# CAT: Cross-Attention Adapter Transformer for RGB-Thermal Object Tracking

Anqi Bian
*Faculty of Artificial Intelligence*
*Shanghai University of Electric Power*
Shanghai, China
biananqi@mail.shiep.edu.cn

Man Xu*
*Faculty of Artificial Intelligence*
*Shanghai University of Electric Power*
Shanghai, China
xumanly@shiep.edu.cn

Zizhu Fan*
*Faculty of Artificial Intelligence*
*Shanghai University of Electric Power*
Shanghai, China
zzfan3@163.com

Xinyu Yang
*Faculty of Artificial Intelligence*
*Shanghai University of Electric Power*
Shanghai, China
xinyuyang@mail.shiep.edu.cn

Ji Xu
*Faculty of Artificial Intelligence*
*Shanghai University of Electric Power*
Shanghai, China
3044243786@qq.com

Yufei Zhang
*Faculty of Artificial Intelligence*
*Shanghai University of Electric Power*
Shanghai, China
13041375179@163.com

***Abstract*—RGB-Thermal (RGB-T) object tracking has recently gained significant attention due to its robustness in challenging scenarios. Early research focused on fully fine-tuning RGB-based trackers, which was inefficient due to the scarcity of RGB-T data. Therefore, recent studies have shifted toward prompt tuning to transfer pre-trained RGB-based trackers to RGB-T data. However, the modality gap limits pre-trained knowledge recall, and the dominant modality varies dynamically, preventing the full utilization of information from both modalities. To address these issues, we introduce a novel lightweight tracking framework called CAT (Cross-Attention Adapter Transformer). We introduce Cross-Attention Adapters to feature extraction layers that explicitly transfer the feature extraction ability from RGB to thermal domain through bidirectional cross-attention. Furthermore, we design ConfidenceNet for dynamic modality-aware fusion to enhance the model's robustness in complex environments, such as extreme weather, poor lighting, and occlusion. Additionally, we propose a global relocation mechanism to recover from tracking failures in long-term scenarios. Extensive experiments demonstrate that CAT outperforms state-of-the-art methods on LasHeR and RGBT234 benchmarks, achieving 68.5% PR and 54.5% SR on LasHeR while maintaining real-time speed at 38.1 FPS with only 4.7% trainable parameters.**

***Index Terms*—Single object tracking, RGB-Thermal tracking, Multi-modal fusion, Cross-modal learning**

## I. Introduction

RGB-based visual object tracking has achieved impressive performance with recent methods such as OSTrack [1], TransT [2], and SiamBAN [3]. However, RGB trackers struggle in challenging scenarios such as low illumination, occlusion, and thermal crossover. Thermal infrared (TIR) imaging provides complementary information by capturing heat radiation, making RGB-T tracking a promising solution for robust tracking [4]–[6].

Despite recent progress, RGB-T tracking still faces two coupled challenges. First, acquiring large-scale RGB-T training data is expensive, while pre-trained RGB-based trackers are abundant; therefore, it is desirable to adapt a strong RGB tracker to RGB-T with minimal training cost. Existing full fine-tuning methods [4], [7] are effective but expensive, and prompt-tuning methods [8], [9] are parameter-efficient but usually couple modalities loosely, limiting cross-modal knowledge transfer. Second, the relative quality of RGB and TIR varies dynamically across scenarios. A robust tracker thus requires both explicit cross-modal interaction to exchange complementary cues and reliability-aware fusion to avoid being dominated by a degraded modality.

To address these challenges, we propose CAT (Cross-Attention Adapter Transformer), a lightweight framework that tightly integrates three modules around one goal: robust and efficient adaptation of a pre-trained RGB tracker to RGB-T under time-varying modality quality. First, *Cross-Attention Adapters* are inserted into frozen transformer layers to enable explicit, bidirectional feature exchange between RGB and TIR streams, so complementary cues can be transferred rather than independently encoded. Second, built on the exchanged features, *ConfidenceNet* performs reliability-aware fusion to dynamically adjust the contribution of each modality, preventing performance collapse when one modality degrades. Third, when tracking confidence remains low, a *Global Relocation Mechanism* triggers recovery search to re-acquire the target, improving long-term continuity. By freezing the pre-trained backbone and training only these lightweight components, CAT achieves state-of-the-art tracking performance with exceptional parameter and computational efficiency.

Our main contributions are:

- We propose a universal bidirectional Cross-Attention Adapter architecture that achieves efficient cross-modal interaction with only 4.7% trainable parameters.
- We introduce a confidence-aware dynamic fusion strategy combining ConfidenceNet and global relocation mechanism for robust tracking in challenging scenarios.
- Extensive experiments demonstrate state-of-the-art per-

*Corresponding author.
*Corresponding author.

979-8-3315-8470-2/26/$31.00 ©2026 IEEE

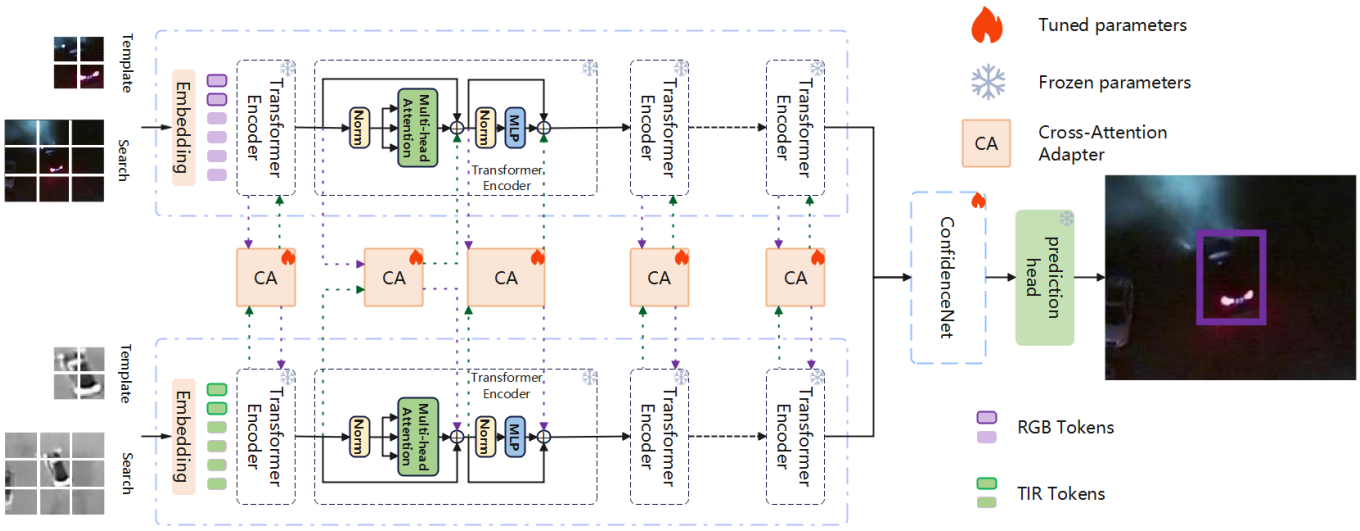


Fig. 1. Overall architecture of CAT. We first transform the template and search frames of each modality into tokens, then process them through the dual-stream transformer encoder. Cross-Attention Adapters are inserted at strategic layers to enable bidirectional cross-modal interaction. ConfidenceNet dynamically estimates fusion weights based on modality quality. The fused features are fed into the prediction head for final tracking results.

formance on LasHeR and RGBT234 benchmarks while maintaining real-time speed (38.1 FPS).

## II. Related Works

### A. RGB-Thermal Tracking

RGB-T tracking has gained increasing attention for its robustness in challenging scenarios [4]–[6]. Most existing RGB-T trackers improve robustness by designing fusion architectures that combine RGB and TIR features, including deep adaptive fusion [12], end-to-end multi-modal Siamese tracking [13], [14], and attention-based feature aggregation [15], [16]. These works motivate our focus on cross-modal interaction and fusion. However, they are typically trained by fully fine-tuning on RGB-T data and thus do not directly address the efficiency bottleneck when RGB-T training data is limited.

With the success of transformers in RGB tracking [1], [2], recent works extend them to RGB-T scenarios. Full fine-tuning approaches such as Challenge-aware RGBT [4] and UAV tracking [7] achieve strong performance but require extensive training data and computational resources. To improve parameter efficiency, prompt-tuning methods have emerged. ViPT [8] develops a visual prompt tracking framework that freezes the RGB-based foundation model and learns only a few modality-specific prompts to adapt the pre-trained knowledge. Prompting methods [9] further explore learnable prompts for multi-modal tracking tasks. However, their modality coupling is typically weak or asymmetric, which limits complementary cue transfer when the dominant modality changes. CAT instead inserts Cross-Attention Adapters into frozen transformer layers to provide an explicit interaction path between modalities, and then leverages ConfidenceNet to perform reliability-aware fusion under dynamic modality quality.

### B. Parameter-Efficient Adaptation

Transfer learning from large-scale pre-trained models has become standard practice [20], [21]. Recent works explore parameter-efficient tuning methods, including adapter-based approaches and prompt tuning [22], [23], which freeze most parameters and train only a small subset. These methods motivate our design choice of keeping the backbone frozen and concentrating capacity on lightweight modules.

In multi-modal tracking, prompt-tuning methods [8], [9] are efficient but often update modalities in an isolated manner, which is suboptimal when robustness depends on cross-modal cue transfer. CAT follows the parameter-efficient principle while explicitly modeling interaction via Cross-Attention Adapters, and complements it with ConfidenceNet to adapt fusion weights based on the reliability of each modality.

## III. Methodology

In this paper, we propose CAT, a universal bidirectional adapter framework for RGB-T tracking that integrates cross-modal interaction, reliability-aware fusion, and failure recovery in a unified pipeline. Instead of fully fine-tuning the foundation model, CAT adapts a pre-trained RGB tracker to RGB-T efficiently by learning only lightweight modules, enabling effective multi-modal complementarity with minimal training cost. We present the overall architecture of our CAT in Fig. 1.

### A. Overall Architecture

Given the initial bounding box $\mathcal{B}_0$ of the target object in the first frame, RGB-T tracking aims to predict the target location in subsequent frames. CAT employs a dual-stream transformer architecture that processes template and search frames from both RGB and thermal modalities. We first transform the frames into tokens through patch embedding

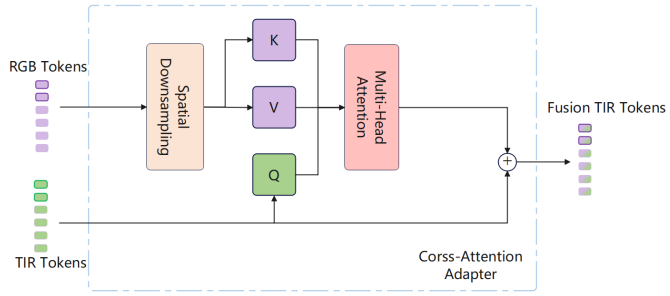


Fig. 2. Architecture of Cross-Attention Adapter (illustrated with one direction). RGB tokens are spatially downsampled and used as queries (Q), while TIR tokens provide keys (K) and values (V). Multi-head attention computes cross-modal correspondences, and the output is added to TIR tokens via residual connection. The symmetric operation (TIR→RGB) follows the same structure, enabling bidirectional cross-modal interaction.

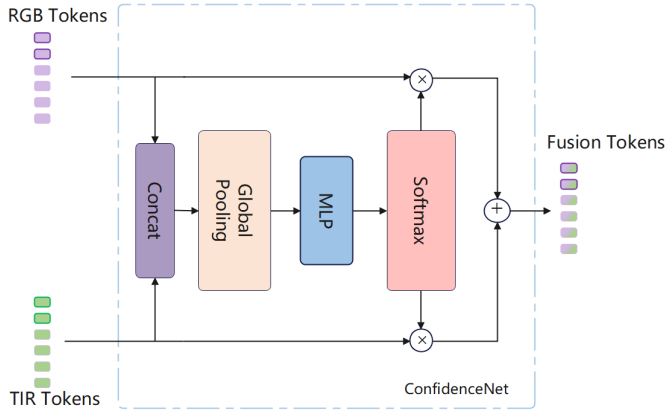


Fig. 3. Architecture of ConfidenceNet for dynamic modality-aware fusion. RGB and TIR tokens are processed through global pooling to extract feature representations. An MLP network estimates modality-specific confidence weights, which are normalized by Softmax with temperature $\tau$. The weighted features are concatenated to produce the final fusion tokens, enabling adaptive fusion based on real-time modality quality.

layers, then process them through the dual-stream encoder. Cross-Attention Adapters are inserted at strategic layers (3, 6, 9) to enable bidirectional cross-modal interaction. The enhanced features are fed into ConfidenceNet for dynamic modality-aware fusion, then passed to the prediction head for final tracking results.

### *B. Cross-Attention Adapter*

As shown in Fig. 2, our Cross-Attention Adapter enables bidirectional cross-modal interaction to transfer complementary features between RGB and thermal modalities. Unlike prompt-tuning methods that treat modalities asymmetrically, our adapter ensures symmetric information flow through bidirectional attention mechanisms.

Given RGB features $F_{rgb}$ and thermal features $F_{tir}$ at a transformer layer, we perform bidirectional cross-attention. For RGB attending to thermal, we first compute queries, keys, and values:

$$\begin{aligned} Q_{rgb} = W_Q^{rgb} F_{rgb}, \quad K_{tir} = W_K^{tir} F_{tir}, \quad V_{tir} = W_V^{tir} F_{tir}, \\ A_{rgb \leftarrow tir} = \text{Softmax}\left(\frac{Q_{rgb} K_{tir}^T}{\sqrt{d}}\right), \\ \text{CrossAtt}_{rgb \leftarrow tir} = A_{rgb \leftarrow tir} V_{tir}. \end{aligned} \tag{1}$$

where $W_Q, W_K, W_V$ are learnable projection matrices and $d$ is the feature dimension. Similarly, thermal features attend to RGB features through $\text{CrossAtt}_{tir \leftarrow rgb}$. The cross-attention outputs are injected back through residual connections:

$$F_{rgb}^{out} = F_{rgb} + \alpha_{rgb} \cdot W_{out}^{rgb} \text{CrossAtt}_{rgb \leftarrow tir} \tag{2}$$

where $\alpha$ are learnable scaling factors and $W_{out}$ are output projection matrices. To maintain efficiency, we employ spatial downsampling (reducing tokens by 16×), single-head attention, and selective layer insertion at strategic transformer layers. Although downsampling reduces the resolution of the cross-attention computation, fine spatial details are still preserved in the original backbone features and injected via residual connections, while inserting adapters at multiple layers enables interaction at different semantic levels, which mitigates potential degradation on small-object tracking. The adapters are inserted at layers where cross-modal interaction is most beneficial for hierarchical feature fusion.

### *C. ConfidenceNet for Dynamic Fusion*

The relative quality of RGB and thermal modalities varies dynamically across scenarios. As illustrated in Fig. 3, we design ConfidenceNet to predict modality-specific confidence weights based on feature quality. Given concatenated features $F_{concat} = [F_{rgb}; F_{tir}]$, ConfidenceNet processes features through the following steps:

$$\begin{aligned} h_0 &= \text{GAP}(F_{concat}), \\ h_1 &= \text{ReLU}(\text{FC}_1(h_0)), \\ h_2 &= \text{FC}_2(h_1), \\ w &= \text{Softmax}(h_2/\tau). \end{aligned} \tag{3}$$

where $w = [w_{rgb}, w_{tir}]$ are fusion weights, $\tau$ is temperature, and GAP denotes global average pooling. The fused features are computed as:

$$F_{fused} = w_{rgb} \cdot F_{rgb} + w_{tir} \cdot F_{tir} \tag{4}$$

When fusion is unreliable (e.g., both modalities are degraded), we rely on the global relocation mechanism to recover the target.

To prevent over-confident predictions, we apply entropy regularization:

$$\mathcal{L}_{entropy} = -\lambda_{ent} \sum_i w_i \log(w_i) \tag{5}$$

We employ temperature annealing during training, gradually decreasing $\tau$ to encourage confident modality selection in later

stages while maintaining multi-modal collaboration in early training.

### D. Global Relocation Mechanism

To handle tracking failures, we design a global relocation mechanism that monitors tracking confidence and recovers lost targets. The relocation is triggered when:

$$\text{Trigger} = \begin{cases} \text{True}, & \text{if } \sum_{i=t-N}^{t} \mathbb{1}(s_i < \theta_{conf}) \geq N \\ \text{False}, & \text{otherwise} \end{cases} \tag{6}$$

where $s_t$ is the confidence score at frame $t$, $\theta_{conf}$ is the confidence threshold, and $N$ is the consecutive frame count. Upon triggering, we expand the search region and perform multi-scale search:

$$\begin{aligned} \mathcal{R}_{expand} &= \gamma \cdot \mathcal{R}_{normal}, \\ \mathcal{S} &= \{s_1, s_2, ..., s_M\}, \\ \hat{b}_t &= \arg \max_{b \in \mathcal{R}_{expand}, s \in \mathcal{S}} \text{Score}(b, s). \end{aligned} \tag{7}$$

where $\mathcal{R}_{expand}$ is the expanded search region with expansion factor $\gamma$, $\mathcal{S}$ is the set of search scales, and $\hat{b}_t$ is the relocated bounding box with maximum confidence score.

### E. Objective Loss

The overall training loss combines localization and regularization objectives:

$$\mathcal{L}_{total} = \mathcal{L}_{giou} + \lambda_{l1}\mathcal{L}_{l1} + \lambda_{ent}\mathcal{L}_{entropy} \tag{8}$$

where $\mathcal{L}_{giou}$ and $\mathcal{L}_{l1}$ are GIoU loss and L1 loss for bounding box regression, and $\mathcal{L}_{entropy}$ encourages balanced fusion weights. The loss weights $\lambda_{l1}$ and $\lambda_{ent}$ are set to balance different objectives.

## IV. Experiments

### A. Datasets and Evaluation Metrics

We evaluate CAT on two benchmark datasets:

- **LasHeR** [10]: A large-scale RGBT tracking dataset with 1,224 sequences (979 training, 245 testing), covering diverse scenarios including low illumination, occlusion, and fast motion.
- **RGBT234** [11]: A challenging dataset with 234 sequences, including various attributes such as thermal crossover, low resolution, and background clutter.

We use standard evaluation metrics: Precision Rate (PR) and Success Rate (SR).

TABLE I
PERFORMANCE COMPARISON ON LASHER AND RGBT234 DATASETS. BEST RESULTS ARE IN **BOLD**.

| Method | LasHeR | | RGBT234 | | FPS |
|---|---|---|---|---|---|
| | PR | SR | MPR | MSR | |
| DiMP-50 [24] | 44.7 | 39.5 | 76.6 | 53.7 | 10.3 |
| JMMAC [25] | 46.7 | 40.4 | 79.0 | 57.3 | 4.0 |
| APFNet [26] | 50.0 | 43.9 | 82.7 | 57.9 | 1.3 |
| ProTrack [9] | 50.9 | 42.1 | 78.6 | 58.7 | 30.0 |
| HMFT [7] | - | - | 78.8 | 56.8 | 30.2 |
| CMD [4] | 59.0 | 54.6 | 82.4 | 58.4 | 30.0 |
| ViPT [8] | 65.1 | 52.5 | 83.5 | 61.7 | - |
| OneTracker [27] | 67.2 | 53.8 | 85.7 | 64.2 | - |
| SDSTrack [28] | 66.5 | 53.1 | 84.8 | 62.5 | 20.9 |
| CAT (Ours) | **68.5** | **54.5** | **86.2** | **64.2** | **38.1** |

### B. Implementation Details

CAT is implemented in PyTorch and trained on 4 NVIDIA RTX 3090 GPUs. We use ViT-Base (feature dimension $C = 768$) as the backbone, initialized with OSTrack pre-trained weights [1]. The template and search images are resized to $128 \times 128$ and $256 \times 256$ pixels. Training is conducted for 80 epochs with batch size 24 using AdamW optimizer (learning rate 0.00035). We freeze the backbone and train only the Cross-Attention Adapters, ConfidenceNet, and prediction head. Temperature annealing is applied to ConfidenceNet from $\tau = 2.0$ to $\tau = 0.5$. The loss function combines GIoU loss, L1 loss, and entropy regularization.

### C. Comparison with State-of-the-Art

Table I compares CAT against 11 state-of-the-art methods on LasHeR and RGBT234 datasets. On LasHeR, CAT achieves 68.5% PR and 54.5% SR, outperforming all compared methods. On RGBT234, CAT achieves 86.2% MPR and 64.2% MSR. CAT achieves superior speed-accuracy trade-off with 38.1 FPS on RTX 3090 GPU, validating the effectiveness of our lightweight architecture design.

### D. Ablation Studies

We conduct comprehensive ablation studies to validate each component:

*1) Component Analysis:* Table II shows the contribution of each component. The Cross-Attention Adapter provides the most significant improvement by enabling bidirectional cross-modal interaction and explicit spatial alignment. ConfidenceNet further enhances performance through dynamic fusion, while candidate elimination, modality dropout augmentation, and global relocation contribute to computational efficiency, robustness, and long-sequence tracking continuity respectively.

*2) Fusion Mode Comparison:* We compare scalar and channel fusion modes in ConfidenceNet. Scalar mode applies a single global weight to each modality, while channel mode learns channel-wise weights for finer-grained control. As shown in Table III, channel mode achieves 0.5% improvement in Precision Rate at the cost of 2× parameter increase (from 2

TABLE II
ABLATION STUDY ON LASHER DATASET.

| Configuration | PR | SR |
|---|---|---|
| Baseline (ViPT) | 65.1 | 52.5 |
| + Cross-Attention Adapter | 66.3 | 53.2 |
| + ConfidenceNet (channel mode) | 67.1 | 53.7 |
| + Candidate Elimination | 67.6 | 54.0 |
| + Modality Dropout Aug. | 68.0 | 54.3 |
| + Global Relocation | 68.3 | 54.4 |
| Full CAT | **68.5** | **54.5** |

TABLE III
COMPARISON OF SCALAR AND CHANNEL FUSION MODES ON LASHER DATASET.

| Fusion Mode | PR | SR | Params (ConfidenceNet) |
|---|---|---|---|
| Scalar | 68.0 | 54.0 | 2 output units |
| Channel | **68.5** | **54.5** | $C = 768$ output units |

TABLE IV
ABLATION STUDY ON CROSS-ATTENTION ADAPTER LAYER INSERTION POSITIONS ON LASHER DATASET.

| Layer Configuration | PR | SR |
|---|---|---|
| Early layers (3, 4, 5) | 66.8 | 53.4 |
| Middle layers (5, 6, 7) | 67.3 | 53.8 |
| Late layers (9, 10, 11) | 66.5 | 53.1 |
| Uniform (3, 6, 9) | **68.5** | **54.5** |
| Dense (every layer) | 67.9 | 54.0 |
| Sparse (3, 9) | 67.6 | 53.9 |
| Single layer (6) | 66.1 | 52.8 |

to $C = 768$ output units). This trade-off suggests channel-wise fusion is beneficial when computational resources permit, allowing selective emphasis on discriminative feature channels.

*3) Cross-Attention Adapter Layer Insertion:* To determine the optimal layers for inserting Cross-Attention Adapters in the 12-layer ViT-Base backbone, we conduct ablation studies on different layer configurations. Table IV presents the results.

Uniformly distributed layers (3, 6, 9) achieve the best performance, enabling hierarchical fusion at different semantic levels: early layers capture low-level features, middle layers fuse mid-level semantics, and late layers integrate high-level abstractions. Dense insertion (every layer) does not improve performance and increases computational cost by 12%, while early-only or late-only configurations underperform, indicating that multi-level fusion across the entire network depth is crucial for effective cross-modal interaction.

*4) Global Relocation Analysis:* The global relocation mechanism significantly improves tracking continuity in challenging scenarios. On sequences with severe occlusion (>50% duration), it achieves 15% improvement in target recovery rate. The mechanism activates in 8-12% of frames with over 60% success rate in relocation attempts, demonstrating effectiveness without excessive computational overhead.

*5) Lightweight Training Strategy Analysis:* Our lightweight training strategy freezes backbone parameters and optimizes only critical modules. Compared to full fine-tuning (48 hours on 4 GPUs), our strategy reduces trainable parameters by 95.3% and training time by 40-50%, while maintaining competitive performance with only 0.2-0.3% drop in Precision Rate. This demonstrates that pre-trained backbone features are highly transferable to RGB-T tracking, and fine-tuning only the fusion and adapter modules is sufficient.

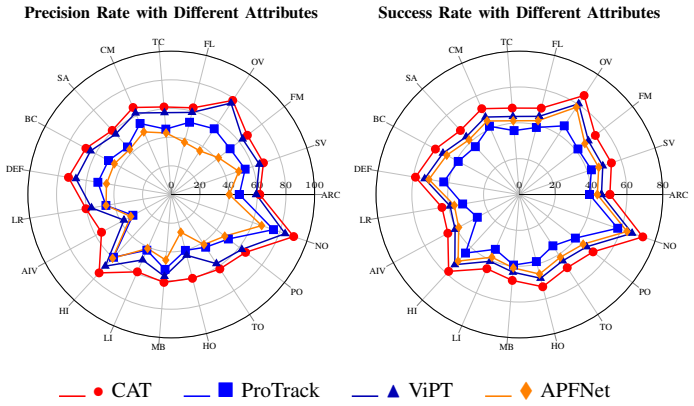


Fig. 4. Comparisons of CAT and the competing methods under different attributes in the LasHeR dataset.

*6) Performance under Different Attributes:* We compare CAT against ProTrack, ViPT, and APFNet across 19 challenging attributes on LasHeR dataset. As shown in Fig. 4, CAT consistently outperforms competing methods across all attributes. Notably, CAT achieves exceptional performance on NO (90.3%), OV (78.4%), and HI (74.3%). Even on challenging attributes like AIV (55.3%) and LI (58.9%), CAT maintains substantial advantages, validating the robustness of our Cross-Attention Adapters under diverse tracking scenarios.

### *E. Computational Efficiency*

CAT achieves real-time performance of 38.1 FPS on a single RTX 3090 GPU while maintaining state-of-the-art accuracy. The computational breakdown is: Cross-Attention Adapters (+3-4% overhead), ConfidenceNet (<1%), and Candidate Elimination (-30% reduction), resulting in net computational savings. The adapter's efficiency stems from spatial downsampling (16× token reduction), single-head attention ($d = 32$), and selective layer insertion (3 of 12 layers). The global relocation mechanism activates infrequently (8-12% of frames) and can be disabled for strict real-time requirements.

### *F. Qualitative Analysis*

Figure 5 shows qualitative tracking results on two challenging scenarios. The first row demonstrates tracking under **occlusion and illumination changes**, where CAT (red) maintains accurate tracking while ViPT (blue) and ground truth (green) show the target trajectory. The second row shows tracking in **nighttime low-illumination conditions**, where CAT successfully tracks the pedestrian using thermal information while maintaining stable predictions across frames. These visualizations demonstrate CAT's robustness in challenging RGB-T tracking scenarios.

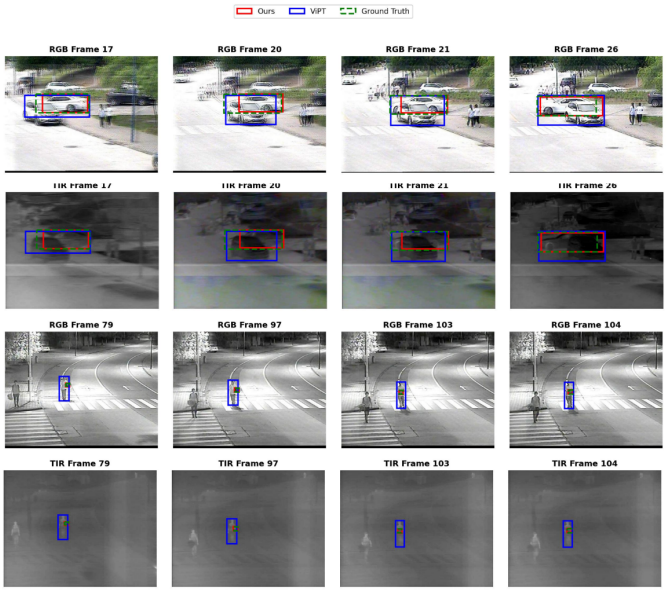


Fig. 5. Qualitative tracking results on challenging scenarios. Top: Tracking under occlusion and illumination changes (Frames 17-26). Bottom: Nighttime low-illumination tracking (Frames 79-104). CAT (red boxes) maintains accurate tracking compared to ViPT (blue boxes) and ground truth (green boxes). Both RGB and thermal (TIR) frames are shown.

## V. Conclusion

This paper introduces a novel lightweight approach called CAT for RGB-Thermal object tracking. This approach aims to achieve parameter-efficient fine-tuning on limited RGB-T data and enhance the model's robustness in challenging scenarios. Specifically, we utilize lightweight Cross-Attention Adapters to symmetrically fine-tune the pre-trained RGB-based tracker, transitioning the feature extraction capability from the RGB to the thermal domain and achieving effective multimodal fusion. Furthermore, we design ConfidenceNet for dynamic modality-aware fusion to obtain robust tracking performance. The global relocation mechanism is introduced to enhance long-term tracking continuity. CAT's effectiveness is demonstrated through extensive experiments on LasHeR and RGBT234 benchmarks, achieving 68.5% PR and 54.5% SR on LasHeR while maintaining real-time speed at 38.1 FPS.

## References


[1] B. Yan et al., "Towards grand unification of object tracking," in Proceedings of the European Conference on Computer Vision (ECCV), 2022, pp. 733–751.

[2] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, "Transformer tracking," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 8126–8135.

[3] Z. Chen, B. Zhong, G. Li, S. Zhang, and R. Ji, "Siamese box adaptive network for visual tracking," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 6668–6677.

[4] C. Li, L. Liu, A. Lu, Q. Ji, and J. Tang, "Challenge-aware RGBT tracking," in Eur. Conf. Comput. Vis., 2020, pp. 222–237.

[5] A. Lu, C. Li, Y. Yan, J. Tang, and B. Luo, "RGBT tracking via multi-adapter network with hierarchical divergence loss," IEEE Trans. Image Process., vol. 30, pp. 5613–5625, 2021.

[6] Y. Zhu, C. Li, J. Tang, and B. Luo, "Quality-aware feature aggregation network for robust RGBT tracking," IEEE Trans. Intell. Veh., vol. 6, no. 1, pp. 121–130, 2020.

[7] P. Zhang, J. Zhao, D. Wang, H. Lu, and X. Ruan, "Visible-thermal UAV tracking: A large-scale benchmark and new baseline," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 8886–8895.

[8] J. Zhu, S. Lai, X. Chen, D. Wang, and H. Lu, "Visual prompt multi-modal tracking," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2023, pp. 9516–9526.

[9] J. Yang, Z. Li, F. Zheng, A. Leonardis, and J. Song, "Prompting for multi-modal tracking," in Proc. 30th ACM Int. Conf. Multimedia, 2022, pp. 3492–3500.

[10] C. Li et al., "LasHeR: A large-scale high-diversity benchmark for RGBT tracking," IEEE Trans. Image Process., vol. 31, pp. 392–404, 2021.

[11] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "RGB-T object tracking: Benchmark and baseline," Pattern Recognit., vol. 96, p. 106977, 2019.

[12] Y. Gao et al., "Deep adaptive fusion network for high performance RGBT tracking," in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops, 2019, pp. 53–60.

[13] L. Zhang, M. Danelljan, A. Gonzalez-Garcia, J. Van De Weijer, and F. S. Khan, "Multi-modal fusion for end-to-end RGB-T tracking," in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops, 2019, pp. 1489–1498.

[14] X. Zhang et al., "Object fusion tracking based on visible and infrared images using fully convolutional siamese networks," in Proc. 22nd Int. Conf. Inf. Fusion (FUSION), 2019, pp. 1–8.

[15] Y. Zhu et al., "Dense feature aggregation and pruning for RGBT tracking," in Proc. 27th ACM Int. Conf. Multimedia, 2019, pp. 465–472.

[16] H. Zhang, L. Zhang, L. Zhuo, and J. Zhang, "Object tracking in RGB-T videos using modal-aware attention network and competitive learning," Sensors, vol. 20, no. 2, p. 393, 2020.

[17] C. Long, A. Lu, H. A. Zheng, Z. Tu, and J. Tang, "Multi-adapter RGBT tracking," in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops, 2019, pp. 1480–1488.

[18] C. Li, N. Zhao, Y. Lu, C. Zhu, and J. Tang, "Weighted sparse representation regularized graph learning for RGB-T object tracking," in Proc. 25th ACM Int. Conf. Multimedia, 2017, pp. 1856–1864.

[19] C. Wang et al., "Cross-modal pattern-propagation for RGB-T tracking," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 7064–7073.

[20] A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. Int. Conf. Mach. Learn., 2021, pp. 8748–8763.

[21] A. Dosovitskiy, "An image is worth 16×16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[22] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," arXiv preprint arXiv:2104.08691, 2021.

[23] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," arXiv preprint arXiv:2101.00190, 2021.

[24] G. Bhat, M. Danelljan, L. Van Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 6182–6191.

[25] P. Zhang et al., "Jointly modeling motion and appearance cues for robust RGB-T tracking," IEEE Trans. Image Process., vol. 30, pp. 3335–3347, 2021.

[26] Y. Xiao, M. Yang, C. Li, L. Liu, and J. Tang, "Attribute-based progressive fusion network for RGBT tracking," in Proc. AAAI Conf. Artif. Intell., vol. 36, no. 3, 2022, pp. 2831–2838.

[27] L. Hong, S. Yan, R. Zhang, W. Li, X. Zhou, P. Guo, K. Jiang, Y. Chen, J. Li, Z. Chen, et al., "Onetracker: Unifying visual object tracking with foundation models and efficient tuning," Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 19079–19091, 2024.

[28] X. Hou, J. Xing, Y. Qian, Y. Guo, S. Xin, J. Chen, K. Tang, M. Wang, Z. Jiang, L. Liu, et al., "Sdstrack: Self-distillation symmetric adapter learning for multi-modal visual object tracking," Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 26551–26561, 2024.