AlignTrack: Top-Down Spatiotemporal Resolution Alignment for RGB-Event Visual Tracking

Chuanyu Sun[1*], Jiqing Zhang[2*], Yang Wang[1], Yuanchen Wang[1], Yutong Jiang[3†], Baocai Yin[4], Xin Yang[1†]

[1] Key Laboratory of Social Computing and Cognitive Intelligence, Dalian University of Technology
[2] Dalian Maritime University
[3] China Northern Vehicle Research Institute
[4] Beijing University of Technology

Abstract

Most existing RGB-Event trackers rely on strictly aligned datasets, overlooking the asynchronous spatio-temporal resolutions common in real-world scenarios. This methodological limitation impedes effective RGB-Event feature alignment and ultimately degrades tracking performance. To overcome this limitation, we propose AlignTrack, a novel tracking framework built upon a Top-Down Alignment (TDA) strategy inspired by the human visual system. Our TDA framework follows an encode-decode-align paradigm: it first encodes multimodal features to generate target-related priors, which are then progressively decoded to guide a subsequent feature alignment pass. Within this framework, we introduce two key innovations: (1) a Cross-Prior Attention (CPA) module that effectively generates and integrates cross-modal priors, and (2) a Cross-Modal Semantic Alignment (CSA) loss that maximizes mutual information to enforce semantic consistency between modalities. Extensive experiments show that AlignTrack achieves state-of-the-art performance on four challenging RGB-Event tracking benchmarks, demonstrating its robustness in both aligned and unaligned scenarios. Ablation studies further validate the significant contribution of each proposed component.

Code — github.com/scy0712/AlignTrack

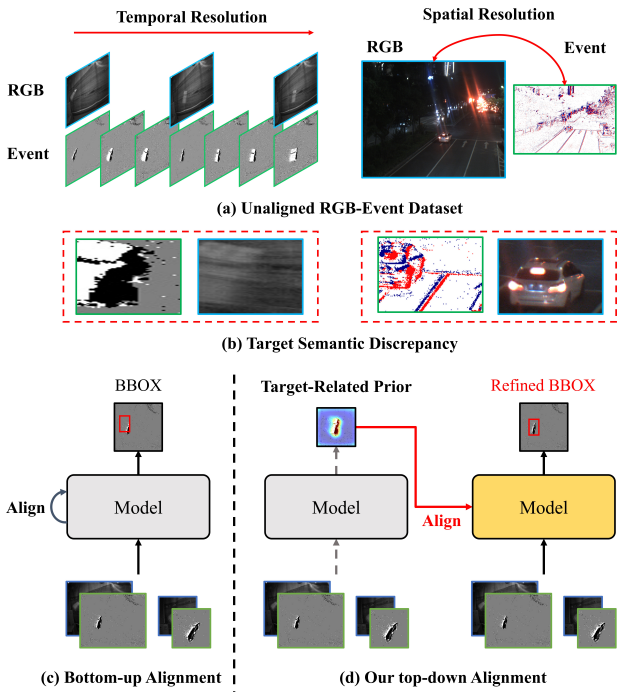

Figure 1: (a) Illustration of temporal and spatial misalignment between RGB and Event data. (b) Examples of target regions cropped from both modalities under temporal misalignment (left) and spatial misalignment (right). (c) Conventional bottom-up alignment framework. (d) Our proposed top-down cross-modal alignment framework.

Introduction

Visual object tracking (VOT) is a fundamental task in computer vision with wide-ranging applications spanning from robotic vision to autonomous driving (Ali et al. 2016; Li et al. 2019). While traditional RGB trackers have shown promise in general settings, which often struggle with challenging scenarios. Owing to the advantages of high temporal resolution and dynamic range offered by Event cameras, there has been a growing trend of integrating them with conventional RGB cameras for robust visual tracking tasks (Zhang et al. 2021; Tang et al. 2022a). These RGBE approaches can significantly enhance tracking performance in challenging conditions, such as fast motion and limited illumination. However, previous RGBE trackers have primarily focused on strictly per-frame aligned datasets, overlooking the inherent discrepancy between RGB and Event sensors in real-world scenarios. As depicted in Fig. 1 (a), RGB cameras typically capture rich appearance and texture details with higher spatial resolution, and event cameras provide fine-grained motion information with superior temporal resolution. Several recent works have identified this discrepancy and introduced alignment strategies designed for particular scenarios. For example, AFNet (Zhang et al.

*These authors contributed equally.
†Co-corresponding authors.

2023) employs deformable convolutions and style transfer networks to align the temporal resolution. CRSOT (Zhu et al. 2024) leverages uncertainty estimation networks to align the spatial information between the two modalities.

However, these methods encounter two key challenges. **(i)** Their meticulously designed alignment modules, tailored for specific misalignment scenarios, exhibit limited generalization to other types of (mis)aligned datasets. **(ii)** Moreover, as illustrated in Fig. 1(b), the spatiotemporal resolution mismatches between RGB and Event modalities lead to a semantic discrepancy, which compromises cross-modal feature alignment.

We attribute the first challenge—limited generalization to the prevailing bottom-up alignment strategies (Fig. 1(c)). These methods rely on low-level features for fusion and alignment, which contradicts the top-down nature of human cognition: "first determining what to look for, then deciding where to look" (Li 2014). Inspired by recent top-down vision frameworks (Anderson et al. 2018; Shi, Darrell, and Wang 2023) that leverage high-level task-related prior knowledge to guide feature extraction, we propose a novel Top-Down Alignment (TDA) strategy to unify spatiotemporal feature alignment (Fig. 1(d)). However, existing top-down pipelines are ill-suited for the unique demands of the RGB-Event tracking task. Therefore, within our TDA framework, we introduce a novel Cross-Modal Prior Attention (CPA) module to generate target-relevant priors and fuse multimodal information to enhance feature alignment. To address the second challenge—semantic discrepancy, we shift focus from model-centric modifications to feature-level optimization, an aspect often overlooked by prior works. Given the significant distributional differences between RGB and Event features, we propose to maximize the Mutual Information (MI) between their representations. This encourages the model to learn modality-invariant yet target-discriminative features, thereby enhancing the robustness of feature extraction, especially under spatio-temporal misalignment.

Extensive experiments demonstrate that our proposed unified framework, termed AlignTrack, achieves state-of-the-art (SOTA) performance on both aligned and unaligned RGB-Event tracking datasets. In summary, our contributions are threefold:

- We propose a novel TDA strategy with a CPA module for RGBE visual tracking. By leveraging high-level target-related priors, our method effectively guides cross-modal fusion and spatiotemporal alignment within a unified framework.
- We design a CSA Loss based on mutual information maximization, which maintains semantic consistency between RGB and Event modalities.
- Our proposed AlignTrack benefiting both aligned and unaligned scenarios, and achieves SOTA performance on four RGBE tracking benchmarks.

Related Work

RGB-Event Object Tracking

Due to the unreliability of RGB-only data in challenging scenarios, event cameras have been incorporated to enhance tracking performance. Recent RGB-Event trackers can be broadly categorized by their feature extraction backbones. The first category employs CNNs, using parallel branches to extract RGB and Event features independently, followed by cross-modal fusion strategies (Zhang et al. 2021, 2022). The second category leverages Transformer-based architectures to capture long-range dependencies and complex interactions between the two modalities (Tang et al. 2022a; Zhang et al. 2024). To mitigate overfitting from limited RGBE datasets, recent Transformer-based works (Cao et al. 2024; Zhu et al. 2023) have introduced Parameter-Efficient Fine-Tuning (PEFT), adapting pretrained RGB trackers to multimodal scenarios with minimal fine-tuning.

Most of these methods focus on strictly aligned RGBE datasets, which do not fully capture real-world conditions or leverage the full potential of multimodal data. As a result, growing research efforts have focused on addressing RGBE misalignment issues to improve practical applications. For example, AFNet (Zhang et al. 2023) introduced deformable convolutions and cross-style transfer modules to align asynchronous RGB and Event frames. CRSOT (Zhu et al. 2024) proposes an uncertainty-aware network to align multimodal information across different spatial resolutions.

However these methods adopt bottom-up fusion strategies tailored to specific misalignment types, leading to limited adaptability. To overcome this limitation, we propose a unified framework that tackles the spatio-temporal misalignment problem from both an architectural and a feature optimization standpoint, which provides a more holistic solution to the RGB-Event alignment challenge.

Top-Down Vision

As a fundamental aspect of human visual processing, Top-down vision, inspired by human visual processing (Li 2014), leverages prior knowledge to focus on task-relevant features, a principle successfully applied in tasks like VQA (Anderson et al. 2018) and object tracking (Lai et al. 2024; Chen et al. 2024). For instance, AbSViT (Shi, Darrell, and Wang 2023) firstly introduced top-down vision to the Transformers architecture (Dosovitskiy et al. 2020). In these tasks, a common implementation is the "encode-decode-encode" strategy: an initial model output is used to generate priors, which are then decoded and fed back to guide a subsequent stage of feature extraction. This enables the model to concentrate on pertinent regions while suppressing irrelevant background distractions, thereby enhancing its discriminative capability. While most top-down methods focus on high-level vision tasks, we extend this concept to the low-level challenge of RGB-Event feature alignment. Our proposed AlignTrack leverages semantic priors to guide the adaptive fusion process, addressing the spatio-temporal discrepancies between the two modalities.

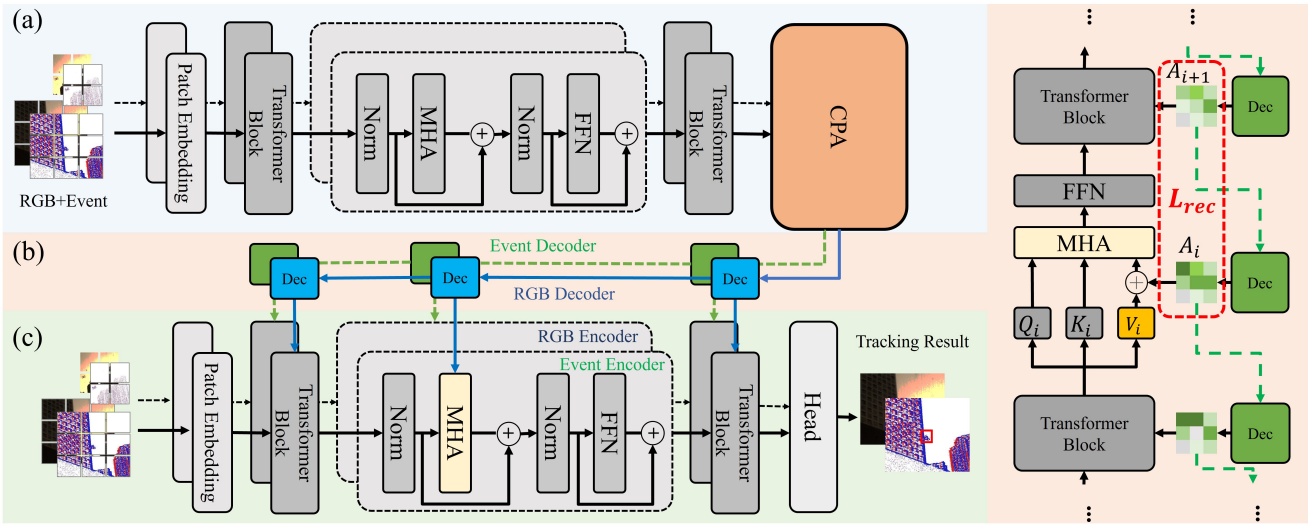

Figure 2: Overview of our Top-Down Alignment (TDA) strategy. (a) Encoding Stage: Extracts multimodal features and generates target-aware priors via the Cross-Modal Prior Attention (CPA) module. (b) Decoding Stage: Progressively decodes these target-aware signals in a layer-by-layer manner. (c) Alignment Stage: Utilizes the priors from one modality to adaptively guide and align the features of the other.

Mutual Information

Mutual Information **(MI)** is a fundamental concept in information theory that quantifies the dependency between two random variables. Given two random variables x and z, the MI between them is defined as:

$$E_{p(x,z)}\left[\log \frac{p(x,z)}{p(x)p(z)}\right] = \mathrm{D_{KL}}(p(x,z) \parallel p(x)p(z)), \quad (1)$$

where $p(x,z)$ represents the joint probability distribution, while $p(x)$ and $p(z)$ denote the corresponding marginal distributions. Here, D_{KL} refers to the Kullback–Leibler divergence (KLD) (MacKay 2003). MI has a long-standing role in feature learning, initially introduced through the Infomax principle (Linsker 1988) that advocates maximizing MI between the input and output representations for enhancing represent learning. Since directly estimating MI from distributions is challenging in practical scenarios, numerous estimators (Belghazi et al. 2018; Oord, Li, and Vinyals 2018; Hjelm et al. 2018) have been proposed to approximate MI based on observed samples. Many recent studies have used neural networks to estimate it. For instance, MIME (Belghazi et al. 2018) optimizes a lower bound of MI using the Donsker-Varadhan representation (Donsker and Varadhan 1976), while InfoNCE (Oord, Li, and Vinyals 2018) employs a Noise-Contrastive Estimation (NCE) bound.

Recent studies have leveraged MI estimation in various visual tasks to enhance feature representation (Liu et al. 2022). Building on its effectiveness in visual tasks, we leverage MI maximization to address semantic inconsistency in multimodal tracking. By increasing the MI between RGB and Event representations, our model learns modality-invariant features that remain robust under spatio-temporal misalignment.

Method

In this section, we present an overview of AlignTrack. The proposed method features a Top-Down Alignment (TDA) framework that adaptively aligns spatiotemporal misalignment, complemented by a cross-modal semantic alignment loss that encourages multimodal semantic consistency.

Overall Architecture

As illustrated in Fig. 2, AlignTrack comprises three main stages: (a) an Encoding Stage that extracts features and generates multimodal target-aware priors, (b) a Decoding Stage that progressively decodes this prior information into high-level guidance signals layer-by-layer, and (c) an Alignment Stage that leverages these signals to adaptively align the two modalities by modulating the attention mechanism. We detail each of these stages below.

Encoding Stage. Our encoding stage begins by processing the RGB-Event template and search frames through a dual-stream Transformer encoder. This yields initial template tokens $X_z^{\mathrm{RGB}}, X_z^{\mathrm{Event}} \in R^{N_z \times D}$ and search tokens $X_x^{\mathrm{RGB}}, X_x^{\mathrm{Event}} \in R^{N_x \times D}$, each with dimension $D = 768$. The number of tokens, N_z and N_x, is determined by the image sizes and a patch size of 16×16. Subsequently, for each modality $m \in \{\mathrm{RGB}, \mathrm{Event}\}$, we concatenate the template and search tokens and pass them through an N-layer Transformer encoder for joint modeling:

$$X_0^m = \mathrm{Concat}(X_z^m, X_x^m), \quad (2)$$

$$X_n^m = \mathcal{F}_n^m(X_{n-1}^m), \quad n = 1, \ldots, N, \quad (3)$$

where $\mathcal{F}_n^m$ is the n-th encoder layer for modality m, and $X_n^m \in R^{(N_z+N_x)\times D}$ is its output feature.

The encoded final layer features, X_N^{RGB} and X_N^{Event}, are then fed into our Cross-modal Prior Attention (CPA) module, which performs prior generation and multimodal fusion

to produce high-level signals A_{zx}^{RGB} and A_{zx}^{Event}, which encapsulate target-relevant semantic tokens from each modality while suppressing background tokens. This process is defined as:

$$A_{zx}^{\text{RGB}}, A_{zx}^{\text{Event}} = \text{CPA}(X_N^{\text{RGB}}, X_N^{\text{Event}}) \tag{4}$$

Decoding Stage. Next, the target-aware signals, A_{N+1}^m, are propagated top-down to guide each encoder layer. Recognizing that encoder layers are hierarchical, with deeper layers focusing on semantics and shallower layers on textures, We therefore employ a cascade of learnable decoder blocks, DEC_n^m (each with two projection layers), to progressively adapt the guidance signals. Each block refines the signal from the above layer A_{n+1}^m into a new signal A_n^m suitable for the granularity of the current layer's features.

$$A_n^m = \text{DEC}_n^m(A_{n+1}^m), \quad m \in \{\text{RGB}, \text{Event}\}. \tag{5}$$

Alignment Stage. The Alignment Stage is designed to adaptively align multimodal features by modulating the attention mechanism. To this end, we re-run the feature extraction process, this time incorporating the high-level semantic signals generated during the Decode Stage. This top-down strategy strengthens target-aware information within each encoder layer by injecting high-level cross-modal guidance. Taking the RGB branch as an example, this guided feature extraction is formulated as:

$$Y_n^{\text{RGB}} = \mathcal{F}_n^{\text{RGB}}(Y_{n-1}^{\text{RGB}}, A_n^{\text{Event}}). \tag{6}$$

Here, Y_n represents the aligned features at the n-th layer. To avoid disrupting the self-attention matrix, we inject the top-down signal from the Event decoder A_n^{Event} directly into the *Value* projection of the RGB branch. This injection is detailed as follows:

$$Y_n^{\text{RGB}} = \text{Attention}(Q_n^{\text{RGB}}, K_n^{\text{RGB}}, V_n^{\text{RGB}}), \tag{7}$$

$$\text{where} \quad \begin{cases} Q_n^{\text{RGB}} = W_n^Q Y_{n-1}^{\text{RGB}}, \\ K_n^{\text{RGB}} = W_n^K Y_{n-1}^{\text{RGB}}, \\ V_n^{\text{RGB}} = W_n^V (Y_{n-1}^{\text{RGB}} + A_n^{\text{Event}}), \end{cases} \tag{8}$$

where $W_n^Q, W_n^K, W_n^V \in R^{d \times d}$ are projection matrices of RGB encoder with $d = 768$. We adopt a multi-head attention mechanism with 8 heads. Y_n^{RGB} represents the aligned RGB features at layer n. The final aligned search features from the last layer Y_x^{RGB} and Y_x^{Event} are then summed and fed into the tracking head for bounding box prediction.

Cross-modal Prior Attention

The human visual system selectively focuses on pertinent information by leveraging prior knowledge. Inspired by this, we aim to explicitly modulate the initial encoder outputs, which implicitly contain high-level semantic responses of the tracked target. To this end, we propose the CPA module to select task-relevant features and fuse multimodal priors in a cross-attention manner.

As detailed in Fig. 3, firstly, the initial outputs from the encoder F_{xz}^{RGB} and F_{xz}^{Event} undergo spatial fusion through a Multi-Head Attention (MHA) block:

$$H_{xz}^{\text{RGB}}, H_{xz}^{\text{Event}} = \text{MHA}(F_{xz}^{\text{RGB}}, F_{xz}^{\text{Event}}), \tag{9}$$

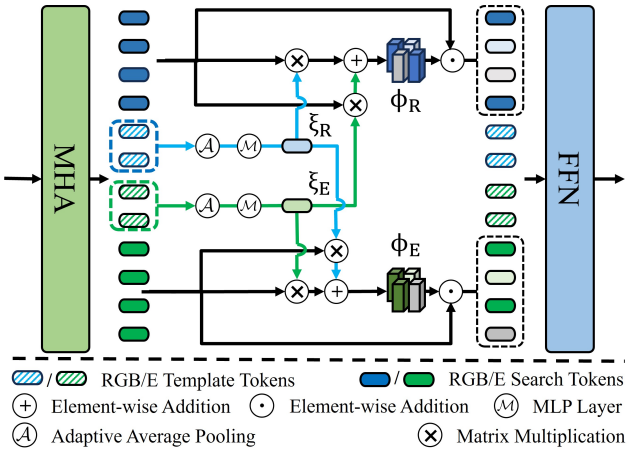

Figure 3: An overview of our proposed CPA module.

here, $H_{xz}^{\text{RGB}} = [H_z^{\text{RGB}}, H_x^{\text{RGB}}]$ represents the fused RGB template and search tokens.

Next, we generate modality-specific priors $\xi^m \in R^{1 \times D}$, which encapsulate target-relevant features within each modality. These priors are then used to compute similarity matrices $\phi^m \in R^{1 \times D}$ between the search tokens H_x^m and the multimodal priors via a cross-fusion strategy. This strategy ensures that the information from one modality is guided by both, effectively highlighting target-salient regions in the search tokens while suppressing background noise. The process is formulated as:

$$\xi^m = \theta^m(\mathcal{P}(H_z^m)), \quad m \in \{\text{RGB}, \text{Event}\}, \tag{10}$$

$$\phi^m = \text{Mul}(H_x^m, \xi^{\text{RGB}}) + \text{Mul}(H_x^m, \xi^{\text{Event}}), \tag{11}$$

$$P^m = \phi^m \odot H_x^m, \tag{12}$$

where $\mathcal{P}(\cdot)$ denotes token-level average pooling and $\theta^m(\cdot)$ is a linear layer. The operator $\text{Mul}(\cdot, \cdot)$ represents matrix multiplication, and $\odot$ denotes element-wise multiplication. Finally, a Feed-Forward Network further refines these fused features P^m to facilitate multimodal channel interaction.

Reconstruction Loss

We optimize the decoder using a layer-wise, VAE-style (Kingma, Welling et al. 2013) reconstruction loss to avoid high-level semantic signals degeneration. Recognizing that strong generative training can impair discriminative feature learning (Shi, Darrell, and Wang 2023), we incorporate a stop-gradient operation to balance these objectives. This allows for effective feature refinement via the decoders without negatively impacting the encoder's learned representations. The total reconstruction loss, $\mathcal{L}_{\text{rec}}$, is computed by summing the layer-wise losses over both the RGB and Event modalities ($m \in \{\text{RGB}, \text{Event}\}$):

$$\mathcal{L}_{\text{rec}} = \sum_m \sum_{n=0}^{12} \left\| \text{sg}(A_n^m) - \text{DEC}_n(\text{sg}(A_{n+1}^m)) \right\|_2^2, \tag{13}$$

where the stop-gradient, $\text{sg}(\cdot)$, isolates the optimization to each decoder layer, DEC_n, preventing gradients from this

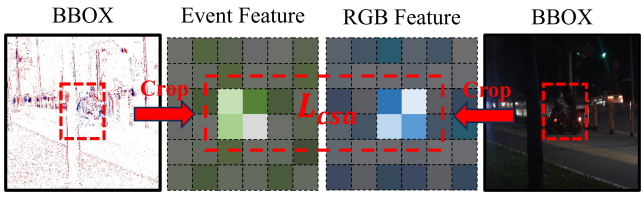

Figure 4: An overview of our proposed L_{csa} loss function and the target-aware cropping operation.

loss from backpropagating into the encoder. This enables rich semantic reconstruction while preserving the discriminative features.

Cross-Modal Semantic Alignment

MI Estimation. Directly estimating MI from high-dimensional distributions is notoriously challenging. To ensure both stability and computational efficiency, we adopt the DeepInfoMax estimator (Hjelm et al. 2018), which provides a reliable lower bound on MI by leveraging the Jensen-Shannon Divergence (JSD) that is formulated as:

$$\begin{aligned}\hat{I}_{\omega}^{(\mathrm{JSD})}(X;Z) := E_{p(x,z)}\left[-\mathrm{sp}(-T_{\omega}(x,z))\right] \\ - E_{p(x)p(z)}\left[\mathrm{sp}(T_{\omega}(x,z))\right],\end{aligned} \tag{14}$$

where $T_{\omega}: \mathcal{X} \times \mathcal{Z} \to R$ is a neural network discriminator parameterized by ω, and $\mathrm{sp}(a) = \log(1+e^{a})$ is the softplus function.

Cross-Modal Semantic Alignment Loss. Spatio-temporal resolution misalignments between RGB and Event data often lead to a semantic discrepancy in their respective feature representations. To mitigate this, we propose a cross-modal semantic alignment loss, $\mathcal{L}_{csa}$, designed to maximize the MI between the two modalities' features. As illustrated in Fig. 4, we first extract the final feature maps from both the RGB and Event branches after the encoding stage. To focus on the target and suppress background noise, we crop the target-specific features using the predicted bounding box and apply a spatial pooling operation to obtain compact semantic representations S^{RGB} and S^{Event}. The loss is then defined as the negative JSD-based MI estimate between these representations:

$$\mathcal{L}_{csa} = -\hat{I}_{\Theta}^{(\mathrm{JSD})}\left(S^{\mathrm{RGB}}, S^{\mathrm{Event}}\right), \tag{15}$$

where Θ represents the parameters of our MI estimator, which is implemented as a three-layer MLP. By maximizing the MI, we encourage the model to learn modality-invariant yet target-discriminative representations.

Head and Loss

Following OSTrack (Ye et al. 2022), we use a convolutional head to predict the target's center and scale. The total loss is a weighted sum of five components:

$$L = \lambda_{cls}L_{cls} + \lambda_{iou}L_{iou} + \lambda_{L_1}L_{L_1} + \lambda_{rec}L_{rec} + \lambda_{csa}L_{csa}, \tag{16}$$

where L_{cls} is the weighted focal loss (Lin et al. 2017); L_{iou} and L_{L_1} are the GIoU (Rezatofighi et al. 2019) and L1 losses for regression; and L_{rec} and L_{csa} are the reconstruction and cross-modal semantic alignment losses. The corresponding weights are $\lambda_{cls} = 1$, $\lambda_{iou} = 2$, $\lambda_{L_1} = 5$, $\lambda_{rec} = 0.1$, and $\lambda_{csa} = 10^{-5}$.

Experiment

Experimental Settings

Implementation Details. We implement AlignTrack using PyTorch and train it on 4 NVIDIA RTX 4090 GPUs with a batch size of 32. We employ the AdamW optimizer with a weight decay of 1×10^{-4} and a learning rate of 1×10^{-4}. We utilize pretrained OSTrack (Ye et al. 2022) as the Transformer encoder. Fine-tuning our method on the RGB-Event training set with 15 epochs. We evaluate the performance of our method on four large-scale RGB-Event single object tracking datasets, including two unaligned datasets, FE240(Zhang et al. 2021) and CRSOT(Zhu et al. 2024), and two aligned datasets, VisEvent(Wang et al. 2023) and COESOT(Tang et al. 2022b).

Evaluation Metrics. We adopt three widely used metrics to quantitatively evaluate tracking performance: Success Rate (SR), Precision Rate (PR), and Normalized Precision Rate (NPR). Specifically, SR measures the proportion of frames where the Intersection over Union (IoU) between the predicted bounding box and the ground truth exceeds a predefined threshold. PR calculates the percentage of frames in which the center location error (CLE) falls within a given threshold. NPR further normalizes the CLE by the size of the ground truth bounding box, ensuring robustness to scale variations and enabling fair comparisons across objects of different sizes and resolutions.

Comparison on Unaligned Datasets

Performance on FE240. The FE240 dataset features an annotation frequency of 20/40 Hz in the frame domain and 240 Hz in the event domain, making it appropriate for assessing the effectiveness of trackers on temporally misaligned multimodal data. As shown in Tab. 1, our method outperforms the previous SOTA Multimodal tracker STTrack by 1.9% in SR and 1.9% in PR, indicating that our approach has superior potential for temporal alignment when addressing frame-event asynchronous data.

Performance on CRSOT. The CRSOT dataset features highly asymmetrical resolutions, with RGB frames at 1440×1080 and event frames at 1280×800. It contains 836 sequences for training and 194 for testing. As shown in Tab. 1, our method surpasses the RGBE tracker CRSOT and the multimodal tracker STTrack by 1.2% in SR, showcasing its superior capability in spatial alignment.

Comparison on Aligned Datasets

To further demonstrate the effectiveness of our AlignTrack, we validate it on two spatiotemporally aligned RGBE multimodal datasets.

Performance on VisEvent. The VisEvent dataset (Wang et al. 2023) offers synchronised RGB and event multimodal data at a resolution of 346×260. Following the protocols in (Zhang et al. 2023, 2024), we filter out sequences with missing event data or misaligned timestamps, resulting in 205

Method	Source	Type	FE240		CRSOT			VisEvent			COESOT		
			SR	PR	SR	PR	NPR	SR	PR	NPR	SR	PR	NPR
TransT (Chen et al. 2021)	CVPR21	RGB	49.3	76.2	54.0	65.5	65.9	37.6	55.1	-	60.5	67.9	66.6
STARK (Yan et al. 2021)	ICCV21	RGB	46.2	79.4	-	-	-	41.8	54.6	-	55.7	62.6	-
ToMP (Mayer et al. 2022)	CVPR22	RGB	52.3	83.1	54.6	63.9	66.6	38.1	49.4	-	59.8	67.2	66.0
OSTrack (Ye et al. 2022)	ECCV22	RGB	-	-	55.5	66.1	67.5	56.1	67.7	-	59.0	67.2	66.0
ARTrack (Wei et al. 2023)	CVPR23	RGB	-	-	56.8	68.1	69.3	62.6	78.2	75.0	-	-	-
AQATrack (Chen et al. 2023)	CVPR24	RGB	66.0	92.2	-	-	-	62.8	79.2	75.8	-	-	-
FENet (Zhang et al. 2021)	ICCV21	RGBE	55.6	84.3	-	-	-	44.2	-	58.9	-	-	-
CEUTrack (Tang et al. 2022a)	ArXiv22	RGBE	-	-	-	-	-	53.5	71.8	66.4	62.0	70.5	69.0
AFNet (Zhang et al. 2023)	CVPR23	RGBE	58.4	87.0	-	-	-	44.5	-	59.3	59.2	67.8	-
CSAM (Zhang et al. 2024)	NIPS24	RGBE	-	-	-	-	-	61.5	76.1	72.4	63.6	73.3	70.5
CRSOT (Zhu et al. 2024)	TMM25	RGBE	65.7	92.3	61.8	74.2	74.4	53.4	70.6	63.2	60.8	75.1	-
ViPT (Zhu et al. 2023)	CVPR23	Multi	60.6	87.2	54.6	64.9	66.0	60.8	76.6	73.0	65.7	73.9	72.2
BAT (Cao et al. 2024)	AAAI24	Multi	66.4	92.1	61.0	71.9	72.0	62.8	78.5	75.4	67.8	76.0	74.2
SDSTrack (Hou et al. 2024)	CVPR24	Multi	-	-	-	-	-	62.6	79.3	75.5	-	-	-
OneTrack (Hong et al. 2024)	CVPR24	Multi	-	-	-	-	-	63.2	78.1	75.6	-	-	-
UnTrack (Wu et al. 2024)	CVPR24	Multi	58.9	88.4	61.2	72.0	72.5	61.2	77.2	73.3	66.7	75.1	73.5
STTrack (Hu et al. 2025)	AAAI25	Multi	66.2	92.4	61.8	72.9	73.3	63.2	78.5	75.5	68.4	76.9	74.9
AlignTrack	-	RGBE	68.1	94.3	63.0	74.1	74.6	63.7	80.4	76.8	68.7	77.1	75.3

Table 1: Comparison of three types of SOTA trackers, each designed for different tasks. However, all trackers are trained and evaluated on RGB-Event data. The top three results are highlighted in red, blue, and green, respectively. The notation "_" indicates the tracker originally proposed alongside the corresponding dataset.

#	Models	FE240	CRSOT	VisEvent	Δ
1	RGB Only	44.2	62.4	61.8	-8.8
2	Event Only	61.5	11.2	40.0	-27.4
3	w/o TDA	62.0	61.9	62.7	-2.7
4	TDA w/o CPA	66.7	62.3	63.2	-0.9
5	TDA w/o CF	67.3	62.7	63.6	-0.4
6	TDA w/o L_{rec}	66.2	62.4	63.4	-0.7
7	w/o L_{csa}	65.1	62.6	63.2	-1.3
8	L_{csa} w/o crop	65.6	62.7	63.6	-0.9
9	**Ours**	**68.1**	**63.0**	**63.7**	-

Table 2: Quantitative comparison of different variants of our method in terms of the SR score. The best performance is highlighted in bold; Δ represents the performance change relative to our default setting.

Layer	FE240	CRSOT	VisEvent	Δ
3	63.0	62.3	62.9	-2.2
6	65.5	62.6	63.1	-1.2
9	66.2	62.7	63.2	-0.9
12	**68.1**	**63.0**	**63.7**	-

Table 3: Ablation study of SR on different decoder layers.

#	Strategy	FE240	CRSOT	VisEvent	Δ
A	$Q_n + A_n$	65.7	62.2	63.3	-1.2
B	$K_n + A_n$	66.3	62.2	63.5	-0.9
C	ALL	65.8	62.3	63.2	-1.2
D	$V_n + A_n$	**68.1**	**63.0**	**63.7**	-

Table 4: Quantitative SR comparison of TDA strategies.

sequences for training and 172 for testing. Tab. 1 shows that our method attains new SOTA results with SR, PR, and NPR scores of 63.7%, 80.4%, and 76.8%, respectively, outperforming all existing trackers.

Performance on COESOT. The COESOT dataset (Tang et al. 2022b) comprises 578K RGB-Event pairs, divided into 827 sequences for training and 527 for testing. As shown in Tab. 1, our method achieves an overall SR of 68.7%, PR of 77.1%, and NPR of 75.3%. These results demonstrate the effectiveness of integrating a top-down semantic prior for enhancing modality fusion in the RGBE object tracking task. An interesting observation is that CRSOT attains runner-up performance on the unaligned dataset, while significantly underperforming compared to our method on the aligned dataset, because the heavily parameterized uncertainty estimation module in CRSOT is carefully designed and performs well under modality misalignment. However, it tends to impair the model's generalization ability when the modalities are well aligned.

Ablation Study

Impact of Multimodal Input. We evaluate the effectiveness of multimodal input by utilizing event and RGB inputs independently. Tab. 2 #1 and #2 show that the tracking results with unimodal inputs are significantly worse than those with multimodal inputs. We can also observe that on the FE240 dataset, inputting solely event data yields superior performance compared to using only RGB input, while the reverse is shown on the CRSOT and VisEvent datasets. This is due to the variety and modality bias inherent in various datasets. These results suggest the effectiveness and generalizability of our approach for multimodal fusion.

Effectiveness of the TDA Components. Our TDA module comprises two key components: the CPA module for prior

λ_{rec}	SR(%)	PR(%)	λ_{csa}	SR(%)	PR(%)
1	67.9	93.8	1	67.5	93.5
0.1	**68.1**	**94.3**	0.001	67.9	93.9
0.01	67.4	93.4	**0.00001**	**68.1**	**94.3**

Table 5: Impact of the loss hyperparameters on FE240.

Tracker	FE240 ↑	CRSOT ↑	Speed (FPS) ↑	Param (M) ↓
UnTrack	58.9	61.2	16.2	**98.7**
CRSOT	65.7	61.8	**33.9**	135.8
STTrack	66.2	61.8	24.7	128.2
Ours	**68.1**	**63.0**	30.3	108.6

Table 6: Efficiency and SR scores comparison.

generation and fusion, and the Reconstruction Loss ($\mathcal{L}_{\text{rec}}$) for optimizing the decoder. We conduct an ablation study to validate their effectiveness, with results shown in Tab. 2. Removing the entire TDA module (#3) leads to a significant performance drop of 2.7% in average SR, highlighting its overall importance, especially in unaligned scenarios. Isolating the CPA module's contribution, we replace it with learnable tokens as in (Shi, Darrell, and Wang 2023) (#4), which results in a 0.9% SR decrease. This confirms that our CPA, particularly its cross-fusion operation (validated by the degradation in #5) is critical for effective modality interaction and alignment. Furthermore, removing the reconstruction loss (#6) also degrades performance, underscoring its necessity. We also perform an ablation on its weight, λ_{rec}, in Tab. 5, finding that a moderate value of 0.1 yields the best results and confirms our design choice.

Impact of the Decoder Layer. We investigate the effect of the number of decoder layers as shown in Tab. 3. The results show a clear trend: as the number of decoder layers increases, the model's SR consistently improves. This suggests that a deeper hierarchy of top-down signals provides more effective guidance, enabling the encoder to better align dual modalities. Furthermore, the consistent performance gains across all three datasets demonstrate that our TDA strategy is broadly effective, enhancing both unaligned and aligned scenarios through its robust cross-modal feature alignment capabilities.

Effectiveness of the TDA Strategy. Additionally, we explore different strategies for injecting top-down signals into the encoder. Tab. 4 compares four configurations: injecting top-down signals into the Query (A), Key (B), Value (D), or all three (C) within the Multi-Head Attention module. Among these, D achieves the best performance. This suggests that introducing signals into the Query or Key disrupt the distribution of the attention matrix, whereas the Value component is more suitable for effectively receiving and integrating top-down guidance.

Effectiveness of CSA Loss. As shown in Tab. 2, removing $\mathcal{L}_{csa}$ (#7) results in a 1.3% average SR drop, highlighting its role in mitigating modality discrepancy. Similarly, removing the cropping operation (#8) causes a 0.9% performance decrease, which emphasizes the importance of minimizing background interference. Furthermore, we ablate the loss weight λ_{csa} in Tab. 5, where our setting of 10^{-5} yields the best performance over larger values, validating our design.

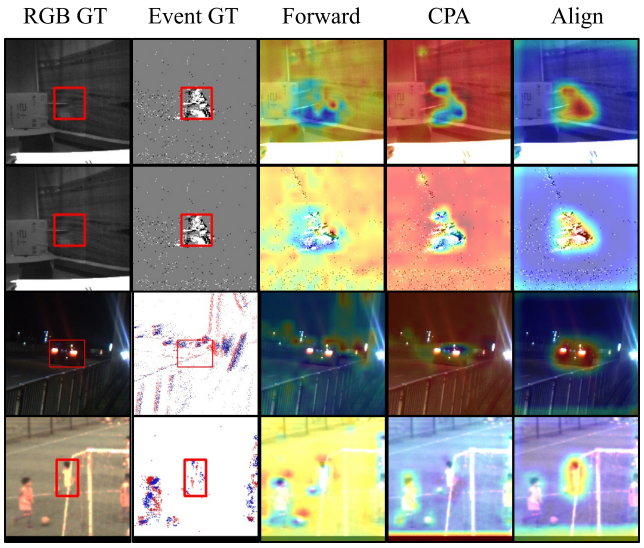

Figure 5: Visualization of attention maps. Row 1 and 2 are visualized on RGB and Event images from FE240, row 3 is from CRSOT, and row 4 is from VisEvent.

Efficiency Comparisons. As shown in Tab. 6, our method achieves an inference speed of 30.3 FPS with 108.6M parameters on a single RTX 4090 GPU. This demonstrates a favorable balance between tracking accuracy and computational efficiency, despite our model's multi-stage feature extraction and layer-wise decoding architecture.

Visualization. Furthermore, we visualize the attention maps from different stages in Fig. 5 to illustrate the inner workings of our TDA strategy. Initially, in the encoding stage, the attention maps are often noisy and diffuse, adversely affected by issues like temporal latency (rows 1-2), spatial misalignment (row 3), or background distractors (row 4), thus hindering accurate target localization. While the CPA module improves focus by applying target priors, its similarity-based nature can fail against similar distractors (e.g., row 4). The final Alignment Stage overcomes this limitation, using top-down guidance to resolve ambiguities and achieve a sharp, precise target localization.

Conclusion

We propose AlignTrack to tackle spatio-temporal misalignment in RGBE tracking through a novel top-down alignment strategy. It leverages a noval CPA module to enhance feature and a CSA loss for semantic consistency. While our method focuses on aligning asynchronous frames, which do not fully exploit the high temporal resolution and polarity information of event streams, future research will focus on developing more advanced alignment strategies between event streams and RGB frames to boost tracking performance.

Acknowledgments

The paper has been supported by National Key Research and Development Program of China (No.2022ZD0210500).

References

Ali, A.; Jalil, A.; Niu, J.; Zhao, X.; Rathore, S.; Ahmed, J.; and Aksam Iftikhar, M. 2016. Visual object tracking—classical and contemporary approaches. *Frontiers of Computer Science*, 10(1): 167–188.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086.

Belghazi, M. I.; Baratin, A.; Rajeswar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, R. D. 2018. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*.

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, L.; Zhong, B.; Liang, Q.; Zheng, Y.; Mo, Z.; and Song, S. 2024. Top-down cross-modal guidance for robust rgb-t tracking. *IEEE Transactions on Circuits and Systems for Video Technology*.

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.

Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; and Lu, H. 2021. Transformer tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8126–8135.

Donsker, M. D.; and Varadhan, S. S. 1976. Asymptotic evaluation of certain Markov process expectations for large time—III. *Communications on pure and applied Mathematics*, 29(4): 389–461.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

Hong, L.; Yan, S.; Zhang, R.; Li, W.; Zhou, X.; Guo, P.; Jiang, K.; Chen, Y.; Li, J.; Chen, Z.; et al. 2024. Onetracker: Unifying visual object tracking with foundation models and efficient tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19079–19091.

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.

Kingma, D. P.; Welling, M.; et al. 2013. Auto-encoding variational bayes.

Lai, S.; Liu, C.; Wang, D.; and Lu, H. 2024. Refocus the Attention for Parameter-Efficient Thermal Infrared Object Tracking. *IEEE Transactions on Neural Networks and Learning Systems*.

Li, K.; He, F.; Yu, H.; and Chen, X. 2019. A parallel and robust object tracking approach synthesizing adaptive Bayesian learning and improved incremental subspace learning. *Frontiers of Computer Science*, 13(5): 1116–1135.

Li, Z. 2014. *Understanding vision: theory, models, and data*. Oxford University Press, USA.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Linsker, R. 1988. Self-organization in a perceptual network. *Computer*, 21(3): 105–117.

Liu, Z.; Feng, R.; Chen, H.; Wu, S.; Gao, Y.; Gao, Y.; and Wang, X. 2022. Temporal feature alignment and mutual information maximization for video-based human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11006–11016.

MacKay, D. J. 2003. *Information theory, inference and learning algorithms*. Cambridge university press.

Mayer, C.; Danelljan, M.; Bhat, G.; Paul, M.; Paudel, D. P.; Yu, F.; and Van Gool, L. 2022. Transforming model prediction for tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8731–8740.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 658–666.

Shi, B.; Darrell, T.; and Wang, X. 2023. Top-down visual attention from analysis by synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2102–2112.

Tang, C.; Wang, X.; Huang, J.; Jiang, B.; Zhu, L.; Zhang, J.; Wang, Y.; and Tian, Y. 2022a. Revisiting color-event based tracking: A unified network, dataset, and metric. *arXiv preprint arXiv:2211.11010*.

Tang, C.; Wang, X.; Huang, J.; Jiang, B.; Zhu, L.; Zhang, J.; Wang, Y.; and Tian, Y. 2022b. Revisiting color-event based tracking: A unified network, dataset, and metric. *arXiv preprint arXiv:2211.11010*.

Wang, X.; Li, J.; Zhu, L.; Zhang, Z.; Chen, Z.; Li, X.; Wang, Y.; Tian, Y.; and Wu, F. 2023. Visevent: Reliable object tracking via collaboration of frame and event flows. *IEEE Transactions on Cybernetics*.

Wei, X.; Bai, Y.; Zheng, Y.; Shi, D.; and Gong, Y. 2023. Autoregressive visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9697–9706.

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.

Yan, B.; Peng, H.; Fu, J.; Wang, D.; and Lu, H. 2021. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10448–10457.

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.

Zhang, J.; Dong, B.; Zhang, H.; Ding, J.; Heide, F.; Yin, B.; and Yang, X. 2022. Spiking transformers for event-based single object tracking. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 8801–8810.

Zhang, J.; Wang, Y.; Liu, W.; Li, M.; Bai, J.; Yin, B.; and Yang, X. 2023. Frame-event alignment and fusion network for high frame rate tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9781–9790.

Zhang, J.; Yang, X.; Fu, Y.; Wei, X.; Yin, B.; and Dong, B. 2021. Object tracking by jointly exploiting frame and event domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13043–13052.

Zhang, T.; Debattista, K.; Zhang, Q.; Han, J.; et al. 2024. Revisiting motion information for RGB-Event tracking with MOT philosophy. *Advances in Neural Information Processing Systems*, 37: 89346–89372.

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.

Zhu, Y.; Wang, X.; Li, C.; Jiang, B.; Zhu, L.; Huang, Z.; Tian, Y.; and Tang, J. 2024. Crsot: Cross-resolution object tracking using unaligned frame and event cameras. *arXiv preprint arXiv:2401.02826*.