CS269 Project Report: A Preliminary Study of BiADT's Domain Adaptation Abilities on Multispectral Images

Kuei-Chun Kao
UCLA
johnson0213@g.ucla.edu

Oliver De Visser
UCLA
oliverdevisser@g.ucla.edu

Yu-Hsin Weng
UCLA
yuhsin1614@g.ucla.edu

Abstract

Multi-spectral infrared object detection, where objects are identified across different infrared wavelengths, is a complex task. While general-purpose cross-domain object detection models like Bidirectional Alignment for Domain Adaptive Detection with Transformers (BiADT) and Adversarial Query Transformers (AQT) show promise on certain datasets, their effectiveness for infrared object detection is unclear. This is because the vast difference in wavelengths between RGB and infrared images creates unique challenges. To investigate this, we conducted experiments on some challenging multispectral datasets, e.g. FLIR and LLVIP to see if existing cross-domain frameworks can handle infrared data. We tested BiADT with different detector backbones to analyze the results. Our findings show that BiADT with the DINO backbone achieved competitive performance in mAP50. However, leveraging multispectral data introduces new challenges. Future research will focus on how to best integrate information from different wavelengths to fully exploit their complementary nature. Additionally, we will explore how to design an effective cross-modality fusion mechanism to further improve performance.

1 Introduction

Infrared thermography, capturing visual data through heat signatures, is widely used in video surveillance, night vision, and various industrial applications. Object detection based on infrared images and videos can automate many tasks. This project explores multi-spectral object detection, a crucial challenge for computer vision. Here, a detector trained on one set of images (source domain) needs to perform well on a different set (target domain). This difference between source and target domains, also known as a "domain gap" or "domain shift," arises because the images have significant variations in appearance and texture, even though they contain the same object classes. Despite advancements in cross-domain object detection, directly applying these methods to multi-spectral images, like infrared and thermal, often leads to poor results. Therefore, we investigate whether existing frameworks can be readily adapted to the multi-spectral scenario.

In unsupervised domain adaptation (UDA) for object detection, most of the works tackle the challenge of training a detector on two different sets of images (source and target). While images from both sets are available, labels are only provided for the source images. To address this, a common approach involves creating a space where the key characteristics extracted from images (features) become similar between the source and target sets.

This paper addresses the challenges of multi-spectral object detection by investigating the effectiveness of the UDA framework BiADT with various backbone detectors. We aim to determine whether multi-spectral object detection can be tackled using existing methods and explore the advantages of different backbones in this context. However, we observe that they nonetheless face significant challenges: (1) Most contemporary methods frequently construct complex fusion modules after a dual-channel feature extraction network, thus increasing computational demands. (2) Attention-based approaches predominantly extract features on a single scale, employing rudimentary fusion techniques, such as concatenation or addition, to either enhance or reduce these features. This method often overlooks the potential for cross-modal information complementarity across various levels. This can lead to incorrect or missing detections of small objects. Extensive evaluations have been conducted on the FLIR and LLVIP datasets. The results demonstrate a key challenge in exploiting multi-spectral data in how to integrate the information from different wavelengths and design an effective cross-modality fusion mechanism to fully leverage their inherent complementarity and achieve maximum performance gains in future work.

2 Related Work

2.1 Transformer-based Object Detection

DETR (DEtection Transformer) (Carion et al., 2020) introduces a novel approach to object detection by directly predicting the final set of detections in parallel. This method leverages a synergistic combination of a convolutional neural network (CNN) for feature extraction and a transformer architecture for processing the extracted features. The CNN extracts high-level features from the input image, which are subsequently passed through an encoder-decoder transformer. The decoder then interacts with a feed-forward network (FFN) to generate the final detection predictions. Notably, during training, DETR employs a bipartite matching technique to establish a correspondence between the predicted detections and the ground truth bounding boxes. This unified framework streamlines the object detection pipeline by eliminating the need for various hand-designed components, leading to a more efficient and potentially more accurate detection process.

Deformable DETR (Zhu et al., 2021) builds on DETR by introducing a novel deformable attention module. Inspired by deformable convolutions, this module focuses on a small set of key points around a reference point, regardless of feature map size. This allows for efficient multi-scale feature aggregation without a Feature Pyramid Network (FPN). Deformable DETR achieves superior performance, particularly for small objects, due to the deformable attention module's ability to capture more precise spatial information. Additionally, replacing standard attention modules with deformable attention modules improves efficiency and accelerates training convergence.

DAB-DETR (Dynamic Anchor Boxes DETR) (Liu et al., 2022) tackles the slow training convergence issue in DETR. It introduces a novel approach by directly using 4D box coordinates (x, y, w, h) as queries in the transformer decoder. These queries are dynamically updated layer-by-layer, incorporating both the position and size of each anchor box. The center coordinates (x, y) guide feature pooling around the center, while the size information (w, h) modulates the cross-attention map, tailoring it to the specific anchor box size. In essence, DAB-DETR modifies the traditional DETR query structure, which combines a content part (decoder self-attention output) and a positional part (often learnable queries). Instead, DAB-DETR utilizes dynamic anchor boxes that are iteratively refined throughout the decoding process.

DN-DETR (DeNoising DETR) (Li et al., 2022) builds upon DAB-DETR by incorporating a denoising training method. This approach tackles the slow convergence issue prevalent in DETR-like methods. The denoising process works by introducing random noise to both bounding boxes and class labels of ground-truth objects for each image. These "noisy queries" essentially act as good initial anchors, having corresponding ground-truth boxes in close proximity. Consequently, the training objective becomes clear: predicting the original bounding box by removing the added noise. This technique offers two key benefits: improved convergence and enhanced stability. By training the model to handle noisy data, DN-DETR reduces the initial distance between predicted positions (anchors) and ground-truth boxes. This effectively localizes the search space, leading to a more efficient training process and faster convergence. Additionally, denoising training stabilizes the matching process, further accelerating convergence and potentially improving performance even with fewer training epochs.

DINO (DETR with Improved deNoising anchOr boxes) (Zhang et al., 2022) pushes the boundaries of end-to-end object detection. This architecture addresses the slow convergence challenge prevalent in DETR-like models through three key innovations. First, contrastive denoising training incorporates both positive and negative samples, guiding the model towards accurate predictions and reducing duplicate detections. Second, mixed query selection bridges the gap between DETR-like and classical two-stage detectors, potentially leading to more accurate initializations. Finally, the look forward twice scheme leverages information from later decoder layers to optimize earlier layers, improving training robustness.

2.2 Unsupervised Domain Adaptation

Unsupervised Domain Adaptation (UDA) addresses the challenge of the "domain gap" between labeled source domain data and unlabeled target domain data. During training, UDA methods leverage both source and target domain images, with the ultimate goal of achieving good performance on the unseen target data. Adversarial learning plays a crucial role in UDA approaches. It involves training a domain classifier to predict the origin (source or target) of an input image. A key element is the gradient reversal layer, which essentially fools the classifier and helps extract features that are invariant to the domain itself. This adversarial learning strategy has become a popular technique in recent cross-domain object detection approaches.

AQT (Adversarial Query Transformers) (Huang et al., 2022) tackles domain adaptation by aligning features through three adversarial tokens. These tokens operate in separate spaces: spatial, channel-wise, and instance-wise. The corresponding attention modules in AQT guide these tokens to achieve alignment between features for the entire image and the entire object sequence. Unlike some methods, AQT does not explicitly separate features into domain-invariant and domain-specific components.

While AQT's three adversarial tokens capture some domain-specific patterns in the overall token sequence, they don't account for individual object variations. BiADT (Bidirectional Alignment for Domain Adaptive Detection with Transformers) (He et al., 2023) seamlessly integrate with standard attention layers. This allows every image/object token to be explicitly split into domain-invariant and domain-specific parts. This fine-grained, token-level domain alignment leads to more precise object detection in cross-domain scenarios. It employs a two-pronged strategy using novel attention mechanisms to reduce the gap between features that should be consistent across domains, and enhance the distinctiveness of domain-specific features. This approach leads to significant performance improvements in cross-domain object detection tasks, consistently outperforming previous methods.

3 Method

Having discussed the domain gap between RGB and multispectral images, in this section, we present a detailed overview of our proposed system. This study aims to achieve two primary objectives: first, to identify the most effective backbone model for learning the domain gap between RGB and multispectral images by fine-tuning the pretrained models, and second, to evaluate the system's performance across various multispectral datasets to ensure accurate capture of latent interactions between RGB and thermal images.

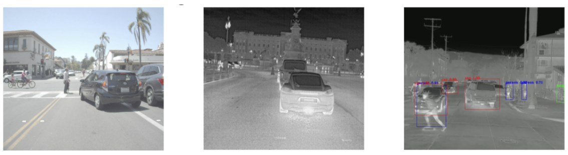


Figure 1: The paired RGB and thermal images of FLIR.

To accomplish these objectives, we have modified the state-of-the-art BiADT model by incorporating different detector backbones. By varying the detector, we can assess the performance of each variant in learning and adapting to the domain gap. This approach enables us to identify the strengths and weaknesses of each model configuration.

Furthermore, we are conducting evaluations using different multispectral datasets. This evaluation is essential for verifying whether the original BiADT model, or its modified versions, can effectively capture interactions across the RGB and thermal image domains. By testing on diverse datasets, we aim to ensure the robustness and generalizability of our proposed method.

4 Experiment Result

Dataset and Metrics

Evaluation is conducted on public benchmark datasets e.g. FLIR and LLVIP with significant domain gaps. For each dataset, we follow the same protocol as in existing work and report the average precision (AP50) of each class.

FLIR ADAS v2 The FLIR ADAS v2 (FLI, 2022) dataset is designed to support research involving visible and thermal spectrum sensors, making it highly relevant for applications in autonomous vehicle safety. This dataset provides a comprehensive set of annotations in the COCO format, covering 15 different categories. the dataset has over 9,700 thermal images with a lot of annotations each, and over 9,200 visible images with more annotation. This dataset is highly suitable for our study for many reasons, such as how it includes multispectral data with both visible (RGB) and thermal images, which is essential for our goal of studying domain adaptation between these two types of data. has a huge amount of annotations with extensive ground truth data that is helpful for training and evaluating object detection models. We cover only three object categories: "person", "car" and "bike" during the evaluation.

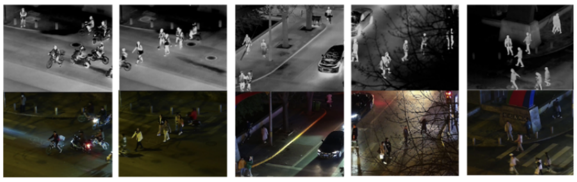


Figure 2: The paired RGB and thermal images of LLVIP.

LLVIP LLVIP (Jia et al., 2023) is a valuable resource for research on low-light pedestrian detection. This dataset features a large collection of visible-infrared image pairs (15,488 pairs from 30,976 total images) that capture various low-light scenarios. The paired nature of the images is crucial for our study, allowing direct comparison and alignment between RGB and thermal data. Furthermore, its focus on low-light pedestrian detection tackles a particularly challenging object detection task, ensuring that our models are tested under difficult conditions. Finally, the dataset's size facilitates robust training and evaluation, enabling us to assess model generalizability across various low-light conditions and measure their real-world applicability. It only covers person object category in this dataset.

Training Details

We conduct with different object detectors e.g. DETR, Deformable-DETR, DAB-DETR, DN-DETR, and DINO with ResNet50 as backbone pretrained on ImageNet. It has 6 encoder layers and 6 decoder layers. In training, we use the same detection loss as in DAB-DETR. The learning rate for the backbone and the transformer are set to 2e-5 and 2e-4, respectively, and the batch size is set to 16. The learning rate decay is set to 0.1, and applied after 40 epochs for the 50-epoch training schedule. All models are trained on NVIDIA A600 GPUs.

Main Results

Table 1 and 2 compare different backbone detectors under BiADT frameworks.

For FLIR ADAS v2 dataset, all architectures predicted "car" category better, we think it is because of the object size of cars is larger than bikes and people, and the dataset is imbalance in the three categories. Specifically, the number of car annotations is larger than the number of person annotations and bike annotations, the number of car annotations is even ten times as large as the number of bike annotations. Starting with DAB-DETR, DAB-DETR, DN-DETR, and DINO exhibited significant performance improvement across all categories. Notably, the "person" category saw a remarkable improvement, highlighting the benefits of dynamic anchor boxes in small objects detection. Finally, DINO achieved the strongest overall performance, though its results were similar to DN-DETR.

This dataset, designed specifically for pedestrian detection, solely includes the "person" category. Consistent with the FLIR ADAS v2 results, all three proposed backbones (DAB-DETR, DN-DETR, and DINO) significantly outperform DETR and Deformable DETR on the LLVIP dataset. Similar to the FLIR results, DINO achieves the strongest overall performance, although closely followed by DN-DETR.

Detector	person	car	bike
DETR	3.1	10.2	1.5
Deformable-DETR	6.3	14.2	5.2
DAB-DETR	36.3	44.1	19.3
DN-DETR	41.4	57.2	28.7
DINO	43.7	61.2	30.1

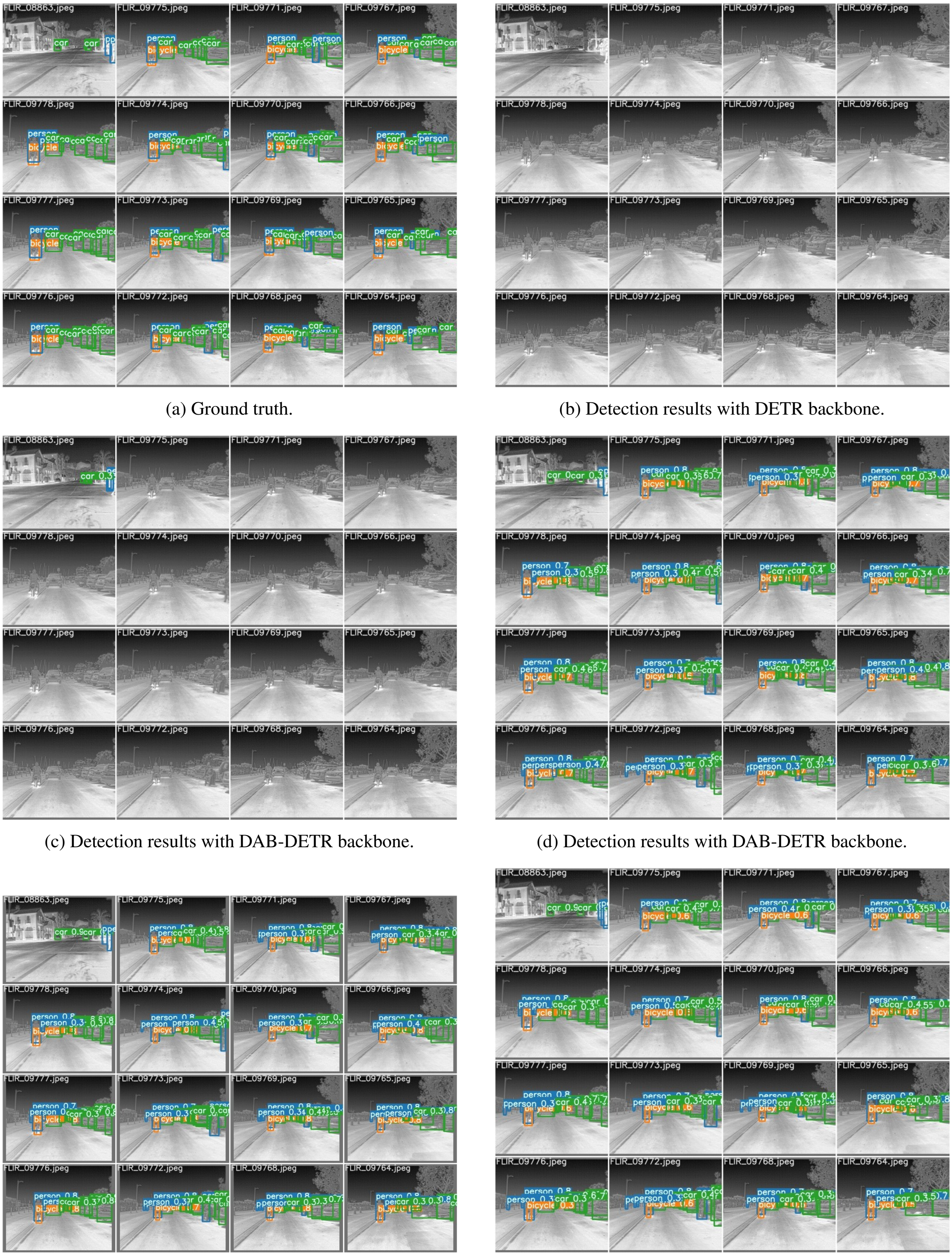
Table 1: Comparisons of performance in Source Dataset: FLIR RGB, Target Dataset: FLIR Thermal in terms of mAP50.

Detector	person
DETR	22.1
Deformable-DETR	25.8
DAB-DETR	57.8
DN-DETR	70.8
DINO	73.6

Table 2: Comparisons of performance in Source Dataset: LLVIP RGB, Target Dataset: LLVIP Thermal in terms of mAP50.

5 Discussion and Analysis

Figure 3b to 3f show the detection results of different backbones. As we have expected, DINO, DN-DETR, and DAB-DETR outperform DETR and Deformable-DETR. While DETR and Deformable-DETR struggle to capture the domain gap, leading to poor object detection, DAB-DETR, DN-DETR, and DINO achieve considerably better results. DAB-DETR excels at detecting small objects due to its dynamic anchor boxes. Furthermore, DN-DETR and DINO push the performance boundary



(a) Ground truth.

(b) Detection results with DETR backbone.

(c) Detection results with DAB-DETR backbone.

(d) Detection results with DAB-DETR backbone.

(e) Detection results with DN-DETR backbone.

(f) Detection results with DINO backbone.

Figure 3: My complicated figure

even further through denoising training and improved query initialization. Though generally, we can conclude that DAB-DETR, DN-DETR, and DINO achieved better performance, there still are some differences between their detection results. For example, we found that DINO and DN-DETR are sensitive to overlapping objects, for images that contains overlapping cars, they sometimes predicted "too many" bounding boxes in the area with overlapping cars. While DAB-DETR tends to generate some small but incorrect bounding boxes

6 Conclusion

This project explored the potential and challenges of unsupervised domain adaptation (UDA) for multi-spectral object detection. We investigated the limitations of current methods in handling the significant domain gap between visible and infrared images. Our work demonstrates that existing UDA frameworks like BiADT, when combined with appropriate backbones, can achieve promising performance in this complex task. However, effectively fusing information from different spectra remains a critical hurdle. Current approaches often rely on computationally expensive fusion modules or miss out on leveraging cross-modal complementarity across various scales.

Addressing these limitations is crucial for unlocking the full potential of multi-spectral object detection. Future research should focus on developing efficient and effective fusion mechanisms that can seamlessly integrate information from different wavelengths. This will enable the creation of robust detectors that can excel in diverse real-world scenarios, such as night vision, autonomous vehicles, and industrial applications where traditional visible-spectrum object detection struggles.

References

2022. Teledyne flir adas thermal dataset v2. https://www.kaggle.com/datasets/samdazel/teledyne-flir-adas-thermal-dataset-v2.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers.

Liqiang He, Wei Wang, Albert Chen, Min Sun, Cheng-Hao Kuo, and Sinisa Todorovic. 2023. Bidirectional alignment for domain adaptive detection with transformers.

Wei-Jie Huang, Yu-Lin Lu, Shih-Yao Lin, Yusheng Xie, and Yen-Yu Lin. 2022. Aqt: Adversarial query transformers for domain adaptive object detection. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 972–979. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, Shengjie Liu, and Wenli Zhou. 2023. Llvip: A visible-infrared paired dataset for low-light vision.

Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, and Lei Zhang. 2022. Dn-detr: Accelerate detr training by introducing query denoising.

Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. 2022. Dab-detr: Dynamic anchor boxes are better queries for detr.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2021. Deformable detr: Deformable transformers for end-to-end object detection.