Tensor Factorization for Leveraging Cross-Modal Knowledge in Data-Constrained Infrared Object Detection

Sharma, Manish; Chatterjee, Moitreya; Peng, Kuan-Chuan; Lohit, Suhas; Jones, Michael J.

TR2023-125 October 02, 2023

Abstract

While state-of-the-art object detection methods have reached some level of maturity for regular RGB images, there is still some distance to be covered before these meth- ods perform comparably on Infrared (IR) images. The primary bottleneck towards accomplishing this goal is the lack of sufficient labeled training data in the IR modality, owing to the cost of acquiring such data. Realizing that object detection methods for the RGB modality are quite robust (at least for some commonplace classes, like person, car, etc.), thanks to the giant training sets that exist, in this work we seek to leverage cues from the RGB modality to scale object detectors to the IR modality, while preserving model performance in the RGB modality. At the core of our method, is a novel tensor decomposition method called TensorFact which splits the convolution kernels of a layer of a Convolutional Neural Network (CNN) into low-rank factor matrices, with fewer parameters than the original CNN. We first pre-train these factor matrices on the RGB modality, for which plenty of training data are assumed to exist and then augment only a few trainable parameters for training on the IR modality – to avoid over-fitting, while encouraging them to capture complementary cues from those trained only on the RGB modality. We validate our approach empirically by first assessing how well our TensorFact decomposed net- work performs at the task of detecting objects in RGB images vis-á-vis the original network and then look at how well it adapts to IR images of the FLIR ADAS v1 dataset. For the latter, we train models under scenarios that pose challenges stemming from data paucity. From the experiments, we ob- serve that: (i) TensorFact shows performance gains on RGB images; (ii) further, this pre-trained model, when fine-tuned, outperforms a standard state-of-the-art object detector on the FLIR ADAS v1 dataset by about 4% in terms of mAP 50 score.

IEEE International Conference on Computer Vision Workshops (ICCV) 2023

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Tensor Factorization for Leveraging Cross-Modal Knowledge in Data-Constrained Infrared Object Detection

Manish Sharma[1*] Moitreya Chatterjee[2*] Kuan-Chuan Peng[2] Suhas Lohit[2] Michael Jones[2]
[1]Rochester Institute of Technology, NY 14623, USA
[2]Mitsubishi Electric Research Laboratories, Cambridge, MA 02139, USA
ms8515@rit.edu metro.smiles@gmail.com kpeng@merl.com slohit@merl.com mjones@merl.com

Abstract

While state-of-the-art object detection methods have reached some level of maturity for regular RGB images, there is still some distance to be covered before these methods perform comparably on Infrared (IR) images. The primary bottleneck towards accomplishing this goal is the lack of sufficient labeled training data in the IR modality, owing to the cost of acquiring such data. Realizing that object detection methods for the RGB modality are quite robust (at least for some commonplace classes, like person, car, etc.), thanks to the giant training sets that exist, in this work we seek to leverage cues from the RGB modality to scale object detectors to the IR modality, while preserving model performance in the RGB modality. At the core of our method, is a novel tensor decomposition method called TensorFact *which splits the convolution kernels of a layer of a Convolutional Neural Network (CNN) into low-rank factor matrices, with fewer parameters than the original CNN. We first pre-train these factor matrices on the RGB modality, for which plenty of training data are assumed to exist and then augment only a few trainable parameters for training on the IR modality – to avoid over-fitting, while encouraging them to capture complementary cues from those trained only on the RGB modality. We validate our approach empirically by first assessing how well our* TensorFact *decomposed network performs at the task of detecting objects in RGB images vis-á-vis the original network and then look at how well it adapts to IR images of the FLIR ADAS v1 dataset. For the latter, we train models under scenarios that pose challenges stemming from data paucity. From the experiments, we observe that: (i)* TensorFact *shows performance gains on RGB images; (ii) further, this pre-trained model, when fine-tuned, outperforms a standard state-of-the-art object detector on the FLIR ADAS v1 dataset by about* 4% *in terms of mAP 50 score.*

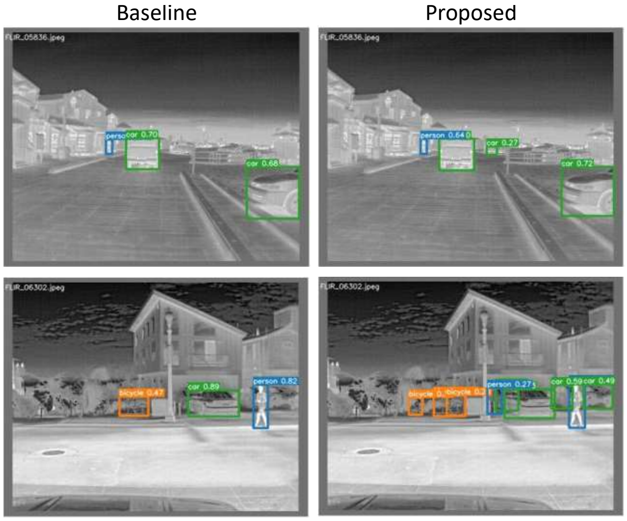

Figure 1: Qualitative comparison of object detections by a state-of-the-art object detector (denoted as baseline) [51] and our TensorFact method on IR images. The orange, cyan, green boxes denote bicycle, person, and car classes respectively, while the associated numbers denote the confidence score of the prediction. The visualizations show that our proposed approach is better at capturing more objects, especially those that are of a smaller size, with higher precision.

1. Introduction

The success of deep neural networks in core computer vision tasks, such as image denoising [42], image classification [44], object detection [8, 40], *etc.* can at least in part be attributed to the availability of large-scale labeled training data [46], which allows these models (with lots of parameters) to avoid over-fitting [11]. This has resulted in wide-ranging applicability of these methods in tasks such as pedestrian detection in vehicles [35], face detection [48], vehicle counting [57], *etc.*

One key element that made such large-scale data available,

[*] Equal Contributions.

is the ubiquity of good quality RGB cameras, which come at throwaway prices. This coupled with the popularity of online platforms for sharing content widely, including social media sites such as YouTube or Meta, meant that sharing such images at a large-scale became commonplace.

However, from the standpoint of certain applications, such as autonomous driving, regular RGB images fall short on some important counts. For instance, while RGB images can provide clear visualization of the surroundings during the day, at night, RGB images are only useful if there is sufficient street lighting, *etc*. In scenarios where the ambient light is insufficient, passive thermal Infrared (IR) cameras come in handy for tasks such as pedestrian detection, as thermal IR sensors capture scenes at wavelengths beyond the visible spectrum and are sensitive to warm objects, such as the human body [4]. Nonetheless, one catch that remains is that IR cameras are not as cheap as their RGB counterparts and are thus not as ubiquitous. This poses a major hurdle in acquiring the profuse amounts of images needed to train deep networks that could operate on IR images at performance levels similar to their RGB counterparts. In such conditions, an overparameterized model results in overfitting, which has an impact on model generalisation and performance. Therefore, a reduction in the number of parameters may be needed for improved performance. Low-rank factorization methods are among the most popular methods towards this end and are utilized for different deep learning applications [22, 23, 41].

While the success of deep neural networks today spans several computer vision tasks, the task of object detection is of particular interest in this paper. The task entails localizing the pixels which an object occupies in an image as well as labeling the cluster of pixels with the class to which the said object belongs. Solving this task is crucial, since it permits acquiring a greater understanding of what an image contains and is often a first step towards understanding the scene [32]. Given the importance of IR images, as a modality for the task of scene understanding, designing effective object detection models that work on such data becomes critical. Nonetheless, the paucity of sufficient training data (*i.e*., datasets with lots of IR images) continues to present a challenge to this end.

In this work, we leverage the observation that while sufficient training data in the IR modality may be difficult to find, such data for the RGB modality is easily available. The key idea in our approach then, is to train an object detection model in the RGB modality and to then transfer the common cross-modal cues to the IR modality where only a few parameters can be trained to capture the complementary cues necessary for successfully detecting objects in the IR image space. Concretely, we devise a novel method called *TensorFact*, which splits the convolution kernel weights of a CNN layer into low-rank factor matrices, with fewer trainable parameters. These factor matrices can be trained to capture the common cues for detecting objects, across modalities, by leveraging the RGB data. These weights can then be augmented with only a few, new learnable parameters to capture the cues specific to the IR modality. This design allows us to train only the relatively small number of IR modality-specific weights when training with IR images, allowing us to prevent over-fitting. Note that naïvely applying domain adaptation methods [1] to transfer from RGB to IR modality fails because here the modality itself switches between the source (RGB) and the target (IR) which represents a big shift in the data distribution.

We conduct experiments on the FLIR ADAS v1 dataset [49] of IR images to empirically validate the efficacy of our method. To derive the common object detection cues from RGB images, we use the FLIR Aligned RGB [13] images. Our experiments show that *TensorFact* decomposition assists with achieving better object detection performance both on RGB and IR images, even when the latter has few training samples. In particular, in the IR dataset (FLIR ADAS v1), our method outperforms a competing state-of-the-art object detection model [51] by 4% on mAP 50, underscoring the efficacy of our method. Figure 1 contrasts detections obtained by our method in comparison to a recent state-of-the-art detection baseline, YOLOv7 [51], on the FLIR ADAS v1 dataset. From the figure, we see that our approach is more capable of detecting objects of different sizes, compared to the state-of-the-art approach.

We summarize below the core contributions of our work.

- We present *TensorFact*, a novel tensor decomposition-based method that can leverage both modality-specific and cross-modal cues for effective object detection in the IR modality, where acquiring sufficient training data is a challenge.
- Our experiments reveal that our proposed method outperforms competing approaches at the task of object detection in a data sparse IR modality, with only 62 training images, by 4% on mAP 50.
- Our formulation also offers a supplementary contribution to the RGB modality, yielding a compressed neural network that improves object detection in this modality.

2. Related works

In this section, we discuss relevant prior works to our paper and present the distinction between these approaches and our method.

Object detection approaches in IR images: The journey of object detection in RGB images, using deep learning, has come a long way [36, 38, 41, 51]. The inception of a two-stage object detection process involving proposal generation and object class prediction, initiated by the work of Girshick *et al*. [16] for RGB images, laid the foundation for the field. However, the computational intensity of the process necessitated faster successors [15, 18, 38, 47, 50]. However, porting these approaches to the realm of IR image object detection, has posed certain challenges. The study by Ghose *et al*. [14] and Devagupta *et al*. [7] sought to enhance infrared image features using saliency maps and multimodal Faster R-CNN, respectively. These efforts, however, encountered challenges such as slow inference speed, non end-to-end multitask training, and a lack of general applicability across different datasets.

To overcome the limitations of two-stage detectors, the work by Redmon and Farhadi [36] introduced a one-stage detector, YOLO, which considered each image cell as a proposal for object detection and achieved end-to-end real-time detection. YOLO's evolution into YOLOv3 [37], YOLOv4 [3], and its subsequent variants, as documented by Kristo *et al*. [26], has accelerated the detection of objects both in RGB and IR images, though issues of omission of small-scale objects and low detection accuracy persist.

Innovative modifications like the SE block in SE-YOLO [27] and the attention module, CIoU loss, improved Soft-NMS, and depthwise separable convolution in YOLO-ACN [31] were proposed to improve detection accuracy, but they still grapple with challenges like large parameter sizes and applicability to embedded settings.

Other one-stage models have been explored, including ThermalDet [5] and TIRNet [6], each of which offers different solutions to the aforesaid problems but falls short when tested in real-world, non-curated datasets. Song *et al*. [45] proposed a multispectral feature fusion network based on YOLOv3, showing promise for smaller-sized images.

The YOLO series has shown considerable potential for IR object detection and several variants to it have been proposed. This includes the network of Shuigen *et al*. [43], an attention mechanism-infused YOLOv3 [14], and a YOLOv3 enhanced with a category balance loss term [30]. Further refinements in object detection have been achieved by using the SAF architecture [34] and the YOLO-FIRI model [29], which incorporate optimization parameters, introduce dilated convolutional block attention modules, and enable the detection of smaller IR targets. Zhao *et al*. [58] and Du *et al*. [10] have contributed to the field by improving the fusion method of YOLOv3 and leveraging YOLOv4 to enhance IR target features, respectively, paving a promising path for future IR object detection research. While we consider these models for designing the backbone of our proposed approach but none of them provide a way to mitigate the data paucity issue in the IR modality which we address front and center.

Domain adaptation methods: The community has explored domain adaptation methods to overcome the challenges associated with less training data in certain domains. Towards this end, several works have been proposed [17,39,53,54,56], which include those that progressively transition from one domain to another [21], or transition through multiple levels of granularity [59], or use semi-supervised [9,52] or unsupervised learning [28, 55] techniques for the same. Nonetheless, these approaches tackle scenarios which represent reasonably minor shifts in the domain of the input data, say from clear RGB images to foggy RGB images [12] and so on. However, our task, deals with much larger-scale shifts in the type of input, in particular from RGB to IR modalities. The change is so stark that certain objects are visible in a given modality, only under specific scenarios. For instance, warm-bodied, dimly lit objects are visible only in the IR images but are very difficult to see in RGB images. This prevents us from trivially adapting these approaches for our task. While some more recent methods have looked into domain adaptation techniques for IR detection tasks, these are fairly limited in scope [20, 24] and focus mostly on detecting people, not other classes. Importantly, none of these approaches simulate the training data paucity scenario, for the IR modality, something we consider in this work.

3. Proposed approach

In this work, we propose *TensorFact* – a novel tensor decomposition-based method designed to tackle the paucity of labeled training data in the IR modality. It effectively leverages knowledge learned from the RGB modality, where training data is abundant, and efficiently transfers this knowledge to the IR modality, overcoming the data scarcity challenge. Initially, we learn two trainable low-rank factor-matrices, the product of which yields the weights for each layer of the CNN and task them with detecting objects in the source RGB modality. This representation cuts down on the number of learnable parameters in the network and facilitates the training of a more generalizable network (due to less over-fitting) on the RGB modality. Following this, in order to facilitate object detection in the IR modality, we enhance the network's capability by a minor expansion of the number of trainable parameters. This is achieved by increasing the number of the columns/rows of the factor matrices. The factor matrices that emerge from the increased columns/rows effectively serve as a parallel trainable branch, enabling the network to leverage the complementary information gleaned from the RGB modality for object detection in the IR modality. In this way, *TensorFact* affords us a practical solution to the challenge of limited training data in the IR modality, demonstrating how robust and transferable features can be effectively extracted and utilized across different modalities.

3.1. Notation

In this paper, we utilize the following conventions: lowercase letters such as x denotes scalar variables, vectors are symbolized by boldface lowercase letters like $\mathbf{x}$, and matrices are depicted by boldface uppercase letters such as $\mathbf{X}$. Tensors, on the other hand, are indicated by calligraphic

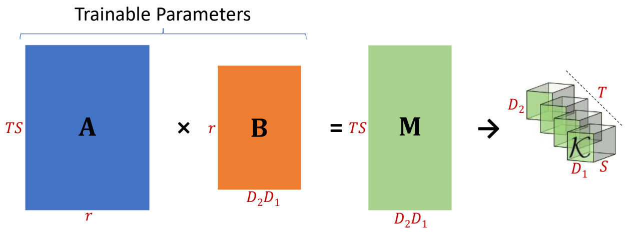

Figure 2: Decomposed convolutional layer.

uppercase letters (for instance, $\mathcal{X}$). $\mathbb{R}$ denotes the set of real numbers. To illustrate a component of a vector, matrix, or tensor, we adopt the $[\cdot]_i$ notation, where i represents a set of indices for that component.

3.2. Decomposed convolution layer

The weights of a convolutional layer in a CNN, denoted by $\mathcal{K} \in \mathbb{R}^{T \times S \times D_2 \times D_1}$, is a 4-way tensor, where D_1 and D_2 represent the width and height, respectively, of the spatial window of the convolution kernels, while S and T denote the number of input channels of the input to the layer and the number of kernels learned in the layer. The number of trainable parameters in a standard convolutional layer is then given by $P = TSD_2D_1$.

For a decomposed convolutional layer, we commence with two trainable factors $\mathbf{A} \in \mathbb{R}^{TS \times r}$ and $\mathbf{B} \in \mathbb{R}^{r \times D_2 D_1}$ with r serving as their inner dimension, as shown in Figure 2, denoting the rank of the original weight matrix (prior to decomposition). These combine to form the intermediate matrix $\mathbf{M} = \mathbf{AB}$, as follows:

$$[\mathbf{M}]_{p,q} = \sum_{c=1}^{r} [\mathbf{A}]_{p,c} [\mathbf{B}]_{c,q}, \tag{1}$$

where, $p = 1, \ldots, TS$ and $q = 1, \ldots, D_2D_1$. This matrix $\mathbf{M}$, operates on the input to the layer. The convolutional filter $\mathcal{K}$, is derived from $\mathbf{M}$ as:

$$[\mathcal{K}]_{t,s,d_2,d_1} = [\mathbf{M}]_{(t-1)S+s,(d_2-1)D_1+d_1}, \tag{2}$$

where, $t = 1, \ldots, T$, $s = 1, \ldots, S$, $d_2 = 1, \ldots, D_2$, and $d_1 = 1, \ldots, D_1$. Therefore, the number of trainable parameters in the decomposed convolutional layer formulation, P_{fac}, is a function of r, resulting in $P_{fac} = r(TS + D_2D_1)$ trainable parameters. The value of r can be altered to adapt to the necessary CNN complexity but typically $r \leq \text{rank}(\mathbf{M})$. Since CNNs are known to be over-parameterized [11], one could choose r such that the number of learnable parameters is fewer than that in $\mathbf{M}$, to avoid the risk of over-fitting.

3.3. Capacity augmentation

To augment the network capacity to accommodate the new modality, we increase r by Δr (where $\Delta r > 0$) for both matrices $\mathbf{A}$ and $\mathbf{B}$, thereby producing $\mathbf{A}' \in \mathbb{R}^{TS \times (r+\Delta r)}$ and $\mathbf{B}' \in \mathbb{R}^{(r+\Delta r) \times D_2 D_1}$ with $r + \Delta r$ serving as their new inner dimension. Now, $\mathbf{A}'$ and $\mathbf{B}'$ can be interpreted as $\mathbf{A}' = \left[\mathbf{A}\, || \Delta\mathbf{A}\right]$ and $\mathbf{B}' = \left[\mathbf{B}\, || \Delta\mathbf{B}\right]^T$ such that $\Delta\mathbf{A} \in \mathbb{R}^{TS \times \Delta r}$ and $\Delta\mathbf{B} \in \mathbb{R}^{\Delta r \times D_2 D_1}$ and $||$ denotes concatenation. Subsequently, $\mathbf{A}'$ and $\mathbf{B}'$ merge to form $\mathbf{M}' = \mathbf{A}'\mathbf{B}' = \mathbf{M} + \Delta\mathbf{M}$, where $\Delta\mathbf{M} = \Delta\mathbf{A}\Delta\mathbf{B}$, as shown in Figure 3. Similar to Equation 2, $\Delta\mathcal{K} \in \mathbb{R}^{T \times S \times D_2 \times D_1}$ can be derived from $\Delta\mathbf{M}$. Hence, increasing r by Δr results in a parallel architectural branch, as depicted in Figure 4. Therefore, the increase in the number of trainable parameters in a decomposed convolutional layer after capacity augmentation is given by $\Delta P_{fac} = \Delta r(TS + D_2D_1)$. We seek to augment as few parameters as possible to ensure the detection network does not suffer from challenges related to over-fitting in the new modality. In particular, we ensure that the total number of network parameters (considering those trained using only RGB and the augmented set) of our proposed framework, is less than the original unfactorized network.

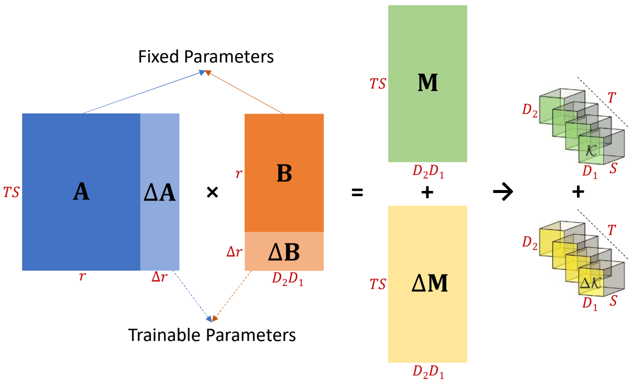

Figure 3: Decomposed convolutional layer with capacity augmentation.

3.4. Training

For an object detector CNN with L convolutional layers, let $\mathbf{A}_l$ and $\mathbf{B}_l$ represent the left and right factor matrices, respectively, for the l^{th} decomposed convolutional layer, with r_l representing their inner-dimension and $l = 1, \ldots, L$. When training for the data-rich source RGB modality, the network weights for the decomposed convolutional layers are SVD-initialized, leading to orthogonal column and row vectors in $\mathbf{A}_l$ and $\mathbf{B}_l$, respectively, with $r_l = \lfloor \alpha r_l^{max} \rfloor$. Here, $r_l^{max} = \min(TS, D_2D_1)_l$ and $\alpha \in (0, 1)$ controls the number of the trainable parameters across layers. With $\alpha \leq 1$, the training process is straightforward and similar to a typical object detector network, leading to the learning of both generic and modality-specific features for the RGB data.

Next, to train for the data-scarce IR modality, we augment

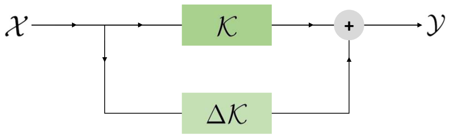

Figure 4: The flow of data in our proposed *TensorFact* approach in every layer. The input $\mathcal{X}$ convolves with $\mathcal{K}$ (top branch) and $\Delta\mathcal{K}$ (bottom branch) and results in output $\mathcal{Y}$ after summation.

the network capacity by increasing the value of α, which introduces new trainable parameters and creates a parallel path for each decomposed convolutional layer. During this training phase, we freeze the trainable parameters learned during the training of the RGB modality, thereby architecturally promoting the learning of complementary features for the IR modality branch. Akin to skip-connections in ResNets [19], which permits the learning of residual mapping, our proposed method leverages cross-modal cues and promotes the learning of features specific to the IR modality that were not learned during RGB modality training. As the factor matrices trained on the RGB data capture several cues essential for object detection, only a small percentage of augmented capacity is required for capturing the facets of object detection in the IR modality. This is an essential requirement to train the model without over-fitting in a data-scarce modality. Additionally, to explicitly capture complementary cues between the RGB and IR modalities, we maximize the L_2 or L_1 distances between the feature activation maps that are output from each branch (RGB and IR) of a layer and have it as an additional term in the training objective for the task. This can be implemented by the following loss L_c:

$$L_c = -||\mathcal{K} * \mathcal{X} - \Delta\mathcal{K} * \mathcal{X}||_p, \tag{3}$$

where $p = \{1, 2\}$ and $*$ denotes convolution. Note that the dimensions of $\mathcal{K}$ and $\Delta\mathcal{K}$ are the same. The final loss function L_f of *TensorFact* can be written as follows:

$$L_f = L_d + \omega_c L_c, \tag{4}$$

where L_d is the object detection loss used in YOLOv7 [51], and ω_c is the weight of L_c. We minimize this loss using the ADAM optimizer [25].

4. Experiments

In this section we layout the empirical evaluation that we conducted to validate the efficacy of our proposed approach.

4.1. Experimental setup

Datasets: In our object detection experiments, we make use of two datasets: FLIR Aligned [13] and FLIR ADAS v1 [49]. The FLIR Aligned dataset contains RGB images, with ground-truth comprising bounding-box coordinates around objects in the image as well as class labels. This dataset includes 4129 images for training and 1013 images for validation, and features three classes: person, bicycle, and car, with the distribution of instances provided in Table 1.

Class	Training Instances	Validation Instances
Person	8987	4107
Bicycle	2566	360
Car	20608	4124

Table 1: Distribution of class instances for training and validation sets for FLIR Aligned RGB dataset [13].

Class	Training Instances	Validation Instances
Person	161	4611
Bicycle	24	842
Car	351	8472

Table 2: Distribution of class instances for training (1%) and validation sets for FLIR ADAS v1 IR dataset [49].

The FLIR ADAS v1 dataset is a dataset of IR images. The ground-truth for this dataset, includes bounding-box coordinates around objects in the image and their class labels, from among: person, bicycle, and car. The dataset includes 7859 images for training and 1360 images for validation. However, for fair comparative studies, we split the training set into train and validation splits in an 80:20 train-validation ratio to create new, randomly selected train and validation sets consisting of 6287 and 1572 images, respectively. To mimic a data-scarce environment, we use randomly selected 62 images (1% of images) from the training set. Table 2 details the distribution of the FLIR ADAS v1 IR dataset classes, as used in our experiment.

Baseline network and evaluation metrics: We use YOLOv7 [51], a state-of-the-art object detector with over 37M trainable parameters as our baseline network. To determine appropriate anchor box sizes for the detector, we use K-Means++ method [2].

In evaluating the performance of our object detection model, we employed the Mean Average Precision (mAP), a widely-used and robust metric in the field. mAP considers both precision and recall, ensuring a balance between detecting as many objects as possible and minimizing false positives. This is achieved by generating Precision-Recall (PR) curves for each object class in two different settings. In the first, the Intersection Over Union (IoU) between predicted and ground truth bounding boxes is set to larger than 0.5, for it to be counted as a true positive prediction, while in the second setting, multiple evaluations are performed with increasing thresholds from 0.5 to 0.95 in increments of 0.05. The Average Precision (AP) is then calculated as

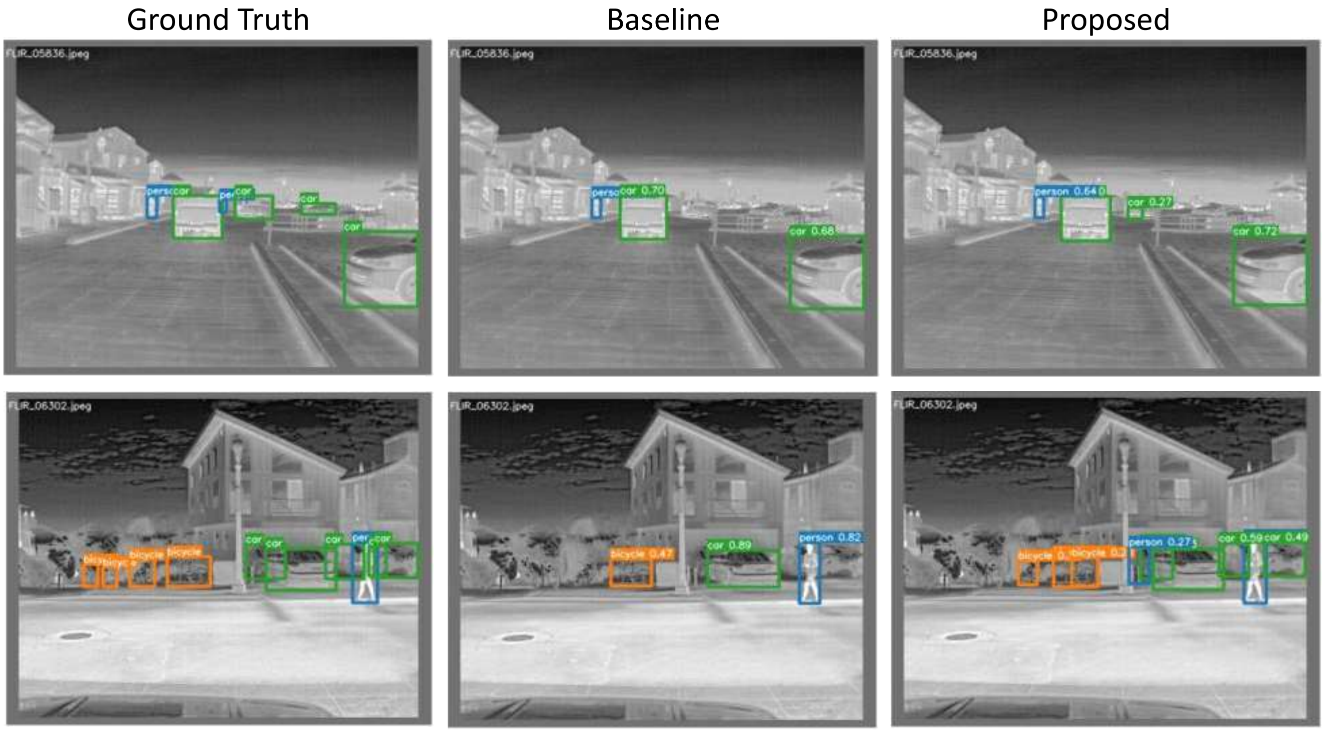

Figure 5: Comparison of object detection results between the state-of-the-art YOLOv7 [51] and our proposed approach. We show the ground truth (left column), baseline (middle column), and proposed method's ($\alpha = 0.1$, right column) detections as rectangular bounding boxes. We show detections on two different images from the FLIR ADAS v1 IR validation dataset, one in each row. The orange, cyan, green boxes denote bicycle, person, and car classes respectively, while the associated numbers denote the confidence score of the prediction.

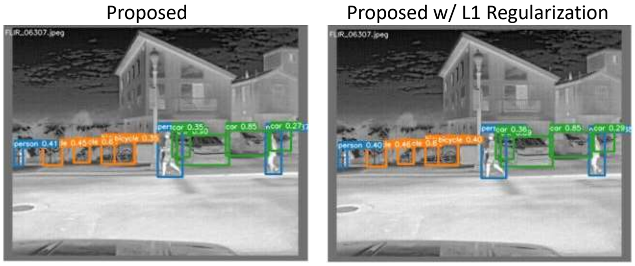

Figure 6: Comparison of object detection results for the proposed method without (left column) and with (right column) L_1 regularization. The orange, cyan, green boxes denote bicycle, person, and car classes respectively, while the associated numbers denote the confidence score of the prediction. We obtain better object detections using the L_1 regularization, as compared to the vanilla model, as manifested by the higher confidence scores for the predicted bounding boxes.

the area under each PR curve for every class, under each of these settings. We then take the mean of these APs, across the different classes, to get the mAP. The IoU threshold of 0.5, is used for the mAP 50 metric, while the range of IoU thresholds from 0.5 to 0.95 (in steps of 0.05) is used for the mAP 50-95 metric.

Implementation details: We train all models for 200 epochs, with mini-batch size of 40 images, where the gradients are accumulated over 2 mini-batch iterations prior to parameter update. We use the ADAM optimizer [25] with a learning rate of 10^{-5} when training on the RGB modality and 10^{-3} when training on the IR modality. We use “reduce on plateau" as the learning rate scheduler, that reduces learning rate by a factor of 0.1 if the validation loss does not improve over 10 epochs. Rather than initializing the RGB network from scratch, we initialize it with the pre-trained weights for detecting objects in the MS-COCO dataset [33]. When explicitly encouraging complementarity between the RGB and IR branches, we set the weight $\omega_c = 0.01$ in Eq. 4 such that both terms have comparable range.

4.2. Results and analysis

In Table 3, we present the evaluation results of the proposed and baseline methods for the FLIR Aligned RGB validation dataset for the task of object detection in RGB images. From the table, we observe that our proposed method demonstrates comparable, if not superior, performance to

Model	# Parameters↓	Compression (%)↑	mAP 50↑	mAP 50-95↑
Baseline	37,205,480	0	0.6826	**0.3173**
TensorFact ($\alpha = 0.9$)	35,400,800	4.8506	**0.6948**	0.3162
TensorFact ($\alpha = 0.8$)	**33,594,257**	**9.7062**	0.6879	0.3168

Table 3: Results for FLIR Aligned RGB validation dataset.

Model	# Trainable Params↓	Compression (%)↑	mAP 50↑	mAP 50-95↑
Baseline	37,205,480	0	0.5849	0.2807
TensorFact ($\alpha = 0.1$)	**1,856,343**	**95.01**	0.6205	**0.2807**
TensorFact ($\alpha = 0.2$)	3,662,886	90.16	**0.6213**	0.2794

Table 4: Results for FLIR ADAS v1 IR validation dataset.

Regularization	mAP 50↑	mAP 50-95↑
none	0.6205	0.2807
L_1	**0.6234**	**0.2823**
L_2	0.6222	0.2815

Table 5: Results for *TensorFact* with explicit complementarity regularization for $\alpha = 0.1$ on FLIR ADAS v1 IR validation dataset.

the baseline model in terms of both mAP 50 and mAP 50-95 evaluation metrics, across varying values of α. Interestingly, while reduction in the value of α leads to significant compression in the model's size, our proposed method successfully maintains, and in certain instances enhances, model performance. We hypothesize that this is because the decrease of the trainable parameters reduces the chance of over-fitting.

Table 4 presents the comparison results between the baseline and the proposed *TensorFact* method on the FLIR ADAS v1 IR validation dataset. For the proposed *TensorFact* method, we employ two different α configurations, 0.1 and 0.2, such that the ratios $\{\Delta r_l : r_l\}_{l=1}^{L}$ are $1 : 9$ and $1 : 4$, respectively, for $l = 1, 2, \ldots, L$. We observe that both proposed model configurations outperform the baseline on the mAP 50 evaluation metric, with only a few additional trainable parameters in the IR branch. These results underscore the potential of our proposed method to efficiently learn and generalize with significantly fewer trainable parameters in a data-scarce environment like the IR modality, while leveraging cross-modal cues from the data-rich RGB modality.

Lastly, in Table 5, we present results for augmenting the training objective with an explicit complementarity criterion, for $\alpha = 0.1$ on the FLIR ADAS v1 IR validation dataset to determine the impact of regularization to promote learning of complementary features for the IR modality beyond the pre-trained RGB modality. We observe that both L_1 and L_2 regularization methods show slight improvements in detection performance compared to the model without explicit regularization.

Qualitative results: In Figure 5, we compare the object detection results using the ground truth (left column), baseline (middle column), and proposed methods ($\alpha = 0.1$, right column). The results are displayed vertically for two different images from the FLIR ADAS v1 IR validation dataset. We observe that the baseline method fails to detect small, distant objects and objects with backgrounds of similar texture as the foreground, whereas the proposed method accurately detects them. This shows that the proposed method is more robust against false negatives relative to the baseline. Next, in Figure 6, we compare the object detection results for the proposed method without (left column) and with (right column) L_1 regularization and observe that this explicit regularization leads to more confident bounding box detections.

5. Conclusions

In this work, we proposed *TensorFact* – a novel approach for object detection to be able to capture cross-modal cues so as to generalize better to modalities with scarce training data. *TensorFact* benefits from pre-training on modalities where plenty of training data is available (such as RGB), mitigating the challenges posed by the target modality (such as IR). In our formulation, at first, the data-rich RGB modality is used to learn the common cross-modal cues using low-rank tensor factorization of the network weights. We then use the IR modality training data to only learn the cues complementary to the RGB modality (either explicitly or implicitly), thereby requiring fewer trainable parameters. We empirically validate the efficacy of our method on the task of object detection in IR images by pre-training our network on RGB object detection datasets and show that *TensorFact* yields performance boosts for object detection, in both RGB and IR images without an increase in the total number of network parameters.

References

[1] Sk Miraj Ahmed, Suhas Lohit, Kuan-Chuan Peng, Michael J Jones, and Amit K Roy-Chowdhury. Cross-modal knowledge transfer without task-relevant source data. In *European Conference on Computer Vision*, pages 111–127. Springer, 2022. 2

[2] David Arthur and Sergei Vassilvitskii. K-means++ the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, 2007. 5

[3] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 3

[4] Angela L Campbell, Rajesh R Naik, Laura Sowards, and Morley O Stone. Biological infrared imaging and sensing. *Micron*, 33(2):211–225, 2002. 2

[5] Yu Cao, Tong Zhou, Xinhua Zhu, and Yan Su. Every feature counts: An improved one-stage detector in thermal imagery. In *2019 IEEE 5th International Conference on Computer and Communications (ICCC)*, pages 1965–1969. IEEE, 2019. 3

[6] Xuerui Dai, Xue Yuan, and Xueye Wei. TIRNet: Object detection in thermal infrared images for autonomous driving. *Applied Intelligence*, 51:1244–1261, 2021. 3

[7] Chaitanya Devaguptapu, Ninad Akolekar, Manuj M Sharma, and Vineeth N Balasubramanian. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3

[8] Mayur Dhanaraj, Manish Sharma, Tiyasa Sarkar, Srivallabha Karnam, Dimitris G Chachlakis, Raymond Ptucha, Panos P Markopoulos, and Eli Saber. Vehicle detection from multi-modal aerial imagery using YOLOv3 with mid-level fusion. In *Big data II: learning, analytics, and applications*, volume 11395, pages 22–32. SPIE, 2020. 1

[9] Jeff Donahue, Judy Hoffman, Erik Rodner, Kate Saenko, and Trevor Darrell. Semi-supervised domain adaptation with instance constraints. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 668–675, 2013. 3

[10] Shuangjiang Du, Baofu Zhang, Pin Zhang, Peng Xiang, and Hong Xue. FA-YOLO: An improved YOLO model for infrared occlusion object detection under confusing background. *Wireless Communications and Mobile Computing*, 2021:1–10, 2021. 3

[11] Abhimanyu Dubey, Moitreya Chatterjee, and Narendra Ahuja. Coreset-based neural network compression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 454–470, 2018. 1, 4

[12] Özgür Erkent and Christian Laugier. Semantic segmentation with unsupervised domain adaptation under varying weather conditions for autonomous vehicles. *IEEE Robotics and Automation Letters*, 5(2):3580–3587, 2020. 3

[13] FLIR aligned. FLIR Aligned Dataset, 2020. Accessed: August 20, 2022. 2, 5

[14] Debasmita Ghose, Shasvat M Desai, Sneha Bhattacharya, Deep Chakraborty, Madalina Fiterau, and Tauhidur Rahman. Pedestrian detection in thermal images using saliency maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3

[15] Ross Girshick. Fast R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 3

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 2

[17] Dayan Guan, Jiaxing Huang, Aoran Xiao, Shijian Lu, and Yanpeng Cao. Uncertainty-aware unsupervised domain adaptation in object detection. *IEEE Transactions on Multimedia*, 24:2502–2514, 2021. 3

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[20] Christian Herrmann, Miriam Ruf, and Jürgen Beyerer. CNN-based thermal infrared person detection by domain adaptation. In *Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything*, volume 10643, pages 38–43. SPIE, 2018. 3

[21] Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 749–757, 2020. 3

[22] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021. 2

[23] Siddhartha Rao Kamalakara, Acyr Locatelli, Bharat Venkitesh, Jimmy Ba, Yarin Gal, and Aidan N Gomez. Exploring low rank training of deep neural networks. *arXiv preprint arXiv:2209.13569*, 2022. 2

[24] My Kieu, Andrew D Bagdanov, Marco Bertini, and Alberto Del Bimbo. Task-conditioned domain adaptation for pedestrian detection in thermal imagery. In *European Conference on Computer Vision*, pages 546–562. Springer, 2020. 3

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5, 6

[26] Mate Krišto, Marina Ivasic-Kos, and Miran Pobar. Thermal object detection in difficult weather conditions using YOLO. *IEEE access*, 8:125459–125476, 2020. 3

[27] M e Li, Tao Zhang, and W Cui. Research of infrared small pedestrian target detection based on YOLOv3. *Infrared Technol*, 42:176–181, 2020. 3

[28] Shuai Li, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Category dictionary guided unsupervised domain adaptation for object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1949–1957, 2021. 3

[29] Shasha Li, Yongjun Li, Yao Li, Mengjun Li, and Xiaorong Xu. YOLO-FIRI: Improved YOLOv5 for infrared image object detection. *IEEE access*, 9:141861–141875, 2021. 3

[30] Wei Li. Infrared image pedestrian detection via yolo-v3. In *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, volume 5, pages 1052–1055. IEEE, 2021. 3

[31] Yongjun Li, Shasha Li, Haohao Du, Lijia Chen, Dongming Zhang, and Yao Li. YOLO-ACN: Focusing on small target and occluded object detection. *IEEE access*, 8:227288–227303, 2020. 3

[32] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE international conference on computer vision*, pages 1261–1270, 2017. 2

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence

Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[34] Samah AF Manssor, Shaoyuan Sun, Mohammed Abdalmajed, and Shima Ali. Real-time human detection in thermal infrared imaging at night using enhanced Tiny-yolov3 network. *Journal of Real-Time Image Processing*, pages 1–14, 2022. 3

[35] Wanli Ouyang and Xiaogang Wang. Joint deep learning for pedestrian detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2056–2063, 2013. 1

[36] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2, 3

[37] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 3

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2, 3

[39] Adrian Lopez Rodriguez and Krystian Mikolajczyk. Domain adaptation for object detection via style consistency. *arXiv preprint arXiv:1911.10033*, 2019. 3

[40] Manish Sharma, Mayur Dhanaraj, Srivallabha Karnam, Dimitris G Chachlakis, Raymond Ptucha, Panos P Markopoulos, and Eli Saber. YOLOrs: Object detection in multimodal remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:1497–1508, 2020. 1

[41] Manish Sharma, Panos P Markopoulos, and Eli Saber. Yolors-lite: A lightweight cnn for real-time object detection in remote-sensing. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 2604–2607. IEEE, 2021. 2

[42] Manish Sharma, Panos P Markopoulos, Eli Saber, M Salman Asif, and Ashley Prater-Bennette. Convolutional autoencoder with tensor-train factorization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 198–206, 2021. 1

[43] Wei Shuigen, Wang Chengwei, Chen Zhen, Z Congxuan, and Z Xiaoyu. Infrared dim target detection based on human visual mechanism. *Acta Photonica Sinica*, 50(1):0110001, 2021. 3

[44] Saurav Singh, Manish Sharma, Jamison Heard, Jesse D Lew, Eli Saber, and Panos P Markopoulos. Multimodal aerial view object classification with disjoint unimodal feature extraction and fully-connected-layer fusion. In *Big Data V: Learning, Analytics, and Applications*, volume 12522, page 1252206. SPIE, 2023. 1

[45] Xiaoru Song, Song Gao, and Chaobo Chen. A multispectral feature fusion network for robust pedestrian detection. *Alexandria Engineering Journal*, 60(1):73–85, 2021. 3

[46] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 1

[47] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse R-CNN: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14454–14463, 2021. 3

[48] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang. DeepID3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015. 1

[49] Teledyne Technologies Incorporated. FLIR ADAS v1 Dataset, 2020. Accessed: August 20, 2022. 2, 5

[50] Evgeniya Ustinova and Victor Lempitsky. Learning deep embeddings with histogram loss. *Advances in neural information processing systems*, 29, 2016. 3

[51] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7464–7475, 2023. 1, 2, 5, 6

[52] Yan Wang, Junbo Yin, Wei Li, Pascal Frossard, Ruigang Yang, and Jianbing Shen. SSDA3D: Semi-supervised domain adaptation for 3D object detection from point cloud. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2707–2715, 2023. 3

[53] Xing Wei, Shaofan Liu, Yaoci Xiang, Zhangling Duan, Chong Zhao, and Yang Lu. Incremental learning based multi-domain adaptation for object detection. *Knowledge-Based Systems*, 210:106420, 2020. 3

[54] Xingxu Yao, Sicheng Zhao, Pengfei Xu, and Jufeng Yang. Multi-source domain adaptation for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3273–3282, 2021. 3

[55] Fuxun Yu, Di Wang, Yinpeng Chen, Nikolaos Karianakis, Tong Shen, Pei Yu, Dimitrios Lymberopoulos, Sidi Lu, Weisong Shi, and Xiang Chen. Unsupervised domain adaptation for object detection via cross-domain semi-supervised learning. *arXiv preprint arXiv:1911.07158*, 2019. 3

[56] Dan Zhang, Mao Ye, Yiguang Liu, Lin Xiong, and Lihua Zhou. Multi-source unsupervised domain adaptation for object detection. *Information Fusion*, 78:138–148, 2022. 3

[57] Shanghang Zhang, Guanhang Wu, Joao P Costeira, and José MF Moura. FCN-rLSTM: Deep spatio-temporal neural networks for vehicle counting in city cameras. In *Proceedings of the IEEE international conference on computer vision*, pages 3667–3676, 2017. 1

[58] Xiaofeng Zhao, Yebin Xu, Fei Wu, Wei Cai, and Zhili Zhang. IYOLO: Multi-scale infrared target detection method based on bidirectional feature fusion. In *Journal of Physics: Conference Series*, volume 1873, page 012020. IOP Publishing, 2021. 3

[59] Wenzhang Zhou, Dawei Du, Libo Zhang, Tiejian Luo, and Yanjun Wu. Multi-granularity alignment domain adaptation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9581–9590, 2022. 3