Task-conditioned Domain Adaptation for Pedestrian Detection in Thermal Imagery

My Kieu[0000−0002−7813−5744], Andrew D. Bagdanov[0000−0001−6408−7043], Marco Bertini[0000−0002−1364−218X], and Alberto del Bimbo[0000−0002−1052−8322]

Media Integration and Communication Center - University of Florence, Italy
{firstname.lastname}@unifi.it

Abstract. Pedestrian detection is a core problem in computer vision that sees broad application in video surveillance and, more recently, in advanced driving assistance systems. Despite its broad application and interest, it remains a challenging problem in part due to the vast range of conditions under which it must be robust. Pedestrian detection at nighttime and during adverse weather conditions is particularly challenging, which is one of the reasons why thermal and multispectral approaches have been become popular in recent years. In this paper, we propose a novel approach to domain adaptation that significantly improves pedestrian detection performance in the thermal domain. The key idea behind our technique is to adapt an RGB-trained detection network to simultaneously solve two related tasks. An auxiliary classification task that distinguishes between daytime and nighttime thermal images is added to the main detection task during domain adaptation. The internal representation learned to perform this classification task is used to condition a YOLOv3 detector at multiple points in order to improve its adaptation to the thermal domain. We validate the effectiveness of task-conditioned domain adaptation by comparing with the state-of-the-art on the KAIST Multispectral Pedestrian Detection Benchmark. To the best of our knowledge, our proposed task-conditioned approach achieves the best single-modality detection results.

Keywords: object detection, pedestrian detection, thermal imagery, task-conditioned, domain adaptation, conditioning network, thermal imagery

1 Introduction

Object detection and, in particular, pedestrian detection is one of the most important problems in computer vision due to its central role in diverse practical applications such as safety and security, surveillance, and autonomous driving. The detection problem is particularly challenging in many common contexts such as limited illumination (nighttime) or adverse weather conditions (fog, rain, dust) [22, 19]. In such conditions the majority of detectors [4, 27, 40] using visible spectrum imagery can fail.

For these reasons, detectors exploiting thermal imagery have been proposed as suitable for robust pedestrian detection [19, 38, 25, 20, 5, 22, 23, 14]. A

growing number of works have also investigated multispectral detectors that combine visible and thermal images for robust pedestrian detection [36, 1, 29, 24, 38, 25, 20, 39, 5, 14, 22, 23].

However, multispectral detectors, in order to make the most out of both modalities, typically need to resort to additional (and expensive) annotations, and are usually based on far more complex network architectures than single-modality methods (see table 3). Moreover, due to the cost of deploying multiple aligned sensors (thermal and visible) at inference time, multispectral models can have limited applicability in real-world applications. Aside from the technical and economic reasons, the privacy-preserving affordances offered by thermal imagery are also a motivation for prefering thermal-only detecion [19]. Because of this, several recent works do not use visible images, but focus only on thermal images for pedestrian detection [18, 16, 3, 7, 19, 15]. They typically yield lower performance than multispectral detectors since robust pedestrian detection using only thermal data is nontrivial and there is still potential for improvement.

In this paper we propose a task-conditioned network architecture for domain adaptation of pedestrian detectors to thermal imagery. Our key idea is to augment a detector with an auxiliary network that solves a simpler classification task and then to exploit the learned representation of this auxiliary network to inject conditioning parameters into strategically chosen convolutional layers of the main detection network. The resulting, adapted network operates entirely in the thermal domain and achieves excellent performance compared to other single-modality approaches.

The contributions of this work are:

- we propose a novel task-conditioned network architecture based on YOLOv3 [32] that uses the auxiliary task of day/night classification to aid adaptation to the thermal domain;
- we conduct extensive ablative analyses probing the effectiveness of various task-conditioning architectures and adaptation schedules;
- to the best of our knowledge, our task-conditioned detection networks outperform all single-modality detection approaches the KAIST Multispectral Pedestrian Detection Benchmark [17]; and
- exploiting only thermal imagery, we outperform many state-of-the-art multispectral pedestrian detectors on the KAIST benchmark at nighttime.

The rest of the paper is organized as follows. In the next section we review the scientific literature related to our proposed domain adaptation approach. In section 3 we describe our approach to conditioning thermal domain adaptation on the auxiliary task of day/night discrimination. We report in section 4 on an extensive set of experiments performed to evaluate the effectiveness of task-conditioning, and in section 5 we conclude with a discussion of our contribution.

2 Related work

Pedestrian detection has attracted much attention from the scientific community over the years because of its usefulness in many applications. Thanks to the

reduction of costs and availability of thermal cameras, many recent works have investigated how to perform it in multispectral and thermal domains.

2.1 Pedestrian detection in the visible spectrum

The main challenges to robust pedestrian detection in the visible spectrum arise from a variety of environmental conditions such as occlusion, changing illumination, and variation of viewpoint and background [29]. In [36] discriminative detectors are learned by jointly optimizing them along with semantic tasks such as pedestrian and scene attributes detection; in [29] joint estimation of visibility of multiple pedestrians and recognition of overlapping pedestrians is done using a mutual visibility deep model; in [5] semantic segmentation is used as an additional supervision to improve the simultaneous detection. In [40] the Region Proposal Network (RPN) originally proposed in Faster R-CNN is used for standalone pedestrian detection; dealing with multiple scales using a specialized sub-networks based on Fast R-CNN is proposed in [24]; prediction of pedestrian centers and scales in one pass and without anchors was recently proposed in [27].

2.2 Multispectral pedestrian detection approaches

Many recent works have used both thermal and RGB images to improve detection results [38, 25, 20, 39, 22, 23], combining visible and thermal images for training and testing. The authors of [38] investigated two types of fusion networks to exploit visible and thermal image pairs. Four different network fusion approaches (early, halfway, late, and score fusion) for the multispectral pedestrian detection task were also introduced in [25]. The cross-modality learning framework including a Region Reconstruction Network (RRN) and Multi-Scale Detection Network (MDN) of [39] used thermal image features to improve detection results in visible data.

Because the combination of visible and thermal images works well in two-stage network architectures, most of top-performing multispectral pedestrian detection are based on the approach originally used in Fast-/Faster R-CNNs. For instance, the Faster R-CNN detector was used to perform multispectral pedestrian detection in Illumination-aware Faster R-CNN (IAF R-CNN) [23]. The authors in [20] detected persons in multispectral video with a combination of a Fully Convolutional RPN and a Boosted Decision Trees Classifier (BDT). The generalization ability of RPN was also investigated in [10], evaluating which multispectral dataset results in better generalization. MSDS-RCNN [22] is a fusion of a multispectral proposal network (MPN) and a multispectral classification network (MCN). In [41] an Aligned Region CNN is proposed to deal with weakly aligned multispectral data. Box-level segmentation via a supervised learning framework was proposed in [6], eliminating the need of anchor boxes.

Approaches based on one-stage detectors have also been investigated. The authors in [37] used YOLOv2 [32] as a fast single-pass network architecture for multispectral detection. A deconvolutional single-shot multi-box detector (DSSD) was also leveraged by authors in [21] to exploit the correlation between

visible and thermal features. The work in [43] adopted two Single Shot Detectors (SSDs) to investigate the potential of fusing color and thermal features with Gated Fusion Units (GFU).

2.3 Pedestrian detection in thermal imagery.

A few works have addressed pedestrian detection using thermal (IR) imagery only. Adaptive fuzzy C-means for IR image segmentation and CNN for pedestrians detection were proposed in [18]. A combination of Thermal Position Intensity Histogram of Oriented Gradients (TPIHOG) and the additive kernel SVM (AKSVM) was proposed by [3] for nighttime-only detection in thermal imagery. Thermal images augmented with saliency maps used as attention mechanism have been used to train a Faster R-CNN detector in [12]. In [16] several video preprocessing steps are performed to make thermal images look more similar to grayscale images converted from RGB, then a pre-trained and fine-tuned SSD detector is used. Recently, the authors in [7] used Cycle-GAN for image-to-image translation of thermal to pseudo-RGB data, using it to fine-tune to a multimodal Faster-RCNN detector. Instead, the authors in [15] used a GAN to transform visible images to synthetic thermal images, as a data augmentation processing to train a pedestrian detector to work on thermal-only imagery. Another recent work dealing with domain adaptation is the Top-down and Bottom-up Domain Adaptation approaches proposed in [19] for pedestrian detection in thermal imagery. In this work, bottom-up adaptation obtains state-of-the-art single-modality results at nighttime on KAIST dataset [17].

2.4 Task-conditioned networks

There are a few task-conditioning approaches, such as conditional generative models like those based on adversarial networks [28], and the seminal work in [31] that proposed architecture guidelines for training Deep Convolutional GANs. In particular, our approach is inspired by the general conditioning layer called Feature-wise Linear Modulation (FiLM) proposed in [30] for conditioning visual reasoning tasks.

In this paper we perform pedestrian detection on thermal imagery only. Our method is based on the single-stage detector YOLOv3 [33], whose computational efficiency makes it particularly well-suited to practical applications with real-time requirements. We extend the YOLOv3 architecture by integrating conditioning layers to better specialize the network to deal with day- and nighttime images. We evaluate conditioning of residual groups, detection heads, and their combination during domain adaptation.

3 Task-conditioned domain adaptation

In this section we describe our approach to conditioning a detector during adaptation to the thermal domain. Our central idea is that robust pedestrian detection naturally depends on low-level semantic qualities of input images – for

example whether an image is captured during the day or at night. This auxiliary information should be useful for learning representations upon which we can condition the adaptation internal representations used for the primary detection task. In the next section we describe the architecture of an auxiliary classification network that is connected to the main detection network, and in section 3.2 we describe the conditioning layers that can be strategically inserted into the network to modify internal representation. We describe two alternative conditioning architectures for YOLOv3 in section 3.3, and in section 3.4 we put everything together into a description of the combined adaptation loss.

3.1 Auxiliary classification network

Let $D_{\Theta_d}(\mathbf{x})$ represent the detector network (YOLOv3 in our case) parameterized by Θ_d, and let $F_i(\mathbf{x})$ represent the output of the i^{th} convolutional layer of the detection network. We define an auxiliary classification network as follows. The output of an early convolutional layer (e.g. $F_4(\mathbf{x})$ as in Fig. 1), is average pooled to form a feature that is then fed to two fully-connected layers of size C with ReLU activations. The resulting feature representation is then passed to a final fully connected layer with a single output and a sigmoid activation. We denote the output of this auxiliary network $A_{\Theta_a}(\mathbf{x})$.

During training we use the following loss attached to the output of the auxiliary network:

$$\mathcal{L}_a(\mathbf{x}_i, y_i; \Theta_a) = \left[y_i \cdot \log f(x_i) + (1 - y_i) \cdot \log(1 - f(x_i)) \right], \tag{1}$$

where for all training images $\mathbf{x}_i$ we associate an auxiliary training label y_i. Since we experiment on the KAIST dataset, which distinguishes daytime and nighttime images in its annotations and evaluation protocol, we define $y_i = 0$ if $\mathbf{x}_i$ was captured during the day, and $y_i = 1$ if $\mathbf{x}_i$ was captured at night. In this case the auxiliary network has the task of classifying images as daytime or nighttime.

3.2 Conditioning layers

Our idea to use the internal, C-dimensional representation learned in the auxiliary classification network (i.e. the representation after the two fully-connected layers used for classification) rather than its output. See Figure 1 for a schematic representation of the conditioning process. This representation is task-specific: in our experiments it is learned to capture the salient information *useful* for determining whether an image was captured during the day or at night. At strategic points in the main detection network we will use this representation to generate *conditioning parameters* that condition a convolutional feature map using the representation learned by the auxiliary network.

Consider an arbitrary convolutional output $F_i(\mathbf{x})$ of the main detector network D_{Θ_d}, and let d_i be the number of convolutional feature maps in $F_i(\mathbf{x})$. We generate conditioning parameters γ_i and β_i:

$$\gamma_i = \text{ReLU}[W_\gamma^i A(\mathbf{x}) + b_\gamma^i]$$
$$\beta_i = \text{ReLU}[W_\beta^i A(\mathbf{x}) + b_\beta^i],$$

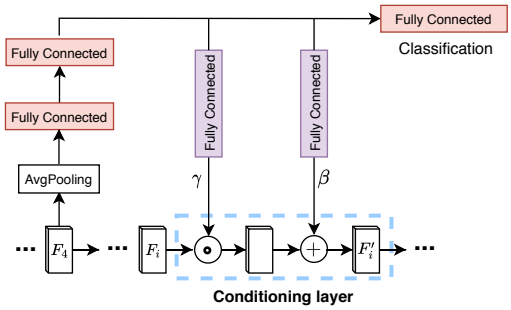

Fig. 1. Conditioning layer and auxiliary classification network. The auxiliary network learns an internal representation used to solve a classification task. This representation is then leveraged by conditioning layers to adjust internal convolutional feature maps in the detection network.

where $W^i_\gamma, W^i_\beta \in \mathbb{R}^{d_i \times C}$ and $b^i_\gamma, b^i_\beta \in \mathbb{R}^{d_i}$ are the weights and biases, respectively, of two new fully connected layers of D units added to the network (purple layers in Fig. 1). These new layers are responsible for generating the parameters used for conditioning F_i.

F_i is substituted by the conditioned version:

$$F'_i(\mathbf{x}) = \text{ReLU}[(1 - \gamma_i) \odot F_i(\mathbf{x}) \oplus \beta_i],$$

where $\odot$ and $\oplus$ are, respectively, the elementwise multiplication and addition operations *broadcasted* to cover the spatial dimensions of the feature maps $F_i(\mathbf{x})$. In this way, the generated γ_i parameters can scale feature maps independently and the β_i parameters independently translate them.

3.3 Conditioned network architectures

YOLOv3 is a very deep detection network with three detection heads for detecting objects at different scales [33]. In order to investigate the effectiveness of conditioning YOLOv3 during domain adaptation, we experimented with two different strategies for injecting conditioning layers into the network. In section 4.3 we report on a series of ablation experiments performed to evaluate these different architectural possibilities for conditioning the network.

Conditioning residual groups (TC Res Group). YOLOv3 uses a 52-layer, fully-convolutional residual network as its backbone. The network is coarsely structured into five residual *groups*, each consisting of one or more residual blocks of two-convolutional layers with residual connections adding the input of each block to the output.

A natural conditioning point is at each of these residual groups. This strategy is illustrated in figure 2; the figure reports also the size of the layers of the conditioning network ($C = 1024$). After each group of residual blocks, we insert

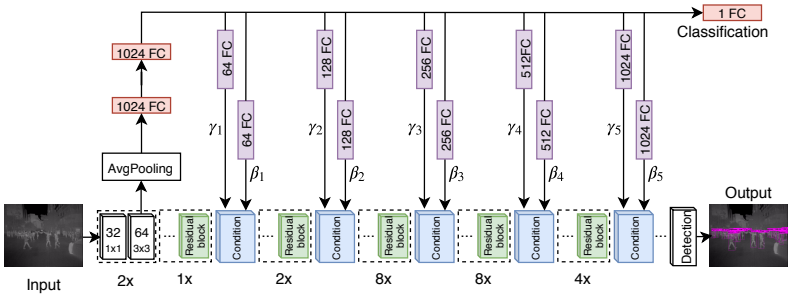

Fig. 2. TC Res Group: Conditioning residual *groups* of YOLOv3. The pre-ReLU activations of the last layer of each convolutional group are modified by parameters γ_i and β_i. Conditioning is done before the final residual connection of each group.

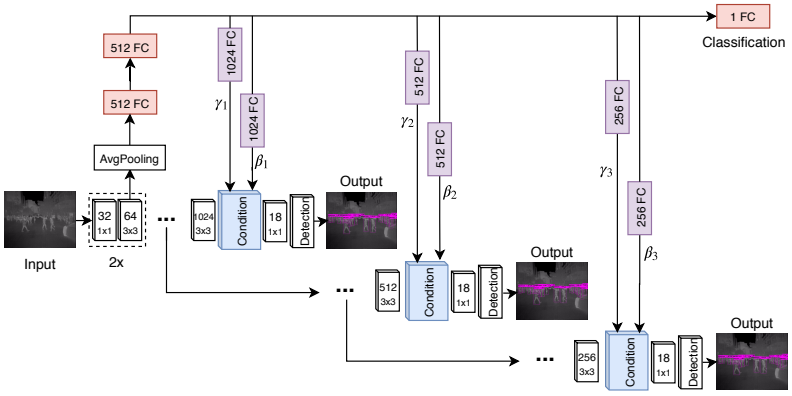

Fig. 3. TC Det: Conditioning the detection heads of YOLOv3. Feature maps used for detection are conditioned using the internal representation of the auxiliary network.

a conditioning layer after the last convolutional layer and *before* the final residual connection of the group.

Conditioning detection heads (TC Det). A natural alternative to conditioning residual groups is to condition each of the three detection heads branching off of the YOLOv3 backbone. The intuition here is to condition the network closer to where the actual detections are being made.

Detection heads in YOLOv3 consist of one convolutional block for the large-scale detection head, and three convolutional blocks for the other two. We insert the conditioning layer after the last convolution of these blocks and before the final 1×1 convolutional layer producing the detection head output. Figure 3 gives a schematic illustration of detection head conditioning architecture, and reports the size of the layers of the conditioning network ($C = 512$).

3.4 Adaptation loss

The final loss function used for domain adaptation is:

$$\mathcal{L}(\mathbf{x}_i, \mathbf{y}_i, y_i; \Theta_D, \Theta_A) = \mathcal{L}_d(\mathbf{x}_i, \mathbf{y}_i) + \mathcal{L}_a(\mathbf{x}_i, y_i),$$

where $\mathbf{x}$ is a training thermal image, $\mathcal{L}_d$ is the standard detection loss based on the structured target detections $\mathbf{y}_i$, and $\mathcal{L}_a$ is the auxiliary classification loss defined in equation (1).

When we backpropagate error from the auxiliary loss $\mathcal{L}_a$ we are improving the internal representation of the auxiliary network A_{Θ_a}, making it better for classifying day/night. When we backpropagate error from the detection loss, we simultaneously improve the generated conditioning parameters (γ_i, β_i) and the internal representation in the YOLOv3 backbone. Our intuition is that this adapts feature maps to be *conditionable* on based on the representation learned in the auxiliary classification network.

4 Experimental results

In this section we report results of a number of experiments we performed to evaluate the effectiveness of task-conditioned domain adaptation. In section 4.1 we describe the characteristics of the KAIST Multispectral Pedestrian Detection benchmark, and in section 4.3 we present two ablation studies we conducted to evaluate the various architectural parameters of our approach. In section 4.4 we compare with state-of-the-art single- and multimodal pedestrian detection approaches.

4.1 Dataset and evaluation metrics

Our experiments were conducted on the KAIST Multispectral Pedestrian Benchmark dataset [17]. KAIST is the only large-scale dataset with well-aligned visible/thermal pairs [7], and it contains videos captured both during the day and at night.

The KAIST dataset consists of 95,328 aligned visible/thermal image pairs split into 50,172 for training and 45,156 for testing. As is common practice, we use the *reasonable* setting [9, 17, 22, 25], and use the improved training annotations from [22] and test annotations from [25]. We sample every two frames from training videos and exclude heavily occluded and small person instances (< 50 pixels). The final training set contains 7,601 images. The test set contains 2,252 image pairs sampled every 20 frames. Figure 4 shows some example images with our detection results on KAIST.

We used standard evaluation metrics for object detection, namely miss rate as a function of False Positives Per Image (FPPI), and log-average miss rate for thresholds in the range of $[10^{-2}, 10^{0}]$. For computing miss rates, an Intersection over Union (IoU) threshold of 0.5 is used to calculate True Positive (TP), False Positives (FP) and False Negatives (FN).

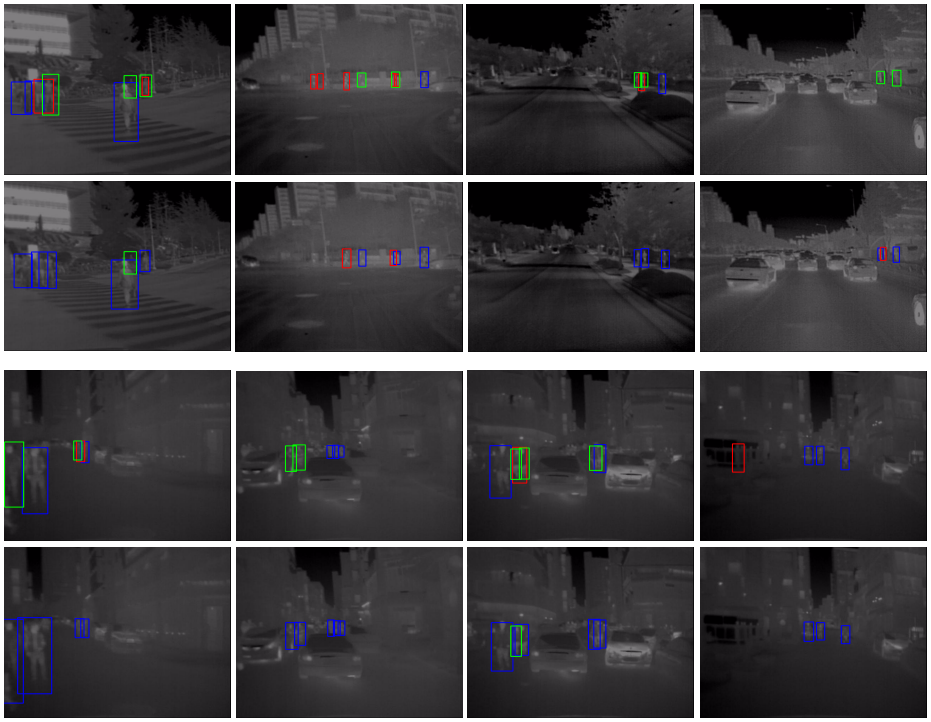

Fig. 4. Examples of KAIST thermal images with detections. The first two rows are daytime images and the last two are nighttime. The first and the third rows show detection results without conditioning, and the second and last rows detections with our **TC Det** detector. Blue boxes are true positive detections, green boxes are false negatives, and red boxes indicate false positives. See section 4.3 for detailed analysis.

4.2 Implementation and training

All of our networks were implemented in PyTorch and source code and pretrained models are available.[1] During training, at each epoch we set aside 10% of the training images for validation for that epoch. We use the same hyperparameter settings of the original YOLOv3 model [33] and use weights pretrained on MS COCO as a starting point. We use Stochastic Gradient Descent (SGD) with an initial learning rate of 0.0001. When the validation performance no longer improves, we reduce the learning rate by a factor of 10. Training is halted after decreasing the learning rate twice in this way. All models were trained for a maximum of 50 epochs with a batch size of 8 and input image size 640×512. For most cases, training stops at around 30 epochs and requires about 12 hours on an NVIDIA GTX 1080.

[1] `https://github.com/mrkieumy/task-conditioned`

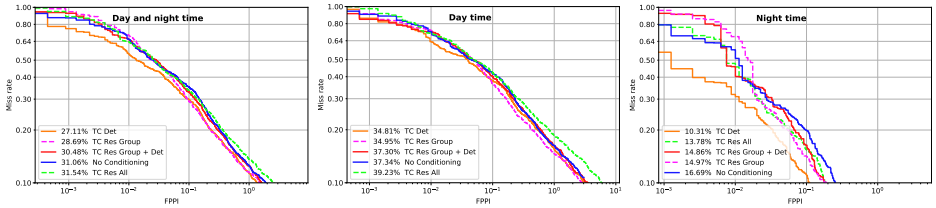

Fig. 5. Ablation study of different conditioning points. Plots report miss rate as a function of false positives per image, and log-average miss rates are given in the legends.

4.3 Ablation studies

In this section we report on a series of experiments we conducted to explore the design space for task-conditioned adaptation of a pretrained YOLOv3 detector to the thermal domain. We first consider the *where*-aspect of task-conditioning (i.e. at which points in the YOLOv3 architecture task-conditioning is most effective), and then consider the *when*-aspect of task conditioning by exploring the many possibilities of conditioning adaptation phases.

Comparison of conditioning points. YOLOv3 is a very deep network which presents many options for intervening with conditioning layers. It has 23 residual blocks, each consisting of two convolutional layers and one residual connection. These 23 residual blocks are organized into five groups as illustrated in figure 2. Inspired by the paper [30], in which the authors demonstrate that conditioning residual blocks can be effective, we performed an architectural ablation on *where* to condition the network by considering conditioning of all residual blocks versus conditioning each residual group. We investigate also conditioning of the three detection heads, both alone and in combination with residual group conditioning.

The configurations investigated are:

- **No Conditioning** (direct fine-tuning on thermal): the YOLOv3 network pretrained on MSCOCO is directly fine-tuned on KAIST thermal images.
- **TC Res Group** (conditioning of residual groups): the conditioning scheme described in section 3.3 and illustrated in figure 2. We insert conditioning layers into all residual groups at the final residual block.
- **TC Res All** (conditioning of all residual blocks): similar to group conditioning, but conditioning all residual blocks of the YOLOv3 network.
- **TC Det** (conditioning of detection heads): the scheme described in section 3.3 and illustrated in figure 3.
- **TC Res Group + Det** (conditioned residual groups and detection heads): a combination of **TC Res Group** and **TC Det**.

In figure 5 we plot the miss rate as a function of False Positive Per Image (FPPI) for the five different conditioning options. Note that *all* task-conditioned networks result in improvement over the **No Conditioning** network trained

Table 1. Ablation on adaptation schedules for **TC Det**. Results are on KAIST in terms of log-average miss rate (lower is better). **NC** indicates the modality is used for adaptation with no conditioning, **C** indicates the modality is used with conditioning of detection heads, and ✗ indicates the modality is not used during adaptation.

Training		Testing		Miss Rate		
visible	thermal	visible	thermal	all	day	night
NC	✗	✓	✗	36.67	32.83	45.00
C	✗	✓	✗	34.73	**29.53**	46.09
✗	NC	✗	✓	31.06	37.34	16.69
NC	NC	✗	✓	30.50	37.45	15.73
C	NC	✗	✓	28.48	35.86	12.97
✗	C	✗	✓	29.95	38.16	12.61
NC	C	✗	✓	28.53	36.59	11.03
C	C	✗	✓	**27.11**	34.81	**10.31**

with standard fine-tuning. **TC Det** performs best overall and performs especially well at nighttime with a miss rate of only 10.31% – an improvement of 6.38% over the **No Conditioning** network.

While conditioning residual groups (**TC Res Group**) is also effective compared to fine-tuning, adding more conditioning layers results in worse performance. One reason for this might be that conditioning layers add parameters to the network, and depending on the size of the feature maps being conditioned could be leading to overfitting on the KAIST training set.

In figure 4 we give example detections from the **TC Det** and **No Conditioning** detectors. **TC Det** yields more true positive and fewer false positive detections with respect to simple fine-tuning. On daytime images (first two columns of figure 4), the detector without conditioning (top row) produces a number of false positives and missed detections which **TC Det** does not. The difference is even more pronounced at nighttime (second two columns of figure 4).

This ablation analysis indicates that conditioning *only* detection layers (**TC Det**) is most effective when compared to conditioning of residual blocks – answering the *where* of task-conditioning. In all of the following experiments we consider only the **TC Det** task-conditioned network.

Comparison of conditional adaptation schedules. In this set of experiments we compare the many options of conditioning when adapting a pretrained detector from the visible to the thermal domain. Starting from a pretrained detector, we can fine-tune (with or without conditioning) on KAIST RGB images, then fine-tune (again with or without conditioning) on KAIST thermal images. In table 1 we give results of an ablation study considering all these possibilities. Adapting first using RGB images, rather than going directly to thermal, is generally useful. In fact, the best adaptation schedule is to fine-tune a conditioning network on visible spectrum images, and then fine-tune that conditioned network on thermal imagery.

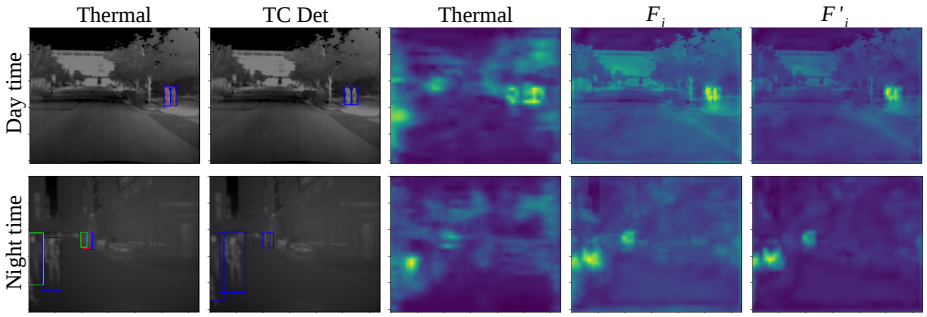

Fig. 6. The effects of conditioning during daytime and nighttime. The first two columns show results for a thermal detector without conditioning and with conditioning. Blue boxes are true positive detections, green boxes are false negatives, and red boxes indicate false positives. See text detailed analysis.

Visualizing the effects of conditioning. Figure 6 illustrates the effect conditioning has on the feature maps of YOLOv3. The heatmaps in this figure were generated by averaging the convolutional feature maps input to the medium-scale detection head of YOLOv3 and superimposing this on the original thermal image. The third column is the average feature map of a non-conditioned thermal detector (TD), and the fourth and fifth columns are, respectively, the average feature maps before and after conditioning.

From the heatmaps in figure 6 we note that pedestrians show more contrast with the background in the task-conditioned feature maps for both daytime and nighttime. Also, the thermal detector without conditioning misses several pedestrians and produces one false positive at nighttime, while TC Det correctly detects these and does not produce false positive detections. Task-conditioning also helps eliminate one false positive in the daytime image.

4.4 Comparison with the state-of-the-art

In this section we compare our approaches with the state-of-the-art on KAIST. Since our approach focuses on detection only in thermal images at test time, we first compare with state-of-the-art single-modality detectors using use only visible or only thermal images. Then, we compare our approaches with state-of-the-art multispectral detectors using both visible and thermal images.

Comparison with single-modality detectors. Table 2 compares our approaches with the single-modality detectors using thermal-only or visible-only at training and testing time. TC Det obtains the best results with 27.11% missrate in all modalities and 10.31% missrate at nighttime. Our results outperform all existing single-modality methods by a large margin in all conditions (day, night, and all). To the best our knowledge, our detectors outperform all state-of-the-art single-modality approaches on KAIST dataset.

Table 2. Comparison with state-of-the-art single-modality approaches on KAIST in term of log-average miss rate (lower is better). Best results highlighted in **underlined bold**, second best in **bold**.

Detectors		MR all	MR day	MR night	Test images
FasterRCNN-C	[25]	48.59	42.51	64.39	RGB
RRN+MDN	[39]	49.55	47.30	54.78	RGB
FasterRCNN-T	[25]	47.59	50.13	40.93	thermal
TPIHOG	[3]	-	-	57.38	thermal
SSD300	[16]	69.81	-	-	thermal
Saliency Maps	[12]	-	**30.40**	21.00	thermal
VGG16-two-stage	[15]	46.30	53.37	31.63	thermal
ResNet101-two-stage	[15]	42.65	49.59	26.70	thermal
Bottom-up	[19]	35.20	40.00	20.50	thermal
Ours TC Visible		34.73	**29.53**	46.09	RGB
Ours TC Thermal		**28.53**	36.59	**11.03**	thermal
Ours TC Det		**27.11**	34.81	**10.31**	thermal

Table 3. Comparison with state-of-the-art multimodal approaches in terms of log-average miss rate on KAIST dataset (lower is better). All approaches use both visible and thermal spectra at training and test time, while ours use only thermal imagery for testing. Results for Methods indicated with * were computed using detections provided by the authors. Best results highlighted in **underlined bold**, second best in **bold**.

Method		MR all	MR day	MR night	Detector Architecture	
KAIST baseline	[17]	64.76	64.17	63.99	ACF	[8]
Late Fusion	[38]	43.80	46.15	37.00	RCNN	[13]
Halfway Fusion	[25]	36.99	36.84	35.49	Faster R-CNN	[34]
RPN+BDT	[20]	29.83	30.51	27.62	VGG-16 + BF	[35, 2]
IATDNN+IAMSS	[14]	26.37	27.29	24.41	VGG-16 + RPN	[35, 20]
IAF R-CNN*	[23]	20.95	21.85	18.96	VGG-16 + Faster R-CNN	[35, 34]
MSDS-RCNN	[22]	**11.63**	**10.60**	13.73	VGG-16 + RPN	[35]
MSDS sanitized*	[22]	**10.89**	**12.22**	**7.82**	VGG-16 + RPN	[35]
YOLO_TLV	[37]	31.20	35.10	22.70	YOLOv2	[32]
DSSD-HC	[21]	34.32	-	-	DSSD	[11]
GFD-SSD	[43]	28.00	25.80	30.03	SSD	[26]
Ours Thermal		31.06	37.34	16.69	YOLOv3	[33]
Ours TC Res Group		28.69	34.95	14.97	YOLOv3	[33]
Ours TC Det		27.11	34.81	**10.31**	YOLOv3	[33]

Comparison with multimodal detectors. Table 3 compares our detectors with state-of-the-art multimodal approaches. Some multispectral methods using both visible and thermal images for training and testing such as MSDS [22], IAF [23] or IATDNN+IAMSS [14] are superior in terms of combined day/night miss rate (all). This is due to the advantage they have in exploiting both visible and thermal imagery, affecting in particular pedestrian detection during the day. In fact, the authors in MSDS [22] proposed a set of manually "sanitized" annotations for KAIST that correct problems in the original annotations and

their sanitized results at night-time (indicated by *) are better than the original results due to misalignment correction. Another key difference is that most state-of-the-art multispectral approaches use more complex, two-stage detection architectures like Faster RCNN (last column of table 3). Note, however, that both TC Res Group and TC Det, surpass many multimodal techniques, while TC Det performs second-best at night.

We note that recent advances in the state-of-the-art on KAIST have been made by augmenting and/or correcting the original dataset annotations. For example, the authors of AR-CNN [42] completely re-annotated the KAIST dataset, correcting localization errors, adding relationships, and labeling unpaired objects, resulting in significantly improved performance. Use of additional manual annotations, however, renders their results impossible to compare with those of other approaches and are thus excluded from our comparison.

Speed analysis. The average inference time for YOLOv3 is 28.57 milliseconds per image (∼35 FPS). Our **TC Det** network requires 33.17 milliseconds per image (∼30 FPS), and **TC Res Group** 35.01 milliseconds per image (∼29 FPS). Thus, task conditioning does not significantly increase the complexity of the network – in fact our **TC Det** network requires less than five milliseconds more for single-image inference compared to the original YOLOv3 detector.

5 Conclusions

In this paper we proposed a task-conditioned architecture for adapting visible-spectrum detectors to the thermal domain. Our approach exploits the internal learned representation of an auxiliary day/night classification network to inject conditioning parameters at strategic points in the detector network. Our experiments demonstrate that task-based conditioning of the YOLOv3 detection network can significantly improve thermal-only pedestrian detection performance.

Task-conditioned networks preserve the efficiency of the single-shot YOLOv3 architecture and perform respectably even compared to some multispectral detectors. However, they are outperformed by more complex, two-stage multispectral detectors such as MSDS [22]. We think, however, that our task-conditioning approach can also be fruitfully applied to such detectors by conditioning both region proposal and classification subnetworks.

Acknowledgments

The authors thank NVIDIA for the generous donation of GPUs. This work was partially supported by the project ARS01_00421: “PON IDEHA - Innovazioni per l’elaborazione dei dati nel settore del Patrimonio Culturale.”

References

1. Angelova, A., Krizhevsky, A., Vanhoucke, V., Ogale, A., Ferguson, D.: Real-time pedestrian detection with deep network cascades. In: Proc. of British Machine Vision Conference (BMVC) (2015)
2. Appel, R., Fuchs, T., Dollár, P., Perona, P.: Quickly boosting decision trees–pruning underachieving features early. In: International conference on machine learning. pp. 594–602 (2013)
3. Baek, J., Hong, S., Kim, J., Kim, E.: Efficient pedestrian detection at nighttime using a thermal camera. Sensors **17**(8), 1850 (2017)
4. Benenson, R., Omran, M., Hosang, J., Schiele, B.: Ten years of pedestrian detection, what have we learned? In: Proc. of European Conference on Computer Vision (ECCV) (2014)
5. Brazil, G., Yin, X., Liu, X.: Illuminating pedestrians via simultaneous detection & segmentation. In: Proc. of IEEE International Conference on Computer Vision (ICCV) (2017)
6. Cao, Y., Guan, D., Wu, Y., Yang, J., Cao, Y., Yang, M.Y.: Box-level segmentation supervised deep neural networks for accurate and real-time multispectral pedestrian detection. ISPRS Journal of Photogrammetry and Remote Sensing **150**, 70–79 (2019)
7. Devaguptapu, C., Akolekar, N., M Sharma, M., N Balasubramanian, V.: Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W) (2019)
8. Dollár, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. IEEE transactions on pattern analysis and machine intelligence **36**(8), 1532–1545 (2014)
9. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **34**(4), 743–761 (2011)
10. Fritz, K., König, D., Klauck, U., Teutsch, M.: 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 (2019)
11. Fu, C.Y., Liu, W., Ranga, A., Tyagi, A., Berg, A.C.: Dssd: Deconvolutional single shot detector. arXiv preprint arXiv:1701.06659 (2017)
12. Ghose, D., Desai, S.M., Bhattacharya, S., Chakraborty, D., Fiterau, M., Rahman, T.: Pedestrian detection in thermal images using saliency maps. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W) (2019)
13. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580–587 (2014)
14. Guan, D., Cao, Y., Yang, J., Cao, Y., Yang, M.Y.: Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. Information Fusion **50**, 148–157 (2019)
15. Guo, T., Huynh, C.P., Solh, M.: Domain-adaptive pedestrian detection in thermal images. In: Proc. of IEEE International Conference on Image Processing (ICIP) (2019)

16. Herrmann, C., Ruf, M., Beyerer, J.: 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 (2018)
17. Hwang, S., Park, J., Kim, N., Choi, Y., Kweon, I.: Multispectral pedestrian detection: Benchmark dataset and baseline. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
18. John, V., Mita, S., Liu, Z., Qi, B.: 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). pp. 246–249 (2015)
19. Kieu, M., Bagdanov, A.D., Bertini, M., Del Bimbo, A.: Domain adaptation for privacy-preserving pedestrian detection in thermal imagery. In: Proc. of International Conference on Image Analysis and Processing (ICIAP) (2019)
20. Konig, D., Adam, M., Jarvers, C., Layher, G., Neumann, H., Teutsch, M.: Fully convolutional region proposal networks for multispectral person detection. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W) (2017)
21. Lee, Y., Bui, T.D., Shin, J.: 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) (2018)
22. Li, C., Song, D., Tong, R., Tang, M.: Multispectral pedestrian detection via simultaneous detection and segmentation. In: Proc. of British Machine Vision Conference (BMVC) (2018)
23. Li, C., Song, D., Tong, R., Tang, M.: Illumination-aware faster R-CNN for robust multispectral pedestrian detection. Pattern Recognition **85**, 161–171 (2019)
24. Li, J., Liang, X., Shen, S., Xu, T., Feng, J., Yan, S.: Scale-aware fast R-CNN for pedestrian detection. IEEE Transactions on Multimedia (TMM) **20**(4), 985–996 (2017)
25. Liu, J., Zhang, S., Wang, S., Metaxas, D.N.: Multispectral deep neural networks for pedestrian detection. arXiv preprint arXiv:1611.02644 (2016)
26. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
27. Liu, W., Liao, S., Ren, W., Hu, W., Yu, Y.: High-level semantic feature detection: A new perspective for pedestrian detection. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
28. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
29. Ouyang, W., Zeng, X., Wang, X.: Learning mutual visibility relationship for pedestrian detection with a deep model. International Journal of Computer Vision (IJCV) **120**(1), 14–27 (2016)
30. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.C.: FiLM: Visual reasoning with a general conditioning layer. In: Proc. of AAAI Conference on Artificial Intelligence (AAAI) (2017)
31. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. CoRR **abs/1511.06434** (2015)
32. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
33. Redmon, J., Farhadi, A.: YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767 **abs/1804.02767** (2018)

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. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
36. Tian, Y., Luo, P., Wang, X., Tang, X.: Pedestrian detection aided by deep learning semantic tasks. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
37. Vandersteegen, M., Van Beeck, K., Goedemé, T.: Real-time multispectral pedestrian detection with a single-pass deep neural network. In: Proc. of International Conference Image Analysis and Recognition (ICIAR) (2018)
38. Wagner, J., Fischer, V., Herman, M., Behnke, S.: Multispectral pedestrian detection using deep fusion convolutional neural networks. In: Proc. of European Symposium on Artificial Neural Networks (ESANN) (2016)
39. Xu, D., Ouyang, W., Ricci, E., Wang, X., Sebe, N.: Learning cross-modal deep representations for robust pedestrian detection. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
40. Zhang, L., Lin, L., Liang, X., He, K.: Is faster R-CNN doing well for pedestrian detection? In: Proc. of European Conference on Computer Vision (ECCV) (2016)
41. Zhang, L., Liu, Z., Chen, X., Yang, X.: The cross-modality disparity problem in multispectral pedestrian detection. arXiv preprint arXiv:1901.02645 (2019)
42. Zhang, L., Zhu, X., Chen, X., Yang, X., Lei, Z., Liu, Z.: Weakly aligned cross-modal learning for multispectral pedestrian detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5127–5137 (2019)
43. Zheng, Y., Izzat, I.H., Ziaee, S.: GFD-SSD: Gated fusion double SSD for multispectral pedestrian detection. arXiv preprint arXiv:1903.06999 (2019)