Early feature distributions alignment in visible-to-thermal unsupervised domain adaptation for object detection

Adrien Maglo, Romaric Audigier

Early Feature Distributions Alignment in Visible-to-Thermal Unsupervised Domain Adaptation for Object Detection

Adrien Maglo and Romaric Audigier

Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
{adrien.maglo,romaric.audigier}@cea.fr

Abstract. Infrared or thermal images are used in many civilian and military applications to detect objects due to the heat they emit, especially when environmental conditions such as nighttime or adverse weather prevent the use of visible images. To train an object detector based on a deep neural network, a significant amount of annotated data is required to achieve good detection performance. However, annotations for infrared images are often unavailable and costly to obtain. Besides, the trained model may show poor robustness against the change of thermal sensor. Therefore, unsupervised domain adaptation (UDA) methods have been proposed to train an object detector with annotated visible images, which are easily available, and unannotated infrared images. This paper presents a new visible-to-thermal UDA approach for object detection based on Deformable-DETR with hybrid matching. Our approach aims to establish common features between visible and thermal images at the earliest stages of the backbone network. The feature distributions extracted from visible and thermal images are aligned thanks to discriminator networks and adversarial learning. Gradient images are also used as a domain translation of input images to ease the alignment. Detection performance is further improved by randomly masking tokens at the input of the transformer. Experiments on public datasets demonstrate that our method consistently outperforms previous works.

Keywords: unsupervised domain adaptation · object detection with transformers · thermal imaging · feature alignment

1 Introduction

Thermal infrared images are used in many applications in both military and civilian domains. They enable the detection of people and objects by capturing the heat they emit, which is particularly useful in nighttime or adverse weather conditions. The detection task involves generating bounding boxes around objects and classifying them. Many detectors have been proposed in the literature based on deep learning approaches, achieving good performance across a wide range of objects. However, they require a significant amount of training data. Large detection datasets in the visible domain, such as MS-COCO [30], have

been proposed to train these models. Despite efforts to release large-scale infrared datasets [21, 11], annotated datasets in the infrared domain are much less common than in the visible domain. Additionally, using different sensors (response, quality, sensitivity, etc.) under varying weather conditions may result in thermal images with different distributions.

The challenges related to thermal data collection and annotation have led to the development of unsupervised domain adaptation (UDA) methods from the visible to the thermal domain, allowing the utilization of knowledge from the more readily available visible domain. The model is trained using both visible and thermal images. However, annotations are only provided in the visible domain. Consequently, the model learns the task in the thermal domain through pseudo-labeling or feature distribution alignment. Some previous works have focused on visible-to-thermal UDA for classification and segmentation tasks, while only Marnissi et al. proposed an approach for the object detection task [31]. The UDA for the detection task from one visible domain to another visible domain has been widely studied in the literature [32]. However, UDA from the visible domain to the thermal domain presents a different challenge. Approaches must contend with two distinct domains that possess very distinctive features. Additionally, thermal images have a single component, while RGB images have three.

We therefore propose a new visible-to-thermal UDA detection framework that aims to early align the distribution of the features extracted from both domains. Our framework is based on the Deformable-DETR detector with hybrid matching (H-Deformable-DETR) [22]. Many previous works align the distribution of features after the backbone, at the detection stage of the model. We demonstrate that early alignment of the features within the backbone can be beneficial for the visible-to-thermal domain adaptation task. Our detection model takes multi-scale backbone features as input. We propose to align the distribution of these features from the two domains using discriminator networks and adversarial training. Furthermore, we align the visible and thermal images by using gradient images as a common translated input modality for the model. Gradient images extracted from visible and thermal images are indeed much more similar than the original images. Finally, we apply token masking to the input of the detector transformer to improve its robustness.

The remainder of the paper is organized as follows. In the second section, we introduce previous work about UDA for detection and visible-to-thermal domain adaptation. We describe our method in the third section. Experimental results on two public datasets are provided in the fourth section.

2 Related work

Unsupervised domain adaptation (UDA) involves training a model with annotated data in the source domain and unannotated data in the target domain. This technique enables the training of a model adapted to a target domain without requiring annotation for the target data. In the literature, various UDA methods have been proposed for classification tasks, segmentation tasks, and

detection tasks. In this section, we will first review previous work related to UDA for object detection. Subsequently, we will delve into the specific case of visible-to-thermal UDA.

2.1 Unsupervised domain adaptation for object detection

The UDA methods for object detection can be classified into three main categories [32]: pseudo-labeling, domain invariant feature learning and image-to-image translation.

Pseudo-labeling frameworks generate annotations for the target images using confident detections obtained by a model trained on the source data. Soft labeling is employed in the framework proposed by RoyChowdhury et al. [35] to mitigate the risk of incorrect pseudo-labels. In the approach by Khodabandeh et al. [24], bounding boxes are generated by the detection model trained on labeled source data while pseudo-label classes are provided by an additional image classifier. Kim et al. [25] proposed an algorithm that mines positive samples and weak-negative samples for each class of pseudo-labels. Zhao et al. [47] introduced a method that aligns features by minimizing the discrepancy between the Faster R-CNN region proposal network and the region proposal classifier. Other approaches utilize a mean-teacher architecture where the teacher model generates pseudo-labels to train the student. The teacher weights are then updated from the student model using exponential moving average (EMA). Cai et al. [2] perform random augmentations on a target image to obtain two images, ensuring the consistency of student predictions between them. In the recent MIC approach [18], the student network is trained by matching the pseudo-labels it generates on masked target images with those generated by the teacher. The Harmonious Teacher method [10] focuses on improving the consistency between classification scores and the Intersection-Over-Union between predicted and real object bounding boxes.

Domain invariant feature learning methods focus on aligning the features extracted by the model between the source and target domains. This is often achieved by adding discriminator modules to the original detector, which learn to classify whether the images come from the source domain or the target domain. Thus, the objective for the detector is to generate common features between the two domains. Therefore, a gradient reversal layer [13] is often added between the feature outputs and the discriminator in order to achieve this contradictory goal. The approach of Chen et al. [6] aligns the features produced by a Faster R-CNN model [34] at both the instance level and global image level. Its extension [7] integrates a feature pyramid network to independently align the image and object features of each scale. In the framework of Saito et al. [36], local image-based features and global instance-based features are extracted and aligned at two different levels of the network. Hsu et al. [19] align the instance features at the center of the object proposals. MeGA-CDA [39] aligns the features with a discriminator at the global level and a discriminator for each category. Since the object categories are unknown for the target images, memory-guided attention maps redirect the features to each discriminator. Li et al. [29] use a

mean-teacher architecture and integrate a discriminator to align the distribution of features generated by the student network.
Graph reasoning techniques model the relations between objects and categories in the source and target images as graphs. The framework proposed by Xu et al. [42] aligns the detected object proposals by merging them. It also aligns the object classes between domains by improving the compactness of each class and its separability from others. Similarly, I3Net [5] follows the same alignment objectives. It weights the target samples based on adaptation difficulty, boosts foreground objects, suppresses redundant background information, and aligns category features between domains using consistency regularization. SIGMA [28] transforms model features into graphs and employs graph matching theory to align class feature distributions.

Image-to-image translation methods involve using a model to convert images from one domain to another. Chen et al. [4] utilize CycleGAN [48] for generating synthetic samples and enhancing the training of the adversarial domain discriminator. Similarly, CycleGAN is employed by Hsu et al. [20] to create synthetic annotated images. Subsequently, their features are aligned with target image features using an adversarial discriminator. Deng et al. [9] use images translated from the source to the target domain with CycleGAN to mitigate the bias of the teacher and student networks towards their trained domain.

The methods listed above are based on convolutional detectors, with the most frequent one being Faster R-CNN. However, recent approaches have also been proposed for Deformable-DETR detectors [49] based on transformers [38]. MTTrans [43] utilizes a mean teacher approach for pseudo-label generation. The method proposed by Wang et al. employs a feature alignment strategy [40]. DA-DETR [46] adds feature fusion modules to enable information communication across channels. These approaches do not directly align features at multiple output levels of the backbone. While multiresolution feature alignment has been studied for Faster R-CNN detectors [17, 41], it has never been used with DETR architectures. Visible and thermal images have very different characteristics. We believe that early feature alignment is important for efficient visible-to-thermal UDA. Multiresolution feature alignment in the backbone network can achieve this objective.

2.2 Visible-to-thermal domain adaptation

The unsupervised domain adaptation from visible-to-thermal images has received less attention in the literature. For the semantic segmentation task, MS-UDA [26] performs UDA from a large visible dataset to a smaller unlabeled visible and thermal paired dataset using pseudo-label generation. Gan et al. [12] employ domain-specific attention maps for segmentation and classification tasks. The network is trained with adversarial learning and fine-tuned with pseudo-labels. Akkaya et al. [1] select high-confidence pseudo-labels that fool a trained domain discriminator. Regarding the detection task, Lee et al. proposed a GAN-based visible to thermal image translation method [27] that focuses on preserving the edges. It is trained on a combination of large visible and thermal datasets.

They conducted thermal detection experiments by training a VFNet detector [44] on a synthetic dataset [23] translated using their method. To our knowledge, only Marnissi et al. [31] have attempted visible-to-thermal UDA for detection. Their UDAT framework, based on Faster R-CNN, requires annotations only in the visible domain. It also aligns features at four different feature map levels and instance levels. Given the distinct characteristics of visible and thermal images, our approach proposes to align thermal and visible features at shallower levels of the network so that deeper levels can benefit from features with common distributions. It also aligns visible and thermal images by using gradient images, as a common translation domain, at the input of the model.

3 Our method

3.1 Overview

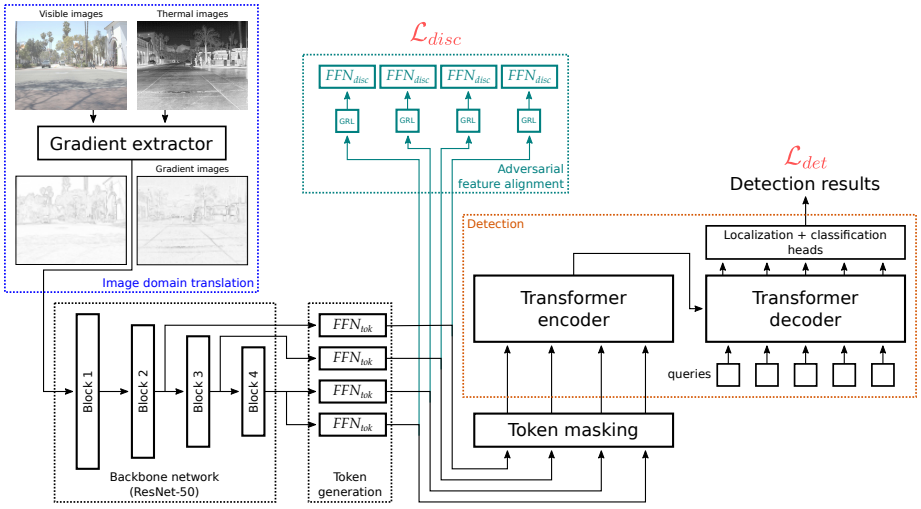

Fig. 1. Illustration of the training of our visible-to-thermal unsupervised domain adaptation method for object detection based on H-Deformable-DETR. The model is trained with a supervised detection loss $\mathcal{L}_{det}$ on visible images. The distribution of features extracted from visible and thermal images are aligned with adversarial learning. The discriminative loss $\mathcal{L}_{disc}$ trains the discrimination networks FFN_{disc} connected to the token generators FFN_{tok} of the backbone output levels through gradient reversal layers GRL (details in section 3.3). Gradient images are also used as a common modality for backbone inputs (details in section 3.4). Token masking improves the detection robustness (details in section 3.5).

Our framework is based on the H-Deformable-DETR detector [22]. In order to detect objects in the thermal domain, for which we do not use annotated

training data, we simultaneously train our detector in a supervised manner with annotated training data from the visible domain and align the distribution of features extracted at several levels of the backbone network. We refer to this strategy as "early alignment" (EA) because the prioritized feature distribution alignments occur at the shallowest output levels of the backbone network. This alignment is performed using discriminator networks connected to the model through gradient reversal layers. We use the image gradient operation as an image domain translation method. The gradient images, extracted from visible and thermal images, serve as inputs to the model. They reduce the domain gap between the visible and thermal images. Finally, we apply token masking at the output of the backbone to increase the model's generalization.

3.2 Baseline detector

The H-Deformable-DETR detector belongs to the *DETR* family of detectors [3] based on transformers [38]. With the *Deformable-DETR* detector [49], the input images are first processed by a ResNet-50 backbone [16] that extracts features at a single resolution. These features are then transformed into tokens by the feedforward networks FFN_{tok}. The shallowest output layer of the ResNet-50 backbone is discarded, and an additional output layer is artificially added by applying another FFN_{tok} to the last output layer. Positional embedding is then added to the tokens. They are processed by a transformer encoder and then by a transformer decoder. The decoder takes as additional inputs object queries that are learned parameters. The model predicts for each decoder output token a class score and a bounding box. During training, each ground-truth object is associated with a decoder query using Hungarian matching based on class scores and bounding box IoU criteria. Deformable-DETR replaces the transformer attention modules with multi-scale deformable attention modules. Instead of computing an attention map for all input feature locations, the deformable attention module is trained to sample only a few significant points around the reference point. This sampling is done at different feature-map scales. It speeds up the model training and improves the detection of small objects. The hybrid matching of *H-Deformable-DETR* increases performance by employing a second round of ground truth and object query matching. In this round, each ground truth can be assigned to multiple decoder queries from a second set of queries. We utilize the two-stage variant of H-Deformable-DETR [49]. The encoder generates object proposals, and the proposals with the highest scores are selected to be refined by the transformer decoder. Their bounding boxes are fed to the transformer decoder as positional embeddings of the decoder object queries. The model is trained in a supervised way with the images and the annotations of the visible domain. The supervised detection losses are the same as with the original H-Deformable-DETR. We call their sum $\mathcal{L}_{det}$.

3.3 Early feature distribution alignment

Our main objective is to build a detector that has a high performance in the thermal domain. Consequently, we want our backbone to generate domain agnostic tokens. Our model should produce features with the same distribution for the thermal or visible images. The features learnt with annotated visible images should also be a good representation of thermal images. To this end, we add at each output of the backbone, gradient reversal layers GRL followed by discriminator networks FFN_{disc}. The role of the discriminators is to classify the tokens: they determine whether tokens come from a thermal image or a visible image. Each discriminator is composed of 5 linear layers with the same dimension as the transformer. The first 4 layers are followed by a ReLU activation function. The output dimension of the last layer is 1. Our discriminator learns to classify tokens coming from either thermal or visible images. However, we aim for backbone features from both domains to be indistinguishable. Therefore, the GRL inverses the signs of the gradients to enable the adversarial learning between the backbone and the discriminators. We use a cross entropy loss to train the discriminator networks:

$$\mathcal{L}_{disc} = -\sum_{l} w_l \sum_{t} y_{l,t} \times log(x_{l,t}) + (1 - y_{l,t}) \times log(1 - x_{l,t})$$

where l is the output layer of the backbone network, w_l is a weight for the layer l, t is the token, $x_{l,t}$ is the output of the discriminator network for the layer l and token t and $y_{l,t}$ is its target value. Features extracted by the backbone network at the shallowest layers are more of spatial nature while features extracted at the deepest levels of the network are more semantic, so less dependent from the input domain. As we want an early alignment of the features, we set much higher weight w_l to the shallower layer outputs than to the deeper ones.

During training, we build mini-batches with one half of the images coming from the visible domain and the other from the thermal domain. The adversarial discriminative loss $\mathcal{L}_{disc}$ applies to both thermal and visible images. The feature alignment task and the detection tasks have two objectives that may disturb each other. The feature alignment task may want to generate completely uniform distribution of features so the discriminator is unable to determine whether they come from visible or thermal images. To balance the importance of the feature alignment task relative to the detection task, we dynamically weight $\mathcal{L}_{\text{disc}}$ with the coefficient α based on the value of $\mathcal{L}_{\text{det}}$, ensuring that a constant ratio r_{loss} between the two losses is maintained:

$$\frac{\alpha \mathcal{L}_{disc}}{\mathcal{L}_{det}} = r_{loss}$$

where r_{loss} is a constant positive parameter set for the entire training. At each iteration, $\mathcal{L}_{det}$ and $\mathcal{L}_{disc}$ are first computed on the total of the mini-batch of images. Then α is determined with the formula:

$$\alpha = \frac{r_{loss} \mathcal{L}_{det}}{\mathcal{L}_{disc}}.$$

No gradient is backpropagated before the determination of α. It becomes a scaling constant for the computation of the total loss:

$$\mathcal{L}_{tot} = \mathcal{L}_{det} + \alpha \mathcal{L}_{disc}.$$

Notice that α is forced to zero for the first epoch in order to bootstrap the detector without the discriminative loss. This mechanism improves the stability of the training.

3.4 Input gradient images

Fig. 2. Visible and thermal images from the Free FLIR dataset [11] (top) and their respective Sobel gradient intensity images (bottom). Despite the fact that the visible and thermal gradient images do not outline the same visual features, the domain gap appears to be narrower than with the original images.

We use the gradient images as a common modality to reduce the domain gap between the visible and thermal. The Sobel [37] and Prewitt [33] image gradients have the advantage of being quick to compute. Their intensity images have a similar appearance between visible and thermal images, as depicted in Figure 2. The gradient outlines edges in the input images, which is crucial for detecting objects in both domains. We compute the gradients for each axis with the following convolutions:

$$\mathbf{G_{Prewitt_x}} = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix} * \mathbf{I} \quad ; \quad \mathbf{G_{Prewitt_y}} = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ +1 & +1 & +1 \end{bmatrix} * \mathbf{I}$$

$$\mathbf{G_{Sobel_x}} = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} * \mathbf{I} \quad ; \quad \mathbf{G_{Sobel_y}} = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} * \mathbf{I}$$

Gradient images are then obtained by taking the L^2 norms of the gradients:

$$\mathbf{G_{Sobel}} = \sqrt{\mathbf{G_{Sobel_x}}^2 + \mathbf{G_{Sobel_y}}^2} \quad ; \quad \mathbf{G_{Prewitt}} = \sqrt{\mathbf{G_{Prewitt_x}}^2 + \mathbf{G_{Prewitt_y}}^2}$$

The gradient images are then normalized between 0 and 1. During training, we randomly switch between the Sobel and Prewitt gradients to artificially augment the amount of training data. During inference, only the Sobel gradient is used.

3.5 Detector token masking

Image masking in the pixel space has shown its effectiveness for UDA [18]. Instead, we propose to leverage this mechanism by randomly masking the transformer tokens from all input levels. Token masking has also been shown to be beneficial for pretraining Vision Transformer models [15]. In our case, it aims at forcing the model to rely on features from all the input levels of the transformer by reducing the overfitting. Token masking is performed by randomly selecting a random ratio α_m of token at the input of the transformer encoder and setting their value to 0. Gradient backpropagation is halted for the masked tokens.

4 Experiments

4.1 Dataset

To run our experiments, we use the Free FLIR "aligned" dataset [45], version derived from the original version 1.3 [11]. It provides annotations for 4,129 well-aligned thermal and visible image pairs for training and 1,013 image pairs for testing. However, in this UDA work no alignment is used: the visible and thermal images from the training set are used in an unpaired way during training. The testing is performed on the thermal images from the test set. In addition, only the person, bicycle and car classes are considered.

Experiments are also performed on the KAIST dataset [21]. We use the "sanitized" annotations and the image sets provided in the latest release of the dataset, selecting one out of every four images for the train set and one out of every twenty images for the test set. In line with previous evaluation protocols on this dataset [45, 31], only instances annotated with the class "person", "person?"

or "people" are kept and grouped in a common "pedestrian" class. Bounding boxes with the minimum of the width and height inferior to 50 pixels or flagged as occluded are discarded. At the end, only images with at least a valid bounding box are used for training and testing. This resulted in a dataset of 4,110 image pairs with 7,908 instances in visible images for training and 859 thermal images with 1,846 instances for testing.

4.2 Implementation details

We built our framework on top of the implementation of H-Deformable-DETR [14] based on the Pytorch framework. Our model utilizes a ResNet-50 backbone [16] pretrained on ImageNet [8], with the remaining parts of the model initialized with random weights. As base configuration for the H-Deformable-DETR, we chose the two-stage configuration from the official implementation that performs the best on the MS-COCO dataset [30]. Thus, the number of queries for the one-to-one matching is set to 300. For the one-to-many matching, each ground truth is set to one of 1500 queries. The weight of the one-to-many matching loss is set to 1. The mixed selection is used. The dimension inside the transformer is set to 256 and its feed-forward network dimension is set to 2048. The data augmentation techniques of Deformable-DETR are used: random horizontal flip, crop and resize. In our experiments, all the model layers are trained during 12 epochs with the AdamW optimizer on two NVIDIA RTX A5500 GPUs with 24 GB or RAM. The learning rate is set to 2×10^{-5} for the backbone network and 2×10^{-4} for the rest of the network. It is divided by 10 after 11 epochs. The weight decay is set to 10^{-4}. The batch size is set to 4: two random visible images and two random thermal images. The w_l feature distribution alignment weights are set to 10, 1, 10^{-4} and 10^{-5} for shallower to deeper layers, respectively. The token masking ratio α_m is set to 0.2. The ratio between the discrimination and the detection loss r_{loss} was fine-tuned to 0.32 after a grid-search on the Free FLIR dataset. The same value of r_{loss} is used for the experiments on the KAIST dataset. We observed that the concurrency between the supervised detection loss $\mathcal{L}_{det}$ and the discrimination loss $\mathcal{L}_{disc}$ can lead to training instabilities and catastrophic detection performance. Disabling $\mathcal{L}_{disc}$ for the first training epoch removed this issue in our experiments.

4.3 Results

Experimental results on the Free FLIR dataset are reported in Table 1. We use the mAP metric with an IoU of 0.5. We compare our method to existing UDA state-of-the-art approaches generaly evaluated on visible-to-visible benchmarks [6, 7, 36, 4, 40, 18, 10]. Only UDAT [31] is specialized in visible-to-thermal UDA. Some experimental results of previous work have been originally reported by Marnissi et al. [31]. For methods we trained and evaluated, we provide mean and standard deviation values over four different runs. Our method outperforms all previous works in terms of mAP. It reaches an average mAP of 68.4 % on the Free FLIR dataset, about 4.9 percentage points (pp) higher than the SOTA

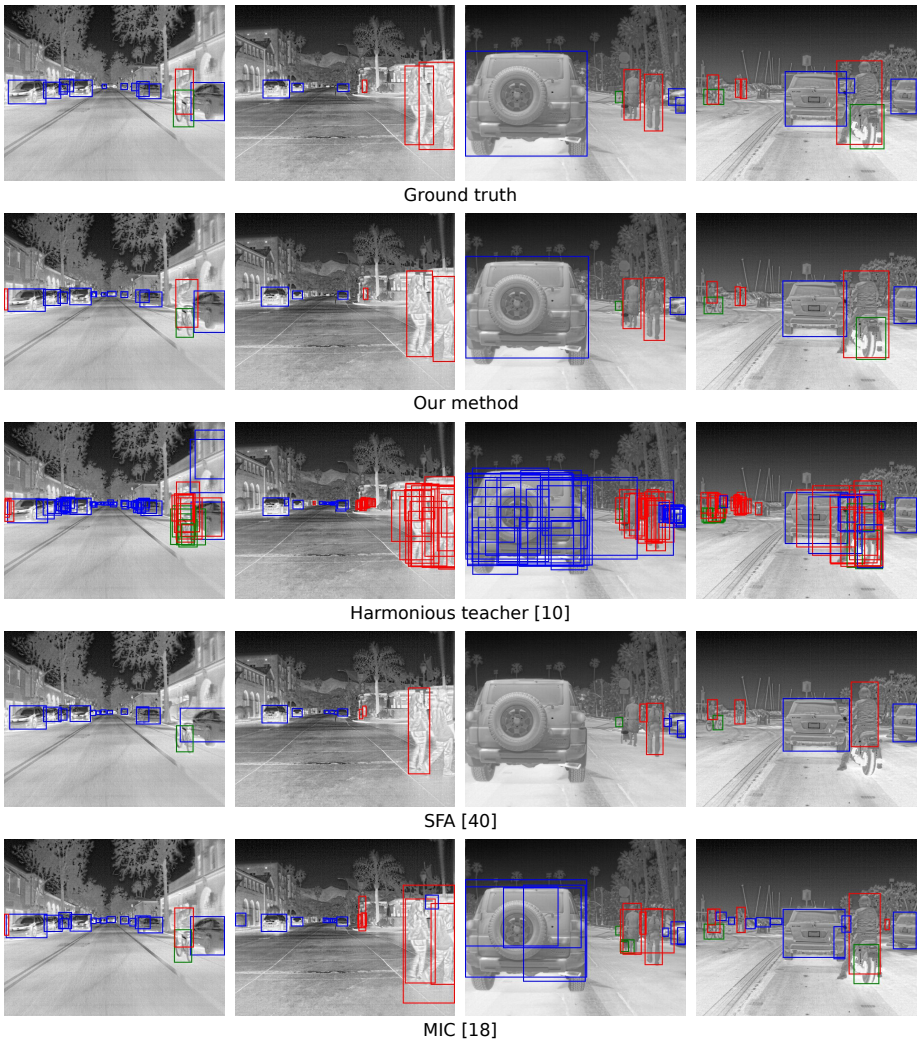

Fig. 3. Qualitative detection results on the thermal images of Free FLIR dataset. The red, blue and green bounding boxes correspond respectively to the person, car and bicycle classes. The score threshold is set to 0.4 for all the methods. No Non-Maximum Suppression is used.

Table 1. Performance in mAP (%) on the Free FLIR aligned dataset without using alignment. Results marked by * have been originally reported by Marnissi et al. [31]. The others show mean and standard deviation of the mAP we obtained by training on 4 runs each. EA stands for the early alignment of features distributions; grad. img. for the use of gradient images; mask. token. for the random masking of the transformer encoder input tokens.

Method	Car	Bicycle	Person	Average mAP
DA-faster [6]*	59.90	24.30	26.60	36.93
SWDA [36]*	58.96	32.02	32.32	41.40
HTCN [4]*	56.37	37.95	33.17	42.49
SA-DA-faster [7]*	70.38	33.30	47.27	50.30
UDAT [31]*	66.83	49.34	43.41	53.19
MIC [18]	67.89 ±5.81	48.45 ±5.53	57.20 ±6.85	57.85 ±5.96
SFA [40]	77.33 ±1.37	45.14 ±3.66	55.58 ±2.23	59.35 ±1.44
Harmonious teacher [10]	78.36 ±1.25	45.67 ±0.71	66.46 ±1.43	63.50 ±0.79
Baseline	68.11 ±2.09	49.91 ±2.01	45.15 ±2.90	54.38 ±2.25
EA	75.29 ±1.16	53.35 ±0.94	50.46 ±1.56	59.70 ±1.04
EA + grad. img.	**82.81** ±0.39	51.27 ±2.14	67.47 ±1.25	67.18 ±1.00
EA + grad. img. + mask. tok.	82.76 ±0.47	**54.55** ±1.31	**67.92** ±0.69	**68.42** ±0.32

method Harmonious teacher [10]. In order to demonstrate the performance improvements brought by each of our components, we conducted an ablation study. The results are provided in Table 1. The "Baseline" method corresponds to the H-Deformable-DETR detector trained on visible images and tested on thermal images without any adaptation. The early alignment of feature distributions improves the mAP by approximately 5.3 pp. The use of gradient images results in an additional improvement of around 7.5 pp. Finally, random token masking enhances the mAP by 1.2 pp. Some qualitative detection results on the Free FLIR dataset are provided in Figure 3.

Experimental results on the KAIST dataset are reported in Table 2. We compare our method with the previous works that performed best on the Free FLIR dataset. Surprisingly, the Harmonious teacher did not perform so well in this benchmark, whereas our approach outperforms the best SOTA method, MIC [18], by about 4.1 pp. We also conducted an ablation study on this dataset, which shows performance increase for each of the components of our approach.

4.4 Discussion

Our approach consistently outperforms previous works on the Free FLIR and KAIST datasets. It uses less computational resources during training than mean teacher approaches [10, 18] that must store, at least, two versions of the model in memory. Our method is easily implemented on top of the efficient H-Deformable-DETR detector that has available source code [14]. It uses the same initial 48 million parameters with only 1 million extra parameters for the domain discriminators FFN_{disc} during training.

Table 2. Performance in mAP (%) on the KAIST dataset.

Method	mAP
Harmonious teacher [10]	32.79 ±2.08
SFA [40]	37.86 ±0.80
MIC [18]	41.95 ±3.38
Baseline	26.45 ±3.69
Early alignment	34.04 ±3.35
Early alignment + gradient images	42.65 ±2.62
Early alignment + gradient images + masked tokens	**46.04** ±1.94

5 Conclusion

We present in this paper a new visible-to-thermal unsupervised domain adaptation method based on an efficient H-Deformable-DETR detector. We demonstrate that early feature distribution alignment combined with image domain translation through gradient images is key to achieving good detection performance in the thermal domain. For future work, we aim to study the performance of our method on thermal images captured by various sensors under different weather and temperature conditions. Additionally, we plan to explore the applicability of the early alignment and gradient translation principles to segmentation approaches.

Acknowledgements This work benefited from a government grant managed by the French National Research Agency (ANR-22-ASTR-0010-02) and the FactoryIA supercomputer financially supported by the Ile-de-France Regional Council.

References

1. Akkaya, I.B., Altinel, F., Halici, U.: Self-training guided adversarial domain adaptation for thermal imagery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4322–4331 (2021)
2. Cai, Q., Pan, Y., Ngo, C.W., Tian, X., Duan, L., Yao, T.: Exploring object relation in mean teacher for cross-domain detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11457–11466 (2019)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
4. Chen, C., Zheng, Z., Ding, X., Huang, Y., Dou, Q.: Harmonizing transferability and discriminability for adapting object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8869–8878 (2020)
5. Chen, C., Zheng, Z., Huang, Y., Ding, X., Yu, Y.: I3net: Implicit instance-invariant network for adapting one-stage object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12576–12585 (2021)
6. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster R-CNN for object detection in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3339–3348 (2018)

7. Chen, Y., Wang, H., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Scale-aware domain adaptive faster R-CNN. International Journal of Computer Vision **129**(7), 2223–2243 (2021)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255 (2009)
9. Deng, J., Li, W., Chen, Y., Duan, L.: Unbiased mean teacher for cross-domain object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4091–4101 (2021)
10. Deng, J., Xu, D., Li, W., Duan, L.: Harmonious teacher for cross-domain object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23829–23838 (2023)
11. Free teledyne flir thermal dataset for algorithm training. https://www.flir.com/oem/adas/adas-dataset-form/, accessed: 2024-03-08
12. Gan, L., Lee, C., Chung, S.J.: Unsupervised rgb-to-thermal domain adaptation via multi-domain attention network. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 6014–6020 (2023)
13. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of machine learning research **17**(59), 1–35 (2016)
14. Official implementation of the paper "DETRs with hybrid matching". https://github.com/HDETR/H-Deformable-DETR, accessed: 2024-04-05
15. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
16. 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)
17. He, Z., Zhang, L.: Multi-adversarial faster-rcnn for unrestricted object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6668–6677 (2019)
18. Hoyer, L., Dai, D., Wang, H., Van Gool, L.: Mic: Masked image consistency for context-enhanced domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11721–11732 (2023)
19. Hsu, C.C., Tsai, Y.H., Lin, Y.Y., Yang, M.H.: Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 733–748. Springer (2020)
20. 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)
21. Hwang, S., Park, J., Kim, N., Choi, Y., So Kweon, I.: Multispectral pedestrian detection: Benchmark dataset and baseline. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1037–1045 (2015)
22. Jia, D., Yuan, Y., He, H., Wu, X., Yu, H., Lin, W., Sun, L., Zhang, C., Hu, H.: DETRs with hybrid matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19702–19712 (2023)
23. Johnson-Roberson, M., Barto, C., Mehta, R., Sridhar, S.N., Rosaen, K., Vasudevan, R.: Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? arXiv preprint arXiv:1610.01983 (2016)

24. Khodabandeh, M., Vahdat, A., Ranjbar, M., Macready, W.G.: A robust learning approach to domain adaptive object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 480–490 (2019)
25. Kim, S., Choi, J., Kim, T., Kim, C.: Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6092–6101 (2019)
26. Kim, Y.H., Shin, U., Park, J., Kweon, I.S.: MS-UDA: Multi-spectral unsupervised domain adaptation for thermal image semantic segmentation. IEEE Robotics and Automation Letters **6**(4), 6497–6504 (2021)
27. Lee, D.G., Jeon, M.H., Cho, Y., Kim, A.: Edge-guided multi-domain RGB-to-TIR image translation for training vision tasks with challenging labels. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 8291–8298 (2023)
28. Li, W., Liu, X., Yuan, Y.: Sigma: Semantic-complete graph matching for domain adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5291–5300 (2022)
29. Li, Y.J., Dai, X., Ma, C.Y., Liu, Y.C., Chen, K., Wu, B., He, Z., Kitani, K., Vajda, P.: Cross-domain adaptive teacher for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7581–7590 (2022)
30. 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: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
31. Marnissi, M.A., Fradi, H., Sahbani, A., Essoukri Ben Amara, N.: Feature distribution alignments for object detection in the thermal domain. The Visual Computer **39**(3), 1081–1093 (2023)
32. Oza, P., Sindagi, V.A., Sharmini, V.V., Patel, V.M.: Unsupervised domain adaptation of object detectors: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
33. Prewitt, J.M., et al.: Object enhancement and extraction. Picture processing and Psychopictorics **10**(1), 15–19 (1970)
34. 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** (2015)
35. RoyChowdhury, A., Chakrabarty, P., Singh, A., Jin, S., Jiang, H., Cao, L., Learned-Miller, E.: Automatic adaptation of object detectors to new domains using self-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 780–790 (2019)
36. Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Strong-weak distribution alignment for adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6956–6965 (2019)
37. Sobel, I.: An isotropic 3x3 image gradient operator. Presentation at Stanford A.I. Project 1968 (02 2014)
38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
39. Vs, V., Gupta, V., Oza, P., Sindagi, V.A., Patel, V.M.: MeGA-CDA: Memory guided attention for category-aware unsupervised domain adaptive object detec-

tion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4516–4526 (2021)
40. Wang, W., Cao, Y., Zhang, J., He, F., Zha, Z.J., Wen, Y., Tao, D.: Exploring sequence feature alignment for domain adaptive detection transformers. In: Proc. of the 29th ACM International Conference on Multimedia. pp. 1730–1738 (2021)
41. Xie, R., Yu, F., Wang, J., Wang, Y., Zhang, L.: Multi-level domain adaptive learning for cross-domain detection. In: Proceedings of the IEEE/CVF international conference on computer vision workshops (2019)
42. Xu, M., Wang, H., Ni, B., Tian, Q., Zhang, W.: Cross-domain detection via graph-induced prototype alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12355–12364 (2020)
43. Yu, J., Liu, J., Wei, X., Zhou, H., Nakata, Y., Gudovskiy, D., Okuno, T., Li, J., Keutzer, K., Zhang, S.: Mttrans: Cross-domain object detection with mean teacher transformer. In: European Conference on Computer Vision. Springer (2022)
44. Zhang, H., Wang, Y., Dayoub, F., Sunderhauf, N.: Varifocalnet: An IoU-aware dense object detector. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8514–8523 (2021)
45. Zhang, H., Fromont, E., Lefevre, 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 (2020)
46. Zhang, J., Huang, J., Luo, Z., Zhang, G., Zhang, X., Lu, S.: DA-DETR: Domain adaptive detection transformer with information fusion. In: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23787–23798 (2023)
47. Zhao, G., Li, G., Xu, R., Lin, L.: Collaborative training between region proposal localization and classification for domain adaptive object detection. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16. pp. 86–102. Springer (2020)
48. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
49. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)