Hindawi
Mobile Information Systems
Volume 2021, Article ID 1837894, 6 pages
https://doi.org/10.1155/2021/1837894

Research Article

Unsupervised Image-Generation Enhanced Adaptation for Object Detection in Thermal Images

Peng Liu,[1] Fuyu Li,[2] Shanshan Yuan,[1] and Wanyi Li[2]

[1]*China National Institute of Standardization, Beijing, China*
[2]*Institute of Automation, Chinese Academy of Sciences, Beijing, China*

Correspondence should be addressed to Wanyi Li; wanyi.li@ia.ac.cn

Received 3 November 2021; Accepted 10 December 2021; Published 27 December 2021

Academic Editor: Hye-jin Kim

Object detection in thermal images is an important computer vision task and has many applications such as unmanned vehicles, robotics, surveillance, and night vision. Deep learning-based detectors have achieved major progress, which usually need large amount of labelled training data. However, labelled data for object detection in thermal images is scarce and expensive to collect. How to take advantage of the large number labelled visible images and adapt them into thermal image domain is expected to solve. This paper proposes an unsupervised image-generation enhanced adaptation method for object detection in thermal images. To reduce the gap between visible domain and thermal domain, the proposed method manages to generate simulated fake thermal images that are similar to the target images and preserves the annotation information of the visible source domain. The image generation includes a CycleGAN-based image-to-image translation and an intensity inversion transformation. Generated fake thermal images are used as renewed source domain, and then the off-the-shelf domain adaptive faster RCNN is utilized to reduce the gap between the generated intermediate domain and the thermal target domain. Experiments demonstrate the effectiveness and superiority of the proposed method.

1. Introduction

Thermal cameras capture passively the infrared radiation emitted by all objects with a temperature above absolute zero [1]. Vision systems using thermal cameras can eliminate the illumination problems of normal greyscale and RGB cameras. Object detection in thermal images is a very important computer vision task and has many applications including unmanned vehicles, robotics, surveillance, night vision, industrial, and military.

Deep learning-based detectors, such as faster RCNN [2], SSD [3], and YOLO [4], have achieved major progress in visible domain, which usually need large amount of labelled training data. However, labelled thermal images for training object detectors are scarce and expensive to collect, while there are large amount of labelled visible images. Thus, it is expected to make use of these annotated visible images and adapt them into thermal image domain for object detection. This problem is referred as domain adaptive object detection from visible to thermal.

The research on object detection in thermal images under domain adaptation context is not as developed as that with color, including only several methods. Herrmann et al. [5] proposed to transform the thermal IR data as close as possible to the RGB domain via basic image processing operations and fine-tune the pretrained CNN-based detector on preprocessed data. Guo et al. [6] presented an approach to pedestrian detection in thermal infrared images with limited annotations. The authors tackled the domain shift between thermal and color images by learning a pair of image transformers to convert images between the two modalities, jointly with a pedestrian detector. For general domain adaptive object detection, [7] is the first work to deal with the domain adaptation problem for object detection. The authors conducted adversarial training on features and designed three adaptation components to deal with domain shift, i.e., image-level adaptation, instance-level adaptation, and consistency check. Existing deep domain adaptive object detection (DDAOD) works can be mainly categorized adversarial-based, reconstruction-based, and hybrid. Detailed review can be found in [8].

Comparing to the abovementioned works, to our best knowledge, this paper is the first work to deal with unsupervised adaptive object detection from visible-to-thermal domain. The contributions of this work mainly consist of the following three aspects:

(1) We propose an unsupervised image-generation enhanced adaptation method for object detection in thermal images, in which an image-generation module and a readaptation module are included.

(2) To reduce the gap between visible domain and thermal domain, an image-generation process is designed. The image-generation process consists of a CycleGAN-based image-to-image translation and an intensity inversion transformation.

(3) We conduct extensive experiments to compare the proposed methods with other methods, where it yields notable performance gains.

2. Proposed Method

In this section, we present details of our proposed unsupervised image-generation enhanced domain adaptive thermal object detector. Figure 1 shows the overview framework. It consists of two modules, image generation and readaptation. The image-generation module generates simulated fake thermal images by a CycleGAN image translation process and an intensity inversion transformation. The readaptation module firstly takes the generated fake thermal images as renewed source domain and the real thermal as target domain and then conducts an off-the-shelf domain adaptive faster RCNN for object detection. Trained detector can be applied to the thermal target domain. More details are provided in the following subsections.

2.1. Image Generation. To reduce the gap between the visible source domain and the thermal target domain, we design an image-generation module to generate simulated images that are similar to target images. The module consists two steps, a CycleGAN [9] step for translating visible image to thermal style, and an intensity inversion step to diversify the appearance of generated fake thermal images.

2.1.1. Image Translation via CycleGAN [9]. CycleGAN is an unpaired image-to-image translation method. In this paper, the goal of CycleGAN [9] is to learn a mapping $G_T: V \longrightarrow T$ such that the distribution of images from $G_T(V)$ is indistinguishable from the distribution T using an adversarial loss. Because this mapping is highly underconstrained, G_T is coupled with an inverse mapping $G_V: T \longrightarrow V$ and introduces a cycle consistency loss to enforce $G_V(G_T(V)) \approx V$ (and vice versa). V represents the color visible domain and T represents the thermal domain. The objective of CycleGAN to minimize is shown as follows:

$$\begin{aligned}\mathscr{L}(G_T, G_V, D_V, D_T) = {} & \mathscr{L}_{\text{GAN}}(G_T, D_T, V, T) \\ & + \mathscr{L}_{\text{GAN}}(G_V, D_V, T, V) \\ & + \lambda \mathscr{L}_{cyc}(G_T, G_V).\end{aligned} \tag{1}$$

In equation (1), $\mathscr{L}_{GAN}(G_T, D_T, V, T)$ and $\mathscr{L}_{GAN}(G_V, D_V, T, V)$ are the adversarial losses of mapping function G_T and G_V, respectively; $\mathscr{L}_{cyc}(G_T, G_V)$ is the cycle consistency loss. λ denotes the relative importance of the adversarial losses and cycle consistency loss. The optimization problem to solve is

$$G_T^*, G_V^* = arg \min_{G_T, G_V} \max_{D_V, D_T} \mathscr{L}(G_T, G_V, D_V, D_T). \tag{2}$$

Translated fake thermal images for demonstration are shown in Figure 2. Images of the left column are from color visible domain, of the middle column are generated fake thermal images, and of the right column are real ground truth thermal images.

2.1.2. Intensity Inversion. The generated fake thermal images and the real ground truth thermal images are compared in Figures 2(b) and 2(c). It is likely that the generated fake thermal images are with the contents of the color visible domain images and with the style of the thermal domain images. However, the intensity of specific target object region is opposite, such as person region. From Figures 2(b) and 2(c), it is shown that the intensity of person region in fake images is low while that of real thermal images is high. We argue that if we train detectors using only images similar to Figure 2(b), the detector will miss the objects with inverse intensity. This argument is shown in our experiments; details can be found in the ablation study, i.e., Section 3.3.

Based on the above analysis, we propose to augment the generated fake thermal images by an intensity inversion transformation. The augmentation is expected to diversify the appearance of labelled training data and improve the performance of the object detector. The proposed intensity inversion transformation is defined as follows:

$$f_{\text{inv}}: T_{\text{inv}} = 255 - T. \tag{3}$$

In equation (3), the invert function f_{inv} corresponds to the intensity inversion transformation, T denotes the fake thermal image to invert which is an eight bit image, and T_{inv} denotes the inverted image.

Examples of intensity inversion transformation are shown in Figure 3. The appearance of object region in inverted images becomes similar to that of real thermal images.

2.2. Readaptation. After doing the image-generation module, we take the union of generated fake thermal images and inverted fake thermal images as renewed source domain, which is defined as

$$S_{\text{renewed}}: \left\{D_{S_{\text{renewed}}}, B_{D_V} | D_{S_{\text{renewed}}} = D_{FT} \bigcup D_{FT_{Inv}}\right\}, \tag{4}$$

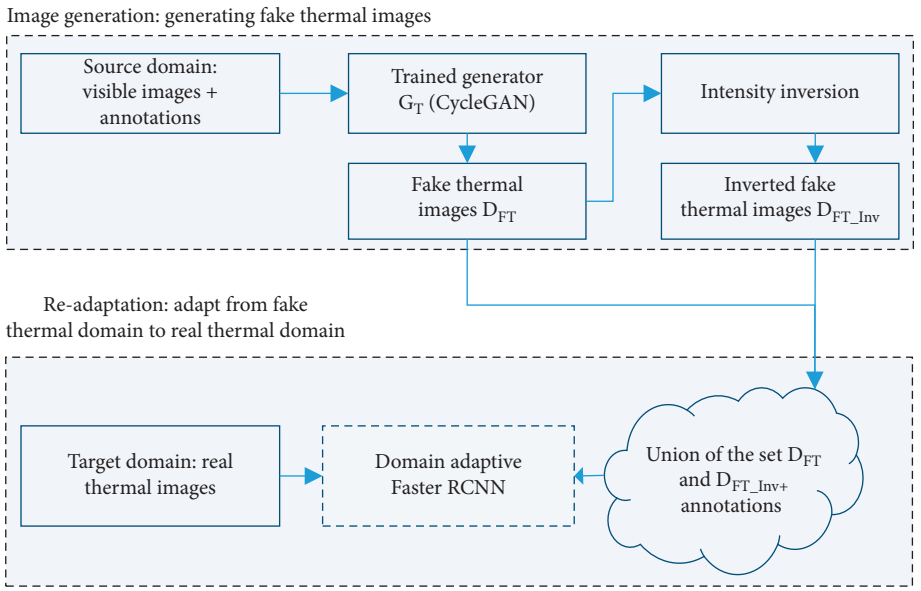

Figure 1: Overview framework of our proposed unsupervised image-generation enhanced adaptive thermal object detector.

Figure 2: Generated fake thermal images for example. (a) Color visible images; (b) generated fake thermal images; and (c) real ground truth thermal images. The color images from top to down are 000492.png and 000505.png in RGB folder of multispectral object detection dataset [10].

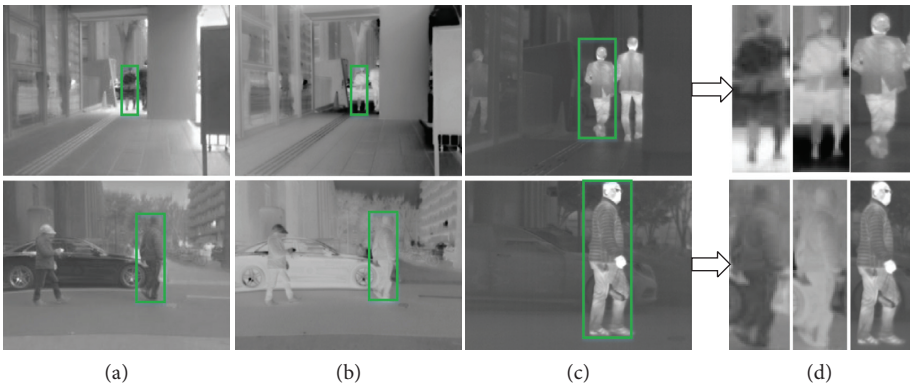

Figure 3: Illustration of intensity inversion transformation. (a) Generated fake thermal images; (b) inverted fake thermal images; (c) real ground truth thermal images, and (d) cropped object regions. The images from top to down are 000492.png and 000505.png of multispectral object detection dataset [10].

where S_{renewed} denotes the renewed source domain, which consists of the generated image set $D_{S_{\text{renewed}}}$ and the annotations B_{D_V}, $D_{S_{\text{renewed}}}$ is the union of the generated fake thermal image set D_{FT} and the inverted fake thermal image set $D_{FT_{Inv}}$, D_V denotes the image set of color visible domain V, and B_{D_V} are the annotations of D_V, noted that the renewed source domain $\mathrm{D}_{S\_\text{renewed}}$ is with double number of D_V and with annotations transferred from D_V.

Intuitively, we can train detector on annotated $\mathrm{D}_{S\_\text{renewed}}$ directly and apply it to target domain T. However, there still exists gap between D_{S_renewed} and T. Thus, we utilize an off-the-shelf domain adaptive faster RCNN [7] (referred as DAF) to conduct a readaptation from D_{S_renewed} to T.

DAF [7] uses H-divergence to measure the divergence between data distribution of source domain and target domain. The authors formulate the object detection as a posterior learning problem in a probabilistic perspective, that is, $P(C, B|I)$, where I is the image, B is the bounding box of an object, and C is the category of the object. Based on the H-divergence measure and the probabilistic formulation, three adaptation components are proposed, i.e., image-level adaptation, instance-level adaptation, and consistency regularization. Three adaptation components are trained jointly with adversarial learning.

3. Experiments

In this section, various experiments are conducted to evaluate the effectiveness of the proposed method. In Section 3.1, we introduce the experiments setup including dataset, evaluation metric, and implementation. In Section 3.2, we compare the proposed method with the state-of-the-art methods in accuracy performance. Finally, in Section 3.3, we analyze and discuss the impact of each module in ablation study.

3.1. Setup

3.1.1. Dataset. In order to evaluate the proposed method, we conduct experiments on multispectral object detection dataset [10]. The multispectral object detection dataset [10] is collected for autonomous vehicles. It consists of RGB, NIR, MIR, and FIR images and added ground truth labels. There are total 7,512 images (3,740 taken at daytime and 3,772 taken at night time). Bounding box coordinates and labels are consisted in the ground truth. Four different images are simultaneously captured and each object is annotated in the spectral images. In this dataset, five class objects (*bike*, *car*, *car_stop*, *color_cone*, *and person*) are labelled. In our experiments, the RGB images with annotations are set as source domain, and the FIR, i.e., thermal images, are set as target domain. The annotations of thermal images are not used during the training process.

3.1.2. Evaluation Metric. To assess the performance of object detector, we adopt the widely used mean average precision (mAP) as evaluation criteria, which is calculated by recall and precision.

Recall (R) and precision (P) are used to get AP value of each class. The mAP means the mean value of AP for all categories. They are defined as follows:

$$AP = \int_0^1 P(R)dR\ , \tag{5}$$

$$mAP = \frac{1}{N_{cls}} \sum_{i=1}^{N_{cls}} AP_i, \tag{6}$$

where N_{cls} represents the number of categories.

3.1.3. Implementation Details. Our experiments are implemented on PyTorch [11] platform. For CycleGAN, the open source PyTorch version [12] is used. The CycleGAN is trained with 200 echoes. For the readaptation part, we use an open source PyTorch implementation [13]. Faster RCNN and DAF are both trained with 20 echoes and parameters are set as default.

3.2. Comparison with the State-of-the-Art Methods. In this section, we evaluate the detection performance quantitatively and qualitatively. In quantitative part, mAP of the faster RCNN [2] trained on source data, the baseline and also the state-of-the-art method domain adaptive faster RCNN [7] (referred as DAF), and our proposed method are compared. In qualitative part, we compare the proposed method to the state-of-the-art method DAF [7].

3.2.1. Quantitative Evaluation. Table 1 summarizes the experimental results of different methods. We compare the proposed method with faster RCNN [2] trained on source data and domain adaptive faster RCNN [7] (referred as DAF). The DAF is trained on annotated source data and unlabeled target data. The proposed method is trained on generated images with annotations of original color visible domain. Faster RCNN trained on annotated target samples is taken as oracle. The proposed method achieved the mAP of 26.5%, while faster RCNN (nonadapted) achieved 1.4%, and DAF achieved 19.4%. Our method outperforms DAF with 8.8%.

3.2.2. Qualitative Evaluation. Some qualitative results are shown in Figures 4 and 5. As shown in Figure 4, faster RCNN cannot detect the person in the middle of the image; DAF can only detect part of the car in the left. In Figure 5, faster RCNN cannot detect the person on the left and two small cars in the middle; the DAF recognizes two legs as persons and misses the right car. While our method detects well. The qualitative results demonstrate that our proposed method detects more objects correctly than faster RCNN and DAF.

3.3. Ablation Study. In this subsection, we conduct an ablation study to analyze the effect of each proposed component of the whole pipeline on performance.

Table 1: Comparison of the performance of the compared methods.

	Bike	Car	Car_stop	Color_cone	Person	mAP
Faster RCNN [2]	1.5	0.7	0.5	0	4.1	1.4
DAF [7]	1.04	3.1	0.71	0.03	39.0	8.8
Ours	**20.3**	**30.8**	**11.7**	**12.9**	**56.9**	**26.5**
Oracle	67.9	81.3	52.6	64.5	76.6	68.6

Ground Truth (a) Faster RCNN (b) DAF (c) Ours (d)

Figure 4: Qualitative results of the compared methods. The example image is 000726.png from FIR folder of multispectral object detection dataset [10]. (a) Ground truth. (b) Faster RCNN. (c) DAF. (d) Ours'.

Ground Truth (a) Faster RCNN (b) DAF (c) Ours (d)

Figure 5: Detection results of the compared methods. The displayed image is 000885.png from FIR folder of multispectral object detection dataset [10]. (a) Ground truth. (b) Faster RCNN. (c) DAF. (d) Ours'.

Table 2: Ablation study for each proposed component.

Image trans	Int-Inv[1]	R-A[2]	Bike	Car	Car_stop	Color_cone	Person	mAP
No	—		1.5	0.7	0.5	0	4.1	1.4
No	—	√	1.04	3.1	0.71	0.03	39.0	**8.8**
Gray			0.94	4.1	1.2	2.03	18.3	5.3
Gray	√		3.9	12.2	5.0	7.5	52.5	16.2
Gray		√	1.5	2.1	2.5	7.2	50.7	12.8
Gray	√	√	3.0	7.3	4.7	17.0	61.2	**18.6**
CycleGAN			12.8	16.4	3.5	4.0	26.1	12.6
CycleGAN	√		18.7	31.6	6.0	6.6	49.0	22.4
CycleGAN		√	13.3	14.4	6.3	7.3	41.5	16.6
CycleGAN	√	√	20.3	30.8	11.7	12.9	56.9	**26.5**

Note. [1]Int-Inv indicates intensity inversion; [2]R-A means readaptation.

Table 2 provided the ablation performance of different configuration of each proposed component. Comparing configurations with CycleGAN-based image translation to those with gray translation, it seems that configs with CycleGAN perform better. For example, config in the 7th row obtains mAP 12.6% while the 1st row obtains 1.4% and the 3rd row obtains 5.3%. Comparing configs with both image translation (gray or CycleGAN) and intensity inversion to configs with only image translation, those with intensity inversion yield obvious gain. For example, config in the 8th row obtains mAP 22.4% while the 7th row obtains 12.6%. Finally, configs with readaptation perform better than those without readaptation. For example, config in the 10th row obtains mAP 26.5% while the 8th row obtains 22.4%. From the above analysis, it is clear that three proposed components, i.e., CycleGAN-based image translation, intensity inversion, and readaptation, are all necessary and yield performance gain.

4. Conclusions

In this paper, we proposed an unsupervised image-generation enhanced adaptation method for object detection in thermal images. Two modules are included. The image-generation module is to generate simulated fake thermal images that are similar to the target images, and the readaptation module is to reduce the gap between generated intermediate domain and the thermal target domain. The presented experimental results demonstrate that the proposed method outperforms the state-of-the-art greatly.

Based on the proposed adaptive detection framework, some future works can be extended, such as generating more similar thermal images from color visible images, integrating the merits of different category domain adaptation methods and applying to the visual-to-thermal domain adaptive object detection, and studying compact end-to-end models.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was financially supported by the Science and Technology Plan Project of State Administration for Market Regulation (2020 MK 162), the National Natural Science Foundation of China (No. 61771471), the Central Foundational Research Funding Project (562020Y-7482), and the National Natural Science Foundation of China (Nos. 61401463, U1613213, and 91748131). A preprint has previously been published [14].

References

[1] R. Gade and T. B. Moeslund, "Thermal cameras and applications: a survey," *Machine Vision and Applications*, vol. 25, no. 1, pp. 245–262, 2014.

[2] S. Ren, K. He, R. Girshick, J. Sun, and R.-C. N. N. Faster, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[3] W. Liu, D. Anguelov, D. Erhan et al., "Ssd: single shot multibox detector," in *Proceedings of the European Conference on Computer Vision*, pp. 21–37, Amsterdam, The Netherlands, October 2016.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, Las Vegas, NV, USA, June 2016.

[5] C. Herrmann, M. Ruf, and J. Beyerer, "CNN-based thermal infrared person detection by domain adaptation," in *Proceedings of the Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything*, Article ID 1064308, Orlando, FL, USA, April 2018.

[6] T. Guo, C. P. Huynh, and M. Solh, "Domain-adaptive pedestrian detection in thermal images," in *Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP)*, pp. 1660–1664, Taipei, Taiwan, September 2019.

[7] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster R-CNN for object detection in the wild," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 3339–3348, Salt Lake City, UT, USA, June 2018.

[8] W. Li, F. Li, Y. Luo, and P. Wang, "Deep domain adaptive object detection: a survey," in *Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1808–1813, Canberra, Australia, December 2020.

[9] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232, Venice, Italy, October 2017.

[10] K. Takumi, K. Watanabe, Q. Ha, A. Tejero-De-Pablos, Y. Ushiku, and T. Harada, "Multispectral object detection for autonomous vehicles," in *Proceedings of the on Thematic Workshops of ACM Multimedia*, pp. 35–43, Mountain View, CA, USA, October 2017.

[11] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, and G. Chanan, "PyTorch: an imperative style, high-performance deep learning library," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 8024–8035, Vancouver, Canada, August 2019.

[12] Image-to-Image translation in PyTorch. Available: https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix.

[13] (2017, 2020/1/3). An unofficial implementation of 'domain adaptive faster R-CNN for object detection in the wild '. Available: https://github.com/tiancity-NJU/da-faster-rcnn-PyTorch.

[14] P. Liu, F. Li, and W. Li, "Unsupervised image-generation enhanced adaptation for object detection in thermal images," 2021, https://arxiv.org/abs/2002.06770.