Journal of Intelligent Computing and Networking
https://www.ffspub.com/index.php/jicn/index
ISSN 3079-9228 (print)
ISSN 3107-2798 (online)
E-mail: jicn.office@ffspub.com

Article

Hierarchical Domain Adaptation with Multi-Level Attention for RGB-Thermal Object Detection

Yunan Liu[1], Mingrong Gong [2], Chaoqi Chen[1,†]
[1] *College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China*
[2] *College of Computing and Data Science, Nanyang Technological University, Singapore 639798, Singapore*
[†]*E-mail: chencq@szu.edu.cn*

Received: June 13, 2026 / Revised: July 6, 2026 / Accepted: July 21, 2026 / Published online: July 30, 2026

Abstract: Cross-modal domain adaptation for RGB-to-Thermal object detection (DAOD) remains a challenging task due to the large modality gap, feature distribution mismatch, and limited supervision on the thermal domain. Conventional domain adaptation detectors often fail to capture consistent semantic representations across modalities, resulting in unstable adversarial learning and suboptimal detection accuracy. To address these challenges, we propose HDAM (Hierarchical Domain Adaptation with Multi-Level Attention), an enhanced domain-adaptive detection framework built upon Faster R-CNN. HDAM progressively aligns cross-domain features from low- to high-level representations through a hierarchical attention mechanism. Specifically, it comprises three key components: (1) Entropy-Guided Cross-level Attention (ECA), which leverages discriminator-derived entropy responses to reweight pixel- and mid-level features for more stable cross-modal encoding; (2) Multi-branch Global Semantics and Discriminator (MGSD), which performs global adversarial alignment and integrates multi-scale spatial information through attention-enhanced branches, while introducing an auxiliary image-level semantic consistency loss; and (3) Instance-Context Alignment (ICA), which fuses ROI-level and contextual embeddings and applies dual instance-level adversarial losses for precise fine-grained adaptation. Extensive experiments on benchmark RGB-Thermal datasets demonstrate that HDAM effectively bridges the modality gap and achieves improved cross-domain detection performance compared to recent DAOD methods.

Keywords: Domain Adaptation; RGB-Thermal Object Detection; Hierarchical Attention; Adversarial Learning; Faster R-CNN
https://doi.org/10.64509/jicn.23.127

1 Introduction

Deep learning has achieved remarkable success in object detection, driven by large-scale annotated datasets and powerful convolutional or transformer-based architectures. However, these models typically assume identical data distributions between training and deployment domains. In real-world applications, such as autonomous driving or surveillance, domain shifts [1] frequently occur due to changes in illumination, weather, or sensing modality. This domain discrepancy significantly degrades detection performance when models trained on RGB images are applied to thermal imagery [2–4]. To mitigate this challenge, Unsupervised Domain Adaptation (UDA) [5–7] has emerged as a promising solution by transferring knowledge from labeled source data to unlabeled target data.

Typical UDA methods seek to learn domain-invariant representations by aligning feature distributions between source and target domains. These approaches have achieved significant and consistent progress in both classification and detection tasks [8–19]. Broadly, they can be categorized into two major directions: (1) adversarial alignment methods [8, 11–13, 20, 21], which introduce domain discriminators to encourage indistinguishable feature representations across domains, and (2) discrepancy-based methods [9, 10, 14, 22, 23], which minimize statistical distance metrics (e.g., MMD [24], CORAL [10]) to reduce the inter-domain gaps [25].

For object detection, several studies have extended these essential UDA principles to region-based frameworks such as Faster R-CNN [26]. Existing works typically incorporate domain discriminators at image or instance levels to perform adversarial feature alignment. For example, some approaches align global scene-level features to achieve coarse

† Corresponding author: Chaoqi Chen
*Academic Editor: Xueyang Fu

adaptation [11, 27], while others focus on instance-level or ROI-level alignment to further refine category separation [12, 20]. These methods collectively demonstrate that multi-granularity adaptation is crucial for effective and reliable domain transfer in detection [28, 29].

However, most existing UDA-based detectors are primarily designed for intra-modal adaptation—such as weather or style shifts within RGB domains—and their performance degrades notably when facing cross-modal discrepancies like RGB and Thermal [3, 18, 30, 31]. In heterogeneous settings, the difference in sensor characteristics introduces large semantic and structural divergence, including severe global semantic drift and contextual mismatch, making simple adversarial alignment alone insufficient for achieving robust cross-domain transfer. Moreover, cross-modal adaptation often suffers from unstable and inconsistent training due to the absence of consistent semantic cues and strong modality-specific noise, which further limits the generalization ability of existing approaches [15, 30].

To address these challenges, we propose HDAM (Hierarchical Domain Adaptation with Multi-Level Attention), a hierarchical domain adaptation framework for RGB→Thermal object detection. Built upon Faster R-CNN [26] with a SWDA-based baseline that includes instance alignment, HDAM progressively bridges domain gaps from low-level textures to high-level semantics through multi-level attention and hierarchical alignment. Specifically, it integrates three complementary components: Entropy-Guided Cross-level Attention (ECA), which reweights early and mid-level features according to discriminator responses; Multi-branch Global Semantics and Discriminator (MGSD), which performs global semantic alignment through multi-branch contextual aggregation to counter semantic drift; and Instance-Context Alignment (ICA), which enforces fine-grained instance consistency by jointly modeling ROI-level and contextual embeddings with gradient decoupling. Together, these components enable HDAM to achieve stable and discriminative cross-modal adaptation from RGB to Thermal imagery with improved generalization.

In summary, our main contributions are as follows:

- We propose a Hierarchical Domain Adaptation Framework (HDAM) that systematically mitigates cross-modal discrepancies across pixel, semantic, and instance levels through progressive alignment.
- We design three complementary modules—ECA, MGSD, and ICA—which collaboratively enhance feature robustness and cross-domain generalization through entropy-guided reweighting, multi-branch fusion, and context-aware instance adaptation.
- Extensive experiments on the Visible→Thermal FLIR benchmark show that HDAM achieves a mAP of 57.2% and remains competitive with recent domain adaptation detectors across multiple RGB-to-thermal settings.

These results support the effectiveness of HDAM for cross-modal object detection where labeled data on the target domain is scarce.

2 Related Work

Unsupervised Domain Adaptation for Object Detection

Unsupervised domain adaptation (UDA) aims to bridge the gap between labeled source and unlabeled target domains, enabling models to generalize across different distributions. In object detection, early works leverage adversarial learning to align features between domains. DA-Faster R-CNN [11] introduces image- and instance-level discriminators to reduce discrepancies, while SWDA [12] applies strong-weak alignment to enhance adaptation via reliable target predictions. Later methods integrate category-level discriminators [20], pseudo-label self-training [32], or graph-based reasoning [33] to further refine category boundaries and mitigate negative transfer.

These approaches excel in intra-modal scenarios (e.g., weather or style shifts in RGB) [34], but struggle with heterogeneous cross-modality shifts.

Cross-Modality Object Detection

Cross-modality adaptation, especially RGB-thermal, often faces severe sensor discrepancies in textures and semantics. Early methods explicitly disentangle features into modality-specific and shared parts [35, 36], or use attention-based fusion to effectively enhance thermal features with RGB guidance [30].

Recent UDA integrations for RGB-thermal combine adversarial training with contrastive or consistency regularization [15]. Multi-domain attention networks attend to informative regions [37], while hierarchical alignments progress from low-level textures to high-level semantics [3]. However, unstable training and reduced discriminability persist due to overlooked modality gaps and noise.

Discussion and Motivation

Prior methods advance cross-domain detection on benchmarks like FLIR, but rely on static alignments ignoring hierarchical discrepancies, leading to semantic drift and discriminability loss.

This motivates our HDAM, which uses dynamic attention for progressive alignment: entropy-guided feature reweighting at low/mid-levels, multi-branch global discriminators against drift, and context-aware instance refinement with gradient decoupling, balancing invariance and discriminability in RGB-thermal adaptation (as shown in Figure 1).

3 Method

As illustrated in Figure 2, the proposed HDAM framework follows a hierarchical four-stage adaptation pipeline that progressively bridges the domain gap from low-level textures to high-level semantics. Specifically, HDAM consists of three key components: ECA: low- and mid-level Entropy-Guided Cross-level Attention that adaptively reweights early features using discriminator-derived entropy signals; MGSD: high-level Multi-Branch Global Semantics and Discriminator for semantic alignment and overall global consistency; ICA: Instance-Context Alignment for fine-grained instance adaptation and contextual robustness. All discriminators are trained adversarially via Gradient Reversal Layers (GRL) [8]. The adversarial supervision is training-only, while the attention

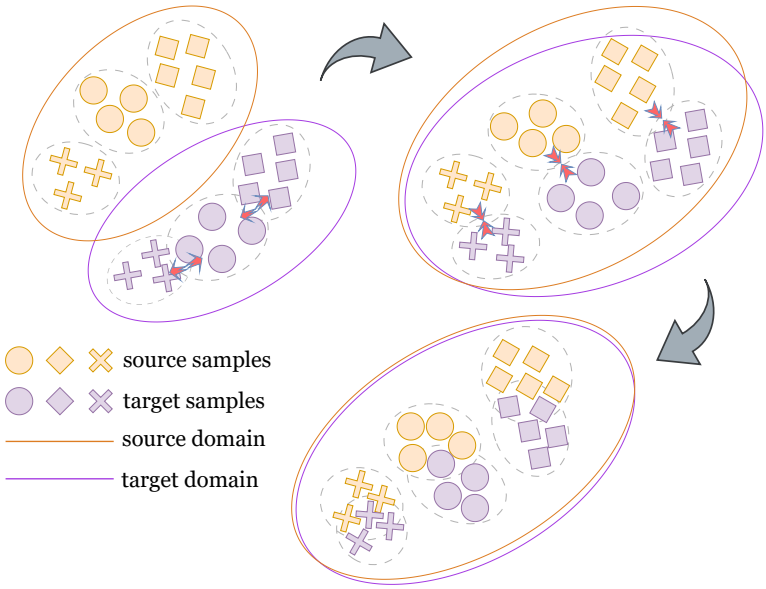

Figure 1: Motivation for hierarchical domain adaptation in HDAM. The figure illustrates the feature distribution shift between the source domain and target domain. Our approach progressively aligns these distributions through hierarchical feature alignment, resulting in a more unified and consistent representation.

and context branches remain part of the full forward model at inference time.

Baseline Setup

Our framework is built upon SWDA [12] (domain-adaptive Faster R-CNN). To strengthen instance-level alignment, we further integrate an additional instance alignment mechanism [11] similar to [3, 20, 38]. In the experiments, we denote this hybrid variant as Baseline; it is therefore stronger than the original SWDA reported in prior work and should be interpreted accordingly in the comparison tables.

3.1 Problem Setup and Training Protocol

Let the labeled source domain (RGB) be denoted by $D_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$, and the unlabeled target domain (Thermal) by $D_t = \{x_j^t\}_{j=1}^{N_t}$. During each training iteration, the network receives mini-batches sampled from both D_s and D_t. Source samples contribute supervised detection losses and domain losses; target samples contribute only unsupervised domain-related losses.

Denote the multi-level backbone feature maps as $\{F^{(1)}, F^{(2)}, F^{(3)}\}$, corresponding to different-level feature extractors respectively, and the final top feature (after base4 / layer4 and GAP) as $F^{(4)}$. Gradient Reversal Layers (GRL) are applied before each domain discriminator to implement adversarial alignment [8, 39].

3.2 Entropy-Guided Cross-Level Attention (ECA)

Low- and mid-level features carry modality-specific textures, edges, and thermal radiation patterns that are often unreliable for direct feature alignment. Following the mainstream uncertainty-based attention paradigm [36, 40], ECA adaptively reweights these features according to domain discriminator responses, so that the backbone learns more discriminative and modality-invariant representations.

Pixel-Level Attention

A pixel-wise discriminator D_{pix} is applied to $F^{(1)}$ right after GRL. The per-pixel confidence $d \in (0,1)^{H\times W}$ is then converted to an entropy map

$$H(d_{h,w}) = -\big(d_{h,w}\log d_{h,w} + (1-d_{h,w})\log(1-d_{h,w})\big), \quad (1)$$

and the reweighting factor $w = 1 - H(d)$ is multiplied channel-wise as described in

$$\widetilde{F}^{(1)}_{c,h,w} = (1 + w_{h,w}) \cdot F^{(1)}_{c,h,w}. \quad (2)$$

This simple reweighting assigns larger responses to lower-entropy, more domain-confident structural regions, which helps preserve stable edges and contours for robust cross-modal generalization.

Mid-level Attention

At $F^{(2)}$, a mid-level discriminator outputs 2-D logits followed by softmax. The self-entropy

$$\bar{H}(p) = -\sum_{c=1}^{2} p_c \log p_c \quad (3)$$

is computed once per image and rescales the entire feature map according to

$$\widetilde{F}^{(2)} = (1 + \bar{H}(p)) \cdot F^{(2)}. \quad (4)$$

This global reweighting enhances mid-level structures (corners, local patterns) that often differ across modalities.

Forward Flow

The reweighted features $\widetilde{F}^{(1)}$ and $\widetilde{F}^{(2)}$ are fed forward and, crucially, supply the contextual vectors ϕ_{pix} and ϕ_{mid} used by ICA (Section 3.4). No additional supervision is introduced at this stage.

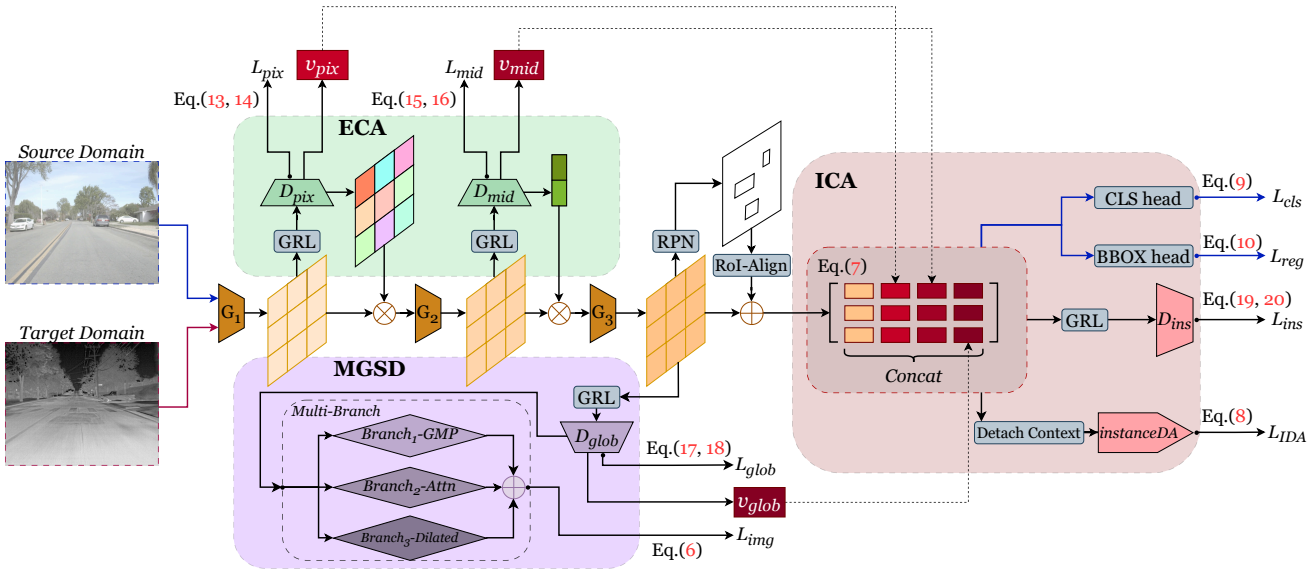

Figure 2: The overall framework of HDAM. G_1, G_2 and G_3 denote feature extractors at different levels. **Detach Context** indicates that the context branch is detached before concatenation.

3.3 Multi-Branch Global Semantics and Discriminator (MGSD)

RGB→Thermal transfer exhibits a layered domain discrepancy: at low–mid levels, modalities differ mainly in texture, edge responses, and thermal radiation patterns; at high levels, object appearance and scene context can change drastically across modalities; at the instance level, isolated RoI features often lack the surrounding cross-modal context required for robust and accurate recognition (especially for small, occluded, or boundary objects). Conventional domain-adaptive detectors typically apply a single global discriminator to $F^{(3)}$ or pooled global features [11]. Such a single-branch discriminator primarily captures coarse distributional statistics and is prone to confusing semantically-similar objects that present different low-level characteristics across modalities, i.e., it tends to learn one global "summary' and therefore may treat the same object class in RGB and Thermal as different domains when their high-level appearance or local cues differ. This naturally leads to global semantic drift and harms cross-modal consistency.

To explicitly counteract global semantic drift and provide complementary global cues for instance alignment, MGSD replaces the single global head with a *multi-branch* semantic extractor and discriminator, inspired by the design philosophy of multi-branch semantics in [15]. The central idea is to force the model to reason about global semantics from multiple perspectives simultaneously: (i) dominant global response, (ii) spatially salient regions, and (iii) multi-scale contextual patterns. By jointly supervising these complementary views and fusing them, the discriminator and the semantic head are discouraged from relying on any single, potentially misleading statistic.

MGSD operates on the top-level feature $F^{(3)}$ and first produces a shared intermediate feature map $G \in \mathbb{R}^{C_g \times H_g \times W_g}$ via a small convolutional head. From G we compute three complementary branches:

- **Branch 1 (global context / max response).** Apply adaptive max pooling to G to obtain $z_1 \in \mathbb{R}^{C_g}$. This branch effectively captures the strongest global activations and is robust to local noise.
- **Branch 2 (spatial attention / salient regions).** Compute a 1×1 convolution followed by a spatial softmax to obtain an attention map $A = \mathrm{softmax}\big(\mathrm{Conv}_{1\times1}(G)\big)$, and aggregate G with A to produce $z_2 = \sum_{h,w} A_{h,w}\, G_{:,h,w}$. This branch highlights spatially salient regions (e.g., objects or discriminative parts) that are likely to be semantically important across modalities.
- **Branch 3 (multi-scale context / dilated conv).** Apply a dilated 3×3 convolution $\mathrm{Conv}^{\mathrm{dil}=2}_{3\times3}$ on G to enlarge the receptive field, then pool to obtain $z_3 \in \mathbb{R}^{C_g}$. This branch captures wider contextual cues and multi-scale dependencies, which help disambiguate objects whose local appearance differs across modalities.

Fuse these complementary semantic vectors:

$$z_{\mathrm{glob}} = z_1 + z_2 + z_3 \in \mathbb{R}^{C_g}, \tag{5}$$

so that the ensuing global signal encodes strong-response, salient-region and multi-scale context jointly. The fused vector z_{glob} is supervised by an image-level semantic loss to regularize category-preserving global features:

$$L_{\mathrm{img}} = \mathrm{BCEWithLogits}(W_{\mathrm{img}}^{\top} z_{\mathrm{glob}} + b_{\mathrm{img}}, y_{\mathrm{img}}), \tag{6}$$

where $(W_{\mathrm{img}}, b_{\mathrm{img}})$ is a small linear classifier and y_{img} denotes the source image-level label. This additional semantic supervision explicitly encourages z_{glob} to be discriminative for object categories while remaining domain-agnostic.

The multi-branch fusion prevents the global discriminator from collapsing onto a single statistic that may be highly domain-dependent. By forcing the model to minimize adversarial loss and semantic loss jointly on z_{glob}, MGSD reduces

erroneous domain separation of same-class instances and mitigates global semantic drift. A derived global feature from this branch is concatenated into instance-level embeddings as ϕ_{glob}. This provides scene-level context to each RoI so that instance discriminators and classifier heads see both the local RoI and the global semantic summary, reducing instance-context mismatch (detailed in Section 3.4).

In practice, MGSD can optionally generate a spatial mask (from an intermediate activation) for target images and apply it multiplicatively to the discriminator inputs prior to pooling. This mask-based modulation explicitly focuses adversarial alignment on semantically confident and highly informative regions, and further reduces the harmful alignment of unreliable background areas.

3.4 Instance-Context Alignment (ICA)

Although MGSD alleviates the global semantic inconsistency between modalities, the instance-level features still suffer from contextual mismatch. Specifically, pure RoI features ϕ_{roi} extracted from the detector tend to lose semantic stability under challenging cross-modality conditions, leading to confidence collapse at object boundaries, occlusions, and small targets. This is mainly because ϕ_{roi} focuses solely on the object region while neglecting its immediate surrounding cues and overall scene context [38].

To effectively address this important issue, we propose the **Instance-Context Alignment (ICA)** module, which enhances instance-level adaptation through context concatenation and gradient decoupling [40].

Context Concatenation

We explicitly aggregate multi-level contextual features from different stages of the adaptation pipeline to complement the RoI representation. As shown in our implementation, the final instance feature is constructed as:

$$\phi_{\text{det}} = [\phi_{\text{roi}}; \phi_{\text{pix}}; \phi_{\text{mid}}; \phi_{\text{glob}}] \in \mathbb{R}^{d_{\text{det}}}, \tag{7}$$

where ϕ_{pix}, ϕ_{mid}, and ϕ_{glob} are extracted from the outputs of ECA and MGSD, respectively. ϕ_{pix} originates from the pixel-level ECA branch, providing fine-grained edge and texture cues that stabilize features at boundaries. ϕ_{mid} comes from the mid-level entropy representation, encoding local structural information to enrich spatial context. ϕ_{glob} is derived from MGSD's high-level semantic branch, offering scene-level semantics that guide instance recognition across modalities. Through this explicit concatenation, ICA ensures the instance discriminator perceives both the target region and its surrounding and global contexts, rather than an isolated RoI feature.

Gradient Decoupling

To prevent back-propagated gradients from the instance discriminator from contaminating the contextual feature generators, we adopt a gradient isolation strategy during feature fusion. Specifically, we apply an auxiliary sigmoid-based InstanceDA head on a detached version $\tilde{\phi}$ and supervise it with BCE:

$$L_{\text{IDA}} = \text{BCE}(\text{InstanceDA}(\tilde{\phi}), y_{\text{dom}}). \tag{8}$$

This careful design guarantees that ECA and MGSD modules remain fully focused on generating stable and highly discriminative representations, while ICA independently learns to align instance-level semantics. Consequently, the overall adaptation of the detector no longer compromises the quality of the contextual features.

Finally, the detection head performs classification and bounding-box regression on the original RoI feature ϕ_{roi} via

$$L_{\text{rcnn-cls}} = \text{CE}(\text{cls}(\phi_{\text{roi}}), y_{\text{roi}}), \tag{9}$$

$$L_{\text{rcnn-reg}} = \text{SmoothL1}(\text{bbox}(\phi_{\text{roi}}), t_{\text{roi}}). \tag{10}$$

3.5 Losses and Optimization

The total training objective integrates detection and multi-level adaptation losses:

$$\begin{aligned} L = L_{\text{det}} &+ \lambda_{\text{pix}}(L^s_{\text{pix}} + L^t_{\text{pix}}) + \lambda_{\text{mid}}(L^s_{\text{mid}} + L^t_{\text{mid}}) \\ &+ \lambda_{\text{glob}}(L^s_{\text{glob}} + L^t_{\text{glob}}) + \lambda_{\text{ins}}(L^s_{\text{ins}} + L^t_{\text{ins}}) \\ &+ \lambda_{\text{IDA}} L_{\text{IDA}} + \lambda_{\text{img}} L_{\text{img}}. \end{aligned} \tag{11}$$

Here the supervised detection loss is

$$L_{\text{det}} = L_{\text{rpn-cls}} + L_{\text{rpn-reg}} + L_{\text{rcnn-cls}} + L_{\text{rcnn-reg}}. \tag{12}$$

We detail each domain loss below.

Pixel-Level Losses

Following prior work [11], we use a per-pixel MSE-style objective that encourages $D_{\text{pix}}(F^{(1)})$ to predict source=1, target=0:

$$L^s_{\text{pix}} = \frac{1}{HW} \sum_{h,w} \left(D_{\text{pix}}(F^{(1)}_s)\right)^2, \tag{13}$$

$$L^t_{\text{pix}} = \frac{1}{HW} \sum_{h,w} \left(1 - D_{\text{pix}}(F^{(1)}_t)\right)^2. \tag{14}$$

Mid-Level Losses

For mid-level discriminator (softmax + CE), we use standard cross-entropy:

$$L^s_{\text{mid}} = \text{CE}(D_{\text{mid}}(F^{(2)}_s), 0), \tag{15}$$

$$L^t_{\text{mid}} = \text{CE}(D_{\text{mid}}(F^{(2)}_t), 1). \tag{16}$$

Global Losses

For the global discriminator we adopt a robust classification loss (Focal Loss or CE) [41]:

$$L^s_{\text{glob}} = \text{FL}(D_{\text{glob}}(F^{(3)}_s), 0), \tag{17}$$

$$L^t_{\text{glob}} = \text{FL}(D_{\text{glob}}(F^{(3)}_t), 1), \tag{18}$$

where FL denotes the focal loss (or CE if preferred).

Instance Losses

Instance discriminator losses are similarly defined per-proposal:

$$L_{\text{ins}}^{s} = \text{FL}(D_{\text{ins}}(\phi_{\text{det},s}), 0), \tag{19}$$

$$L_{\text{ins}}^{t} = \text{FL}(D_{\text{ins}}(\phi_{\text{det},t}), 1). \tag{20}$$

Auxiliary Instance DA

The auxiliary InstanceDA loss (Equation (8)) uses BCE on detached features, providing a stable auxiliary signal:

$$L_{\text{IDA}} = \text{BCE}(\text{InstanceDA}(\tilde{\phi}), y_{\text{dom}}).$$

Image-Level Semantic Loss

As in Equation (6), L_{img} enforces image-level semantic consistency using source image-level labels.

Hyper-Parameters and Dataset-Dependent Weighting

Weighting coefficients λ_{pix}, λ_{mid}, λ_{glob}, λ_{ins}, λ_{IDA}, λ_{img} are carefully tuned per dataset. Empirically, we found that small mid/image weights are particularly beneficial for Cityscapes→FLIR transfer and accordingly adjusted λ_{ins} based on scene density (reduced for crowded city scenes).

4 Experiments

4.1 Datasets

Visible-FLIR → Thermal-FLIR

This setting considers a one-to-one cross-modal adaptation scenario, where the domain gap primarily stems from the heterogeneous image modalities between visible and infrared sensors. The FLIR [42] dataset contains paired RGB and thermal images captured in both daytime and nighttime conditions, covering three object categories: person, car, and bicycle. It consists of 4,129 training images and 1,013 testing images, offering a compact yet representative benchmark for evaluating RGB-Thermal object detection.

Cityscape → Thermal-FLIR

This setting evaluates the model's ability to adapt across both modality and geographic scene discrepancies. The Cityscape [43] dataset provides RGB images captured by vehicle-mounted cameras in urban driving environments, containing 2,975 training and 500 validation images annotated with eight common object categories. When adapted to the Thermal-FLIR target domain, this scenario introduces additional challenges due to differences in weather conditions, illumination, and scene layout between datasets.

Pascal VOC → Thermal-FLIR

This setting investigates adaptation from a general-purpose dataset with diverse daily-life scenes to the thermal domain, emphasizing differences in object appearance and background context. The Pascal VOC [44] dataset includes 16,551 images from 20 object categories, providing rich object diversity but with large domain gaps from the thermal imagery in FLIR. This scenario is particularly useful for analyzing the robustness and generalization capacity of the proposed approach across heterogeneous and semantically disjoint domains.

4.2 Implementation Details

Our model is implemented based on the Faster R-CNN [26] framework with a ResNet-101 [45] backbone pretrained on ImageNet [46], following the typical setting in [3, 12, 37]. The backbone is divided into three main stages (conv1–layer3) for feature extraction, and layer4 serves as the detection head with global average pooling. ROIAlign [47] is adopted for proposal pooling.

We attach four domain discriminators at different levels: pixel-level, mid-level, global-level, and instance-level on ROI features. Each discriminator is preceded by a gradient reversal layer (GRL, $\lambda = 0.1$) [8] and composed of stride-2 convolutions with Batch Normalization and Dropout. To further enhance feature robustness, local and mid-level attention are introduced by reweighting features with entropy-derived importance signals from discriminator responses. A multi-branch global head integrates three complementary pathways (global pooling, spatial attention, and dilated convolution) to compute a fused representation supervised by BCEWithLogits loss.

The model is optimized using stochastic gradient descent (SGD) with a momentum of 0.9 and weight decay of 10^{-4}. The initial learning rate is set to 10^{-3} and decayed by 0.1 at the 6th epoch. Input images are resized to have a shorter side of 800 pixels. The loss weights are set to $\lambda_{\text{pix}} = 0.5$, $\lambda_{\text{mid}} = 0.15$, $\lambda_{\text{glob}} = 0.5$, $\lambda_{\text{ins}} = 0.5$, $\lambda_{\text{IDA}} = 0.1$, and $\lambda_{\text{img}} = 0.1$, with dataset-specific adjustments. The reported HDAM results were obtained after repeated independent training, indicating that the observed performance was reproducible across runs. Compared with the Baseline, HDAM introduces approximately 8.95M additional parameters and increases both training time and inference time by about 15%, which we regard as a moderate trade-off for the observed accuracy gains.

4.3 Performance on the FLIR Benchmark

We compare our **HDAM** with DA-Faster [11], SA-DA-Faster [33], SWDA [12], SCL [48], VDD [49], CR-DA-DET [22], HTCN [37], UDAT [50], TIA [51] and UGA [3]. For several competitors, we report the benchmark numbers published in prior work, especially the UGA protocol used for FLIR-based comparison; therefore, the tables should be interpreted as benchmark-level references rather than as strict same-codebase re-implementations for every method.

Table 1 shows the results of adaptation from Visible-FLIR to Thermal-FLIR. HDAM achieves **57.2%** mAP, exceeding the previous best reported result (UGA 56.0%) by **+1.2%**. Notably, our Baseline already reaches 51.7%, providing a stronger starting point than original SWDA; HDAM further improves upon it through hierarchical attention and context-aware alignment.

Table 2 reports the detailed results of Cityscape → Thermal-FLIR. HDAM attains **53.5%** mAP, slightly improving over UGA (53.0%) and outperforming our Baseline (47.8%). This result suggests that ECA and MGSD remain beneficial under larger scene-style shifts, although the gain over the strongest prior method is modest.

Table 1: Results of Visible-FLIR → Thermal-FLIR (%).

Methods	Car	Bicycle	Person	mAP
DA-Faster [11]	59.9	24.3	26.6	36.9
SA-DA-Faster [33]	70.3	33.3	47.2	50.3
SWDA [12]	58.9	32.0	32.3	41.4
SCL [48]	63.8	43.6	36.9	48.1
VDD [49]	66.6	49.2	45.8	53.9
CR-DA-DET [22]	66.4	41.7	43.2	50.4
HTCN [37]	56.3	37.9	33.1	42.4
UDAT [50]	66.8	49.3	43.4	53.1
TIA [51]	67.8	44.8	48.8	53.8
UGA [3]	68.5	48.0	51.6	56.0
Baseline	65.8	45.2	44.0	51.7
HDAM (ours)	**68.7**	**49.7**	**53.2**	**57.2**

Table 2: Results of Cityscape → Thermal-FLIR (%).

Methods	Car	Bicycle	Person	mAP
DA-Faster [11]	59.7	31.6	46.0	45.8
SWDA [12]	58.3	38.6	44.9	47.3
SCL [48]	61.2	39.6	48.2	49.7
VDD [49]	53.5	35.9	43.1	44.2
CR-DA-DET [22]	58.3	37.6	46.3	47.4
HTCN [37]	55.3	36.2	41.9	44.4
TIA [51]	57.6	40.5	50.4	49.5
UGA [3]	63.6	42.7	52.8	53.0
Baseline	55.8	38.9	48.8	47.8
HDAM (ours)	**64.6**	39.9	**55.9**	**53.5**

The results of Pascal VOC → Thermal-FLIR (%) are reported in Table 3. HDAM reaches **45.0%** mAP, improving over UGA (43.1%) and over the Baseline (35.4%) by **+9.6%**, indicating stronger robustness when everyday images are transferred to thermal night scenes.

Table 3: Results of Pascal VOC → Thermal-FLIR (%).

Methods	Car	Bicycle	Person	mAP
DA-Faster [11]	44.7	10.9	26.2	27.2
SWDA [12]	45.6	21.8	43.2	36.9
SCL [48]	46.3	23.6	47.1	39.0
VDD [49]	44.3	24.0	38.8	35.7
CR-DA-DET [22]	50.5	20.8	43.5	38.3
HTCN [37]	37.8	23.6	33.3	31.6
TIA [51]	34.7	19.5	28.7	27.6
UGA [3]	50.8	29.8	48.5	43.1
Baseline	46.8	20.7	38.6	35.4
HDAM (ours)	**55.9**	29.6	**49.6**	**45.0**

Figure 3 further summarizes the cross-task comparison. Compared with the strongest prior method, HDAM improves mAP by +1.2, +0.5 and +1.9 points on Visible-FLIR, Cityscape and Pascal VOC transfers, respectively. Taken together, these results suggest that the proposed hierarchical adaptation strategy remains effective under both paired cross-modal transfer and larger scene-level domain shifts, while the gains vary by task difficulty.

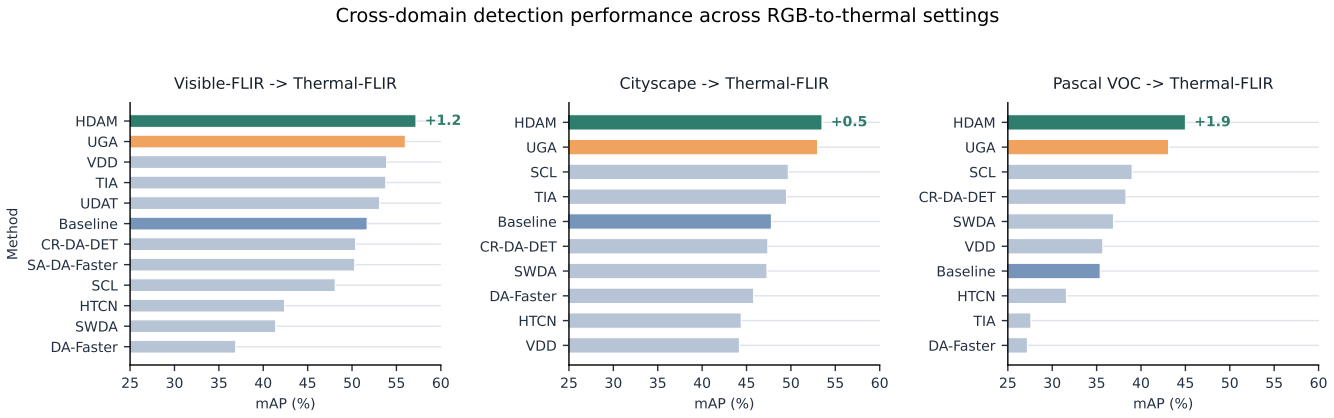

Figure 3: Overall mAP comparison across three RGB-to-thermal adaptation settings. HDAM consistently achieves the best performance, and the annotated margins indicate its improvement over the strongest competing method in each setting.

4.4 Ablation Study

To evaluate the contribution of each proposed component, we perform ablation experiments on the Visible→Thermal FLIR task. The results are summarized in Table 4. Starting from the Baseline, we progressively integrate the ECA, MGSD and ICA modules. From Table 4, several important observations can be made:

Effectiveness of ECA. Removing ECA causes a notable drop in mAP (-1.1) and degrades performance on texture-sensitive categories such as Bicycle. This demonstrates that entropy-guided attention effectively strengthens low- and mid-level adaptation across modalities.

Role of MGSD. Excluding MGSD reduces the mAP from 57.2 to 54.7, mainly due to weaker performance on the Bicycle and Person classes. The global semantic discriminator in MGSD enhances category-level alignment and stabilizes feature distribution across the two domains.

Impact of ICA. Without ICA, the mAP noticeably drops by 2.1 points compared to the full model. This indicates that

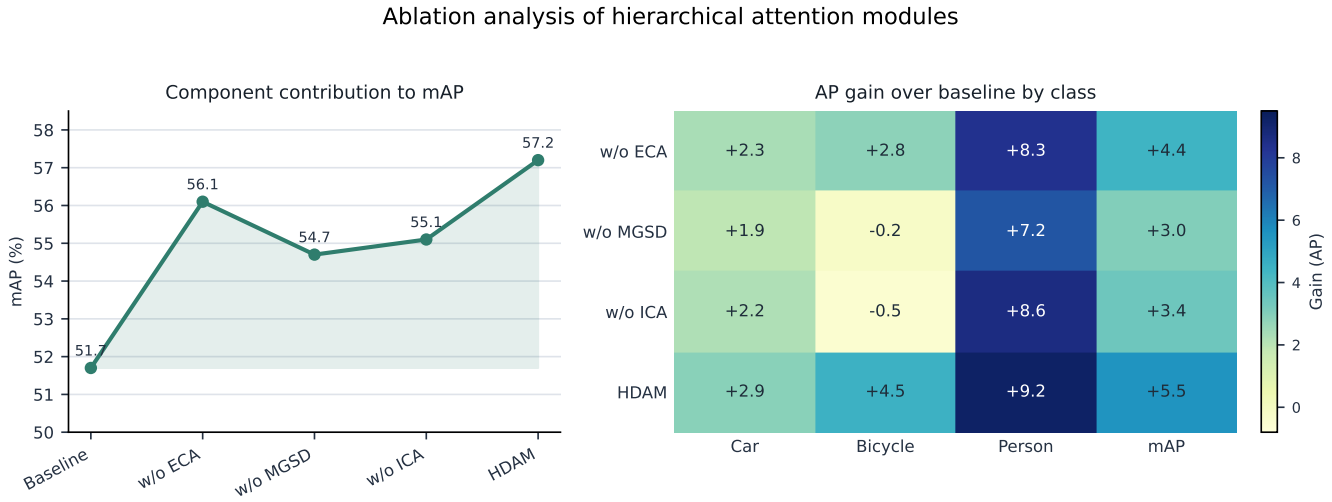

Figure 4: Ablation visualization on the Visible→Thermal FLIR task. The left panel reports mAP changes, while the right panel shows class-wise AP gains over the baseline.

integrating contextual features into instance-level representations significantly improves detection precision, particularly for heavily occluded or small targets.

Overall Synergy. The full HDAM model achieves the highest mAP of 57.2, surpassing the Baseline by +5.5 points. As visualized in Figure 4, the complete model also produces stable class-wise gains, especially for Person (+9.2 AP) and Bicycle (+4.5 AP). These results indicate that HDAM yields complementary benefits, progressively narrowing both global and local modality gaps.

Table 4: Ablation study of HDAM on the Visible→Thermal FLIR task (%). Each module contributes to performance improvement, and their combination yields the best result.

Method	Car	Bicycle	Person	mAP
Baseline	65.8	45.2	44.0	51.7
HDAM (w/o ECA)	68.1	48.0	52.3	56.1
HDAM (w/o MGSD)	67.7	45.0	51.2	54.7
HDAM (w/o ICA)	68.0	44.7	52.6	55.1
HDAM (full)	**68.7**	**49.7**	**53.2**	**57.2**

4.5 Visualization of ROI-Level Feature Distributions

Figure 5 presents the t-SNE visualization of ROI-level embeddings for the Baseline and HDAM on the Visible-FLIR → Thermal-FLIR task. The Baseline features still exhibit noticeable domain separation and looser class structures, especially for Person and Bicycle. In contrast, HDAM produces more compact category clusters while improving the overlap between source and target samples within the same class. This observation provides direct evidence that the proposed hierarchical alignment strategy reduces cross-modal discrepancy without sacrificing instance-level discriminability.

4.6 Visualization of Mid-Level Features

Figure 6 visually compares the heat-maps of mid-level features: Baseline responses are noisy and domain-specific, whereas HDAM yields sharper, object-centric activations across both modalities, supporting the role of entropy-guided feature reweighting.

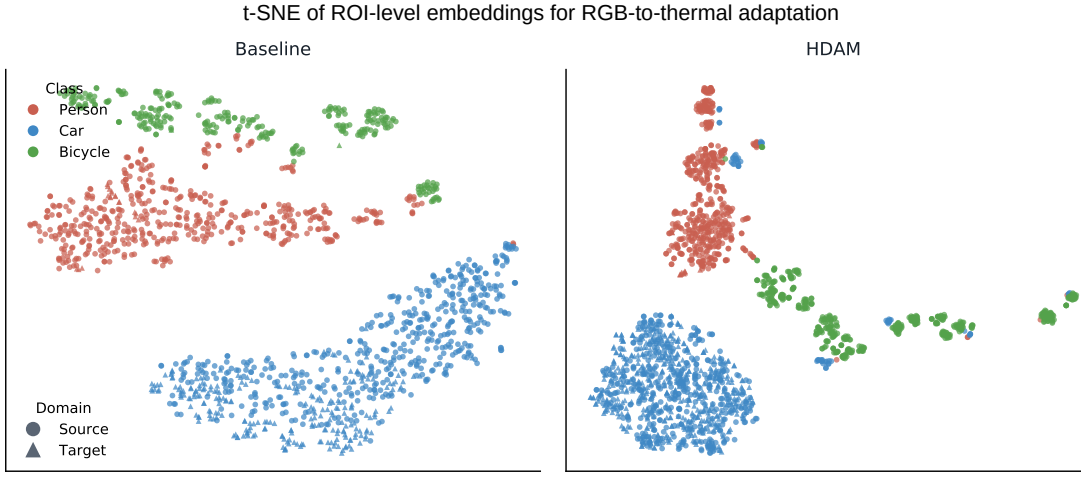

Figure 5: t-SNE visualization of ROI-level embeddings on the Visible-FLIR → Thermal-FLIR task. Different colors denote object categories, while marker shapes indicate source and target domains. Compared with the baseline, HDAM yields more compact class-wise clusters and better cross-domain overlap.

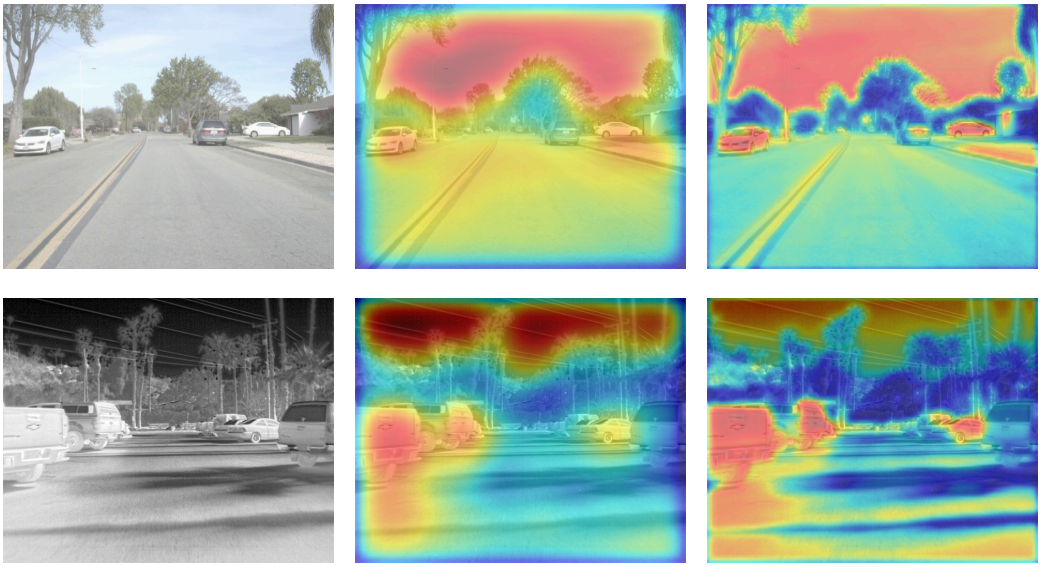

Figure 6: Illustration of the heat-maps of mid-level features on the Visible-FLIR → Thermal-FLIR task.

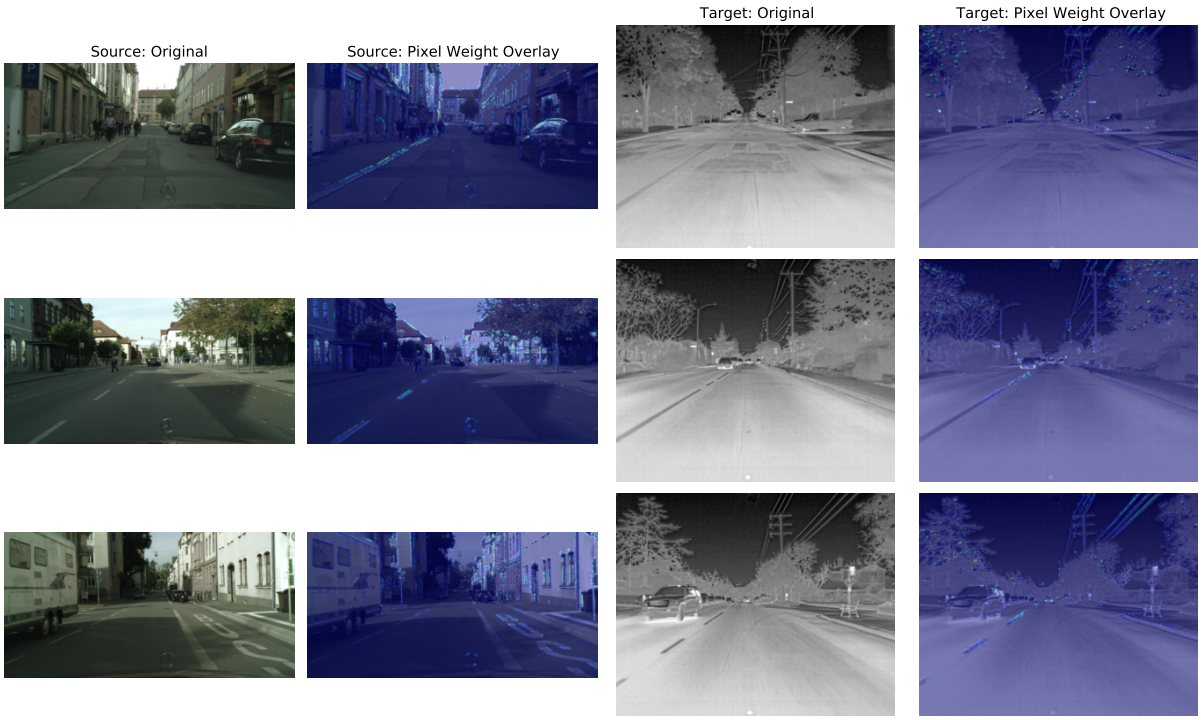

Figure 7: Pixel-level attention visualization on the Cityscape → Thermal-FLIR task. For each source-target pair, the left image is the original input and the right image is the corresponding pixel-weight overlay. HDAM consistently emphasizes domain-shared structural cues, such as road boundaries, vehicle contours, and lane markings, across both modalities.

4.7 Visualization of Pixel-Level Attention

Figure 7 visualizes the pixel-level attention learned by ECA. The overlays show that HDAM focuses on domain-shared local structures, such as road boundaries, object contours, and salient foreground-background transitions, in both source RGB and target thermal images. This behavior indicates that the entropy-guided weighting suppresses modality-specific noise and facilitates more robust early-stage cross-modal alignment.

4.8 Example of Detection Results

Figure 8 presents qualitative detection results on the Pascal VOC to Thermal-FLIR task. As shown, HDAM generally outperforms both the Baseline and UGA [3]. In several examples, HDAM captures finer details, as evidenced by tighter bounding boxes around individuals, and more effectively identifies distant or less distinct targets. These qualitative differences are consistent with the role of hierarchical adaptation and attention mechanisms in improving cross-modal feature alignment for thermal object detection.

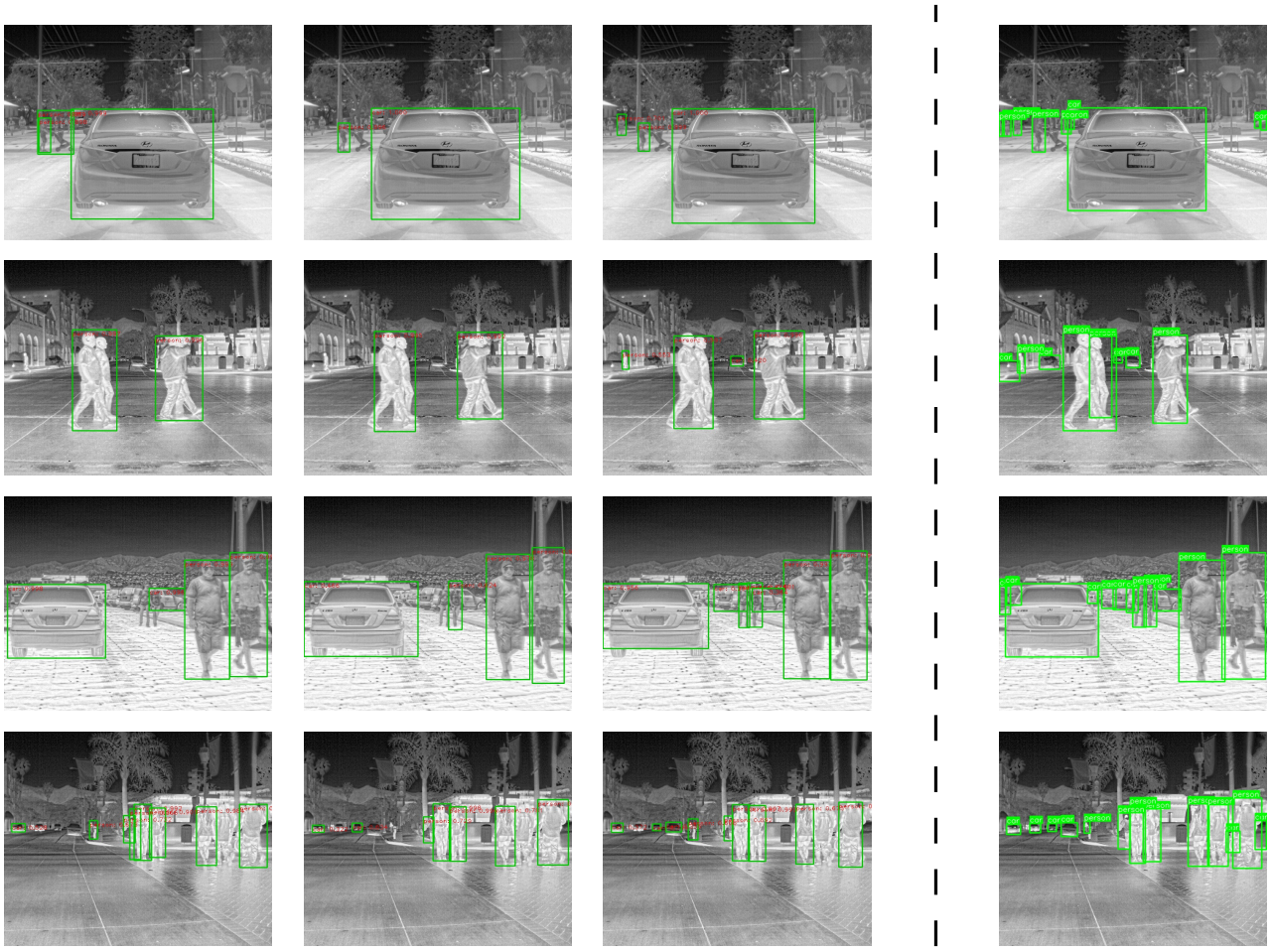

Figure 8: Illustration of the detection results on the Pascal VOC → Thermal-FLIR task.

5 Conclusion

In this paper, we introduce HDAM, a hierarchical domain adaptation framework for visible-to-thermal object detection. By aligning features progressively from low-level textures to high-level semantics, HDAM mitigates both global semantic drift and instance-context mismatch. Our framework integrates multi-level attention and multi-branch semantic alignment to enhance cross-modal feature discriminability. Experiments on benchmark datasets show that HDAM achieves improved performance over strong baselines and competitive recent methods, supporting its usefulness for cross-modal detection tasks.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author contributions

Y.L. conducted the experiments, implemented the method, prepared the figures, and drafted the manuscript. M.G. contributed to methodology discussion, experimental analysis, and manuscript revision. C.C. supervised the research, guided the study design, and revised the manuscript. All authors reviewed and approved the final manuscript.

Conflict of Interest

The authors declare that they have no competing interests.

Data Availability

The datasets used in this study are publicly available. Code is available upon reasonable request.

References

[1] Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In Conference on Computer Vision and Pattern Recognition (CVPR) 2011, pp. 1521–1528 (2011). https://doi.

org/10.1109/CVPR.2011.5995347

[2] Dai, Y., Wu, Y., Zhou, F., Barnard, K.: Attentional local contrast networks for infrared small target detection. IEEE transactions on geoscience and remote sensing **59**(11), 9813–9824 (2021). https://doi.org/10.1109/TGRS.2020.3044958

[3] Shi, C., Zheng, Y., Chen, Z.: Domain adaptive thermal object detection with unbiased granularity alignment. ACM Transactions on Multimedia Computing, Communications and Applications **20**(9), 1–23 (2024). https://doi.org/10.1145/3665892

[4] Berjawi, J., Dupas, Y., Cérin, C.: Towards a Generalizable Fusion Architecture for Multimodal Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2192–2200 (2025). https://doi.org/10.1109/ICCVW69036.2025.00231

[5] Pan, S.J., Yang, Q.: A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering **22**(10), 1345–1359 (2010). https://doi.org/10.1109/TKDE.2009.191

[6] Oza, P., Sindagi, V.A., VS, V., Patel, V.M.: Unsupervised domain adaptation of object detectors: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(6), 4018–4040 (2023). https://doi.org/10.1109/TPAMI.2022.3217046

[7] Wilson, G., Cook, D.J.: A survey of unsupervised deep domain adaptation. ACM Transactions on Intelligent Systems and Technology (TIST) **11**(5), 1–46 (2020). https://doi.org/10.1145/3400066

[8] Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In Proceedings of the 32nd International Conference on Machine Learning, pp. 1180–1189 (2015)

[9] Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In Proceedings of the 32nd International Conference on Machine Learning, pp. 97–105 (2015)

[10] Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In European Conference on Computer Vision, pp. 443–450 (2016). https://doi.org/10.1007/978-3-319-49409-8_35

[11] 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). https://doi.org/10.1109/CVPR.2018.00352

[12] 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). https://doi.org/10.1109/CVPR.2019.00712

[13] Kim, Y., Cho, D., Han, K., Panda, P., Hong, S.: Domain adaptation without source data. IEEE Transactions on Artificial Intelligence **2**(6), 508–518 (2021). https://doi.org/10.1109/TAI.2021.3110179

[14] Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1406–1415 (2019). https://doi.org/10.1109/ICCV.2019.00149

[15] Do, D.P., Kim, T., Na, J., Kim, J., Lee, K., Cho, K., Hwang, W.: D3t: Distinctive dual-domain teacher zigzagging across rgb-thermal gap for domain-adaptive object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 23313–23322 (2024). https://doi.org/10.1109/CVPR52733.2024.02200

[16] Li, H., Zhang, R., Yao, H., Zhang, X., Hao, Y., Song, X., Peng, S., Zhao, Y., Zhao, C., Wu, Y., *et al.*: SEEN-DA: SEmantic ENtropy guided Domain-aware Attention for Domain Adaptive Object Detection. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 25465–25475 (2025). https://doi.org/10.1109/CVPR52734.2025.02371

[17] Zhou, Y., Zhang, H., Sun, Y., Liu, Y., Kung, S.-Y.: Illumination Guided Domain Adaptation Object Detection in Thermal Imagery. Neurocomputing **653**, 131237 (2025). https://doi.org/10.1016/j.neucom.2025.131237

[18] Yu, H., Deng, J., Li, W., Duan, L.: Towards unsupervised model selection for domain adaptive object detection. In Proceedings of the 38th International Conference on Neural Information Processing Systems, pp. 58423–58444 (2024)

[19] Rivadeneira, R.E., Sappa, A.D., Wang, C., Jiang, J., Zhong, Z., Chen, P., Wang, S.: Thermal Image Super-Resolution Challenge Results - PBVS 2024. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 3113–3122 (2024). https://doi.org/10.1109/CVPRW63382.2024.00317

[20] 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). https://doi.org/10.1109/CVPR42600.2020.01237

[21] Medeiros, H.R., Belal, A., Muralidharan, S., Granger,

E., Pedersoli, M.: Visual Modality Prompt for Adapting Vision-Language Object Detectors. In 2025 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2172–2182 (2025). https://doi.org/10.1109/ICCV51701.2025.00210

[22] Xu, C.-D., Zhao, X.-R., Jin, X., Wei, X.-S.: Exploring Categorical Regularization for Domain Adaptive Object Detection. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11721–11730 (2020). https://doi.org/10.1109/CVPR42600.2020.01174

[23] Zellinger, W., Grubinger, T., Lughofer, E., Natschläger, T., Saminger-Platz, S.: Central moment discrepancy (CMD) for domain-invariant representation learning. arXiv preprint arXiv:1702.08811 (2017). https://doi.org/10.48550/arXiv.1702.08811

[24] Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. The journal of machine learning research **13**(1), 723–773 (2012)

[25] Na, J., Jung, H., Chang, H.J., Hwang, W.: FixBi: Bridging Domain Spaces for Unsupervised Domain Adaptation. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1094–1103 (2021). https://doi.org/10.1109/CVPR46437.2021.00115

[26] Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(6), 1137–1149 (2017). https://doi.org/10.1109/TPAMI.2016.2577031

[27] He, Z., Zhang, L.: Multi-Adversarial Faster-RCNN for Unrestricted Object Detection. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6667–6676 (2019). https://doi.org/10.1109/ICCV.2019.00677

[28] Jiao, L., Wei, H., Pan, Q.: Region and Sample Level Domain Adaptation for Unsupervised Infrared Target Detection in Aerial Remote Sensing Images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **18**, 11289–11306 (2025). https://doi.org/10.1109/JSTARS.2025.3561737

[29] Neuwirth–Trapp, M., Bieshaar, M., Paudel, D.P., Van Gool, L.: Incremental Object Detection with Prompt-Based Methods. In 2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 5224–5232 (2025). https://doi.org/10.1109/ICCVW69036.2025.00544

[30] Li, T., Ye, M., Wu, T., Li, N., Li, S., Tang, S., Ji, L.: Pseudo Visible Feature Fine-Grained Fusion for Thermal Object Detection. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6710–6719 (2025). https://doi.org/10.1109/CVPR52734.2025.00629

[31] Peng, H., Hu, Y., Yu, B., Zhang, Z.: TCAINet an RGB T salient object detection model with cross modal fusion and adaptive decoding. Scientific Reports **15**(1), 14266 (2025). https://doi.org/10.1038/s41598-025-98423-z

[32] Roy, S., Krivosheev, E., Zhong, Z., Sebe, N., Ricci, E.: Curriculum Graph Co-Teaching for Multi-Target Domain Adaptation. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5347–5356 (2021). https://doi.org/10.1109/CVPR46437.2021.00531

[33] 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). https://doi.org/10.1007/s11263-021-01447-x

[34] Chen, H.-Y., Chao, W.-L.: Gradual domain adaptation without indexed intermediate domains. In Proceedings of the 35th International Conference on Neural Information Processing Systems, pp. 8201–8214 (2021)

[35] Wu, Z., Wang, X., Gonzalez, J., Goldstein, T., Davis, L.: ACE: Adapting to Changing Environments for Semantic Segmentation. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2121–2130 (2019). https://doi.org/10.1109/ICCV.2019.00221

[36] Guan, D., Huang, J., Xiao, A., Lu, S., Cao, Y.: Uncertainty-Aware Unsupervised Domain Adaptation in Object Detection. IEEE Transactions on Multimedia **24**, 2502–2514 (2022). https://doi.org/10.1109/TMM.2021.3082687

[37] 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). https://doi.org/10.1109/CVPR42600.2020.00889

[38] Zhang, Y., Wang, Z., Li, J., Zhuang, J., Lin, Z.: Towards effective instance discrimination contrastive loss for unsupervised domain adaptation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11388–11399 (2023). https://doi.org/10.1109/ICCV51070.2023.01046

[39] Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7167–7176 (2017). https://doi.org/10.1109/CVPR.2017.316

[40] Sun, T., Lu, C., Ling, H.: Local context-aware active domain adaptation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 18634–18643 (2023). https://doi.org/10.1109/

ICCV51070.2023.01708

[41] Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal Loss for Dense Object Detection. In 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2999–3007 (2017). https://doi.org/10.1109/ICCV.2017.324

[42] 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). https://doi.org/10.1109/ICIP40778.2020.9191080

[43] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes Dataset for Semantic Urban Scene Understanding. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3213–3223 (2016). https://doi.org/10.1109/CVPR.2016.350

[44] Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010). https://doi.org/10.1007/s11263-009-0275-4

[45] He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

[46] 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). https://doi.org/10.1109/CVPR.2009.5206848

[47] He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017). https://doi.org/10.1109/TPAMI.2018.2844175

[48] Shen, Z., Maheshwari, H., Yao, W., Savvides, M.: Scl: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses. arXiv preprint arXiv:1911.02559 (2019). https://doi.org/10.48550/arXiv.1911.02559

[49] Wu, A., Liu, R., Han, Y., Zhu, L., Yang, Y.: Vector-Decomposed Disentanglement for Domain-Invariant Object Detection. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9322–9331 (2021). https://doi.org/10.1109/ICCV48922.2021.00921

[50] 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). https://doi.org/10.1007/s00371-021-02386-x

[51] Zhao, L., Wang, L.: Task-specific inconsistency alignment for domain adaptive object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14217–14226 (2022). https://doi.org/10.1109/CVPR52688.2022.01382