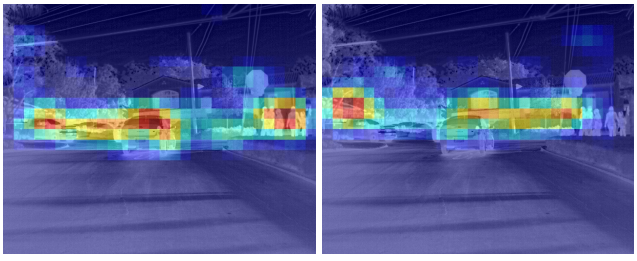
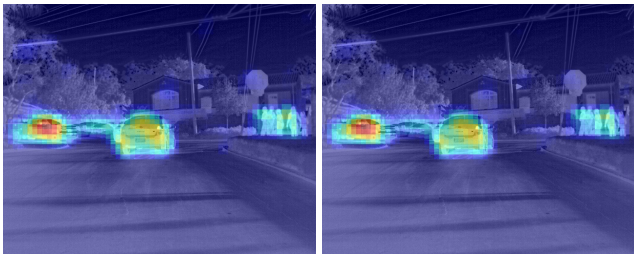
D3T: Distinctive Dual-Domain Teacher Zigzagging Across RGB-Thermal Gap for Domain-Adaptive Object Detection

Supplementary Material



(a) (b)



(c) (d)

Figure 1. CAM generated by dual teachers at various stages of training: (a) and (b) represent CAM from the RGB and thermal teacher models, respectively, during the early stages of training, whereas (c) and (d) are CAM from the same models at later stages of training.

1. Class Activation Mapping

Class Activation Map (CAM) [8] is an algorithm that generates heatmaps highlighting the critical regions in an image for a particular class. Grad-CAM [7] is an improved version of CAM that can be applied across various network architectures. To illustrate the distinctions between the two teacher models in the initial stages, we employ Eigen-CAM [5], an advanced version of Grad-CAM.

As shown in Fig. 1 (a) and (b), during the early stages of training, the RGB teacher and the Thermal teacher show different activation maps. In the later training steps, as shown in Fig. 1 (c) and (d), the activation mappings become similar and focus more on the object. This demonstrates the effectiveness of our proposed method in narrowing the gap between the RGB and Thermal domains, as well as enhancing the accuracy of object recognition.

2. Performance Changes Across Iterations

We plot performance changes of RGB teacher, thermal teacher, and a student across iterations, representing an optimization trend in Fig. 2. The zig-zag patterns are changed at 20k and 30k iterations. In early stages, RGB teacher outperforms thermal teacher, because we provide more training chances to RGB teacher with labels, then reverses in the middle stages. In the final, their performances converge to similarity, indicating a reduced domain gap. We also note fluctuations in the student's performance, ultimately showing improvement in the upper-right direction. This result demonstrates the stability of our method and we only use thermal teacher for inference.

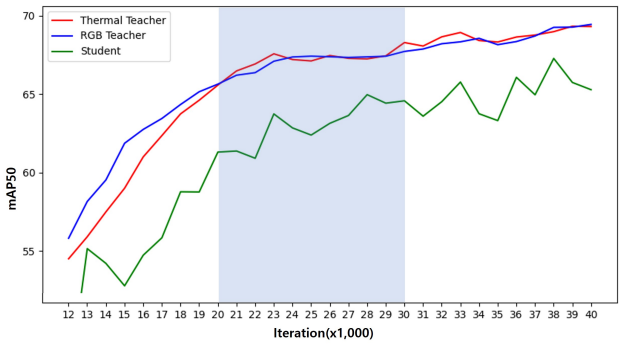


Figure 2. Performance changes.

HT	AT	Ours
50.4	50.9	51.7

Table 1. Cityscapes [1]→Foggy Cityscapes [6].

Dual-Teachers	Zigzag-Learn	Incor-Know	mAP
✓		✓	67.92
✓	✓	✓	69.30

Table 2. Ablation test for components of the proposed method.

3. RGB DA Benchmark

We present a new experiment in Table 1.Our performance has higher mAP than the latest work (e.g., HT [3], AT [4]). *Note that our primary focus is on adapting from RGB domain to thermal domain, rather than within RGB domain itself.*

4. Effect of Zigzag-Learn

We conduct the ablation test for checking the performance improvement (+1.38%) of Zigzag-Learn in Table 2.

AT [4]	Ours
61.90	69.30

Table 3. RGB→Thermal in FLIR.

Figure 3. KAIST RGB and thermal images illustrating disparities between the two domains. Evaluation is conducted without the use of this paired information.

5. Novel Approach in Domain Adaptation with Dual Teachers

The existing methods [2, 4] are all based on a single teacher and a single student network for DA. However, we use two teachers, with a student sequentially adopting to each teacher from RGB and thermal domains. *Note that a simple extension from a single to dual teachers does not guarantee optimal performance.* To solve this concern, we propose our novel zigzag-learn and incor-know components. We also conduct an ablation test, comparing ours with AT [4] in RGB→thermal DA test. Table 3 shows that a single teacher-based method fails to achieve satisfactory performance.

6. Pseudo-label Selection

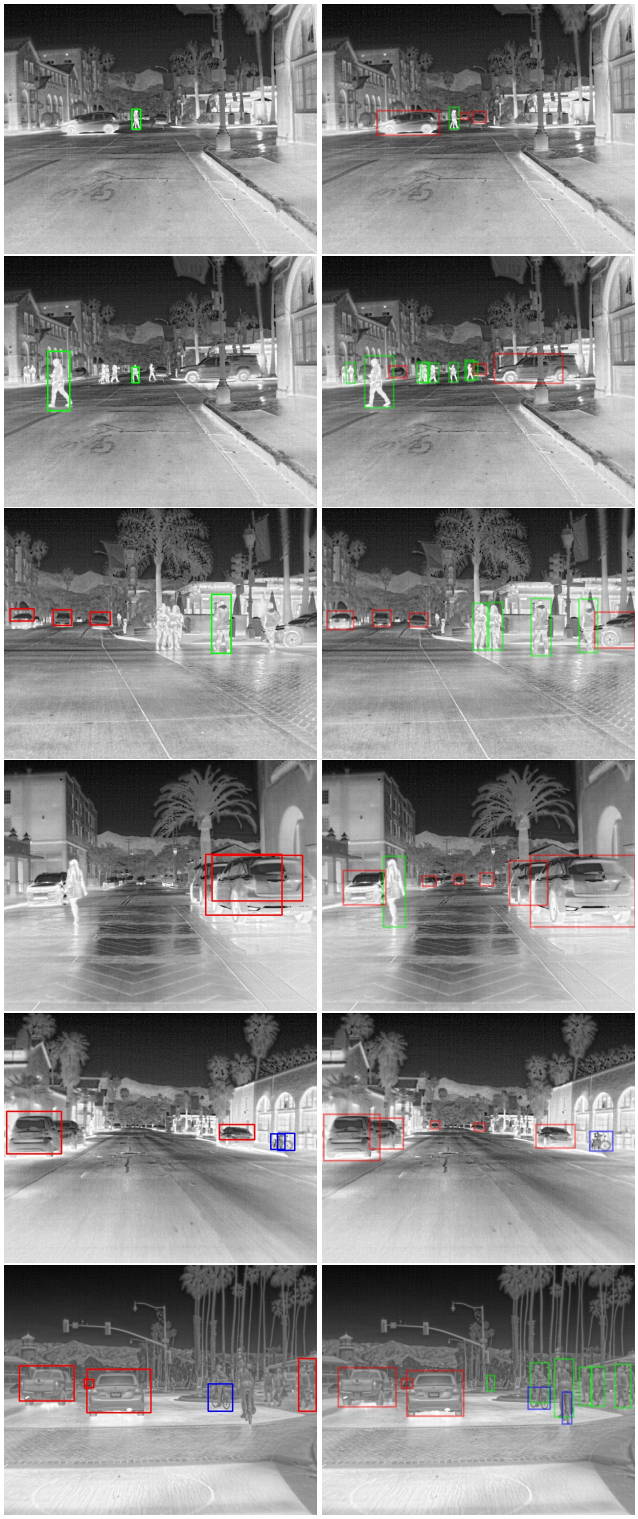
To solve bad impact of false positives/negatives on pseudo-label, we only employ pseudo-labels in later iterations, once the model has attained stability with dynamically changing λ values. We then select only top 1% pseudo-labels based on confidence values for training the model. As shown in Fig. 3, each modality exhibits unique characteristics, and there is a possibility that labels given from RGB may not align perfectly with thermal labels. This is the primary rationale for employing pseudo-labels to improve the performance in this paper.

7. Changing λ

From iteration 10k to 20k, λ gradually increases from 0 to 1 according to the equation (Iteration - 10k)/10k. λ equals 1 for iterations greater than 20k and 0 for iterations less than 10k.

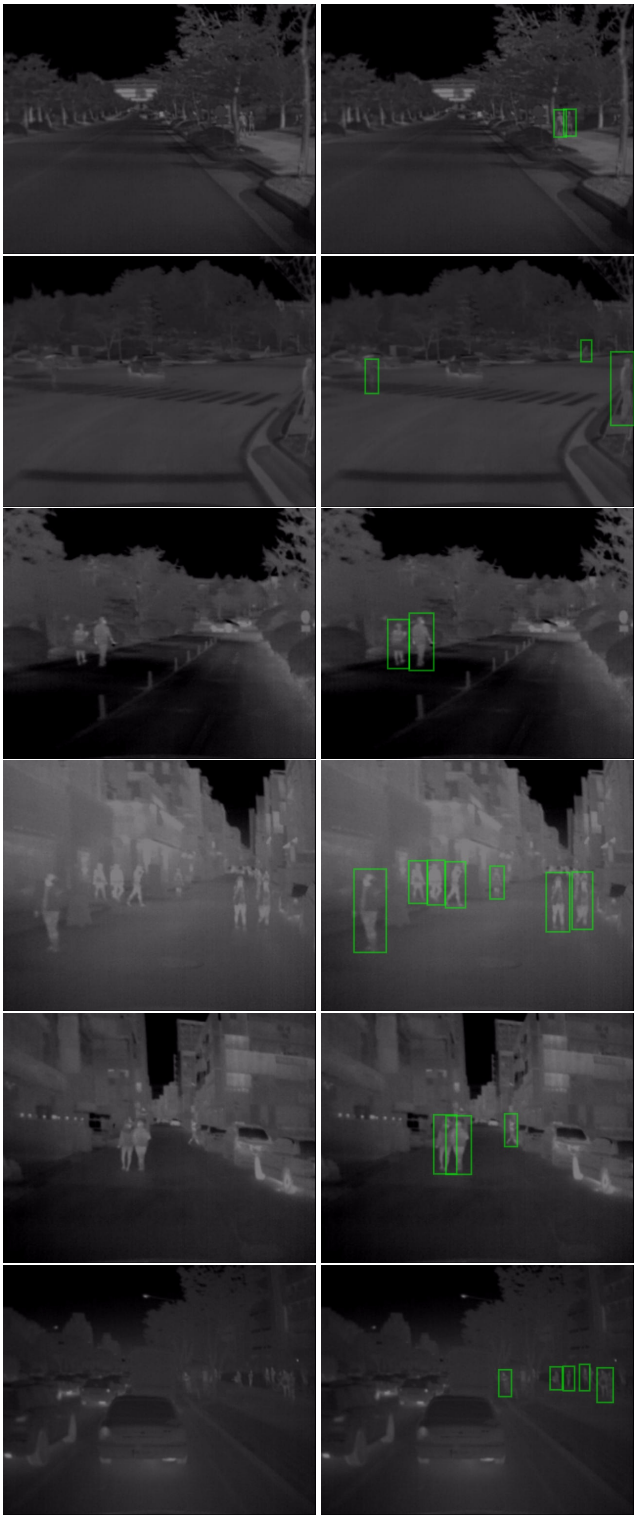
8. Visualization

To offer a clearer understanding of the D3T algorithm's effectiveness, we present further results using images from both the source-only model and the D3T model. We present results using both the FLIR and KAIST datasets. The results indicate that our algorithm significantly outperforms the source-only model, which does not utilize the UDA.



(a) Source only (b) D3T (Ours)

Figure 4. Visualization of UDA results for object detection models on the FLIR dataset for the RGB to thermal domain: Source-only model and our D3T model. The green, blue and red boxes represent the classes of person, bicycle and car.



(a) Source only (b) D3T (Ours)

Figure 5. Visualization of UDA results for object detection models on the KAIST dataset for the RGB to thermal domain: Source-only model and our D3T model. The green boxes represent the classes of person.

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1

[2] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4101, 2021. 2

[3] Jinhong Deng, Dongli Xu, Wen Li, and Lixin Duan. Harmonious teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23829–23838, 2023. 1

[4] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7581–7590, 2022. 1, 2

[5] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–7. IEEE, 2020. 1

[6] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018. 1

[7] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1

[8] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 1