Unaligned UAV RGBT Tracking: A Largescale Benchmark and A Novel Approach

Yun Xiao[1,2,3], **Yuhang Wang**[3], **Jiandong Jin**[4], **Wankang Zhang**[3], **Chenglong Li**[1,2,3*]

[1]State Key Laboratory of Opto-Electronic Information Acquisition and Protection Technology, Hefei, 230601, Anhui, China
[2]Anhui Provincial Key Laboratory of Multimodal Congnitive Computation, Hefei, 230601, Anhui, China
[3]School of Artificial Intelligence, Anhui University, Hefei, 230601, China
[4]School of Computer Science and Technology, Anhui University, Hefei, 230601, China
xiaoyun@ahu.edu.cn,yhwang1022@163.com,jdjinahu@foxmail.com,wankangahu@163.com,lcl1314@foxmail.com

Abstract

With the rapid development of the low-altitude economy, multi-modal visual tracking in UAV scenarios has attracted extensive attention. UAVs are typically equipped with independent visible (RGB) and thermal infrared (TIR) sensors, resulting in an inherent spatial misalignment between the two modalities. However, existing RGBT tracking methods generally rely on spatially aligned data inputs, making them unsuitable for unaligned RGBT tracking task in UAV scenarios. In this work, we introduce a new task called unaligned UAV RGBT tracking and construct the first largescale unaligned RGB and TIR video dataset to promote the research and development in this field. The dataset contains 1,453 pairs of UAV-captured RGBT sequences with precise dual-modal bounding box annotations, and covers 42 object categories, 22 typical challenge attributes, and diverse spatial misalignment scales to simulate real-world challenging scenarios better. To address the limitations of existing methods that fail to handle the spatial misalignment issue in UAV scenarios, we propose the novel RGBT tracking approach. In particular, we design a mixture of shift estimation experts module to adaptively estimate the spatial shifts across two modalities at different scales, along with a cross-modal alignment and fusion module to correct feature shifts, compensate for nonlinear deformations, and integrate multi-modal information. Extensive experiments on the created dataset demonstrate that the proposed tracker significantly outperforms existing state-of-the-art tracking methods, validating its practicality and robustness in real-world unaligned UAV tracking scenarios.

Code&Datasets — https://github.com/NOP1224/Unaligned_RGBT_Tracking

Introduction

With the rapid development of low-attitude economy, unmanned aerial vehicles (UAVs) have attracted a great deal of attention due to their low cost, high efficiency, and flexibility. Recently, UAV RGBT Tracking has also become a research focus, while the task aims to fuse complementary information from RGB and TIR modalities, significantly improving robustness to visual tracking in UAV scenarios.

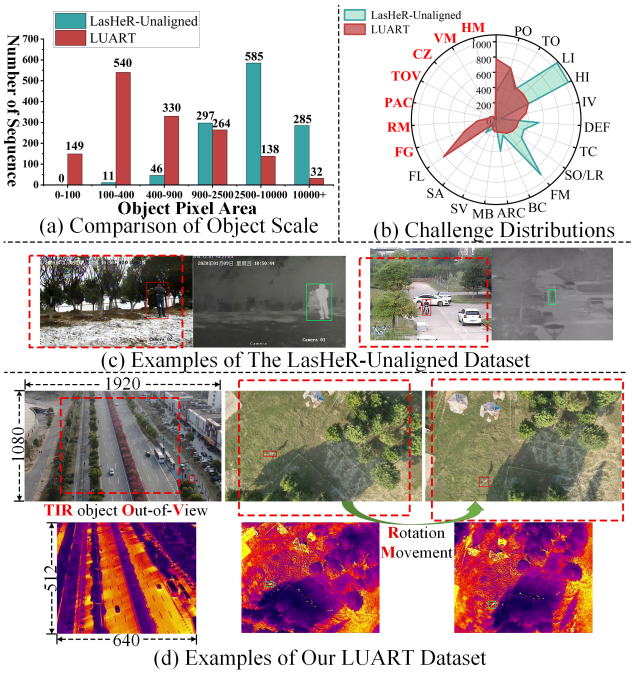

Figure 1: Comprehensive Comparison between our LUART and LasHeR-Unaligned. Figure (a) shows the distribution of object pixel area of two datasets.The challenges marked in red represent new challenge added by our LUART and those marked in black are common to two datasets in (b). The red dashed boxes in (c) and (d) represent the position of the TIR image within RGB image.

However, most existing UAV RGBT tracking methods (Zhang et al. 2022) rely on strictly manually aligned RGBT sequences, typically achieved by applying rigid transformations to align targets across modalities and providing only a single ground-truth bounding box. It makes the models not suitable for real unaligned RGBT tracking task and thus limits their adaptability to real-world scenarios. For RGB and TIR modalities, UAV-captured RGBT sequences are typically acquired using different sensors separately. Due to the different resolution and position between RGB and TIR sensors, it can inevitable lead to spa-

*Chenglong Li is the corresponding author.

tial shifts like out-of-view objects and rotational movements, and obvious appearance feature difference as shown in Figure 1(d), while the misalignment characteristic of LasHeR-Unaligned(Li et al. 2021) is only manifested as the positional translation of the target between the two modalities as shown in Figure 1(c). Owing to the significant differences between RGB and TIR modalities, we introduce a new task called the unaligned UAV RGBT tracking task, which aims to predict two different target bounding boxes across both modalities using original unaligned RGB and TIR images without any manual post-processing.

To promote research and development in this field, we construct the first unaligned RGB and TIR video dataset, called Largescale Unaligned UAV RGBT Tracking (LUART) benchmark. The dataset contains 1,453 pairs of UAV-captured RGBT sequences with precise dual-modal bounding box annotations, collected under various UAV altitudes and motion patterns, addressing a critical gap in existing datasets for unaligned UAV RGBT tracking domain. In particular, the dataset covers 42 distinct tracking objects in total, ranging from common objects such as pedestrians and vehicles to specialized objects such as animals and industrial machines. And we annotate each sequence in our dataset with comprehensive challenge attributes. The dataset includes 15 attributes shared by previous RGBT tracking datasets such as partial occlusion and 7 distinct challenges uniquely critical to unaligned UAV RGBT tracking, such as the out-of-view of the TIR object and rotational movements as demonstrated in Figure 1. The dataset features a wide range of modality misalignment scales, specifically encompassing cases ranging from minor positional offset (e.g., 0-20 pixels) to substantial offset exceeding 100 pixels. In addition, the distribution of data across these misalignment ranges aligns with real-world patterns, which better simulate real-world challenging scenarios.

Moreover, most existing RGBT Trackers (Lu et al. 2021; Liu et al. 2024) focus on developing a more effective fusion strategy, yet overlook the critical process of spatial alignment of the multi-modal features. Rarely do methods address modality misalignment, such as AMNet (Zhang et al. 2024), mainly predict a single global feature shift and apply this uniform shift to all feature maps. Although these global shifts can manage minor shifts to some extent, they struggle to address the significant spatial misalignment of targets. To address the limitations of existing methods that fail to handle the spatial misalignment issue in UAV scenarios, we propose the novel RGBT trackers, the Spatial-Feature Collaborative Alignment Tracker named SFCATrack. It leverages a cooperative mechanism that integrates image spatial alignment and feature alignment to address the inherent spatial misalignment between two modalities in the unaligned UAV RGBT tracking task.

In particular, we first introduce the Mixture of Shift-Estimation Experts (MSEE) module. This module estimates the spatial shifts between two modalities by leveraging multi-modal features modeled through one shared expert and dynamically selected active experts. Finally, this module leverages the predicted shifts to adaptively warp the search region, thereby achieving spatial registration of the multi-modal images. Since spatial alignment typically relies on homography matrices, it struggles to effectively handle target distortions caused by nonlinear deformations. Such feature misalignment poses significant challenges for multi-modal information fusion. To address this issue, we design a Cross-Modal Alignment and Fusion (CMAF) module to achieve feature alignment during the fusion process. Specifically, the module cascades multiple deformable convolution blocks. These blocks impose nonlinear transformations on multi-modal features and correct feature misalignment by integrating information across modalities. Finally, the aligned features are fused through a lightweight gating network.

The main contributions of this paper are threefold.

- We introduce a new task called unaligned UAV RGBT tracking and construct the first dataset, which comprises millions of annotated samples.
- We propose SFCATrack, which introduces a collaborative alignment tracking architecture that integrates image alignment and feature alignment to achieve precise multi-modal correspondence.
- We perform extensive comparative experiments on the LUART dataset and existing unaligned RGBT datasets, evaluating 14 representative trackers.

Related Work

UAV RGBT Tracking

UAV RGBT tracking aims to utilize RGB and TIR modalities to improve performance. To advance this field, VTUAV (Zhang et al. 2022) proposes the first largescale UAV RGBT dataset, which provides crucial data support. Meanwhile, HiAl (Xiao et al. 2025a) focuses on tracking from high-altitude UAVs and presents a greater challenge due to the prevalence of smaller targets. To address UAV-specific challenges, such as the small target and limited onboard computational resources, researchers focus on optimizing multi-modal fusion, enhancing robustness in dynamic environments, and developing efficient deployment methods. For example, HMFT (Zhang et al. 2022) enhances cross-modal representation through a multi-stage fusion framework, DFANet (Gao et al. 2023) integrates complementary information using a cross-dimensional attention mechanism, IAMTrack (Shi et al. 2025) improves robustness in complex scenarios via a cross-frame token propagation mechanism, and SiamCAF (Xue et al. 2023) designs a compact Transformer design significantly reduces computational demands. However, existing UAV RGBT trackers rely on strictly aligned RGBT sequences, which are unsuitable for unaligned UAV RGBT tracking in real-world scenarios. To address the limitation, we introduce the LUART dataset containing misaligned RGBT pairs and propose SFCATrack, a tracker designed to handle spatial misalignment from two modalities.

Multi-modal Alignment

Multi-modal alignment aims to solve the inherent spatial misalignment in multi-modal images for better information fusion in downstream tasks. Existing multi-modal alignment

Benchmark	Num. Seq.	Avg. Frame	Total Frame	Resolution	Obj. Class	Num. Challenges	Multi-Modal	Drone Plat.	Train Subset	Misaligned Data	Year
UAV123 (Benchmark 2016)	123	915	112K	1280×720	–	12	×	✓	×	×	2016
DTB (Li and Yeung 2017)	70	225	15.7K	1280×720	3	11	×	✓	×	×	2017
UAVDT (Yu et al. 2020)	100	832	139.3K	1024×540	3	10	×	✓	×	×	2018
GTOT (Li et al. 2016)	50	157	7.8K	384×288	9	7	✓	×	×	×	2016
RGBT210 (Li et al. 2017)	210	498	104.7K	630×460	22	12	✓	×	×	×	2017
RGBT234 (Li et al. 2019)	234	498	116.7K	630×460	22	12	✓	×	×	×	2019
LasHeR (Li et al. 2021)	1224	600	734.8K	630×480	32	19	✓	×	✓	✓	2021
VTUAV (Zhang et al. 2022)	500	3329	1.7M	1920×1080	13	13	✓	✓	✓	×	2022
HiAl (Xiao et al. 2025a)	150	-	-	1920×1080	9	12	✓	✓	✓	×	2023
MV-RGBT (Tang et al. 2024)	122	737	89.9K	640×480	36	10	✓	×	×	×	2024
UniRTL (Zhang et al. 2025)	676	701	474.2K	640×480	35	16	✓	×	✓	×	2025
LUART(Ours)	1453	700	1.02M	1920×1080 640×512	42	22	✓	✓	✓	✓	2025

Table 1: Statistics comparison among existing RGBT and UAV tracking datasets.

methods primarily align multi-modal images through predicting deformation fields, deformable or dilated convolutions for dynamic sampling, and predicting offset parameters. For example, C2RF (Tang et al. 2025) utilizes multi-level deformation fields to progressively refine alignment and introduces fusion-driven contrastive learning to align images. CAGTDET (Yuan et al. 2024) dynamically adjusts target bounding boxes by predicting transform parameters along the translation, scale, and rotation. MuIFS-CAP (Li et al. 2025) guides the deformation of TIR images to spatially align with RGB images with a pixel-level relationship matrix. MURF (Xu, Yuan, and Ma 2023) first corrects global rigid misalignment through progressive affine transformations, and then corrects local misalignment using an inverse deformation field. OAFA (Chen et al. 2024) uses predicted offsets as guidance to drive deformable convolution to adaptively adjust feature sampling locations. ReCoNet (Huang et al. 2022) explicitly compensates for geometric distortion via a deformation module and implicitly alleviates artifacts in misaligned regions using an attention mechanism. However, these methods focus on utilizing features to guide the prediction of deformation fields or transformation parameters for warping images to benefit downstream tasks, while struggling to handle target distortions caused by nonlinear deformations. In contrast, we integrate image spatial alignment with feature alignment, which ensures robust performance even under severe spatial misalignment conditions.

Dataset

LUART Benchmark Dataset Construction

Video Collection. We collect data by employing professional drone pilots operating professional UAVs, such as DJI Matrice 300 RTK and DJI Mavic 3T equipped with the Zenmuse H20T camera, to capture RGBT video at different locations and scenarios. These videos captured by UAVs span the four seasons, including spring, summer, autumn, and winter. We preserve the original resolutions of its dual sensors that RGB images at a sharp 1920×1080 and infrared images at 640×512, reflecting real-world UAV setups and introducing the critical challenge of cross-resolution feature learning and alignment. Finally, we obtain the LUART, which comprises 1,453 video pairs, totaling 1.02 million RGBT image pairs, with each video averaging 700 frames, as shown in Table 1.

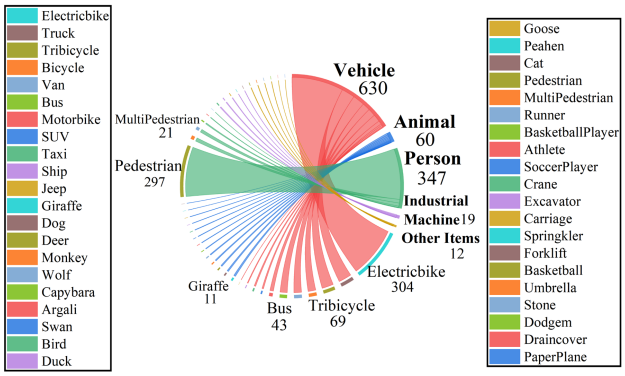

Figure 2: Object main category and subcategory statistics of our LUART dataset.

Annotations. The LUART is annotated by five professional annotators using computer vision annotation tools (Sekachev et al. 2020) for frame-level bounding box annotation conducted, followed by multi-stage reviews to ensure quality. Due to significant resolution and positional offset variations between modalities, annotations are maintained separately for each modality. Notably, the average object size in our LUART is significantly smaller than in LasHeR-Unaligned, as Figure 1(a) shown. Furthermore, sequence-level annotation includes 22 challenge attributes that integrate 15 general challenge attributes included in the previous dataset with 7 specific challenges proposed by our dataset, as shown in Figure 1(b).

Dataset Partition. To facilitate the evaluation of RGBT trackers, LUART is partitioned into training and testing sets at the video sequence level based on criteria including object categories, scene categories, challenge attribute distributions, and positional offset distributions. The training set

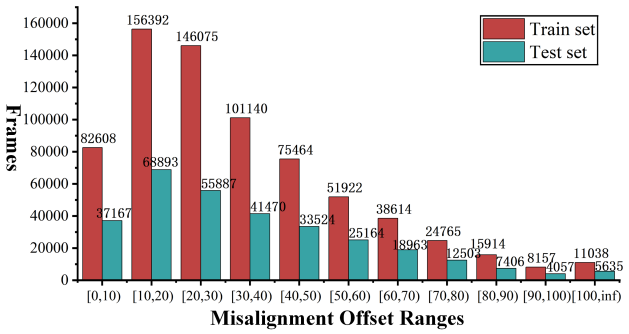

Figure 3: Number of frames at different offset scales in our LUART dataset.

consists of 1,010 sequences totaling 706,689 image pairs, while the test set contains 443 sequences totaling 310,669 image pairs.

Statistics of Our LUART

Diverse Object Classes. Our LUART dataset significantly improves current benchmarks by providing diverse categories and complex scenes. As shown in Figure 2, it includes 42 object classes with hierarchical subcategories (e.g., motor and electric bikes) across various environments.

Diverse Scenarios. We collect these videos in highly varied scenarios, our dataset provides critical data support to enhance the environmental robustness of trackers, as Figure 1(d) shows. Besides, to fully reflect the complexity of the real world, the dataset includes an extensive range of lighting variations and diverse, challenging weather conditions such as rainy, snowy, and foggy.

Modality Misalignment Statistics. We analyze target displacement caused by spatial misalignment between RGB and TIR modalities by quantifying spatial offsets from annotation coordinates. Figure 3 shows a significant negative correlation in the probability density distribution of object spatial offsets, with the number of corresponding video frames decreasing exponentially as the offset increases. The train set and the test set exhibit similar probability distributions across different offset ranges in our dataset.

Methodology

We propose a Spatial-Feature Collaborative Alignment Tracker named SFCATrack. The framework is shown in Figure 4. It is a dual-branch tracking framework based on OSTrack (Ye et al. 2022). We incorporate a combined method of image alignment and feature alignment & fusion to effectively align and fuse multi-modal images, thus improving tracking robustness. The framework consists of two key modules. First, for image alignment, we employ a Mixture of Shift Estimation Experts (MSEE) module, which selects appropriate experts to handle varying degrees of offset, to address diverse scales of modality misalignment. Second, for feature alignment and fusion, we introduce the Cross-Modal Alignment and Fusion (CMAF) module, which builds upon the initial spatial alignment results. It employs progressive deformable convolutions to achieve precise feature alignment and further adopts a gated fusion mechanism to capture complementary cross-modal information. The following sections provide a detailed introduction to these modules.

Mixture of Shift Estimation Experts Module

To align the search regions at the image level, we propose the MSEE module to handle modality-specific shifts at different scales using a set of experts, which are guided by a router that enables adaptive expert selection. Specifically, given a pair of misaligned multi-modal search region images $[\mathbf{X}^V, \mathbf{X}^I]$, MSEE first extracts image features $[\mathbf{F}^V, \mathbf{F}^I]$ using a ResNet (He et al. 2016) backbone:

$$\mathbf{F}^m = Enc^m(\mathbf{X}^m), m \in \{V, I\}, \tag{1}$$

where Enc^m denotes the ResNet backbone assigned to the search region of the input image $\mathbf{X}^m$, V and I are the RGB and TIR modalities. To prevent the model from becoming biased toward any single modality, the ResNet backbone parameters are shared across multi-modal branches. We concatenate the features as $\mathbf{F}^{VI} = [\mathbf{F}^{\mathrm{V}}; \mathbf{F}^{\mathrm{I}}]$ from the two modalities to facilitate subsequent alignment. Here, D indicates the channel dimension of the modality features. To accommodate misalignment at different scales, we introduce multiple scale expert networks, where N experts are designed to model misalignment features at a specific scale separately, thus enhancing the network's capacity to capture complex shift patterns. The modeling process of each expert will be described in detail below.

$$\mathbf{F}_i = Expert_i(\mathbf{F}^{VI}), i = 1, 2, \ldots, N. \tag{2}$$

Here, i denotes the index of the scale expert, $\mathbf{F}_i$ denotes the feature extracted by the i-th expert, which is a tunable hyperparameter. $Expert_i$ forms a SwiGLU (Shazeer 2020) activation unit, whose detailed structure is illustrated at the bottom right corner of Figure 4. In our experimental setup, we set $N = 5$. In addition, we incorporate a shared expert $Expert_{shared}$ with the same structure to obtain the feature map F_S. It is responsible for modeling common misalignment features that are prevalent in multi-modal alignment tasks, thereby improving the generalization capability of the model.

$$\mathbf{F}_s = Expert_{shared}(\mathbf{F}^{VI}). \tag{3}$$

To enable scale-aware adaptive expert selection, we design a scale-aware routing module that takes the joint feature representation F^{VI} as input and outputs an expert selection weight vector $\omega = [\omega_1, \omega_2, ..., \omega_N]$, where each $w_i \in \mathbb{R}$ represents the confidence score assigned to the i-th expert. The expert with the highest confidence score is selected as the active expert. Here, we define i^* as the index of the active expert. Finally, the fused multi-expert feature $\mathbf{F}_m = \mathbf{F}_s + \mathbf{F}_{i^*}$ is fed into the offset prediction head to regress the final output Δp between the target and reference modalities.

$$\Delta p = (\Delta x_1, \Delta y_1, \Delta x_2, \Delta y_2) = Head(\mathbf{F}_m) \in \mathbb{R}^4, \tag{4}$$

where $(\Delta x_1, \Delta y_1, \Delta x_2, \Delta y_2)$ represents the differences in the x_1, y_1, x_2, y_2 coordinates between the ground truth

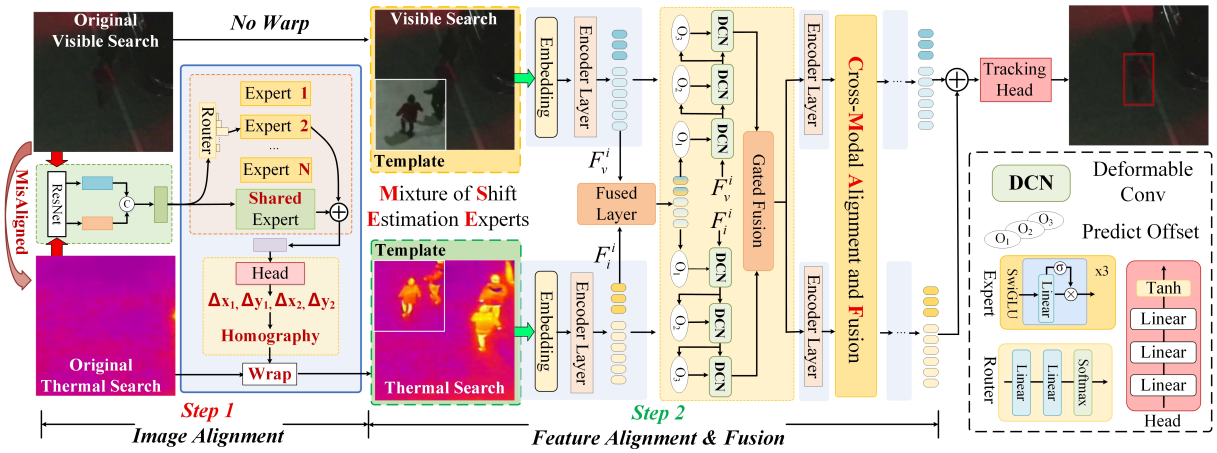

Figure 4: Overall Framework of Our SFCATrack.

bounding boxes of the two modalities. We then utilize these four predicted parameters to construct a homography matrix that is then applied to warp the TIR search region. The training of the MSEE module is divided into three stages to enable scale experts to effectively handle misalignment at varying scales. In the first stage, the feature extractor, the shared expert, and the offset prediction head are trained on the full training set to obtain the model initialization parameters. In the second stage, the parameters of the feature extractor are frozen. The shared expert is used to initialize the scale experts, with the offset scales divided into five intervals: $[0, 20)$, $[20, 35)$, $[35, 50)$, $[50, 75)$, and $[75, +\infty)$. By simulating specific offset ranges on the aligned images, the corresponding scale experts are trained to specialize in their respective offset scales. In the third stage, on the full training set, all experts and the backbone network are frozen, and only the router and offset prediction head are trained to adaptively select the most appropriate scale expert, thereby completing the multi-modal image alignment task.

Cross-Modal Alignment and Fusion Module

After performing spatial alignment of multi-modal search regions using MSEE, the misalignment between multi-modal images is alleviated; however, target region distortions caused by nonlinear deformations still persist, which degrades the effectiveness of multi-modal feature fusion. To address this issue, we introduce the CMAF module to further refine and fuse the spatially aligned RGBT features. The CMAF module applies multiple deformable convolution blocks to impose nonlinear transformations on multi-modal features, thereby reducing feature misalignment and enabling effective multi-modal feature fusion.

Specifically, given the multi-modal template $\mathbf{Z}^m \in \mathbb{R}^{H_z \times W_z \times 3}, m \in \{V, I\}$ and the search region image pairs $\mathbf{X}^m \in \mathbb{R}^{H_x \times W_x \times 3}$ aligned by the MSEE module, where z and x denote the template and search region. We first partition each image into patches of size $P \times P$ and then flatten them to obtain patch sequences $\mathbf{P}_r^m \in \mathbb{R}^{N_r \times (3P^2)}, r \in \{x, z\}$, and $N_r = (H_r \times W_r)/P^2$ represents the number of patches for the corresponding region.

The patch sequences are then passed through an embedding layer $Embed(\cdot)$ with learnable positional encoding $Pos_r \in \mathbb{R}^{N_r \times D}$, D denotes the number of channel dimensions:

$$\mathbf{E}_r^m = Embed(\mathbf{P}_r^m) + Pos_r, r \in \{x, z\}. \quad (5)$$

Subsequently, the initial multi-modal template and search patch embeddings $\mathbf{E}_z^m, \mathbf{E}_x^m$ are concatenated and used as the input $\mathbf{F}_0^m = [\mathbf{E}_z^m; \mathbf{E}_x^m]$ to the tracking backbone:

$$\mathbf{F}_j^m = Layer_j(\mathbf{F}_{j-1}^m) \in \mathbb{R}^{(N_z+N_x)\times D}, j = 1, \ldots, 12. \quad (6)$$

where $\mathbf{F}_j^m$ denotes the multi-modal features if j-th layer, and $Layer_j$ represents the j -th Transformer encoder layer.

We introduce the CMAF module into the forward pass of both modality-specific backbones to further align features and integrate multi-modal information. Specifically, we first employ the multi-modal feature fusion layer $FL(\cdot)$ to project the multi-modal features into a shared feature space within the fusion layer. The feature projection process can be formulated as follows:

$$\mathbf{F}^{fuse} = FL(\mathbf{F}^V, \mathbf{F}^I) \in \mathbb{R}^{(N_z+N_x)\times D}. \quad (7)$$

Here, $\mathbf{F}^{fuse}$ denotes the fused multi-modal features. In this work, the CMAF modules are integrated into all layers. $FL(\cdot)$ denotes the fusion layer, which is composed of three MLPs. We first align the modality-specific features using MLPs applied at both spatial and channel levels, and then concatenate them. The concatenated features are then fused into $\mathbf{F}^{fuse}$ via another MLP. Leveraging the fused features projected into a shared feature space, the multi-modal features are further guided towards spatial alignment using a deformable convolution operation. The alignment process is formulated as follows.

$$\tilde{\mathbf{F}}_k^m = D_k(\tilde{\mathbf{F}}_{k-1}^m, \mathbf{F}^{fuse}), k = 1, \ldots, N, \quad (8)$$

where k denotes the iteration index of the deformable convolution, N means the number of cascade deformable convolution blocks, and $\tilde{\mathbf{F}}_k^m$ represents the offset-corrected features

Tracker	Pub. Info.	PR↑	NPR↑	SR↑
mfDiMP (Zhang et al. 2019)	ICCVW 2019	41.6	40.1	33.5
CAT (Li et al. 2020)	ECCV 2020	42.8	39.8	34.4
ADRNet (Zhang et al. 2021)	IJCV 2021	44.6	43.1	33.0
HMFT (Zhang et al. 2022)	CVPR 2022	44.5	41.5	35.7
SeqTrackv2 (Chen et al. 2023)	CVPR 2023	48.3	45.2	37.5
ViPT (Zhu et al. 2023)	CVPR 2023	52.1	48.6	41.3
TBSI (Hui et al. 2023)	CVPR 2023	52.2	48.5	41.4
BAT (Cao et al. 2024)	AAAI 2024	49.6	45.9	39.5
SDSTrack (Hou et al. 2024)	CVPR2024	50.0	46.3	39.7
UnTrack (Wu et al. 2024)	CVPR2024	53.3	48.8	41.7
CAFormer (Xiao et al. 2025b)	AAAI 2025	52.7	48.8	41.6
STTrack (Hu et al. 2025)	AAAI 2025	53.6	49.6	42.2
SUTrack (Chen et al. 2025)	AAAI 2025	54.7	49.6	42.6
Baseline (Single-modal)	ECCV 2022	45.4	41.7	35.6
Baseline (Multi-modal)	ECCV 2022	48.6	45.3	38.3
SFCATrack(Ours)	-	**57.3**	**51.9**	**44.6**

Table 2: Retraining result of tracking performance on the LUART dataset. The best results are indicated in **bold**, while the second-best results are marked with underlining, evaluated across PR, NPR, and SR metrics.

obtained after the k-th iteration, which are initialized with the multi-modal feature $\mathbf{F}^m$. $D_k(\mathbf{X}, \mathbf{Y})$ denotes the k-th deformable convolution, where $\mathbf{X}$ serves as the input feature to be convolved, while $\mathbf{Y}$ is not convolved but provides spatial offset guidance.

We use the multi-modal feature $\mathbf{F}^m$ in the first layer as the input to be offset and the fused feature $\mathbf{F}^{fuse}$ as the offset guide to perform progressive alignment. Finally, the aligned multi-modal features $\mathbf{F}^m$ are fused with dynamic weights generated by a lightweight gating network $\mathbf{G}$. The aligned and fused features $\mathbf{F}^{\text{fuse}}$ are residually added to the input features $\mathbf{F}^V$ and $\mathbf{F}^I$, and subsequently fed into the next network layer. This feature's multi-modal alignment and fusion process effectively integrates cross-modal representations.

Loss Function

Our method adopts a staged training strategy. First, the MSEE module applies three training stages before its parameters are frozen. Subsequently, on the basis of the frozen MSEE module, both the backbone encoder and the CMAF module are trained. It is noted that distinct loss functions are employed for these two training phases. For training the MSEE module, we employ an L1 loss function to measure the discrepancy between the predicted offsets and relative displacements of the ground truth bounding boxes across both modalities.

$$L_{msee} = L_1(\Delta p, \Delta gt). \tag{9}$$

Here, Δp denotes the predicted offset and Δgt represents the ground-truth relative displacements. For training the encoder and the CMAF module, our loss function adopts the identical formulation scheme as OSTrack (Ye et al. 2022). The loss function can be formulated as:

$$L_{total} = L_{cls} + \lambda_1 L_{iou} + \lambda_2 L_1. \tag{10}$$

Tracker	Pub. Info.	PR↑	NPR↑	SR↑
MANet (Long Li et al. 2019)	ICCVW 2019	32.9	26.6	24.1
CAT (Li et al. 2020)	ECCV 2020	36.3	29.9	25.3
ADRNet (Zhang et al. 2021)	IJCV 2021	34.5	29.2	23.8
APFNet (Xiao et al. 2022)	AAAI 2022	40.3	32.4	29.1
TBSI (Hui et al. 2023)	CVPR 2023	53.5	49.6	41.7
BAT (Cao et al. 2024)	AAAI 2024	58.5	53.4	46.2
CAFormer (Xiao et al. 2025b)	AAAI 2025	58.7	54.0	46.9
SFCATrack(Ours)	-	**60.7**	**55.1**	**47.9**

Table 3: Comparison of tracking performance on LasHeR-Unaligned dataset. The best results are in **bold**, second-best are underlined.

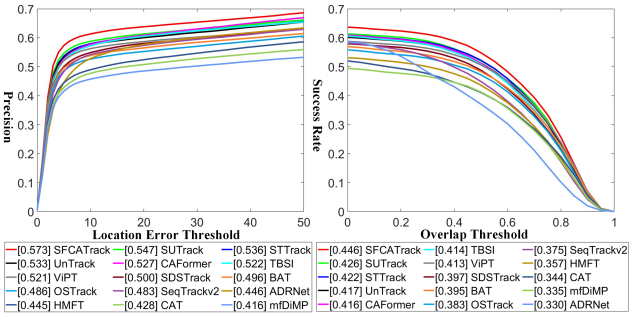

Figure 5: Evaluation result on LUART test set using precision and success plots, where the scores are presented in the legend. All trackers are trained on the LUART training set.

L_{cls} denotes the weighted focal loss for classification, while the generalized IoU loss L_{iou} and L_1 are adopted for bounding box regression, λ_1 and λ_2 are trade-off parameters.

Experiments

Overall Evaluation Results

We evaluate our SFCATrack and 14 existing mainstream RGBT trackers on our LUART testing set after retraining these trackers on our LUART training set. Table 2 demonstrates the state-of-the-art performance of our method with 57.3% PR, 51.9% NPR, and 44.6% SR. Besides, we visualize the PR and SR curves of these 14 trackers on LUART testing set in Figure 5 to demonstrate the advantages of our SFCATrack. In addition, to further validate our methods on various datasets, we perform experiments on the LasHeR-Unaligned (Li et al. 2021). Table 3 shows that our method achieves superior performance with 60.7% PR, 55.1% NPR and 47.9% SR. The results show that our method can handle the spatial misalignment between two modalities.

Challenge-based Evaluation Results

The two challenge attribute radar charts in Figure 6 indicate that our method demonstrates significant robustness in challenges, such as extreme illumination (HI), small object tracking (SO), and cross-modal interference (TC/BC). Importantly, it shows substantial advancements over current state-of-the-art trackers in handling horizontal and vertical

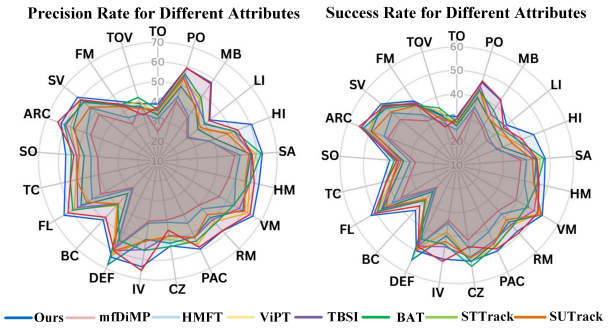

Figure 6: Comparisons of SFCATrack (Ours) and the competing methods under different attributes in our LUART testing set.

Baseline	MSEE	CMAF	PR↑	SR↑
✓			48.6	38.3
✓	✓		53.2	41.5
✓	✓	✓	57.3	44.6

Table 4: Ablation study of different components.

motion scenarios (HM/VM) for UAVs. This highlights our method's robustness to dynamic viewpoint changes in unaligned UAV RGBT tracking scenarios.

Ablation of Component

To assess the effectiveness of the two proposed modules, we perform ablation studies based on the baseline. As presented in Table 4, incorporating MSEE for image spatial alignment yields 53. 2%, 41. 5% on PR and SR, corresponding to an improvement of 4.6% / 3.2%. Furthermore, integrating the CMAF for feature alignment and fusion further improves the results, achieving 8.3%/6.7% compared to the Baseline. The results indicate that the proposed MSEE module and CMAF module make a significant contribution to the feature alignment and fusion.

Ablation Study of Scale-aware Experts

To validate the design rationality of the MSEE module, we conduct ablation studies on its internal components, which consist of a shared expert for capturing cross-modal feature consistency and scale experts for handling multi-modal features under diverse spatial misalignment conditions. As shown in Table 5, incorporating only the shared expert yields a performance gain of 2.5%/1.7% over the baseline. Subsequently, the scale experts further improve 4.6%/3.2% compared to the baseline, achieving an overall performance of 53.2%/41.5% and confirming the need for both components. The results show the collaboration of the shared expert and scale expert in our MSEE module.

Visualization of Search Region Alignment

To demonstrate the effectiveness of the MSEE module, we present visualizations of several representative frames that include the original RGB search region, the TIR search region, and the aligned TIR search region by first frame offset and our MSEE module, as illustrated in Figure 7. The MSEE exhibits improved multi-modal spatial alignment between RGB and TIR targets, especially compared to adjusting the TIR search region by using the first frame offset between bounding boxes of the two modalities.

Baseline	Shared Expert	Scale Experts	PR↑	SR↑
✓			48.6	38.3
✓	✓		51.1	40.0
✓	✓	✓	53.2	41.5

Table 5: Ablation study of the components in MSEE module.

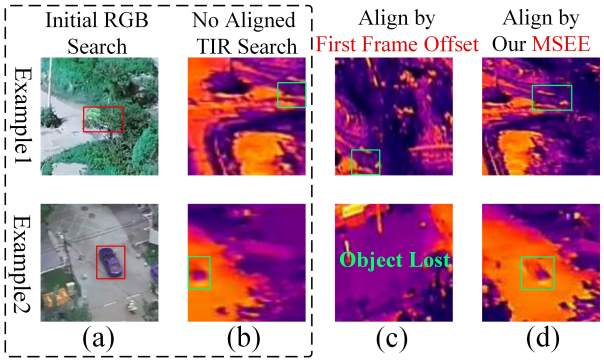

Figure 7: Illustration of the relocation of the search region. (a) Initial RGB search region. (b) No Aligned TIR search region. (c) Aligned TIR search region by first frame offset. (d) Aligned TIR search region by our MSEE.

Conclusion

In this paper, we introduce a new task called unaligned UAV RGBT tracking, which aims to predict bounding boxes using original unaligned RGBT images without manual alignment. Then, we introduce LUART, the first largescale unaligned UAV RGBT tracking dataset specifically designed for UAV applications. The dataset consists of 1,453 sequences across 42 object categories, each annotated with dual-modality bounding boxes and 22 challenge attributes. To better reflect real-world conditions, it incorporates diverse challenges and varying degrees of alignment offsets. To address the issue of spatial misalignment in unaligned UAV RGBT tracking, we propose SFCATrack, which employs a collaborative mechanism based on image spatial alignment and feature alignment. Specifically, a mixture of shift estimation experts module is introduced to adapt to the diverse alignment offsets encountered in the image spatial alignment stage. Furthermore, a cross-modal alignment and fusion module is proposed to leverage complementary information and enable feature alignment. This module achieves progressive bidirectional alignment through multi-layer deformable convolution blocks. Our method outperforms 14 state-of-the-art trackers, demonstrating strong adaptability to unaligned UAV RGBT tracking scenarios.

Acknowledgments

This research is jointly supported by the Anhui Provincial Natural Science Foundation (No. 2408085MF153) and the Young and Middle Age Teachers Training Action Program in Colleges and Universities (No. YQZD2025006). The authors acknowledge GEOVIS Earth Technology Co., Ltd. for providing essential computing resources and data support.

References

Benchmark, U. 2016. A benchmark and simulator for uav tracking. In *European Conference on Computer Vision*, volume 7, 445–461.

Cao, B.; Guo, J.; Zhu, P.; and Hu, Q. 2024. Bi-directional adapter for multimodal tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 927–935.

Chen, C.; Qi, J.; Liu, X.; Bin, K.; Fu, R.; Hu, X.; and Zhong, P. 2024. Weakly misalignment-free adaptive feature alignment for uavs-based multimodal object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26836–26845.

Chen, X.; Kang, B.; Geng, W.; Zhu, J.; Liu, Y.; Wang, D.; and Lu, H. 2025. Sutrack: Towards simple and unified single object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2239–2247.

Chen, X.; Peng, H.; Wang, D.; Lu, H.; and Hu, H. 2023. Seqtrack: Sequence to sequence learning for visual object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14572–14581.

Gao, Z.; Li, D.; Wen, G.; Kuai, Y.; and Chen, R. 2023. Drone based RGBT tracking with dual-feature aggregation network. *Drones*, 7(9): 585.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hou, X.; Xing, J.; Qian, Y.; Guo, Y.; Xin, S.; Chen, J.; Tang, K.; Wang, M.; Jiang, Z.; Liu, L.; et al. 2024. Sdstrack: Self-distillation symmetric adapter learning for multi-modal visual object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26551–26561.

Hu, X.; Tai, Y.; Zhao, X.; Zhao, C.; Zhang, Z.; Li, J.; Zhong, B.; and Yang, J. 2025. Exploiting multimodal spatial-temporal patterns for video object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3581–3589.

Huang, Z.; Liu, J.; Fan, X.; Liu, R.; Zhong, W.; and Luo, Z. 2022. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *European conference on computer Vision*, 539–555. Springer.

Hui, T.; Xun, Z.; Peng, F.; Huang, J.; Wei, X.; Wei, X.; Dai, J.; Han, J.; and Liu, S. 2023. Bridging search region interaction with template for rgb-t tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13630–13639.

Li, C.; Cheng, H.; Hu, S.; Liu, X.; Tang, J.; and Lin, L. 2016. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Transactions on Image Processing*, 25(12): 5743–5756.

Li, C.; Liang, X.; Lu, Y.; Zhao, N.; and Tang, J. 2019. RGB-T object tracking: Benchmark and baseline. *Pattern Recognition*, 96: 106977.

Li, C.; Liu, L.; Lu, A.; Ji, Q.; and Tang, J. 2020. Challenge-aware RGBT tracking. In *European Conference on Computer Vision*, 222–237. Springer.

Li, C.; Xue, W.; Jia, Y.; Qu, Z.; Luo, B.; Tang, J.; and Sun, D. 2021. LasHeR: A large-scale high-diversity benchmark for RGBT tracking. *IEEE Transactions on Image Processing*, 31: 392–404.

Li, C.; Zhao, N.; Lu, Y.; Zhu, C.; and Tang, J. 2017. Weighted sparse representation regularized graph learning for RGB-T object tracking. In *Proceedings of the 25th ACM International Conference on Multimedia*, 1856–1864.

Li, H.; Yang, Z.; Zhang, Y.; Jia, W.; Yu, Z.; and Liu, Y. 2025. MulFS-CAP: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47: 3673–3690.

Li, S.; and Yeung, D.-Y. 2017. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 4140–4146.

Liu, L.; Li, C.; Xiao, Y.; Ruan, R.; and Fan, M. 2024. Rgbt tracking via challenge-based appearance disentanglement and interaction. *IEEE Transactions on Image Processing*, 33: 1753–1767.

Long Li, C.; Lu, A.; Hua Zheng, A.; Tu, Z.; and Tang, J. 2019. Multi-Adapter RGBT Tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2262–2270.

Lu, A.; Li, C.; Yan, Y.; Tang, J.; and Luo, B. 2021. RGBT tracking via multi-adapter network with hierarchical divergence loss. *IEEE Transactions on Image Processing*, 30: 5613–5625.

Sekachev, B.; Manovich, N.; Zhiltsov, M.; Zhavoronkov, A.; Kalinin, D.; Hoff, B.; TOsmanov; Kruchinin, D.; Zankevich, A.; DmitriySidnev; Markelov, M.; Johannes222; Chenuet, M.; a andre; telenachos; Melnikov, A.; Kim, J.; Ilouz, L.; Glazov, N.; Priya4607; Tehrani, R.; Jeong, S.; Skubriev, V.; Yonekura, S.; vugia truong; zliang7; lizhming; and Truong, T. 2020. opencv/cvat: v1.1.0.

Shazeer, N. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Shi, H.; Mu, X.; He, H.; Zhong, C.; Zhang, B.; and Zhao, P. 2025. IAMTrack: interframe appearance and modality tokens propagation with temporal modeling for RGBT tracking. *Applied Intelligence*, 55: 583.

Tang, L.; Yan, Q.; Xiang, X.; Fang, L.; and Ma, J. 2025. C2RF: Bridging Multi-modal Image Registration and Fusion via Commonality Mining and Contrastive Learning. *International Journal of Computer Vision*, 1–19.

Tang, Z.; Xu, T.; Feng, Z.-H.; Zhu, X.; Wang, H.; Shao, P.; Cheng, C.; Wu, X.; Awais, M.; Atito, S.; et al. 2024. Revisiting RGBT tracking benchmarks from the perspective of modality validity: A new benchmark, problem, and method. *CoRR*.

Wu, Z.; Zheng, J.; Ren, X.; Vasluianu, F.-A.; Ma, C.; Paudel, D. P.; Van Gool, L.; and Timofte, R. 2024. Single-model and any-modality for video object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19156–19166.

Xiao, Y.; Cao, D.; Li, C.; Jiang, B.; and Tang, J. 2025a. A benchmark dataset for high-altitude UAV multi-modal tracking. *Journal of Image and Graphics*, 30: 361–374.

Xiao, Y.; Yang, M.; Li, C.; Liu, L.; and Tang, J. 2022. Attribute-based progressive fusion network for rgbt tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2831–2838.

Xiao, Y.; Zhao, J.; Lu, A.; Li, C.; Yin, B.; Lin, Y.; and Liu, C. 2025b. Cross-modulated Attention Transformer for RGBT Tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8682–8690.

Xu, H.; Yuan, J.; and Ma, J. 2023. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(10): 12148–12166.

Xue, Y.; Zhang, J.; Lin, Z.; Li, C.; Huo, B.; and Zhang, Y. 2023. SiamCAF: Complementary attention fusion-based Siamese network for RGBT tracking. *Remote Sensing*, 15: 3252.

Ye, B.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2022. Joint feature learning and relation modeling for tracking: A one-stream framework. In *European Conference on Computer Vision*, 341–357. Springer.

Yu, H.; Li, G.; Zhang, W.; Huang, Q.; Du, D.; Tian, Q.; and Sebe, N. 2020. The unmanned aerial vehicle benchmark: Object detection, tracking and baseline. *International Journal of Computer Vision*, 128: 1141–1159.

Yuan, M.; Shi, X.; Wang, N.; Wang, Y.; and Wei, X. 2024. Improving RGB-infrared object detection with cascade alignment-guided transformer. *Information Fusion*, 105: 102246.

Zhang, L.; Danelljan, M.; Gonzalez-Garcia, A.; Van De Weijer, J.; and Shahbaz Khan, F. 2019. Multi-modal fusion for end-to-end RGB-T tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2252–2261.

Zhang, L.; Wang, L.; Wu, Y.; Chen, M.; Zheng, D.; Cao, L.; Zeng, B.; and Cai, Y. 2025. UniRTL: A universal RGBT and low-light benchmark for object tracking. *Pattern Recognition*, 158: 110984.

Zhang, P.; Wang, D.; Lu, H.; and Yang, X. 2021. Learning adaptive attribute-driven representation for real-time RGB-T tracking. *International Journal of Computer Vision*, 129: 2714–2729.

Zhang, P.; Zhao, J.; Wang, D.; Lu, H.; and Ruan, X. 2022. Visible-thermal UAV tracking: A large-scale benchmark and new baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8886–8895.

Zhang, T.; He, X.; Jiao, Q.; Zhang, Q.; and Han, J. 2024. AMNet: Learning to align multi-modality for RGB-T tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8): 7386–7400.

Zhu, J.; Lai, S.; Chen, X.; Wang, D.; and Lu, H. 2023. Visual prompt multi-modal tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9516–9526.