, Italy
ANDREW D. BAGDANOV, Media Integration and Communication Center (MICC), University of Florence, Italy
MARCO BERTINI, Media Integration and Communication Center (MICC), University of Florence, Italy

Pedestrian detection is a canonical problem for safety and security applications, and it remains a challenging problem due to the highly variable lighting conditions in which pedestrians must be detected. This paper investigates several domain adaptation approaches to adapt RGB-trained detectors to the thermal domain. Building on our earlier work on domain adaptation for privacy-preserving pedestrian detection, we conducted an extensive experimental evaluation comparing top-down and bottom-up domain adaptation and also propose two new bottom-up domain adaptation strategies. For top-down domain adaptation we leverage a detector pre-trained on RGB imagery and efficiently adapt it to perform pedestrian detection in the thermal domain. Our bottom-up domain adaptation approaches include two steps: first, training an adapter segment corresponding to initial layers of the RGB-trained detector adapts to the new input distribution; then, we reconnect the adapter segment to the original RGB-trained detector for final adaptation with a top-down loss. To the best of our knowledge, our bottom-up domain adaptation approaches outperform the best-performing single-modality pedestrian detection results on KAIST, and outperform the state-of-the-art on FLIR©.

CCS Concepts: • **Computing methodologies** → *Artificial intelligence*; Computer vision; Computer vision problems; Object detection;

Additional Key Words and Phrases: domain adaptation, bottom-up adaptation, thermal, pedestrian detection, privacy-preserving, object detection, fine-tuning, layer-wise adaptation, transfer learning

ACM Reference Format:
My Kieu, Andrew D. Bagdanov, and Marco Bertini. 2020. Bottom-up and Layer-wise Domain Adaptation for Pedestrian Detection in Thermal Images. In *ACM Transactions on Multimedia Computing, Communications and Applications.* ACM, New York, NY, USA, 19 pages.

1 INTRODUCTION

Pedestrian detection is one of the essential topics in computer vision, with diverse applications to safety and security such as video surveillance, autonomous driving, robotics, criminal investigation, etc. According to [29], there were an estimated 240 million installed video surveillance cameras worldwide in 2014. The continued need for detection and observation of humans in public spaces and the advent of autonomous driving systems promises to add many more cameras. In the last decade, the majority of existing detectors focused on RGB images [4, 27, 41] and can fail to work under various illumination (nighttime) or adverse weather (fog, dust) conditions [22].

Thermal imagery offers a way to address these challenges since thermal cameras capture the radiated heat of objects. They can image clear object silhouettes independent of lighting conditions and are also less affected by adverse weather conditions. Thus, recent works on pedestrian detection

TOMM, 2020,

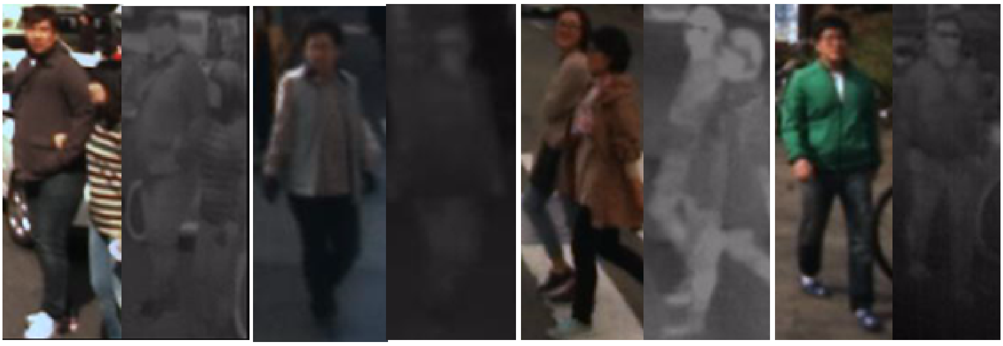

Fig. 1. **Thermal imaging and privacy preservation**. Shown are four cropped images from the KAIST dataset. On the left of each is the RGB image, to the right the crop from the corresponding thermal image. Note how persons are readily identifiable in visible spectrum images, but not in corresponding thermal images. Although identity is concealed, there is still enough information in thermal imagery for detection.

have investigated the use of thermal sensors as a *complementary* signal for visible spectrum images [5, 13, 19, 22, 23, 26, 39]. Approaches such as these aim to combine thermal and RGB image information to obtain the most robust pedestrian detection possible. However, this can limit applicability in real applications due to the cost of deploying multiple, aligned sensors (thermal and visible). Most importantly, using visible spectrum sensors does not offer the same degree of privacy preservation guarantee as using thermal-only for person detection.

On the other hand, citizens are naturally concerned that being observed violates their right to privacy. A variety of solutions on how to mitigate privacy-preservation concerns have been discussed in [1]. For example, developing sensors for privacy-preservation [31], using Kinect or other depth sensors, leveraging Low Resolution (LR) video data, using a joint ActionXPose and Single Shot MultiBox Detector (SSD) for semi-supervised anomaly detection [1]. However, most of the mentioned solutions have limitations due to the high cost of installing multiple cameras, or their performance is limited.

Figure 1 gives an example of cropped pairs of color and thermal images from the KAIST dataset [16]. As we can see from these examples that even in relatively low-resolution color images, persons can be readily identified. Meanwhile, thermal images can retain distinctive image features for detection while preserving privacy. Thermal imagery is privacy-preserving in the sense that person identification is difficult or impossible. Our approaches are motivated by the fact that thermal images can guarantee the balance between security and privacy concerns.

State-of-the-art work on pedestrian detection has mostly concentrated on using multispectral images combining visible and thermal images for training and testing [2, 3, 5, 13, 19, 22–24, 26, 32, 37, 39, 40]. Only a few single-modality detection works focus only on thermal images [3, 7, 15, 17, 18]. Since these works discard the visible spectrum, which contains much more detection information, they are weaker when compared to multispectral detectors. Robust pedestrian detection on only thermal data is a non-trivial task and there is still a large potential to improve the thermal detection performance. Our previous work probed the limits of pedestrian detection using thermal imagery alone [18]. We investigated three top-down domain adaptation approaches and a bottom-up domain adaptation approach, which outperform state-of-the-art pedestrian detection at nighttime in thermal imagery. Exploiting only thermal data is a fundamental advantage of our work (visible data are not employed at both training adaptation and the test phase), this is crucial when deploying surveillance systems under a variety of environmental conditions.

In addition to extending our work on bottom-up domain adaptation for thermal pedestrian detection, in this paper we focus on improving pedestrian detection results on thermal-only data by proposing a new, layer-wise domain adaptation strategy which gradually adapts early convolutional layers of a pre-trained detector to the thermal domain. The main motivation for layer-wise adaptation is that, from previous work on bottom-up adaptation, we recognized that adaptation of *early* layers of the network helped the most to preserve the learned knowledge from the visible domain which is useful for the adaptation to the thermal domain. Our hypothesis is that if we adapt slowly from the bottom up in a layer-wise manner, it should improve adaptation. Through extensive experimental evaluation on two datasets and an analysis and interpretation of the contribution of bottom-up adaptation we show that, though only exploiting thermal imagery at test time, our domain adaptation approaches outperform state-of-the-art thermal detectors on the KAIST Multispectral Pedestrian Dataset[16] and FLIR Starter Thermal Dataset [10]. Moreover, our bottom-up and layer-wise adaptation approaches outperform many state-of-the-art *multispectral* detection approaches which exploit both thermal and visible spectra at test time. To the best of our knowledge, this is the first work that investigates the bottom-up domain adaptation for pedestrian detection in thermal imagery.

The contributions of this work are:

- We propose a new type of bottom-up domain adaptation which adapts the pre-trained detector in a *layer-wise* manner. The result shows that the relatively simple bottom-up strategy better preserves learned features from the visible domain and lead to robust pedestrian detection results in the final detector.
- We give a detailed comparison between three top-down domain adaptation approaches and our two proposed bottom-up adaptation approaches. Our experiments show that our bottom-up and layer-wise adaptation approaches consistently outperform top-down adaptation.
- To the best of our knowledge, we obtain the best detection result on the FLIR© dataset [10], and we are also the best detection result on KAIST dataset [16] compared to all existing single modality approaches [3, 15, 18, 40]. Moreover, by exploiting only thermal imagery on KAIST dataset, we outperform many the state-of-the-art multispectral pedestrian detectors [16, 19, 21, 38, 39, 43] which use both visible and thermal for training and testing.

The rest of this paper is organized as follows. In the next section, we briefly review related work from the computer vision literature on pedestrian detection, thermal imaging, and domain adaptation. In section 3, we describe several approaches to domain adaptation that we apply to the problem of pedestrian detection in thermal imagery. We report on a range of experiments conducted in section 4, and section 5 concludes with discussion of our contribution and future research directions.

2 RELATED WORK

In this section we briefly review the literature on pedestrian detection in RGB, thermal, and multispectral imagery.

Pedestrian detection in RGB images. Pedestrian detection has attracted constant attention from the computer vision research community through the years, and the literature on it is vast [4]. As with other computer vision applications, higher and higher accuracy has been achieved with the advent of deep neural networks [2]. However, pedestrian detection is still challenging due to a variety of environmental conditions such as changing illumination, occlusion, and variation of viewpoint and background [32].

Many works using Convolutional Neural Networks (CNNs) compete for state-of-the-art results on standard benchmark datasets. For example, authors in [37] used a single task-assistant CNN

(TA-CNN) to train multiple tasks from multiple sources to bridge the gaps between different datasets. Their method learns a high-level representation for pedestrian detection by jointly optimizing with semantic attributes. The speed of the proposed model, however, is limited to 5 fps. Along these lines, the estimation of visibility statuses for multiple pedestrians and recognition of co-existing pedestrians via a mutual visibility deep model was proposed in [32]. Fast/Faster R-CNN has become the predominant framework for pedestrian detection, as in the Scale-Aware Fast R-CNN model [24]. This approach incorporates a large sub-network and a small sub-network into a unified architecture which implements a divide-and-conquer strategy. The authors of [41] proposed a combination of Region Proposal Network (RPN) followed by Boosted Forests (BF) for pedestrian detection based on Faster R-CNN. In a different direction, semantic segmentation has also been used for pedestrian localization given its robustness to shape and occlusion [5]. Finally, high-level semantic features were used for anchor-free pedestrian detection task in [27]. The advantages of these techniques is that they demonstrate excellent RGB-domain performance, although they can catastrophically fail under low-illumination conditions.

Multispectral pedestrian detection. The combination of thermal and RGB images has been shown to improve object detection results. For example, thermal image features were used for robust visible detection results in a cross-modality learning framework [40] including a Region Reconstruction Network (RRN) and Multi-Scale Detection Network (MDN). Two types of fusion networks to explore visible and thermal image pairs were investigated by the authors in [39]. The work in [26] also considered four different network fusion approaches (early, halfway, late, and score fusion) for multispectral pedestrian detection task.

Most of the recent top-performing, multispectral pedestrian detection are variations of VGG or Fast-/Faster-RCNN, which leverage two-stage network architectures to investigate the combination of visible and thermal features. For instance, Illumination-aware Faster R-CNN (IAF RCNN) [23] used the Faster R-CNN detector to perform multispectral pedestrian detection. The authors in [19] detected persons in multispectral video with a combination of a Fully Convolutional Region Proposal Network (RPN) and a Boosted Decision Trees Classifier (BDT). The generalization ability of RPN was also investigated for multispectral person detection [11]. Similar to the ideas in [9], a fusion architecture network (MSDS-RCNN [22]), which includes a multispectral proposal network (MPN) and a multispectral classification network (MCN), was proposed multispectral detection. A region feature alignment module was proposed by [42]. Similarly, the authors in [6] suggested box-level segmentation via a supervised learning framework.

The common advantages of these multispectral methods is that they leverage two-stage frameworks which are suitable for learning combined representations of two inputs. However, since they are usually based on far more complex network architectures in order to align modalities at inference time, detection speed is usually under 5 fps.

On the other hand, differing from most of the above, some papers utilized a one-stage detector for multispectral pedestrian detection. For example, a fast RGB single-pass network architecture (YOLOv2 [33]) was used by the authors in [38] for multispectral person detection. Authors in [21] leveraged a deconvolutional, single-shot multi-box detector (DSSD) to exploit the correlation between visible and thermal features. Two Single Shot Detectors (SSDs) were adopted by [43] to investigate the potential of fusing color and thermal features with Gated Fusion Units (GFU).

Most of the aforementioned multispectral methods use both visible and thermal images for training and testing in order to make the most out of both modalities, and they typically need to resort to additional (and expensive) annotations. Aside from the technical and economic motivations for preferring thermal-only sensor deployment over multispectral methods, using visual spectrum images does not guarantee the same privacy-preserving affordances offered by thermal-only detectors.

Pedestrian detection in thermal imagery. There are also a few works that, like ours, focus on pedestrian detection using only thermal imagery. An early example is the work in [17] which uses adaptive fuzzy C-means clustering to segment IR images and retrieve candidate pedestrians, then prunes candidate pedestrians by classifying with CNN. The authors report a significant reduction in computational complexity compared to the sliding window framework. The authors in [12] use a Pixel-wise Contextual Attention network (PiCA-Net) and R3-Net to create saliency maps. Then, Faster R-CNN is trained for pedestrian detection using the original thermal image and another channel containing the saliency map. The work in [3] proposed to use Thermal Position Intensity Histogram of Oriented Gradients (TPIHOG) and the Additive Kernel SVM (AKSVM) for nighttime-only detection in thermal imagery.

A few approaches leverage RGB images as a data augmentation by performing RGB to thermal image translation. For instance, several data preprocessing steps were applied by [15] to make thermal images look more similar to grayscale-converted RGB images, then a fine-tuning step was performed on a pre-trained SSD300 detector. Recently, the Cycle-GAN was used as a preprocessing step for image-to-image translation [7], before feeding the input to a Faster-RCNN multi-modal for pedestrian detector training. The common drawbacks of most of methods mentioned above is that they use many complex preprocessing steps or use hand-crafted features and, as a consequence, their performance suffers (e.g. Fuzzy C-means [17] requires 2.5 seconds per frame and achieves only 34% miss, and TPIHOG [3] needs 40 seconds per frame to reach only 56.8% miss rate).

We also focus on thermal-only detection, but we do not use any complex data preprocessing step to translate images. We tackle the problem of transferring knowledge between domains and adapting the learned knowledge from the previous domain to the new domain. Our domain adaptation approaches are relatively simple because they are based on the single-stage detector YOLOv3 [34], which can be optimized end-to-end and retain its real-time performance.

Domain Adaptation. Domain adaptation has a long history for both supervised and unsupervised recognition in computer vision. Domain adaptation attempts to exploit learned knowledge from a source domain in a new target domain. Many works have proposed domain adaptation techniques to bridge the gap between thermal and visible domains [20, 28, 30]. An early work in thermal infrared person detection is [15] which uses domain adaptation based on feature transformation techniques (inversion, equalization, and histogram stretching) to transform thermal images as close as possible to the color. A similar idea is the Invertible Autoencoder introduced in [36]. One of our approaches described in this work was inspired by the AdapterNet [14], which proposed adding a shallow CNN before the original model that transforms the input image to the target domain. The transformed input is then passed through an unmodified network trained in the source domain. Our domain adaptation approaches leverage the features learned from the source RGB domain and perform detection in the target domain after bottom-up domain adaptation.

Our contribution with respect to the state-of-the-art. Our approach is substantially different from the above mentioned approaches, as our goal is to perform detection using *only* thermal imagery. Our results demonstrate the bottom-up domain adaptation leads results that outperform all single-modality detection approaches (both RGB- and thermal-only) from the literature. Moreover, our thermal-only detectors better preserve information from the source domain and perform comparably to multispectral approaches that leverage both thermal and visible spectra at detection time. Accurate detection from thermal imagery offers an affordable and scalable solution to privacy preserving person detection in a variety of safety and security application scenarios.

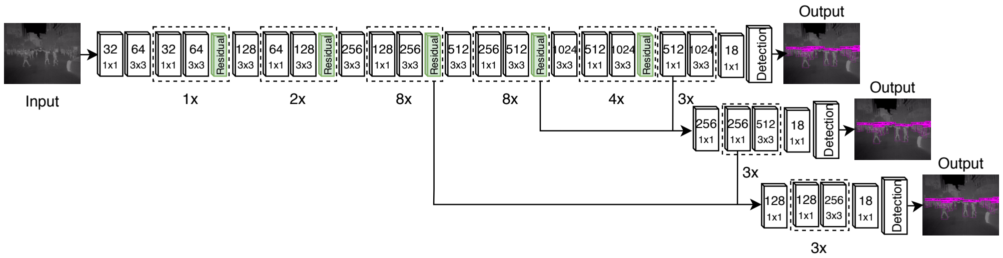

Fig. 2. YOLOv3 architecture

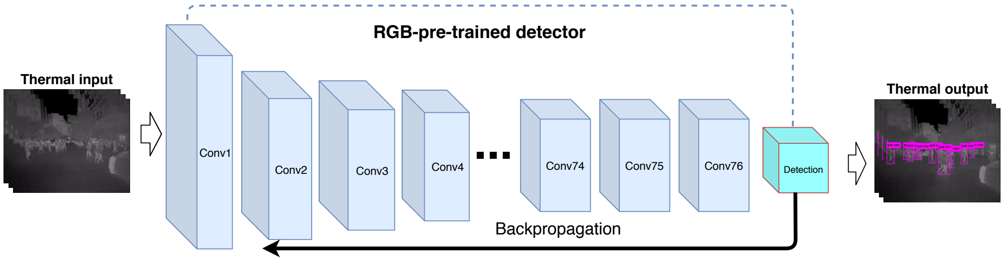

Fig. 3. Top-down domain adaptation refers to using fine-tuning to adapt a detector to a new domain (e.g. thermal imagery). Adaptation to the new input distribution happens only via back-propagated loss from the end of the network (at the top) down to the new input distribution.

3 BOTTOM-UP DOMAIN ADAPTATION

This section describes our proposed approaches, which build upon our earlier work on domain adaptation approaches for pedestrian detection [18]. The goal of our earlier domain adaptation methods showed that relatively simple domain adaptation on only thermal image can outperform many state-of-the-art approaches. Here we first briefly summarize the salient ideas of three top-down domain adaptation and bottom-up adaptation in order to contrast with ours. Then we introduce a simple layer-wise technique that significantly boosts performance of pedestrian detection in thermal imagery.

3.1 Top-down adaptation approaches

The top-down adaptation approaches we consider are based on transfer learning and use one of the fastest and most accurate detectors available today: YOLOv3 [34], which is pre-trained on ImageNet and subsequently fine-tuned on the MS COCO dataset [25]. We adapt the YOLOv3 detector to the new target domain through a sequence of domain adaptation steps. YOLOv3 is a very deep detection network with 106 layers and three detection heads for detecting objects at different scales as illustrated in figure 2. YOLOv3 uses a fully-convolutional residual network as its backbone. The network is coarsely structured into five residual *groups*, each consisting of one or more residual *blocks*. As we see in figure 2, these five *groups* include 23 residual *blocks*, each consisting of two-convolutional layers with residual connections adding the input of each block to the output.

We refer to this as *top-down adaptation* because of the way fine-tuning on the new domain happens only via back-propagation where the supervision signal comes from the loss at the *end* (i.e. the *top* of the network), down to the new input distribution. In figure 3 we illustrate this top-down adaptation, which refers to the fine-tuning approach to the new input distribution (thermal domain in our case). We fine-tune the pre-trained RGB detector to adapt to the new thermal input.

In the descriptions below, we use a notational convention to refer to each technique that indicates which image modalities are used for training and testing. For example, the technique reported as TD(VT, T) is Top-Down domain adaptation, with adaptation on Visible spectrum images, followed by adaptation on Thermal images, and finally tested on Thermal images. The three top-down domain adaptation approaches we consider are ([18]):

- **Top-down visible: TD(V, V)**: This domain adaptation approach directly fine-tunes YOLOv3 on visible images in the target domain for pedestrian detection. Testing was performed on visible spectrum images. This experiment mainly served as the baseline for comparison with single modality techniques on visible imagery.
- **Top-down thermal: TD(T, T)**: This approach directly fine-tunes YOLOv3 on only thermal images by duplicating the thermal image three times, once for each input channel of the RGB-trained detector. Testing was performed only on thermal imagery (no RGB images are available at test time). This experiment served as the baseline for the comparison with single modality techniques and domain adaptation in thermal imagery alone.
- **Top-down visible/thermal: TD(VT, T)**: This approach is a variant of the two top-down approaches described above. First, we adapt YOLOv3 to the visible spectrum pedestrian detection domain, and then we fine-tune that detector on thermal imagery. Testing was performed only on thermal images (no RGB images available). The idea here was to determine if knowledge from the visible spectrum could be retained and exploited after the final adaptation to the thermal domain.

3.2 Bottom-up adaptation

In our previous work we proposed a type of *bottom-up* domain adaptation [18]. *Bottom-up* domain adaptation includes two stages: we first train a bottom-up adapter segment to adapt to the new input distribution, and then we reconnected this adapter segment for the final fine-tuning using a top-down loss. The main components of our bottom-up domain adaptation approach are as follows (see [18] for more details).

Notation. Let $f_\Theta(\mathbf{x})$ represent the detector (YOLOv3 in our case) parameterized by parameters $\Theta = \{\theta_1, \theta_2, \ldots, \theta_N\}$, where θ_i represents the parameters of the ith layer of the network. We use the notation $\Theta_{n:m}$ to denote the parameters of layers n through m of the network f (so $\Theta = \Theta_{1:N}$). Similarly, we use $f_{n:m}$ to represent the forward pass of the network f from layer n through layer m. We assume f to be pre-trained for detection in the RGB domain – in our experiments we start from the network TD(V, V) described above.

Adapter segment training. The first step in bottom-up domain adaptation it to train an *adapter segment* to mimic RGB feature activations when given *thermal* images as input. We create a new network segment f' identical to the first m layers of f (i.e. copying the weights of TD(V, V). Given paired visible/thermal spectrum images $(\mathbf{x}_v, \mathbf{x}_t)$ (such as those available in KAIST), we train the parameters θ' of f' using Stochastic Gradient Descent (SGD) on the following loss:

$$\mathcal{L}_A(\mathbf{x}_v, \mathbf{x}_t; \theta') = ||f_{1:m}(\mathbf{x}_v) - f'(\mathbf{x}_t)||_2^2$$

This encourages the network f' – when given a thermal image as input – to output features close to those generated at the mth layer of f from an input RGB image. The main idea of the adapter segment is to intervene an early stage of the RGB-trained detector network and to train this adapter segment to adapt to the thermal domain. In early experiments we found $m = 10$ to be a good point of intervention for adapter segment training.

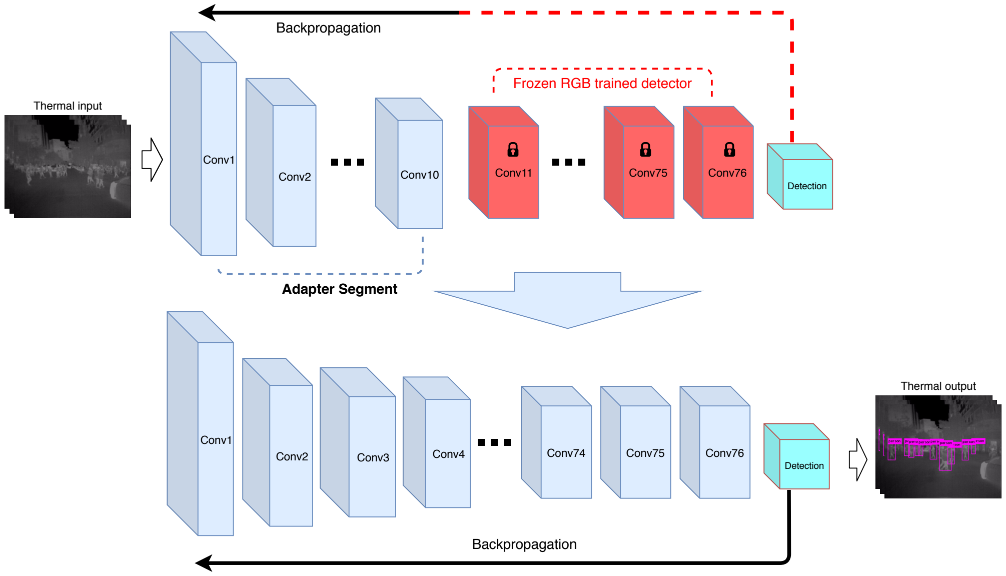

Fig. 4. **Bottom-up domain adaptation**. Starting from an RGB-trained detector, an *adapter segment* is trained to take thermal images as input and produce convolutional features similar to the features of the original network. When the adapter segment training has converged we reconnect the adapter segment to the *RGB-trained detector* for the final fine-tuning.

Fine-tuning of entire detector. After adapter segment training has converged, we reconnect the newly trained adapter segment to the original RGB-trained detector for the final fine-tuning of the whole detector on thermal images (as illustrated at the bottom of figure 4). To do this we train the final detection network $f_{m+1:N}(f'(\mathbf{x}_t))$ using only thermal images $\mathbf{x}_t$ and the original loss of the YOLOv3 network.

3.3 Layer-wise adaptation

We hypothesize that the fine-tuning slowly from the bottom of the network should preserve more knowledge from the original domain. Here we propose a new type of bottom-up domain adaptation, which we call *layer-wise* domain adaptation. It progressively fine-tunes each layer of the network starting from the *bottom* of the network. Layer-wise adaptation also includes two stages: first, layer-wise adaptation to adapt slowly to the new input distribution. Then, a final fine-tuning phase that trains the whole pipeline with a top-down loss. The conceptual schema of our layer-wise domain adaptation approach is given in figure 5.

Layer-wise domain adaptation proceeds as follows:

(1) **Layer-wise adaptation**: We start from the TD(V, V) network described in section 3.1 – i.e. from an RGB-trained detector already adapted to the new domain in the visible spectrum. We gradually train the initial layers of the YOLOv3 network using thermal images from the training set. As illustrated in the upper part of figure 5, the main idea is to adjust the RGB-trained detector network to adapt slowly to the thermal input from bottom of the network up to the top. At epoch i we freeze parameters $\theta_{3i+1:N}$ while fine-tuning. That is, at each epoch another three layers are unfrozen until the entire network is being fine-tuned. In the experiments we denote this approach as BU(VLT, T), the "L" in the "VLT" signifies training with layer-wise adaptation.

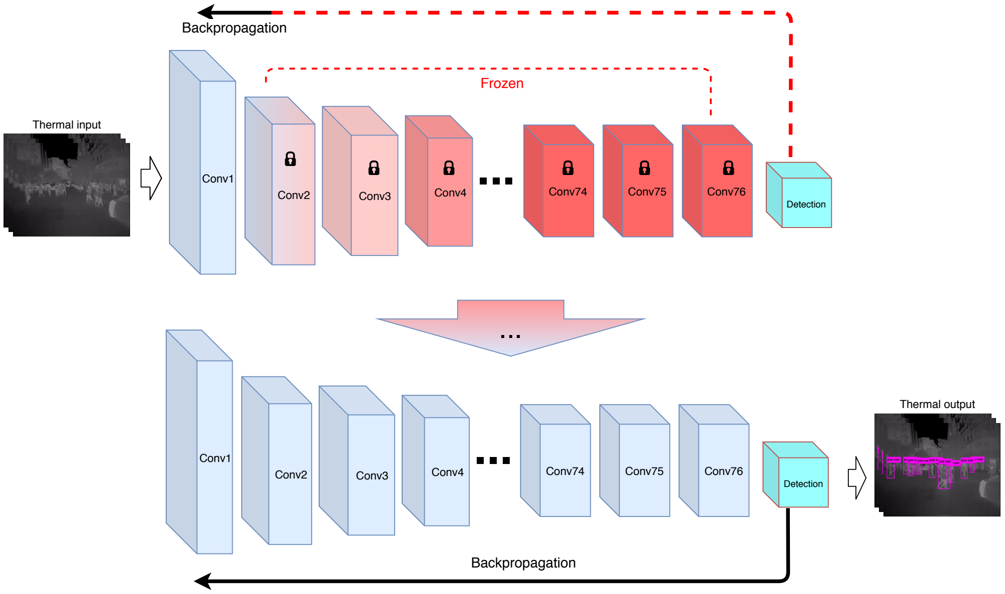

Fig. 5. **Layer-wise domain adaptation**. Instead of adapter segment training, in layer-wise adaptation we *gradually* adapt layers during fine-tuning. This is done by freezing layers during training and progressively unfreezing them. After gradually including all layers in the training a final fine-tuning of the entire detector pipeline on thermal images is performed.

(2) **Fine-tuning of the entire detector**: After adapter segment training has converged, the entire detector trained using end-to-end fine-tuning shown at the bottom of figure 5. Note that no RGB images are used and the whole pipeline is trained using only thermal images.

We also experimented with a variant of layer-wise adaptation that is more similar to the bottom-up adaptation strategy described in [18]. Instead of gradually unfreezing layers during adaptation, we freeze only the parameters $\Theta_{11:N}$ for the first 50 epochs, and then unfreeze them for the remaining 50 epochs to fine-tune the entire network. An overview of this bottom-up strategy is given in figure 4. We refer to this bottom-up variant as BU(VAT, T) in the experimental results.

4 EXPERIMENTAL RESULTS

In this section we report on a number of experiments we performed to evaluate our domain adaptation approaches with respect to the state-of-the-art in single- and multi-modal detection.

4.1 Datasets

All our approaches are evaluated and compared to the state-of-the-art on two public datasets: the KAIST multispectral pedestrian benchmark [16] and the FLIR Starter Thermal Dataset [10]. The KAIST dataset is the only large-scale pedestrian dataset with well-aligned visible/thermal image pairs [7]. Furthermore, over the past four years the annotations of the KAIST dataset have been improved for both training and test set [22]. We also evaluate our approaches on FLIR Starter Thermal Dataset. We chose this dataset because of its consistent annotation and its use in other work [7].

More specifically, the two datasets used in our experimental evaluation are:

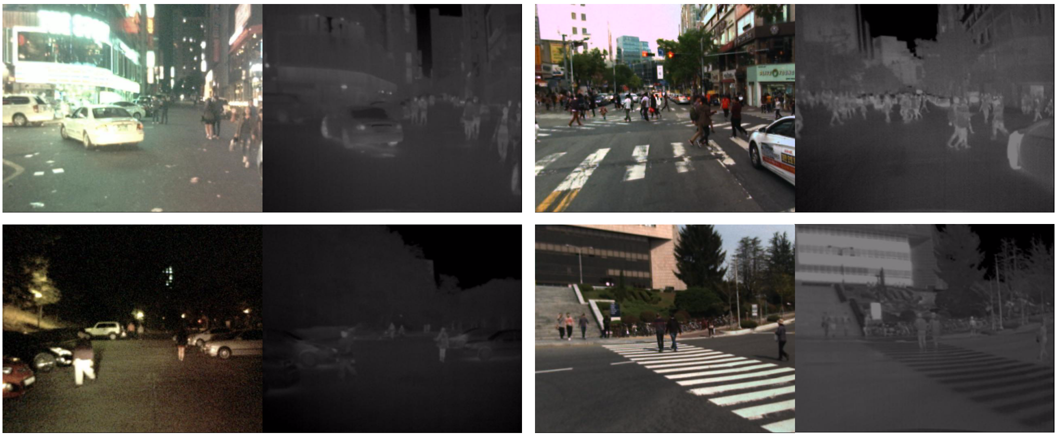

Fig. 6. Example thermal/RGB image pairs from the KAIST dataset.

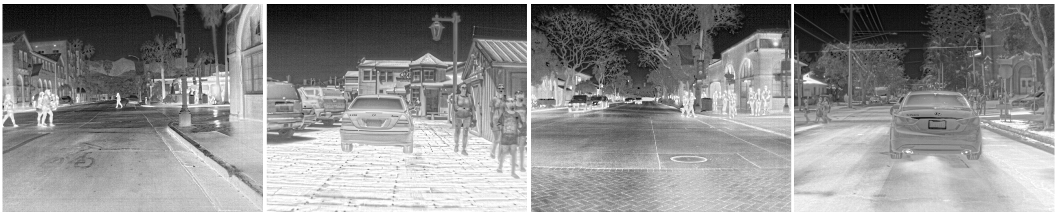

Fig. 7. Examples from FLIR Starter Thermal Dataset

- **The KAIST dataset [16]**: contains 95,328 aligned visible/thermal image pairs in total. The training and test sets include 50,172 and 45,156 image pairs, respectively. Following other methods [22, 26], we sample images every 2 frames from training videos and exclude heavily occluded instances and small instances under 50 pixels. The final training set contains 7,601 training images. The test set contains 2,252 image pairs sampled every 20th frame from videos, of which 797 pairs are captured at night. Figure 6 gives some example thermal/visible image pairs from the KAIST dataset [16].
- **The FLIR dataset [10]**: was released by the thermal camera manufacturer FLIR©. It consists of 10,228 color/thermal image pairs with bounding box annotations for five classes: person, bicycle, car, dog, and other. These images are split into 8,862 for training and 1,366 for testing. However, the color and thermal cameras have different focal lengths and resolutions and are not properly aligned. In order to compare with the state-of-the-art on this dataset, we follow the benchmark procedure described in [7]. We evaluate only on thermal images and three object categories: person (28,151 instances), car (46,692 instances), and bicycle (4,457 instances). Some example images from the FLIR Starter Thermal Dataset are given in figure 7.

4.2 Evaluation metrics

For evaluation, we strictly follow the *reasonable* setting provided by the KAIST benchmark [16]. We used the standard precision and log-average miss rate (MR) evaluation metrics for measuring object detection as defined in [8]. To calculate MR a True Positive (TP) is counted if a detected bounding box is matched to a ground-truth box with an Intersection of Union (IoU) of 50% or greater.

Unmatched detected and ground truth boxes are considered False Positives and False Negatives, respectively. The MR is computed by averaging miss rate (false negative rate) at nine False Positives Per Image (FPPI) rates evenly spaced in log-space.

For consistent comparison with the state-of-the-art, we use mean Average Precision (mAP) on the FLIR dataset, while on the KAIST dataset we plot the MR over false positive per images and precision over recall curves to compare to the state-of-the-art. The results from our previous paper [18] trained on the old annotation and setting from the KAIST baseline [16]. In this paper, we re-experiment all of our methods on the new training annotation, which provided by [22]. The result is significantly improved and comparable with the state-of-the-art. This confirms the critical role of annotation for training the deep neural network.

4.3 Experimental setup

All of our models were implemented in PyTorch and source code and pretrained networks are available.[1] Rather than set apart a fixed validation set, at each epoch we set aside 10% of the training images to use for validation at that epoch. We use mini-batch stochastic gradient descent (SGD) with momentum of 0.9 and do not use the learning rate warm-up strategy of original YOLO model. Training begins with a learning rate of 0.001, and when the training loss no longer decreases and the validation recall no longer improves we decrease the rate by a factor of 10. Training is halted after decreasing the learning rate twice in this way. All models were trained on a GTX 1080 for a maximum of 50 epochs with a batch size of 8. We keep all hyperparameters the same as the original settings of the YOLOv3 model [34].

4.4 Performance on KAIST

The KAIST multispectral pedestrian benchmark is a challenging dataset with both nighttime and daytime video. We divide our comparison between single-modality detectors and multi-modal detectors from the literature.

Comparison with other single-modality methods: Table 1 compares the performance of our approaches with state-of-the-art single-modality approaches (i.e. using only thermal or visible imagery) in terms of miss rate (MR). Our approaches outperform all existing single-modality methods by a large margin in all conditions (day, night, and all). Our layer-wise adaptation obtains the best result with 25.61% MR at all and 10.87% MR at nighttime, improving on the current state-of-the-art by 9.59% at all and 9.63% at night. Our BU(VAT, T) approach reaches 26.26% combined day/night MR and 11.95% MR at nighttime and obtains the second best result compared to the best state-of-the-art of 35.2% MR at all and 20.50% MR at night. Note that we exploit thermal imagery alone for training domain adaptation, while some of the state-of-the-art single-modality methods exploit both color and thermal at training time [13, 40]. Among existing single-modality approaches, our detectors are the best on the KAIST dataset in all conditions.

Comparison with both single- and multi-modality approaches: Table 2 compares our approaches and more than fourteen other single- and multi-modal from the literature. The last two columns indicate the type of imagery used at test time. The first group contains results for multispectral detectors using both visible and thermal imagery for training and testing. The second group contains single-modality approaches, and our results are in the last group.

We draw a number of conclusions from these results. First of all, from the MR combined results (all), we note that multimodal techniques like MSDS [22] or IATDNN+IAMSS [13] are superior to our domain adaptation approaches. This seems to be due to the advantage they have when

[1] https://github.com/mrkieumy/YOLOv3_PyTorch

Table 1. Comparison with state-of-the-art **single-modality approaches** in term of **log-average Miss Rate on KAIST dataset (lower is better)**. Our approaches outperform all others in all conditions (day/night/all).

Detector	**all**	**day**	**night**	**test images**
Color plotted [22]	45.70	36.00	68.30	RGB
RRN+MDN [40]	49.55	47.30	54.78	RGB
Thermal plotted [22]	35.70	40.40	25.20	thermal
TPIHOG [3]	-	-	57.38	thermal
SSD300 [15]	69.81	-	-	thermal
Saliency Maps [12]	-	30.40	21.00	thermal
Bottom-up [18]	35.20	40.00	20.50	thermal
Ours: TD(V, V)	34.75	**29.77**	46.25	RGB
Ours: TD(T, T)	31.10	37.30	16.70	thermal
Ours: TD(VT, T)	30.67	37.42	15.45	thermal
Ours: BU(VAT, T)	26.26	32.84	11.95	thermal
Ours: BU(VLT, T)	**25.61**	32.69	**10.87**	thermal

detecting during the day and thus exploiting visible imagery. Secondly, our domain adaptation approaches, both top-down and bottom-up, outperform all other single-modality techniques and many multimodal techniques. Thirdly, looking at the nighttime results, our bottom-up domain adaptation BU(VLT, V) is the best result with 10.87% MR. This surpasses the all state-of-the-art approaches in both single- and multi-modal detection. This demonstrates the potential of our domain adaptation methods to capture useful information from RGB detectors and adapt them to nighttime.

Of particular note is the fact that performing domain adaptation on visible images before adapting to thermal input is beneficial. This can be seen in the difference between BU(VAT, T) and BU(VLT, T) – both of which start by fine-tuning TD(V, V) on KAIST visible images – and TD(T, T), which directly fine-tunes YOLOv3 on thermal images. This seems to indicate that both bottom-up domain adaptation approaches can retain and exploit domain knowledge acquired when training the detector on visible spectrum imagery. Notably, slow layer-wise adaptation, BU(VLT, T), helps robust pedestrian detection at night and outperforms other bottom-up methods.

The plots in figure 8 provide a more detailed picture of our approaches and the state-of-the-art in terms of precision/recall (left column) and log-average miss rate (right column). The plots also break down results in terms of time-of-day: the first row averaged over day and night, the second row daytime only, and the third row nighttime only. The Intersection of Union (IoU) used is the standard 0.5. The results in the plot are slightly different those originally published because: (1) the authors of [22] said that the number in their official article is calculated by the average of 5 runs; and (2) the authors in [38] used the overlap IoU 0.4 because they said YOLO had trouble with small objects. All results are are generated using the framework provided by the authors [38]. The results reported in the original papers are given in table 2.

The results from figure 8 show that the ranking is similar to that reported in table 2. We are in the top three results during the day and combined (all). We are also the best results at night. Note that the MSDS result plotted here is higher than numbers in their published paper. Importantly, our domain adaptation approaches, both top-down and bottom-up adaptation, surpass all other methods at nighttime and are comparable with the other two multispectral methods for combined day/night (all).

Table 2. **Log-average Miss Rate on KAIST dataset (lower is better)**. The final two columns indicate which image modality is used at *test time*. Our approaches outperform all single-modality techniques from the literature, and outperform all methods at night.

Method		MR all	MR day	MR night	Visible	Thermal
KAIST baseline	[16]	64.76	64.17	63.99	✓	✓
Late Fusion	[39]	43.80	46.15	37.00	✓	✓
Halfway Fusion	[26]	36.99	36.84	35.49	✓	✓
RPN+BDT	[19]	29.83	30.51	27.62	✓	✓
IATDNN+IAMSS	[13]	26.37	27.29	24.41	✓	✓
IAF R-CNN	[23]	15.73	14.55	18.26	✓	✓
MSDS	[22]	**11.63**	**10.60**	13.73	✓	✓
YOLO_TLV	[38]	31.20	35.10	22.70	✓	✓
DSSD-HC	[21]	34.32	-	-	✓	✓
GFD-SSD	[43]	28.00	25.80	30.03	✓	✓
RRN+MDN	[40]	49.55	47.3	54.78	✓	
TPIHOG	[3]	-	-	57.38		✓
SSD300	[15]	69.81	-	-		✓
Bottom-up	[18]	35.20	40.00	20.50		✓
Ours: TD(V, V)		34.75	29.77	46.25	✓	
Ours: TD(T, T)		31.06	37.34	16.69		✓
Ours: TD(VT, T)		30.67	37.42	15.45		✓
Ours: BU(VAT, T)		26.26	32.84	11.95		✓
Ours: BU(VLT, T)		25.61	32.69	**10.87**		✓

Table 3. Comparative performance analysis on the FLIR dataset.

Method	Bicycle	Person	Car	mAP
Baseline	39.7	54.7	67.6	54.0
MMTOD-UNIT [7]	49.4	64.5	70.8	61.5
Our: TD(T,T)	51.9	75.5	86.9	71.4
Our: BU(AT,T)	56.1	76.1	**87.0**	73.1
Our: BU(LT,T)	57.4	75.6	86.5	73.2

4.5 Performance on FLIR

Table 3 compares our approaches with state-of-the-art on the FLIR Starter Thermal Dataset [10]. Results on this dataset are measured using average precision (AP) for each class and the mean Average Precision (mAP) over all classes. Note that the FLIR dataset has five categories, but the baseline and the state-of-the-art results reported n only three of these: person, car and bicycle. From these results we see that our approaches, both top-down and bottom-up, outperform the baseline and the state-of-the-art on all classes and in overall mAP. Our layer-wise adaptation is the best result with 73.2% mAP, improving on the current state-of-the-art by 11.7% mAP. Our bottom-up approach BU(AT,T) also obtains 87.0% precision on cars, advancing the current state-of-the-art by 16.2% average precision.

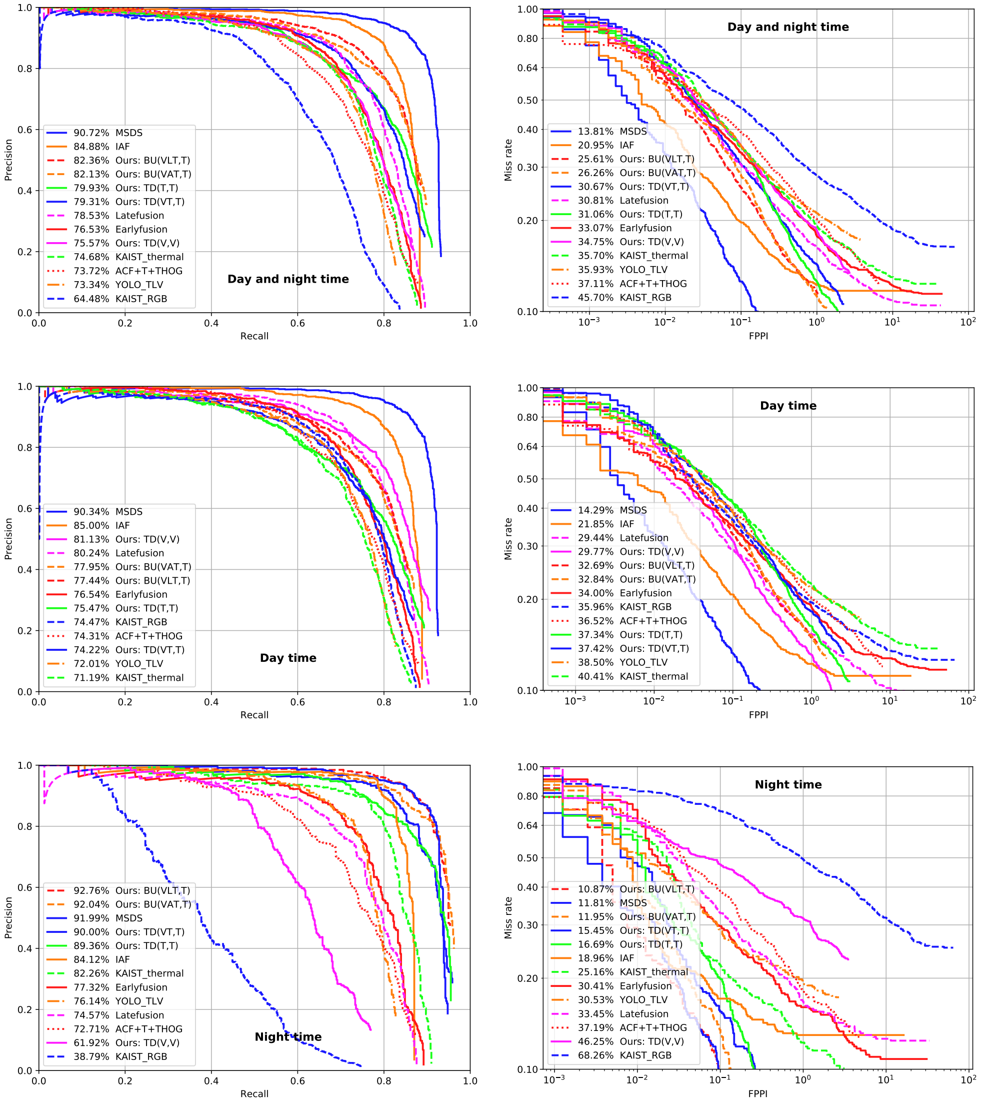

Fig. 8. **Comparative performance analysis**. Precision/Recall (left, higher is better) and Log-average Miss Rate (right, lower is better) of our method and other state-of-the-art papers are given. See text for detailed analysis.

4.6 Qualitative analysis of detector adaptation

In figure 9 we plot the average gradient magnitudes of every layer of the network during the first epoch of fine-tuning for three different adaptation methods: top-down, bottom-up and layer-wise. These plots show the capacity of the methods to adapt network weights during adaptation to the new domain. We see from these plots that the gradient magnitudes for layer-wise adaptation are highest for all layers – especially for the *early* convolutional layers where the network must adapt the most to the new input domain. The gradient magnitudes for bottom-up adaptation are

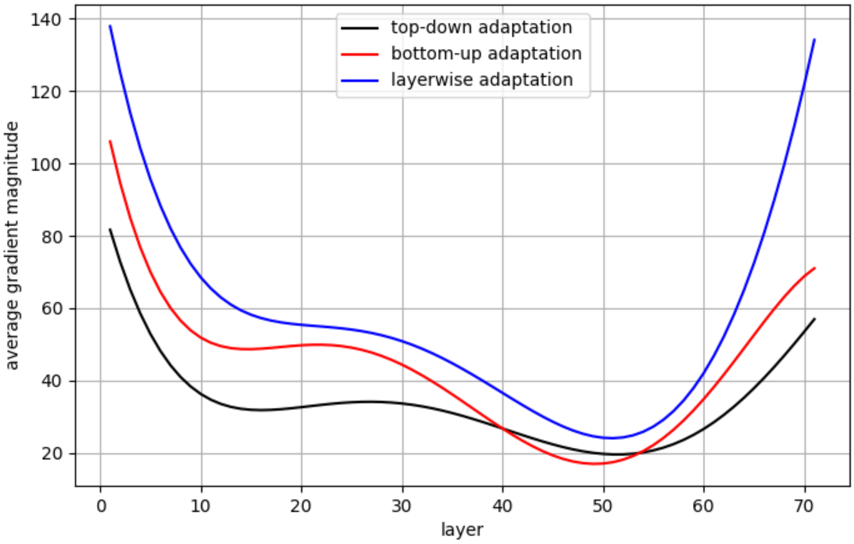

Fig. 9. Average gradient magnitudes per layer for each adaptation method during first epoch of fine-tuning.

also larger than simple fine-tuning (top-down adaptation), which illustrates the positive effect the adapter network has on domain adaptation.

In order to better understand how layer-wise adaptation improves internal feature representations for detection in thermal images we used the Gradient weighted Class Activation Map (Grad-CAM) [35] visualization technique on input images from the KAIST and FLIR datasets for all three adaptation approaches (top-down, bottom-up, and layer-wise). These visualizations are shown in figure 10. Grad-CAM heatmaps were computed at layer 52 of the adapted YOLOv3 using the backpropagated loss from the medium-scale detection head. We see in these visualizations that the layer-wise network has learned to attend to more areas of the image salient to pedestrian detection compared to the bottom-up and top-down adapted networks. This explains how layer-wise adaptation leads to more correct positive detections on average.

Figure 11 shows detection results on three images by two methods. The first row gives results of top-down adaptation (TD(VT,T)), and the second row results of bottom-up adaptation (BU(VAT,T)) on the same images from the KAIST dataset.

As we can see on the figure 11, *bottom-up adaptation* results in more True Positive bounding box detections than *top-down* (in the first and the second images). Similarly, *Top-down adaptation* results in more False Positive detections than the *bottom-up adaptation* on the last image. This is consistent with with our experimental results that show *bottom-up adaptation* is superior to *top-down adaptation*. We believe this is because bottom-up adaptation preserves more knowledge which has been learned from the visible domain before adapting to the thermal domain. Moreover, we note that the *bottom-up adaptation* approach requires significantly less training time than *top-down adaptation*. In only 15 epochs it converges to about 82.13% precision, which is a higher than top-down adaptation after 50 epochs. *Bottom-up adaptation* seems to be an effective way to accelerate *top-down adaptation* through fine-tuning. Last but not least, even though person identification is impossible in thermal images, thermal imagery retains enough information to

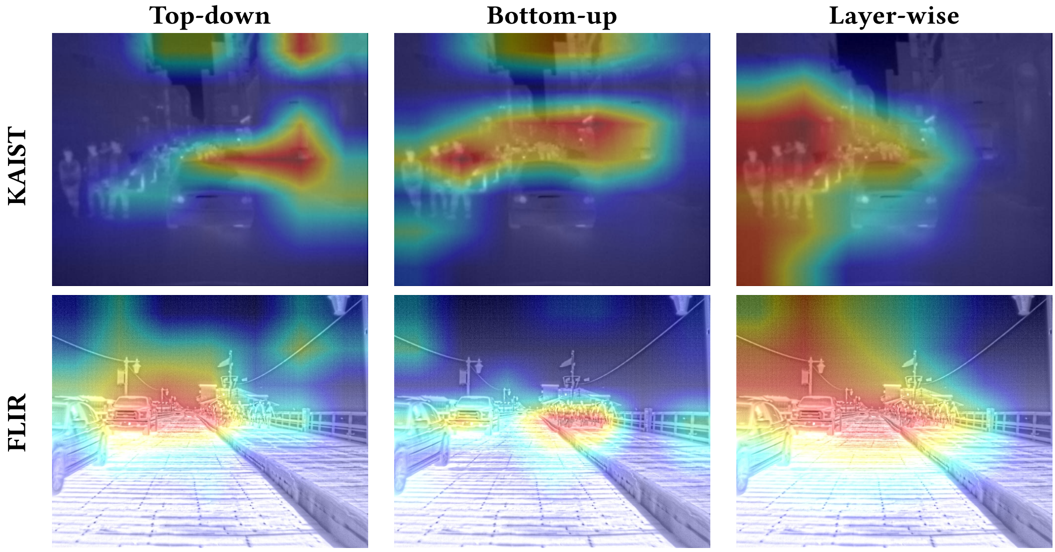

Fig. 10. Grad-CAM visualization of feature importance for three adaptation methods. Red areas indicate which parts of the image contribute most to the features used in the detection heads of the adapted networks. Note how feature importance for layer-wise adaptation is spread across more pedestrians compared to the other two adaptation methods.

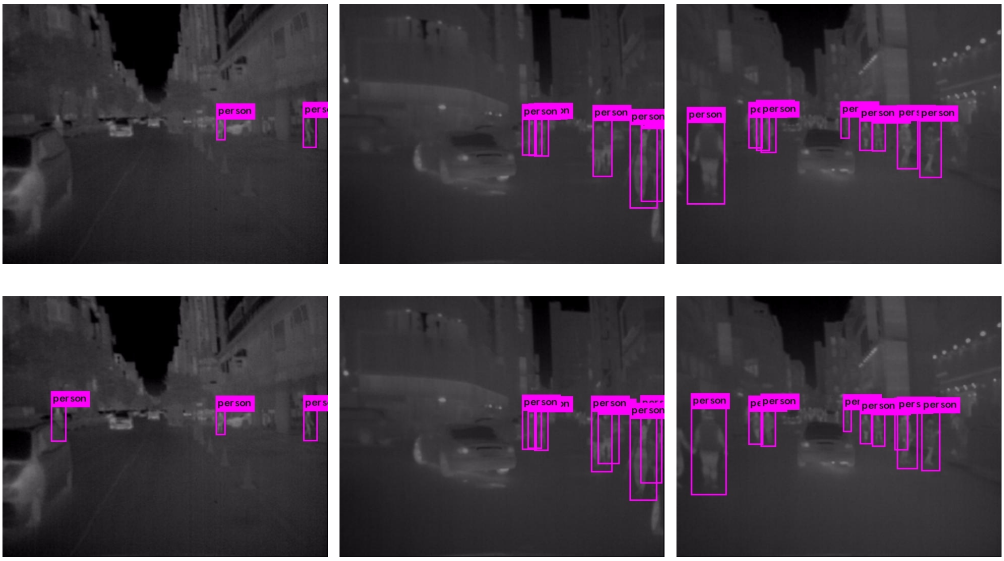

Fig. 11. The first row gives three frames of detections resulting from top-down domain adaptation. The second row gives results of bottom-up adaptation on the same frames. Even though person identification is impossible in all of images, thermal imagery can retain distinctive image features to effectively perform pedestrian detection, and our *bottom-up adaptation* made more True Positive and less False Positive detection result than *top-down adaptation.*

effectively perform pedestrian detection in a privacy-preserving way without using any visible spectrum imagery at detection time.

Looking at our results on KAIST in the last row of table 1 and the table 2, we make some observations:

- Firstly, the technique exploiting visible spectrum images during the day outperforms all our approaches which only use thermal imagery. This is expected as the daytime images are similar to visible images than thermal images.
- The methods TD(VT, T), BU(VAT, T), and BU(VLT, T), which start from TD(V, V), surpass TD(T, T). This shows that the first adaptation on the visible domain helps the network adapt better to the final, thermal target domain. This opens an opportunity to leverage transfer learning from other datasets for robust pedestrian detection.
- Our best thermal detection networks work extremely well at night, but only modestly well on daytime imagery. We believe that daytime and nighttime detection require different features and filters. Thus, there is still plenty of opportunity for improvement in thermal-only detection results.
- Moreover, as we can see our two bottom-up results in tables 1, 2, and 3), The layer-wise adaptation BU(VLT, T) is always superior to BU(VAT, T) on both datasets. This seems to indicate that a slowly adapting from the bottom of the network better preserves visible feature, which helps maintain robust detection at night.

5 CONCLUSIONS

In this paper, we described the potential of three domain adaptation approaches for pedestrian detection tasks, and we also proposed two effective bottom-up domain adaptation strategies for pedestrian detection in thermal imagery. The goal of our research is to close the performance gap between pedestrian detection exploiting only thermal imagery and multispectral approaches using both visible and thermal images for training and testing.

The results on two datasets show that our relatively simple domain adaptation schemes are effective, and our results outperform all state-of-the-art single-modality methods on two datasets. Exploiting only on thermal domain, our detectors perform comparably with the state-of-the-art and outperform many multispectral approaches on KAIST. Furthermore, the results reveal that a preliminary adaptation to visible spectrum images is useful to acquire domain knowledge that can be exploited after the final adaptation to the thermal domain. As far as we know, ours are the best detectors in thermal imagery on both KAIST and FLIR datasets.

Closing the gap between multispectral pedestrian detectors and single modality detectors in thermal imagery is a non-trivial task. There is still an enormous potential to improve pedestrian detection results by using multispectral data to exploit only one domain. In particular, balancing the results between daytime and nighttime is crucial for single-modality models. Thermal imagery is inherently privacy-preserving, and we believe that thermal-only pedestrian detection has great potential for the future if this balance can be found and the gap between single- and multi-modal detection closed.

REFERENCES

[1] Federico Angelini, Jiawei Yan, and Syed Naqvi. 2019. Privacy-preserving Online Human Behaviour Anomaly Detection Based on Body Movements and Objects Positions. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 8444–8448. https://doi.org/10.1109/ICASSP.2019.8683026

[2] Anelia Angelova, Alex Krizhevsky, Vincent Vanhoucke, Abhijit Ogale, and Dave Ferguson. 2015. Real-time pedestrian detection with deep network cascades. In *Proc. of British Machine Vision Conference (BMVC)*. Article 32, 12 pages.

[3] Jeonghyun Baek, Sungjun Hong, Jisu Kim, and Euntai Kim. 2017. Efficient pedestrian detection at nighttime using a thermal camera. *Sensors* 17, 8 (2017), 1850. https://doi.org/10.3390/s17081850

[4] Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. 2014. Ten years of pedestrian detection, what have we learned?. In *Proc. of European Conference on Computer Vision (ECCV)*, Vol. 8926. Springer, 613–627.

[5] Garrick Brazil, Xi Yin, and Xiaoming Liu. 2017. Illuminating pedestrians via simultaneous detection & segmentation. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*. 4950–4959.

[6] Yanpeng Cao, Dayan Guan, Yulun Wu, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. 2019. Box-level segmentation supervised deep neural networks for accurate and real-time multispectral pedestrian detection. *ISPRS Journal of Photogrammetry and Remote Sensing* 150 (2019), 70–79.

[7] Chaitanya Devaguptapu, Ninad Akolekar, Manuj M Sharma, and Vineeth N Balasubramanian. 2019. Borrow from Anywhere: Pseudo Multi-modal Object Detection in Thermal Imagery. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 1029–1038. https://doi.org/10.1109/CVPRW.2019.00135

[8] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. 2011. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 4 (2011), 743–761.

[9] Xianzhi Du, Mostafa El-Khamy, Jungwon Lee, and Larry Davis. 2017. Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection. In *Proc. of IEEE Winter Conference on Applications of Computer Vision (WACV)*, Vol. abs/1610.03466. IEEE, 953–961.

[10] FLIR. 2018. FLIR Starter Thermal Dataset. (2018). https://www.flir.com/oem/adas/adas-dataset-form/

[11] Kevin Fritz, Daniel König, Ulrich Klauck, and Michael Teutsch. 2019. Generalization ability of region proposal networks for multispectral person detection. In *Proc. of Automatic Target Recognition XXIX*, Vol. 10988. International Society for Optics and Photonics, 109880Y.

[12] Debasmita Ghose, Shasvat M Desai, Sneha Bhattacharya, Deep Chakraborty, Madalina Fiterau, and Tauhidur Rahman. 2019. Pedestrian Detection in Thermal Images using Saliency Maps. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Vol. abs/1904.06859.

[13] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. 2019. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion* 50 (2019), 148–157.

[14] Alon Hazan, Yoel Shoshan, Daniel Khapun, Roy Aladjem, and Vadim Ratner. 2018. AdapterNet - learning input transformation for domain adaptation. *CoRR* abs/1805.11601 (2018). arXiv:1805.11601 http://arxiv.org/abs/1805.11601

[15] Christian Herrmann, Miriam Ruf, and Jürgen Beyerer. 2018. CNN-based thermal infrared person detection by domain adaptation. In *Proc. of Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything*, Vol. 10643. International Society for Optics and Photonics, 1064308.

[16] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and Inso Kweon. 2015. Multispectral Pedestrian Detection: Benchmark Dataset and Baseline. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 20.

[17] Vijay John, Seiichi Mita, Zheng Liu, and Bin Qi. 2015. Pedestrian detection in thermal images using adaptive fuzzy C-means clustering and convolutional neural networks. In *Proc. of IAPR International Conference on Machine Vision Applications (MVA)*. IEEE, 246–249.

[18] My Kieu, Andrew D Bagdanov, Marco Bertini, and Alberto Del Bimbo. 2019. Domain Adaptation for Privacy-Preserving Pedestrian Detection in Thermal Imagery. In *Proc. of International Conference on Image Analysis and Processing (ICIAP)*. Springer, 203–213.

[19] Daniel Konig, Michael Adam, Christian Jarvers, Georg Layher, Heiko Neumann, and Michael Teutsch. 2017. Fully convolutional region proposal networks for multispectral person detection. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 49–56. https://doi.org/10.1109/CVPRW.2017.36

[20] Wouter M. Kouw. 2018. An introduction to domain adaptation and transfer learning. *CoRR* abs/1812.11806 (2018). arXiv:1812.11806 http://arxiv.org/abs/1812.11806

[21] Yongwoo Lee, Toan Duc Bui, and Jitae Shin. 2018. Pedestrian Detection based on Deep Fusion Network using Feature Correlation. In *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 694–699.

[22] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. 2018. Multispectral Pedestrian Detection via Simultaneous Detection and Segmentation. In *Proc. of British Machine Vision Conference (BMVC)*. BMVA Press, 225. http://bmvc2018.org/contents/papers/0738.pdf

[23] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. 2019. Illumination-aware faster R-CNN for robust multispectral pedestrian detection. *Pattern Recognition* 85 (2019), 161–171.

[24] Jianan Li, Xiaodan Liang, ShengMei Shen, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. 2017. Scale-aware fast R-CNN for pedestrian detection. *IEEE Transactions on Multimedia* 20, 4 (2017), 985–996.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proc. of European Conference on Computer Vision (ECCV)*. Springer, 740–755.

[26] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N. Metaxas. 2016. Multispectral Deep Neural Networks for Pedestrian Detection. In *Proc. of British Machine Vision Conference (BMVC)*. BMVA Press. http://www.bmva.org/bmvc/2016/papers/paper073/index.html

[27] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. 2019. High-level Semantic Feature Detection: A New Perspective for Pedestrian Detection. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5187–5196.

[28] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. 2015. Learning Transferable Features with Deep Adaptation Networks. In *Proc. of International Conference on Machine Learning (ICML) (ICML'15)*. JMLR.org, 97–105.

[29] IHS Markit. 2019. 245 million video surveillance cameras installed globally in 2014. Web page. https://technology.ihs.com/532501/245-million-video-surveillance-cameras-installed-globally-in-2014 Accessed: May 5, 2019.

[30] Marc Masana, Joost van de Weijer, Luis Herranz, Andrew D Bagdanov, and Jose M Alvarez. 2017. Domain-adaptive deep network compression. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*. 4289–4297.

[31] Shota Nakashima, Yuhki Kitazono, Lifeng Zhang, and Seiichi Serikawa. 2010. Development of privacy-preserving sensor for person detection. *Procedia - Social and Behavioral Sciences* 2 (12 2010), 213–217.

[32] Wanli Ouyang, Xingyu Zeng, and Xiaogang Wang. 2016. Learning mutual visibility relationship for pedestrian detection with a deep model. *International Journal of Computer Vision* 120, 1 (2016), 14–27.

[33] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: better, faster, stronger. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7263–7271. https://doi.org/10.1109/CVPR.2017.690

[34] Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An Incremental Improvement. *CoRR* abs/1804.02767 (2018). arXiv:1804.02767 http://arxiv.org/abs/1804.02767

[35] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*. 618–626.

[36] Yunfei Teng and Anna Choromanska. 2019. Invertible Autoencoder for Domain Adaptation. *Computation* 7, 2 (2019), 20. https://doi.org/10.3390/computation7020020

[37] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Pedestrian detection aided by deep learning semantic tasks. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5079–5087.

[38] Maarten Vandersteegen, Kristof Van Beeck, and Toon Goedemé. 2018. Real-time multispectral pedestrian detection with a single-pass deep neural network. In *Proc. of International Conference Image Analysis and Recognition (ICIAR)*. Springer, 419–426.

[39] Jörg Wagner, Volker Fischer, Michael Herman, and Sven Behnke. 2016. Multispectral Pedestrian Detection using Deep Fusion Convolutional Neural Networks. In *Proc. of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*. 509–514.

[40] Dan Xu, Wanli Ouyang, Elisa Ricci, Xiaogang Wang, and Nicu Sebe. 2017. Learning cross-modal deep representations for robust pedestrian detection. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5363–5371.

[41] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. 2016. Is faster r-cnn doing well for pedestrian detection?. In *Proc. of European Conference on Computer Vision (ECCV)*, Vol. abs/1607.07032. Springer, 443–457.

[42] Lu Zhang, Zhiyong Liu, Xiangyu Chen, and Xu Yang. 2019. The Cross-Modality Disparity Problem in Multispectral Pedestrian Detection. *CoRR* abs/1901.02645 (2019). arXiv:1901.02645 http://arxiv.org/abs/1901.02645

[43] Yang Zheng, Izzat H. Izzat, and Shahrzad Ziaee. 2019. GFD-SSD: Gated Fusion Double SSD for Multispectral Pedestrian Detection. *CoRR* abs/1903.06999 (2019). arXiv:1903.06999 http://arxiv.org/abs/1903.06999