Modality Translation for Object Detection Adaptation Without Forgetting Prior Knowledge

Heitor Rapela Medeiros, Masih Aminbeidokhti, Fidel Alejandro Guerrero Peña, David Latortue, Eric Granger, and Marco Pedersoli

Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA)
Dept. of Systems Engineering, ETS Montreal, Canada
{heitor.rapela-medeiros.1, masih.aminbeidokhti.1,
fidel-alejandro.guerrero-pena.1,
david.latortue.1}@ens.etsmtl.ca,
{eric.granger, marco.pedersoli}@etsmtl.ca

Abstract. A common practice in deep learning involves training large neural networks on massive datasets to achieve high accuracy across various domains and tasks. While this approach works well in many application areas, it often fails drastically when processing data from a new modality with a significant distribution shift from the data used to pre-train the model. This paper focuses on adapting a large object detection model trained on RGB images to new data extracted from IR images with a substantial modality shift. We propose Modality Translator (ModTr) as an alternative to the common approach of fine-tuning a large model to the new modality. ModTr adapts the IR input image with a small transformation network trained to directly minimize the detection loss. The original RGB model can then work on the translated inputs without any further changes or fine-tuning to its parameters. Experimental results on translating from IR to RGB images on two well-known datasets show that our simple approach provides detectors that perform comparably or better than standard fine-tuning, without forgetting the knowledge of the original model. This opens the door to a more flexible and efficient service-based detection pipeline, where a unique and unaltered server, such as an RGB detector, runs constantly while being queried by different modalities, such as IR with the corresponding translations model. Our code is available at: https://github.com/heitorrapela/ModTr.

Keywords: Object Detection · Modality Translation · Infrared · Visual

1 Introduction

Powerful pre-trained models have become essential in the field of computer vision, particularly in object detection (OD) tasks [31,32]. These OD models are typically pre-trained on extensive natural-image RGB datasets, such as COCO [28]. Moreover, the knowledge encoded by these models can be leveraged for various tasks in a zero-shot way or with additional fine-tuning for downstream tasks [43]. However, adding new modalities to these models, such as infrared (IR), without losing the intrinsic knowledge of the detector remains a challenge [29].

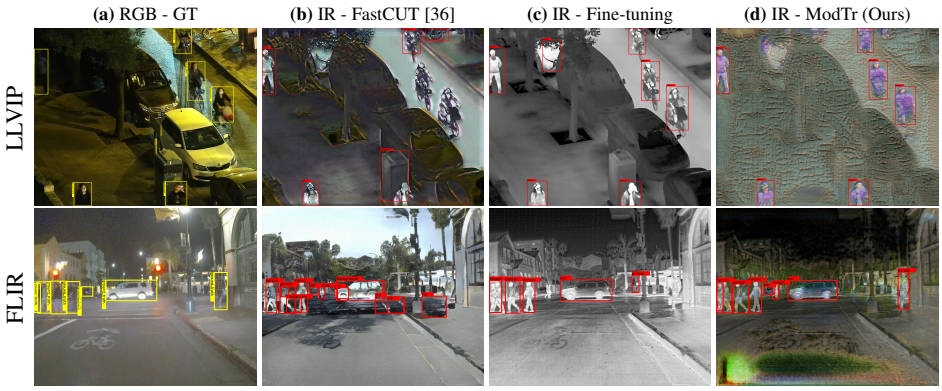

Fig. 1: Bounding box predictions over different adaptations of the RGB detector (Faster R-CNN) for IR images on two benchmarks: LLVIP and FLIR. Yellow and red boxes show the ground truth and predicted detections, respectively. In a) we see the RGB data. In b) FastCUT is an unsupervised image translation approach that takes as input infrared images (IR) and produces pseudo-RGB images. It does not focus on detection and requires both modalities for training. In c) we have fine-tuning, which is the standard approach to adapting the detector to the new modality. It requires only IR data but forgets the original knowledge of the original RGB detector. Finally, in d) is the ModTr, which focuses the translation on detection, requires only IR data and does not forget the original knowledge so that it can be reused for other tasks. Bounding box predictions for other detectors are provided in the supplementary material.

These additional modalities, though not as common as RGB images, are still important in various tasks, like surveillance [5,8], autonomous driving [33,41], and robotics [23, 38], which strive to achieve robust performance in real-world environments, where capture conditions change, such as different illumination conditions [2]. The dominant way to adapt pre-trained detectors to these novel conditions is by fine-tuning the model [29]. However, fine-tuning often results in catastrophic forgetting and can destroy the intrinsic knowledge of the detector [25]. Ideally, we would like to adapt the detector to new modalities without changing the original model. This is most useful for server-side applications, where a single model runs uninterrupted and can be queried by different inputs, ideally on different modalities. The main challenge lies in the significant distribution shift introduced by the new modality. This shift occurs because the pre-trained knowledge, such as the visual information in RGB images, differs markedly from the thermal data in IR images. This shift can degrade model performance when applied directly as input to the model, since the features learned from one modality may not be relevant or present in another. This can ultimately impact the resulting OD performance [44].

Image translation methods [35, 36] have emerged as powerful tools to overcome the downsides of fine-tuning and narrowing the gap between source and target modalities [17]. These methods do not directly work on the weight space of the original detector but rather adapt the input values to reduce the discrepancy between the source and target modalities. However, such methods often require access to source data or

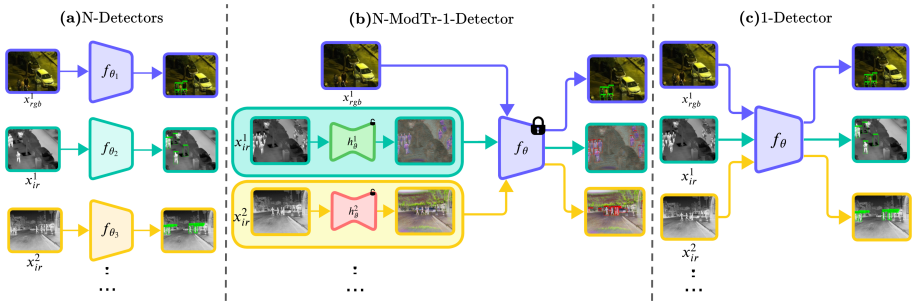

Fig. 2: Different approaches to deal with multiple modalities and/or domains. (a) The simplest approach is to use a different detector adapted to each modality. This can lead to a high level of accuracy but requires storing several models in memory. (b) Our proposed solution uses a single pre-trained model normally trained on the more abundant data (RGB) and then adapts the input through our ModTr model. (c) A single detector is jointly trained on all modalities. This allows using of a single model but requires access to all modalities jointly, which is often impossible, especially when dealing with large pre-trained models.

some statistics about it during training. Furthermore, their primary focus is on image reconstruction quality rather than the final OD task, which can cause a significant drop in performance. For instance, Figure 1 shows different ways to adapt the RGB detector (see the caption for more details).

Our work aims to improve the image translation paradigm while addressing its limitations. Our proposed approach, Modality Translation for OD (ModTr), incorporates the detector's knowledge into the translation module by training directly for the final detection task. Unlike traditional image translation methods, ModTr does not require any source data. It is a conceptually simple approach that can be easily integrated with any detector, be it a one-stage or two-stage detector. A notable application of ModTr is using a pre-trained RGB detector as a server that incorporates different ModTr blocks as input translators for new modalities such as IR. This new detector generates the desired output with performance comparable to full fine-tuning without losing the original knowledge of the pre-trained model. In Fig. 2, we present several options for integrating IR modalities into an RGB system. Fig. 2a illustrates the N-Detectors approach, where each detector is trained for a specific case. This method effectively demands more memory and forgets previously learned information. Fig. 2c shows a single detector trained on combined modalities. This method does not incur additional memory, yet it requires simultaneous access to all modalities, which may not always be feasible. Fig. 2b illustrates our proposed approach, which involves training a specialized translator for each condition without altering the parameters of the original detector. The N-ModTr-1-Detector strikes a balance between the previous methods, addressing their shortcomings by requiring only a single detector. Importantly, it retains the original pre-training knowledge, as it leaves the detector unchanged. In this work, we focus on the effectiveness of our approach for the IR modality, commonly used in surveillance and robotics, and the incremental modality detector server-based application, which is crucial for many settings that require uninterrupted detection predictions.

Our main contributions can be summarized as follows.
(1) We present ModTr, a method for adapting pre-trained ODs from large RGB datasets to new scarce modalities like IR, without requiring access to any source dataset, by translating the input signal.
(2) In contrast to standard fine-tuning, our approach does not modify the original detector weights. This allows the detector to retain the knowledge of the source data while adapting to a new modality. As a result, a single model can be used to handle multiple modalities across various translators. For instance, the same model can be used to process RGB during the daytime and IR at nighttime.
(3) An extensive empirical evaluation of ModTr in several scenarios, showcasing its advantages and flexibility. In particular, with our different proposed fusion strategies, ModTr achieves OD accuracy that is competitive when compared with image translation methods on two challenging RGB/IR datasets (LLVIP and FLIR).

2 Related Work

(a) Object Detection. OD is a computer vision task that aims to provide labels and localization for the objects in the image [47]. Two-stage detectors, exemplified by Faster R-CNN [39], generate regions of interest and then use a second classifier to confirm object presence within those regions. On the other hand, one-stage detectors streamline the detection process by eliminating the proposal generation stage, aiming for end-to-end training and real-time inference speeds. RetinaNet [27] is a one-stage OD model that utilizes a focal loss function to address class imbalance during training. Also, models like FCOS [42] have emerged in this category, eliminating predefined anchor boxes to potentially enhance inference efficiency. The proposed work investigates these three traditional and powerful detectors: Faster R-CNN, RetinaNet, and FCOS. The choice of such detectors was due to the simplicity in implementation and integration among other methods, as well as a different range of pre-trained backbone weights, such as ResNet [13] and MobileNet [15].
(b) Image Translation. Image translation is a pivotal task in computer vision, aiming to map images from a source domain to a target domain while preserving inherent content [35]. The goal is to discover a transformation function such that the distribution of images in the translated domain is aligned with the distribution of images in the target domain. The commonly used approaches for image translation are based on variational autoencoders (VAEs) [24] and generative adversarial network (GANs) [11, 35]. Isola et al. developed the Pix2Pix [21], a method that consists of a generator (based on U-Net) and a discriminator (based on GANs architecture) that work together to generate images based on input data and labels. Then, Zhu et al. proposed a method called CycleGAN [50], which is based on GANs, with the objective of unsupervised domain translation. Even though CycleGAN can produce quite visual results, it's hard to optimize due to the adversarial mechanism and memory footprint needed. In contrast, VAEs are easier to train than GANs but require more constraints in the optimization to produce images of good quality than GAN-based approaches. Recent advancements include diffusion models known for their high-quality image generation, although they may not inherently suit domain translation tasks. To enhance models such as CycleGAN, novel

methods like Contrastive Unpaired Translation (CUT) [36] and FastCUT [36] have been introduced. CUT, in particular, accelerates the image translation process by maximizing mutual information between image patches, achieving competitive results quickly. In the context of RGB/IR modality, InfraGAN presents an image-level adaptation for RGB to IR conversion, prioritizing image quality [34]. This approach is distinct in its focus on optimizing image quality losses. Moreover, Herrmann et al. have explored OD in RGB/IR modality by adapting IR images to RGB using traditional image preprocessing techniques, allowing the use of RGB object detectors without parameter modification [14]. Despite significant advances in image translation, these techniques do not specifically address OD tasks. In our previous work, we introduced HalluciDet [29], which employs an image translation mechanism for OD. However, this approach requires prior access to the source RGB data from the same domain as the target for pre-training the detector.

(c) Adapting Without Forgetting. Catastrophic forgetting (CF) is the idea that a neural network tends to forget knowledge when sequentially trained on a different task and replaces it with knowledge tailored to the new objective [45]. CF can be harmful or beneficial. Researchers identified harmful learning as situations where retaining the original knowledge while adapting to a different task is necessary. In that case, it is imperative to mitigate the risk of CF. However, some CF can also be beneficial, for instance, to prevent privacy leakage from large pre-trained models, to enhance the generalization, or to remove noisy information from the originally, acquired knowledge that is negatively affecting the new tasks. In our case, knowledge-forgetting is harmful. There are different ways to address this issue including simple techniques like decreasing the learning rate [16], use weight decay [4, 49] or mixout regularization [26] during fine-tuning or more complex approaches like Recall and learn [6], Robust Information Fine-tuning [46] or CoSDA [10]. Some adaptation methods use techniques based on replay of the source data or even using the weights of the initial model to keep some prior information [30]. Some of these works focus on adding continually different tasks in an incremental learning setting. However, these methods may still produce a loss of knowledge since the original parameters are not frozen. Furthermore, in adapting without forgetting, an adapter, which adopts a frozen pre-trained backbone to generate a representation followed by a different classifier for each downstream task [45], can be seen as a powerful method to preserve knowledge. Even though our ModTr shares some similarities, we work in the input space to adapt to the new modalities, and address this incremental modality adaptation, optimizing the translation directly for the final OD task.

3 Proposed Method

(a) Preliminary Definitions. The training set for OD is denoted as $\mathcal{D} = \{(x, Y)\}$, where $x \in \mathbb{R}^{W \times H \times C}$ represents an image in the dataset, with dimensions $W \times H$ and C channels. Subsequently, the OD model aims to identify N regions of interest within these images, denoted as $Y = \{(b_i, c_i)\}_{i=1}^{N}$. The top-left corner coordinates and the width and height of the object define each region of interest b_i. Additionally, a classification label c_i is assigned to each detected object, indicating its corresponding

class within the dataset. In this study, the number of input channels for the detector is fixed at three, corresponding to RGB-like inputs. In terms of optimization, the primary goal of this task is to maximize detection accuracy, often measured using the average precision (AP) metric across all classes. An OD is formally represented as the mapping $f_\theta : \mathbb{R}^{W\times H\times C} \rightarrow \hat{Y}$, where θ denotes the parameter vector. To effectively train a detector, a differentiable surrogate for the AP metric, referred to as the detection cost function, $\mathcal{C}_{det}(\theta)$, is employed. The typical structure of such a cost function involves computing the average detection loss over dataset $\mathcal{D}$, denoted as $\mathcal{L}_{det}$, described as:

$$\mathcal{C}_{det}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{(x,Y)\in\mathcal{D}} \mathcal{L}_{det}(f_\theta(x), Y). \tag{1}$$

(b) Modality Translation Module. Our approach primarily consists of an image-to-image translation network responsible for converting the input modality into an RGB-like space intelligible to the detector. These networks typically adopt an encoder-decoder structure to synthesize and reconstruct knowledge in a pixel-wise manner. While we employ U-Net [40] as the translation network, with parameters ϑ, in this work, our framework is general and not limited by the translation architecture. In general terms, this mapping is denoted as $h^d_\vartheta \colon \mathbb{R}^{W\times H\times C} \rightarrow \mathbb{R}^{W\times H\times 3}$, with a translation network assigned to each available input modality d. Unlike the detection network, the number of input channels varies depending on the modality, for instance, $C = 1$ for IR and depth images. It's important to note that, being a pixel-level architecture, the output of such a network retains the spatial resolution of the input. However, the number of output channels is consistently fixed at three, corresponding to RGB-like images ($C = 3$).

Unlike other image-to-image translation approaches, we drive the process using the aforementioned detection cost (Equation (1)). Thus, the underlying optimization problem is formulated as $\vartheta^* = \arg\min \mathcal{L}_{det}(\vartheta)$, incorporating the output of the composition $(f_\theta \circ h^d_\vartheta)(x)$ at the loss function level. To streamline the learning process, we utilize a residual learning strategy in which the function h^d_ϑ focuses on capturing the small variations in the input that are necessary to solve the task. This approach is similar to the one employed on diffusion models, which inspired our work. For the sake of simplicity, we separate the fusion step from the translation mapping in our notation, as various types of fusion are investigated. Consequently, the proposed image-to-image translation loss function is defined as:

$$\mathcal{L}_{\text{ModTr}}(x, Y; \vartheta) = \mathcal{L}_{det}(f_\theta \left(\Phi(h^d_\vartheta(x), x)\right), Y), \tag{2}$$

where $\Phi(.,.)$ is a non-parametric fusion function. Note that the output of $h^d_\vartheta(x)$ is an RGB-like image, whereas x may only consist of a single channel, depending on the input modality. We have chosen this definition to simplify the notation, but appropriate reshaping should be performed during implementation to ensure compatibility.

In addition, note that, while a detection loss is employed to update the translation network, the weight vector θ remains constant. This constraint is consistent with the premise of this study, where a pre-trained detector is solely available on the server side and remains unaltered. An overview of the proposed approach can be seen in Fig. 2 b).

(c) Fusion strategy. As previously mentioned, we utilize a non-parametric fusion of the intermediate representation $h^d_\vartheta(x)$ and the original input x to simplify the learning process of the translation network. In this context, we employ an element-wise product, also known as the Hadamard product, which is particularly interesting for attention mechanisms and has been explored previously for re-calibrating feature maps based on their importance [19]. Although we investigated various fusion mechanisms, the element-wise product yielded the best results. For more details on different fusion strategies, please refer to supplementary materials.

ModTr$_\odot$: The Hadamard product-based fusion serves as a gating mechanism to filter or highlight information from the input image. In this approach, the output of the translation network acts as a weight map for the input, and they are fused using pixel-wise multiplication, $\odot$. Consequently, the translation network tends to highlight information from the input when the pixel value tends toward 1 or discard it when it approaches 0. Additionally, the output translation modality can be interpreted as an attention map, as described by the following Equation (3):

$$\mathcal{L}_{\text{ModTr}_\odot}(x, Y; \vartheta) = \mathcal{L}_{det}(f_\theta \left(h^d_\vartheta(x) \odot x\right), Y). \tag{3}$$

In our design choices, we opt to utilize these straightforward non-parametric functions to assist in optimization while maintaining low inference costs.

4 Results and Discussion

4.1 Experimental Methodology

(a) Datasets: LLVIP: LLVIP is a surveillance dataset composed of $30,976$ images, in which $24,050$ ($12,025$ IR and $12,025$ RGB paired images) are used for training and $6,926$ for testing ($3,463$ IR and $3,463$ RGB paired images) with only pedestrians annotated. **FLIR ALIGNED:** We used the sanitized and aligned paired sets provided by Zhang et al. [48]. It has $10,284$ images, that is $8,258$ for training ($4,129$ IRs and $4,129$ RGBs) and $2,026$ ($1,013$ IRs and $1,013$ RGBs) for test. FLIR images are captured from the perspective of a camera in the front of a car, with a resolution of 640 by 512. It contains the bicycles, dogs, cars, and people classes. It has been found that with FLIR, the "dog" objects are inadequate for training [3], thus we decided to remove them.

(b) Implementation details: In our experiments, we randomly selected 80% of the training set for training and the rest for validation. All results reported are on the test set. As starting pre-trained weights for the detectors, we used Torchvision models with COCO [28] weights and for the U-Net translation network, we used PyTorch Segmentation Models [20] and we changed the last layer for 3-channel (RGB-like) with a Sigmoid function, to be closer to an image with values between 0 and 1, to perform translation instead of traditional segmentation. For the translation network backbones, we explored our default ResNet$_{34}$, and for subsequent studies on reducing parameters, we dive into ResNet and MobileNet-family. All the code is available on GitHub for reproducibility in the experiments. To ensure fairness, we trained the detectors under the library version

and the same experimental design, i.e., data order, augmentations, etc. Furthermore, we trained with PyTorch Lightning [9] training framework, evaluated the APs with TorchMetrics [7], and logged all experiments with WandB [1] logging tool. The different performance measures (e.g., APs) can be found in suppl. materials.

4.2 Comparison with Translation Approaches

In this section, ModTr is compared with different image-to-image translation methods employing different learning strategies. These include basic image processing strategy [14], reconstruction strategies such as CycleGAN [51], CUT [37], and FastCUT [37], which employs a contrastive learning approach, as well as HalluciDet [29], which utilizes a detection-based loss. As outlined in Table 1, we evaluated the methods based on their final detection performance across three commonly used detectors: FCOS, RetinaNet, and Faster R-CNN. The reported results are derived from the IR test set and are averaged over three different seeds, which helps mitigate the impact of randomness across runs and splits of the training and validation datasets.

For each method, we also consider its dependency on the prior knowledge data (RGB) and ground truth bounding boxes (bboxes) on the IR images. Methods that rely on reconstruction techniques do not require bbox annotations on IR images but cannot provide accurate translations for detection purposes. However, HalluciDet and ModTr require bbox annotations to adjust the input image in a discriminative manner. The main difference between HalluciDet and ModTr is the use of source images. HalluciDet requires RGB images for an initial fine-tuning of the model, while our approach can work without that fine-tuning by reusing the detector's zero-shot knowledge.

The proposed ModTr displays robustness across the three detectors and consistently exhibits improvement on two different datasets: LLVIP [22] and FLIR aligned [12]. Note that each algorithm described in Table 1 employs different training supervisions. For instance, CycleGAN employs an adversarial mechanism with both RGB and infrared modalities in an unpaired setting. Similarly, CUT and FastCUT operate with positive and negative patches in an unpaired setting. In contrast, HalluciDet doesn't require the presence of both modalities during training but employs a detection mechanism during training similar to ours. In our approach, we solely require examples from the target modality. In this section, we present the performance of our best approach $\text{ModTr}_{\odot}$. For additional results, refer to suppl. materials.

As reported in Table 1, the detection performance of ModTr over the LLVIP dataset exhibited significant improvements. Specifically, it surpassed HalluciDet, the second best, by more than 29.0 AP with both FCOS and RetinaNet architectures, while obtaining comparable results with Faster R-CNN. Such disparity with the previous technique can be attributed to the loss of previous knowledge inherent in HalluciDet, which necessitates a pre-fine-tuning strategy on the source modality. Although the performance of the FLIR dataset also improved, the dataset's inherent challenges, such as changing the background from a moving car setup, make detection more difficult. Nonetheless, our proposal consistently enhances results, with improvements of more than 11 AP for FCOS and RetinaNet, and over 7 AP for Faster R-CNN. We also observed improvements on the AP_{50} and AP_{75}. Because of the space constraint, we include these in

Table 1: Detection performance (AP) of ModTr versus baseline image-to-image methods to translate the IR to RGB-like images, using three different detectors (FCOS, RetinaNet, and Faster R-CNN). The methods were evaluated on IR test set of LLVIP and FLIR datasets. The RGB column indicates if the method required access to RGB images during training, and Box refers to the use of ground truth boxes during training.

Image translation	RGB	Box	Test Set IR (Dataset: LLVIP)		
			FCOS	**RetinaNet**	**Faster R-CNN**
Histogram Equal. [14]			31.69 ± 0.00	33.16 ± 0.00	38.33 ± 0.02
CycleGAN [51]	✓		23.85 ± 0.76	23.34 ± 0.53	26.54 ± 1.20
CUT [37]	✓		14.30 ± 2.25	13.12 ± 2.07	14.78 ± 1.82
FastCUT [37]	✓		19.39 ± 1.52	18.11 ± 0.79	22.91 ± 1.68
HalluciDet [29]	✓	✓	28.00 ± 0.92	19.95 ± 2.01	57.78 ± 0.97
ModTr$_{\odot}$ (ours)		✓	**57.63 ± 0.66**	**54.83 ± 0.61**	**57.97 ± 0.85**
Image translation	**RGB**	**Box**	**Test Set IR (Dataset: FLIR)**		
			FCOS	**RetinaNet**	**Faster R-CNN**
Histogram Equal. [14]			22.76 ± 0.00	23.06 ± 0.00	24.61 ± 0.01
CycleGAN [51]	✓		23.92 ± 0.97	23.71 ± 0.70	26.85 ± 1.23
CUT [37]	✓		18.16 ± 0.75	17.84 ± 0.75	20.29 ± 0.48
FastCUT [37]	✓		24.02 ± 2.37	22.00 ± 2.73	26.68 ± 2.59
HalluciDet [29]	✓	✓	23.74 ± 2.09	22.29 ± 0.45	29.91 ± 1.18
ModTr$_{\odot}$ (ours)		✓	**35.49 ± 0.94**	**34.27 ± 0.27**	**37.21 ± 0.46**

supplementary materials. These promising results indicate that our proposal can effectively translate images from the original IR modality to an RGB-like representation, sufficiently close to the source data to be usable by the detector.

4.3 Translation vs. Fine-tuning

In this section, we further show that the proposed approach can be trained jointly with both translation and detector, which preserves the detector's knowledge. Here, ModTr is compared to three baselines fully fine-tuning (FT), FT of the head and LoRA [18], and our best ModTr fusion strategy, as shown in Tab. 2.

We conduct LoRA fine-tuning using two settings. In the first, we apply LoRA across all layers; in the second, only to the last layer of detectors. The latter results in superior performance, so we have adopted it as our default LoRA setting. The Tab. 2 shows AP for the LLVIP and FLIR datasets, with a consistent trend across all detectors (FCOS, RetinaNet, and Faster R-CNN). Furthermore, in the case of the FLIR dataset, we observed enhancements of ModTr over the standard detector FT. As demonstrated, our approach surpasses standard fine-tuning while maintaining the detector's performance in the original modality. It is worth noting that our method also improves performance in terms of localization metrics such as AP_{50} and AP_{75} compared to fine-tuning alone, and we provide detailed results in the supplementary materials.

Table 2: Detection performance (AP) of ModTr versus baseline fine-tuning (FT) of the detector, FT of the head and LoRA [18], using three different detectors (FCOS, RetinaNet, and Faster R-CNN. The methods were evaluated on IR test set of LLVIP and FLIR datasets. Results with "-" diverged from the optimization.

Method	Test Set IR (Dataset: LLVIP)		
	FCOS	**RetinaNet**	**Faster R-CNN**
Fine-Tuning (FT)	57.37 ± 2.19	53.79 ± 1.79	**59.62 ± 1.23**
FT Head	49.11 ± 0.70	44.00 ± 0.28	59.33 ± 2.17
LoRA [18]	47.72 ± 0.58	-	54.83 ± 1.30
ModTr$_{\odot}$ (ours)	**57.63 ± 0.66**	**54.83 ± 0.61**	57.97 ± 0.85
Method	**Test Set IR (Dataset: FLIR)**		
	FCOS	**RetinaNet**	**Faster R-CNN**
Fine-Tuning (FT)	27.97 ± 0.59	28.46 ± 0.50	30.93 ± 0.46
FT Head	27.40 ± 0.12	26.78 ± 0.70	33.53 ± 0.36
LoRA [18]	-	-	29.44 ± 0.61
ModTr$_{\odot}$ (ours)	**35.49 ± 0.94**	**34.27 ± 0.27**	**37.21 ± 0.46**

4.4 Different Backbones for ModTr

In this context, we evaluate ModTr and examine the trade-off between performance and parameter cost. It is widely recognized that increasing the number of parameters can enhance performance, but this relationship is not strictly linear. We demonstrated that models with fewer parameters can still achieve good performance; for example, MobileNet$_{v2}$, with fewer parameters than ResNet$_{18}$, sometimes outperformed it. This trade-off highlights the versatility of the model, which can be deployed with MobileNet-based architectures and utilized in low-cost devices. In Table 3, the default number of parameters is successfully reduced from 24.4M (ResNet$_{34}$) to 6.6M using MobileNet$_{v2}$ while maintaining similar performance. For instance, on LLVIP, MobileNet$_{v2}$ achieved a mean AP of 56.15, comparable to 56.35 AP$_{50}$ from ResNet$_{34}$ (others APs and detectors are reported in the supplementary material).

This approach opens up new possibilities, particularly in scenarios where using one translation network and one detector (e.g., one ModTr and one detector for RGB/IR) proves advantageous. This setup requires a total of 44.9M parameters, compared to 83.6M parameters, when employing two detectors—one for each modality (for example, for Faster R-CNN). Similar reductions in parameter costs were observed for FCOS (from 66.4M to 36.3M) and RetinaNet (from 68M to 37.1M) when using one detector for both modalities while preserving the knowledge of the previous modality and incorporating a new one. These numbers are based on MobileNet$_{v3s}$, which strikes a balance between performance and the number of parameters, making it suitable for memory-restricted systems. The complete evaluations for FCOS and RetinaNet are included in the supplementary material.

Table 3: Detection performance (AP) of ModTr with different backbones for the translation networks with different numbers of parameters, using three different detectors (FCOS, RetinaNet, and Faster R-CNN). The methods were evaluated on IR test set of LLVIP and FLIR datasets.

Test Set IR (Dataset: LLVIP)		
Method	**Parameters**	**AP** ↑
Faster R-CNN	41.8 M	
MobileNet$_{v3s}$	+ 3.1 M	54.51 ± 0.28
MobileNet$_{v2}$	+ 6.6 M	56.15 ± 0.51
ResNet$_{18}$	+ 14.3 M	55.53 ± 1.14
ResNet$_{34}$	+ 24.4 M	56.35 ± 0.65
Test Set IR (Dataset: FLIR)		
Faster R-CNN	41.8 M	
MobileNet$_{v3s}$	+ 3.1 M	32.06 ± 0.75
MobileNet$_{v2}$	+ 6.6 M	36.77 ± 0.67
ResNet$_{18}$	+ 14.3 M	36.68 ± 0.22
ResNet$_{34}$	+ 24.4 M	37.21 ± 0.46

4.5 Knowledge Preservation through Input Modality Translation

ModTr is designed to prevent catastrophic forgetting by keeping the weights of the pre-trained detector fixed. In this section, we demonstrate how various adaptation paradigms, shown in Figure 2, effectively solve the final task while preserving intrinsic knowledge. We compare our proposed method, ModTr, with two fine-tuning baseline methods. The first baseline method involves N-detectors, each fine-tuning the target modality individually. The second baseline method employs a single detector trained on the joint modality using balanced sampling. Note that while a copy of the original detector can be used in the N-detectors paradigm, it is unavailable in the 1-detector paradigm because the original modality is assumed to be inaccessible during training.

In all scenarios, we use COCO as the pre-training dataset and LLVIP and FLIR as target domains. Specifically, in the N-detectors scenario (Fig.2a), we fine-tune one detector on each dataset and use a copy of the original detector for the RGB modality. In the 1-detector scenario (Fig.2c), we fine-tune one detector on the combined FLIR and LLVIP datasets. In the N-ModTr-1-Detector scenario (Fig.2b), two translators are trained, one per dataset. To assess catastrophic forgetting, we re-evaluate each scenario on COCO-val.

Table 4 shows the final performance. While all adaptation paradigms achieve relatively similar performance, the 1-detector method completely fails in the zero-shot scenario. The N-detectors method mitigates this by duplicating the detector three times. In contrast, ModTr preserves knowledge using a single detector and three efficient translators, demonstrating its practicality for embedded devices, as it requires less memory. Based on the average performance on all datasets, ModTr obtains the best results.

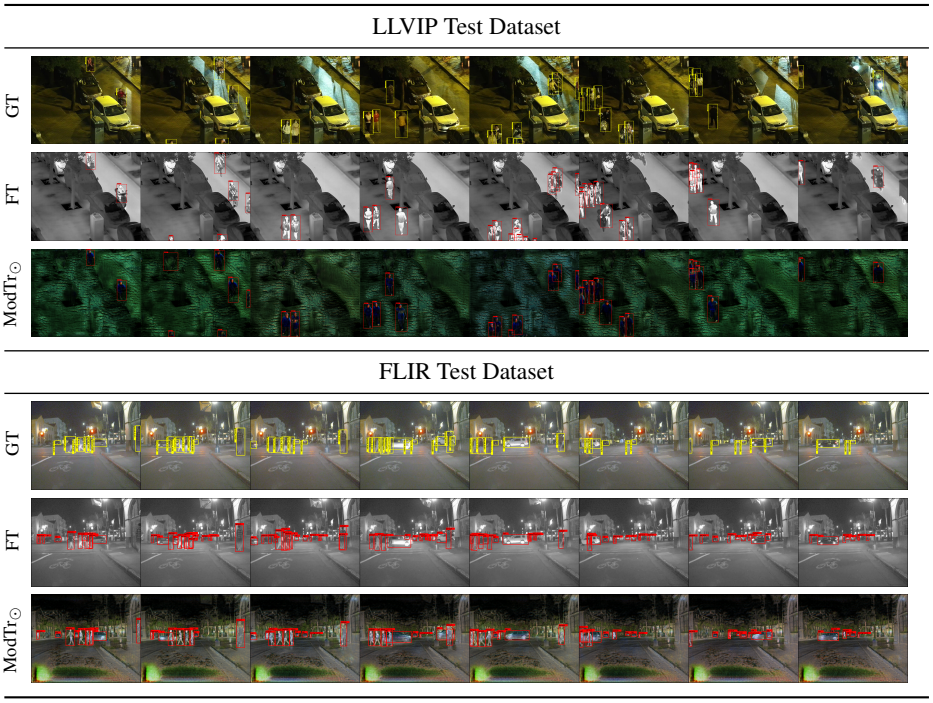

Fig. 3: Illustration of a sequence of 8 images of LLVIP and FLIR dataset for Faster R-CNN. For each dataset, the first row is the RGB modality, followed by the IR modality and different representations created by ModTr. For visualizations of other detectors and variants of ModTr, please refer to the supplementary materials.

4.6 Visualization of ModTr Translated Images

In Figure 3, we present qualitative results for LLVIP and FLIR, alongside a comparison with fine-tuning. Each dataset section includes three rows: the first row displays the ground-truth RGB images, the second row showcases the results of fine-tuning using IR, and the last row features ModTr with a Hadamard product-based fusion over the Faster R-CNN detector. Due to space constraints, additional visualizations for other detectors and fusion strategies are provided in the supplementary materials. Notably, the IR results exhibit some false positives, particularly when detected objects overlap. Our method mitigates this issue effectively. Further insights, provided in the supplementary materials, reveal how our method effectively blurs or removes objects that do not belong to the target classes, thereby enhancing detection accuracy. Although the obtained intermediate representations are not visually pleasant, they prove more efficient for incorporating the knowledge necessary for the OD. Additionally, we conducted experiments with loss function terms aimed at enhancing the visual effects of the image, but they were not conclusive in terms of helping the detection performance.

Table 4: Detection performance (AP) of knowledge preserving techniques N-Detectors, 1-Detector, and N-ModTr-1-Detector, using three different detectors (FCOS, RetinaNet, and Faster R-CNN). The methods were evaluated on COCO and IR test sets of LLVIP and FLIR datasets.

Detector	Dataset	N-Detectors	1-Detector	N-ModTr-1-Det.
FCOS	LLVIP	57.37 ± 2.19	**58.55 ± 0.89**	57.63 ± 0.66
	FLIR	27.97 ± 0.59	26.70 ± 0.48	**35.49 ± 0.94**
	COCO	**38.41 ± 0.00**	00.33 ± 0.04	**38.41 ± 0.00**
	AVG.	41.25 ± 0.92	28.52 ± 0.47	**43.84 ± 0.53**
RetinaNet	LLVIP	53.79 ± 1.79	53.26 ± 3.02	**54.83 ± 0.61**
	FLIR	28.46 ± 0.50	25.19 ± 0.72	**34.27 ± 0.27**
	COCO	**35.48 ± 0.00**	00.29 ± 0.01	**35.48 ± 0.00**
	AVG.	39.24 ± 0.76	26.24 ± 1.28	**41.52 ± 0.29**
Faster R-CNN	LLVIP	59.62 ± 1.23	**62.50 ± 1.29**	57.97 ± 0.85
	FLIR	30.93 ± 0.46	28.90 ± 0.33	**37.21 ± 0.46**
	COCO	**39.78 ± 0.00**	00.40 ± 0.00	**39.78 ± 0.00**
	AVG.	43.44 ± 0.56	30.60 ± 0.54	**44.98 ± 0.43**

4.7 Fine-tuning of ModTr and the Detector

The main reason to use ModTr is to avoid fine-tuning the detector for a specific task so that it can preserve its knowledge and be used for multiple modalities. However, in this section, we consider what would happen if we learn jointly ModTr and the detector weights. Results are reported in Figure 4. We see that fine-tuning the detector can further boost performance. Thus, another application of ModTr could be used to improve the fine-tuning of a detector with a reduced additional computational cost.

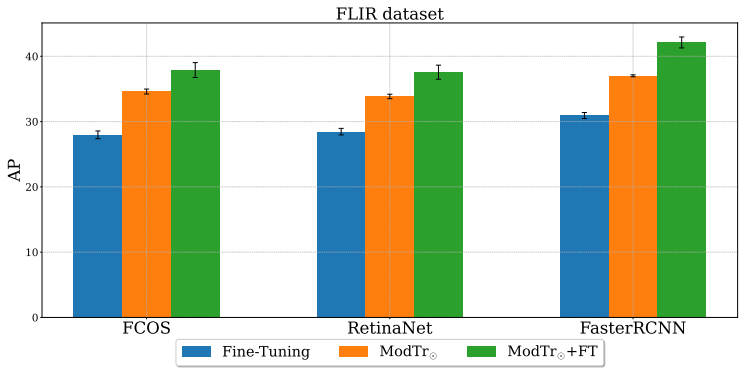

Fig. 4: Comparison of the performance of fine-tuning the ModTr and normal fine-tuning on the FLIR dataset for the three different detectors (FCOS, RetinaNet, and Faster R-CNN). In blue, the Fine-tuning; in orange, the ModTr$_{\odot}$, and in green, ModTr$_{\odot}$ + FT.

5 Conclusion

In this paper, a novel method called ModTr is proposed for adapting RGB object detectors (ODs) for IR modality without changing their parameters. A key advantage of our approach is that it preserves the full knowledge of the detector, allowing the translation network to act as a node that changes the modality for an unaltered detector. This is much more flexible and computationally efficient than having a specialized OD for each modality. Our approach performs well in various settings, outperforming powerful image-to-image models and previous competitors. We evaluated ModTr for different tasks, including detection based on image translation, comparison with traditional fine-tuning, and incremental IR modality application. Experimental results show the high performance and versatility of our method in all these settings.

Additionally, to explore integrating modalities beyond IR, we applied ModTr to Canny edges extracted from IR images as detailed in the supplementary material. While ModTr significantly enhances the performance of zero-shot RGB OD on edges, it still does not match the effectiveness of full fine-tuning on this other modality. We believe this shortfall arises from the limited information provided by edges compared to the richer data in the IR modality, leading to lower initial zero-shot OD performance. A potential solution is to replace the deterministic translator module within ModTr with a generative model. This substitution could enrich modality information by generating the missing data, potentially improving the zero-shot detector's performance. This promising direction will be explored in future research.

Acknowledgments

This work was supported in part by Distech Controls Inc., the Natural Sciences and Engineering Research Council of Canada, the Digital Research Alliance of Canada, and MITACS.

References

1. Biewald, L.: Experiment tracking with weights and biases (2020), software available from wandb.com
2. Bustos, N., Mashhadi, M., Lai-Yuen, S.K., Sarkar, S., Das, T.K.: A systematic literature review on object detection using near infrared and thermal images. Neurocomputing p. 126804 (2023)
3. Cao, Y., Bin, J., Hamari, J., Blasch, E., Liu, Z.: Multimodal object detection by channel switching and spatial attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 403–411 (2023)
4. Chelba, C., Acero, A.: Adaptation of maximum entropy capitalizer: Little data can help a lot. Computer Speech and Language **20**(4), 382–399 (2006). https://doi.org/https://doi.org/10.1016/j.csl.2005.05.005, https://www.sciencedirect.com/science/article/pii/S0885230805000276
5. Chen, J., Li, K., Deng, Q., Li, K., Philip, S.Y.: Distributed deep learning model for intelligent video surveillance systems with edge computing. IEEE Transactions on Industrial Informatics (2019)

6. Chen, S., Hou, Y., Cui, Y., Che, W., Liu, T., Yu, X.: Recall and learn: Fine-tuning deep pretrained language models with less forgetting. CoRR **abs/2004.12651** (2020), https://arxiv.org/abs/2004.12651
7. Detlefsen, N.S., Borovec, J., Schock, J., Jha, A.H., Koker, T., Di Liello, L., Stancl, D., Quan, C., Grechkin, M., Falcon, W.: Torchmetrics-measuring reproducibility in pytorch. Journal of Open Source Software **7**(70), 4101 (2022)
8. Dubail, T., Guerrero Peña, F.A., Medeiros, H.R., Aminbeidokhti, M., Granger, E., Pedersoli, M.: Privacy-preserving person detection using low-resolution infrared cameras. In: European Conference on Computer Vision. pp. 689–702. Springer (2022)
9. Falcon, W., The PyTorch Lightning team: PyTorch Lightning (Mar 2019). https://doi.org/10.5281/zenodo.3828935
10. Feng, H., Yang, Z., Chen, H., Pang, T., Du, C., Zhu, M., Chen, W., Yan, S.: Cosda: Continual source-free domain adaptation (2023)
11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
12. Group, F., et al.: Flir thermal dataset for algorithm training (2018)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Herrmann, C., Ruf, M., Beyerer, J.: Cnn-based thermal infrared person detection by domain adaptation. In: Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything. vol. 10643, p. 1064308. International Society for Optics and Photonics (2018)
15. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
16. Howard, J., Ruder, S.: Fine-tuned language models for text classification. CoRR **abs/1801.06146** (2018), http://arxiv.org/abs/1801.06146
17. Hsu, H.K., Yao, C.H., Tsai, Y.H., Hung, W.C., Tseng, H.Y., Singh, M., Yang, M.H.: Progressive domain adaptation for object detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 749–757 (2020)
18. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9
19. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
20. Iakubovskii, P.: Segmentation models pytorch (2019)
21. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on (2017)
22. Jia, X., Zhu, C., Li, M., Tang, W., Zhou, W.: Llvip: A visible-infrared paired dataset for low-light vision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3496–3504 (2021)
23. Jing, C., Potgieter, J., Noble, F., Wang, R.: A comparison and analysis of rgb-d cameras' depth performance for robotics application. In: 2017 24th International Conference on Mechatronics and Machine Vision in Practice (M2VIP). pp. 1–6. IEEE (2017)
24. Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2022)
25. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)

26. Lee, C., Cho, K., Kang, W.: Mixout: Effective regularization to finetune large-scale pretrained language models (2020)
27. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
28. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
29. Medeiros, H.R., Pena, F.A.G., Aminbeidokhti, M., Dubail, T., Granger, E., Pedersoli, M.: Hallucidet: Hallucinating rgb modality for person detection through privileged information. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1444–1453 (2024)
30. Menezes, A.G., de Moura, G., Alves, C., de Carvalho, A.C.: Continual object detection: A review of definitions, strategies, and challenges. Neural Networks (2023)
31. Minderer, M., Gritsenko, A., Houlsby, N.: Scaling open-vocabulary object detection. Advances in Neural Information Processing Systems **36** (2024)
32. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., et al.: Simple open-vocabulary object detection. In: European Conference on Computer Vision. pp. 728–755. Springer (2022)
33. Natan, O., Miura, J.: End-to-end autonomous driving with semantic depth cloud mapping and multi-agent. IEEE Transactions on Intelligent Vehicles **8**(1), 557–571 (2022)
34. Özkanoğlu, M.A., Ozer, S.: Infragan: A gan architecture to transfer visible images to infrared domain. Pattern Recognition Letters **155**, 69–76 (2022)
35. Pang, Y., Lin, J., Qin, T., Chen, Z.: Image-to-image translation: Methods and applications (2021)
36. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 319–345. Springer (2020)
37. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: European Conference on Computer Vision (2020)
38. Pierson, H.A., Gashler, M.S.: Deep learning in robotics: a review of recent research. Advanced Robotics **31**(16), 821–835 (2017)
39. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28**, 91–99 (2015)
40. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
41. Stilgoe, J.: Machine learning, social learning and the governance of self-driving cars. Social studies of science **48**(1), 25–56 (2018)
42. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9627–9636 (2019)
43. Vasconcelos, C., Birodkar, V., Dumoulin, V.: Proper reuse of image classification features improves object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13628–13637 (2022)
44. Wang, Q., Chi, Y., Shen, T., Song, J., Zhang, Z., Zhu, Y.: Improving rgb-infrared object detection by reducing cross-modality redundancy. Remote Sensing **14**(9), 2020 (2022)
45. Wang, Z., Yang, E., Shen, L., Huang, H.: A comprehensive survey of forgetting in deep learning beyond continual learning (2023)

46. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Lopes, R.G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al.: Robust fine-tuning of zero-shot models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7959–7971 (2022)
47. Zhang, A., Lipton, Z.C., Li, M., Smola, A.J.: Dive into deep learning. arXiv preprint arXiv:2106.11342 (2021)
48. Zhang, H., Fromont, E., Lefèvre, S., Avignon, B.: Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In: 2020 IEEE International Conference on Image Processing (ICIP). pp. 276–280. IEEE (2020)
49. Zhang, T., Wu, F., Katiyar, A., Weinberger, K.Q., Artzi, Y.: Revisiting few-sample bert fine-tuning (2021)
50. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Computer Vision (ICCV), 2017 IEEE International Conference on (2017)
51. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)