Journal of King Saud University Computer and Information Sciences (2025) 37:128
https://doi.org/10.1007/s44443-025-00142-y

ORIGINAL PAPER

Missingness-aware prompting for modality-missing RGBT tracking

Guyue Hu[1,2,3,5,6] · Zhanghuan Wang[1,2,4,7] · Chenglong Li[1,2,3,6] · Duzhi Yuan[6] · Bin He[8] · Jin Tang[1,2,4,7]

Received: 22 January 2025 / Accepted: 24 June 2025

Abstract
RGBT tracking has drawn great attention recently due to its ability to leverage enhancement and complementary information from the RGB and thermal infrared modalities. Nevertheless, RGBT tracking in real-world scenarios inevitably encounters heavy modality-missing challenges caused by substantial environmental factors (such as device overheating, and frame skipping). Existing methods for RGBT tracking are built upon pre-processed missingness-free datasets and suffer significant performance degradation when applied to noisy datasets with random missing modalities. In this paper, we propose a novel missingness-aware prompting framework (MAP) for modality-missing RGBT tracking. It is a lightweight prompting framework consisting of two-stage prompts focusing on compensating essential information for RGBT tracking stage-by-stage. Specifically, prototypical missingness-aware prompts (pMAP) are explored to compensate for modality-specific but instance-agnostic prototypical missing information. Contextual missingness-aware prompts (cMAP) are further designed to compensate for instance-specific detailed missing information. Extensive experiments on three large-scale datasets demonstrate the effectiveness and superiority of the proposed framework for RGBT tracking with random missing modalities.

Keywords RGBT tracking · Modality-missing · Prompt learning · Parameter-efficient tuning

1 Introduction

Given the initial bounding box of the targeted object in the first frame in tracking videos, single-object tracking (SOT) aims at tracking the targeted object in subsequent frames. It has wide applications including robot vision (Chen et al. 2017), autonomous driving (Dai et al. 2021), intelligent security (Marvasti-Zadeh et al. 2021), and so on. Over the past decades, the SOT in visible light (RGB) modality has made significant progress benefiting from the wave-by-wave neural network revolutions (Dosovitskiy et al. 2021; Wu et al. 2021; Gao et al. 2024; Liu et al. 2022; Cui et al. 2021).

In recent years, RGBT tracking has garnered increasing attention for its capability of leveraging enhancement and complementary information from the RGB and thermal infrared (TIR) modalities. Thus multi-modal RGBT tracking becomes more robust in challenging imaging conditions than the single RGB modality, such as low illumination, adverse weather, and foggy conditions (Feng and Su 2024). Some pioneer works (Zhang et al. 2019; Peng et al. 2023) directly concatenate the representations from the RGB and TIR encoders, which are somewhat effective and avoid additional fusion modules simultaneously, but they tend to introduce excessive noise. In addition, some other approaches (Zhu et al. 2019; Gao et al. 2019; Lu et al. 2021; Xiao et al. 2022) select candidate boxes from search frames and employ various attention mechanisms for modality fusion and blend the representations from each candidate box pair. Unfortunately, since these candidate boxes only encompass local features from the search frames, the fusion paradigm has not fully leveraged global information and has failed to exploit the complementary multi-modal information. Nowadays, the mainstream tracking methods (Hui et al. 2023; Zhu et al. 2023) usually leverage the powerful capability of arbitrary-term dependency modeling from Transformer networks to realize global and adaptive modality fusion. They incorporate multi-modal template information with search region information during the fusion process, the background noise can be gradually reduced, significantly enhancing the target information in turn.

Despite such significant progress, most existing methods for RGBT tracking are built upon pre-processed datasets under the ideal missingness-free paradigm, as shown in Fig. 1a. However, RGBT tracking in real-world scenarios inevitably encounters heavy modality-missing issues induced by various environmental factors, such as overheat-

Extended author information available on the last page of the article

Published online: 23 July 2025

ing of the collecting device, frame skipping in multi-modal camera, packet loss during transmission, etc. Directly applying existing off-the-shelf methods to noisy datasets without special consideration of missingness alleviation (i.e. the missingness-unaware paradigm in Fig. 1b) suffers significant performance degradation (see experimental evidence in Table 1 for details). So, it is important to consider missing alleviation during methodology design for real-world RGBT tracking (i.e. the missingness-aware paradigm in Fig. 1c).

Random modality-missing phenomenon in multi-modal learning is common in real-world scenarios and some techniques handling modality-missing have been developed for other computer vision tasks. The most straightforward technical pipeline is to restore the missing data. Typically, Zhao et al. (2021) propose a missing modality imagining network to substitute for the missing modality. Ma et al. (2021) construct a reconstruction network to restore missing modality with prior modality knowledge and the Bayesian meta-learning mechanism. Wei et al. (2023) further utilize the knowledge distillation mechanism to stir rich modality-specific prior information in a teacher network. These approaches for missing alleviation for other tasks mainly focus on recovering missing modality information as fully as possible. However, different from most semantic understanding tasks in computer vision, RGBT tracking usually demands high inference speed but does not pursue accurately recovering tracking-irrelevant details. Due to such particularity, it is unnecessary to recover all details (such as surroundings) of the missing modality since restoring the location and shape information of the tracking target is enough.

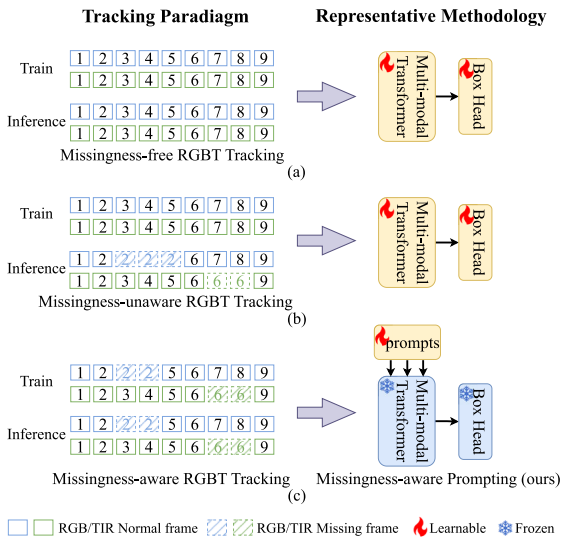

Fig. 1 Comparisons of tracking paradigm and representative methodology for RGBT tracking. (a) Conventional missingness-free RGBT tracking: training and inference on pre-processed clean data without any missing. (b) Missingness-unaware RGBT tracking: training on pre-processed clean data but inference on noisy data with random missing modalities. (c) Missingness-aware RGBT tracking: training and inference on noisy data with random missing modalities and is equipped with special methodology designs for missing alleviation

To move beyond such limitations, we propose a simple yet effective missingness-aware prompting (MAP) framework for modality-missing RGBT tracking. It compensates essential missing information for RGBT tracking via two-stage parameter-efficient prompting instead of recovering all details in the missing frames. Specifically, we first utilize a series of prototypical missingness-aware prompts (pMAP) to approximate the compensation for modality-specific but instance-agnostic modality information, which narrows the gap between padding values and modality prototypes. Then, we apply a series of contextual missingness-aware prompts (cMAP) to perform instance-specific missingness compensation through the context information nearby, which further narrows the information gap between modality prototypes and missing instances.

In summary, the main contributions of this paper could be summarized as follows:

- We highlight the heavy modality-missing issue in real-world RGBT tracking and propose a novel missingness-aware prompting (MAP) framework to alleviate it.
- Prototypical missingness-aware prompts (pMAP) aim to approximate the modality-specific but instance-agnostic missing information, which narrows the information gap between padding values and modality prototypes.
- Contextual missingness-aware prompts (cMAP) are proposed to compensate for instance-specific missing information, which narrows the information gap between modality prototypes and missing instances.
- Our MAP achieves state-of-the-art performance on three large-scale datasets for RGBT tracking, demonstrating its effectiveness in tackling real-world RGBT tracking with random missing modalities.

2 Related work

2.1 RGBT tracking

RGBT tracking refers to tracking objects in complementary visible light and thermal infrared multi-modal data, which is more robust to challenging imaging conditions such as low illumination and adverse weather. Regarding the network architecture, existing RGBT tracking approaches primarily encompass three paradigms: MDNet-based, Siamese-based, and Transformer-based approaches.

MDNet-based methods (Lu et al. 2021; Xiao et al. 2022; Long Li et al. 2019; Li et al. 2020; Zhu et al. 2020) initially generate candidate boxes from search frames and then employ various fusion techniques within these boxes to

Table 1 Comparisons with state-of-the-art RGBT tracking approaches regarding the PR, NPR, and SR metrics on the RGBT210, RGBT234, and LasHeR datasets

Paradigm	Method	RGBT210		RGBT234		LasHeR		
		PR↑	SR↑	PR↑	SR↑	PR↑	NPR↑	SR↑
Missingness-free	TBSI (Hui et al. 2023)	84.1	61.3	85.6	63.7	70.0	66.0	55.9
	BAT (Cao et al. 2024)	85.2	61.3	86.8	64.1	70.2	66.4	56.3
	OSTrack (Ye et al. 2022)*	85.9	62.6	88.0	65.4	68.1	64.2	54.3
	Mixformer (Cui et al. 2022)*	86.5	63.2	86.6	62.9	65.0	60.5	51.6
Missingness-unaware	TBSI (Hui et al. 2023)	75.8	54.2	78.1	56.9	57.4	51.9	46.2
	BAT (Cao et al. 2024)	76.7	54.2	78.5	55.8	59.5	53.5	47.9
	OSTrack (Ye et al. 2022)*	77.4	55.3	80.5	58.6	58.8	52.8	47.1
	Mixformer (Cui et al. 2022)*	78.0	55.4	79.7	57.0	56.0	49.6	44.8
Missingness-aware (ours)	TBSI (Hui et al. 2023)+**MAP**	79.5$_{(+3.7)}$	56.7$_{(+2.2)}$	82.5$_{(+4.4)}$	60.1$_{(+3.2)}$	65.2$_{(+7.8)}$	61.0$_{(+9.1)}$	52.0$_{(+5.8)}$
	BAT (Cao et al. 2024)+**MAP**	80.5$_{(+3.8)}$	57.2$_{(+3.0)}$	82.0$_{(+3.5)}$	58.6$_{(+2.8)}$	63.0$_{(+3.5)}$	59.0$_{(+5.5)}$	50.6$_{(+2.7)}$
	OSTrack (Ye et al. 2022)*+**MAP**	79.9$_{(+2.5)}$	57.3$_{(+2.0)}$	83.1$_{(+2.6)}$	60.8$_{(+2.2)}$	62.9$_{(+4.1)}$	58.6$_{(+5.8)}$	50.2$_{(+3.1)}$
	Mixformer (Cui et al. 2022)*+**MAP**	82.8$_{(+4.8)}$	58.0$_{(+2.6)}$	83.6$_{(+3.9)}$	60.3$_{(+3.3)}$	60.8$_{(+4.8)}$	55.5$_{(+5.9)}$	48.1$_{(+3.3)}$

The values in brackets denote the performance gain obtained from our MAP framework w.r.t various missingness-unaware baseline methods. The largest performance gains in each column are underlined

* denotes our re-implementation that extends original uni-modal trackers to multi-modal RGBT trackers

integrate features from different modalities. Subsequently, binary classification and box regression are performed on the fused multi-modal features. Specifically, to facilitate the information fusion of different modalities, MANet (Long Li et al. 2019) designs a multi-adapter network that extracts modality-shared features, modality-specific features, and instance-level features through three adapters. APFNet (Xiao et al. 2022) decouples the fusion process based on challenge attributes and introduces a novel attribute-based progressive fusion network to fully exploit the advantages of each modality regarding different challenges. The drawback of this paradigm of RGBT tracking methods is that their candidate boxes only cover a portion of the search region, thus lacking global feature modeling and limiting the effectiveness of modality fusion.

In addition, Siamese-based methods (Peng et al. 2023; Zhang et al. 2020) typically extract features from RGB and TIR separately before modality integration. Then, the integrated features are fed into the prediction head for box regression. Although faster than MDNet-based methods, this technical paradigm often lags in accuracy compared to the other two paradigms. With the emerging of large-scale RGBT datasets (Li et al. 2019, 2021) and the rapid development of Transformer networks (Dosovitskiy et al. 2021; Wu et al. 2021; Vaswani et al. 2017; Dordevic et al. 2024), Transformer-based RGBT tracking methods (Hui et al. 2023; Zhu et al. 2023; Feng and Su 2024; Wang et al. 2024) also achieve significant success. For example, the TBSI (Hui et al. 2023) introduces the Transformer backbone into RGBT tracking and designs a TBSI insertion in the transformer backbone for modality fusion. ViPT (Zhu et al. 2023) takes visible light as the primary modality and other modalities are aggregated as prompts into the pre-trained model of a single modality. The BAT (Cao et al. 2024) further employs lightweight bidirectional adapters to adapt off-the-shelf RGB trackers to multimodal scenarios, which parameter-efficiently achieves outstanding tracking performance. Despite such remarkable success, existing methods overlook the heavy modality-missing issue in real-world RGBT tracking scenarios that the multi-modal RGBT datasets inevitably contain substantial random missing modalities. Consequently, these methods will undergo significant performance degradation in real-world scenarios.

2.2 Modality-missing in multi-modal learning

Multi-modal learning aims to utilize the complementary information from multiple modalities to jointly accomplish a task. In practical scenarios, multi-modal data often suffers from inevitable modality-missing due to complex environmental factors. Learning from incomplete multi-modal data is an important and urgent research topic. A simple and straightforward technical pipeline is to impute the missing data. For example, Zhang et al. (2020) introduced an adversarial strategy into the tracking models to handle the missing data, which enhances the model completeness. Zhao et al. (2021) combined the CRA with a cycle consistency loss in an imaginative module to supplement the missing modality information. Lian et al. (2023) simulated the real-world scenario of missing data by randomly discarding modalities and utilized an end-to-end graph completion network to reconstruct the missing modality information. Ma et al. (2021) proposed a Bayesian meta-learning approach to reconstruct the missing modality with prior modality knowledge, demonstrating superior performance in modality-missing scenarios with high missing rates. Besides the imputation-based methods, Ma et al. (2022) investigated the sensitivity of Transformer models to modality dropout and proposed an optimal fusion strategy that searches incompetent input data to enhance the robustness of Transformers. Despite such high success, the existing restore-based technical pipeline of missingness alleviation for other computer vision tasks does not align well with RGBT tracking. Unlike conventional semantic understanding tasks, RGBT tracking usually demands high inference speed but does not pursue accurately recovering tracking-irrelevant details. Therefore, we design a lightweight parameter-efficient prompting framework to compensate for necessary tracking-relevant information while avoiding complete missingness restoration.

2.3 Prompt learning

Prompt Learning is one parameter-efficient tuning technique (Lialin et al. 2023; Hu et al. 2023), which originated from the NLP field and also become popular in the computer vision and multi-modal learning fields nowadays. It can be broadly categorized into two categories. The first type of prompting paradigm (Zhou et al. 2022; Jia et al. 2022; Su et al. 2022; Liu et al. 2024) utilizes soft prompts with a few learnable parameters. Specifically, it freezes learned parameters in original off-the-shelf large models and adds a few learnable tokens to the models for parameter-efficient tuning. For example, CoOp (Zhou et al. 2022) utilizes learnable vectors in continuous space to model context information in soft prompts while freezing parameters in the pre-trained CLIP model (Radford et al. 2021). VPT (Jia et al. 2022) freezes the vision backbone and slightly modifies the Transformer input, achieving better performance than the fully fine-tuning method on various downstream tasks. But this prompting paradigm (Zhou et al. 2022; Chen et al. 2022; Yang et al. 2022; Karimi Mahabadi et al. 2021) only employ inherent learnable vectors that limit the generalization ability. The second type of prompting paradigm is an adapter-based tuning technique that introduces lightweight neural networks into sub-layers of off-the-shelf large models. For example, CoCoOp (Zhou et al. 2022) introduces a lightweight neural network on top

of CoOp to generate conditional vectors for each image and adds them to the original learnable prompts, resulting in a more generalized prompt. AdaptFormer (Chen et al. 2022) introduces lightweight modules into ViT (Dosovitskiy et al. 2021), significantly outperforming fully fine-tuned models on multiple action recognition benchmarks. ProTrack (Yang et al. 2022) proposes a multi-modal prompt tracker to adapt RGB-based trackers to multi-modal tracking without extra training on multi-modal data, effectively alleviating the data deficiency challenge in multi-modal tracking tasks. In this paper, we develop a lightweight parameter-efficient prompting framework to alleviate random modality-missing issues in real-world RGBT tracking scenarios.

3 Method

3.1 Preliminary

3.1.1 Problem setting

Given the initial target bounding box $\boldsymbol{B}_0$ in a pair of spatial-temporal synchronized RGBT video streams, conventional missingness-free RGBT tracking paradigm (Fig. 1a) aims at predicting the subsequent bounding boxes of the targeted object, which could be formulated as follows:

$$Tracker : \{\boldsymbol{X}^t_{RGB}, \boldsymbol{X}^t_{TIR}, \boldsymbol{B}^0\} \rightarrow \boldsymbol{B}^t \tag{1}$$

where $\boldsymbol{X}$ represents the input video frame (image), $\boldsymbol{B}$ represents the bounding box of the targeted object, and t denotes the time stamp. For the RGBT tracking paradigm with random missing modalities (Fig. 1c), continuous frames may be random missing in arbitrary modalities. Following similar notations, it could be formulated as follows:

$$Tracker : \{\boldsymbol{X}^t_{RGB}, \boldsymbol{X}^{t-n}_{TIR}, \boldsymbol{B}^0\} \rightarrow \boldsymbol{B}^t \tag{2}$$

$$Tracker : \{\boldsymbol{X}^{t-n}_{RGB}, \boldsymbol{X}^t_{TIR}, \boldsymbol{B}^0\} \rightarrow \boldsymbol{B}^t \tag{3}$$

where (2) and (3) respectively represent the cases in which the infrared (TIR) and visible (RGB) modalities miss n frames. The missingness-aware RGBT trackers should train and infer on such noisy datasets with random missing modalities to predict the bounding box of the targeted object.

3.1.2 Foundation of tracking models

Typically, an RGBT tracker can be decomposed into three modules: f, g, φ. Specifically, the function $f : \{\boldsymbol{X}_{RGB}, \boldsymbol{X}_{TIR}, \boldsymbol{B}^0\} \rightarrow \{\boldsymbol{H}_{RGB}, \boldsymbol{H}_{TIR}\}$ represents the feature extraction and interaction function. The $g : \{\boldsymbol{H}_{RGB}, \boldsymbol{H}_{TIR}\} \rightarrow \boldsymbol{H}_m$ denotes the function for multi-modal representation fusion. The box prediction head $\varphi : \boldsymbol{H}_m \rightarrow \boldsymbol{B}$ estimates the final tracking box. In this paper, we utilize the powerful Vision Transformer backbone as function f. In detail, the template $\boldsymbol{Z}$ and search $\boldsymbol{X}$ are first embedded into patches and then flattened into the 1D token sequence adding with positional embedding (Dosovitskiy et al. 2021). Then the token sequence will be fed into a linear projection layer to generate the initial query ($\boldsymbol{q}$), key ($\boldsymbol{k}$), and value ($\boldsymbol{v}$) tokens for Transformer layers. The query, key, and value tokens corresponding to the template and search input are $\boldsymbol{q}_Z, \boldsymbol{k}_Z, \boldsymbol{v}_Z$ and $\boldsymbol{q}_X, \boldsymbol{k}_X, \boldsymbol{v}_X$, respectively. In each transformer layer, the search query q_X queries the target feature from the concatenated template and search features, i.e.

$$\boldsymbol{k} = Concat(\boldsymbol{k}_Z, \boldsymbol{k}_X), \boldsymbol{v} = Concat(\boldsymbol{v}_Z, \boldsymbol{v}_X) \tag{4}$$

$$Attention(\boldsymbol{q}_X, \boldsymbol{k}, \boldsymbol{v}) = Softmax(\frac{\boldsymbol{q}_X \times \boldsymbol{k}^T}{\sqrt{d}})\boldsymbol{v} \tag{5}$$

After n layers of the above multi-head attention operations, the tracking networks complete feature extraction and interaction between the template and search. The box prediction head φ is appended to regress the bounding box. For more preliminary details, please refer to the Mixformer (Cui et al. 2022).

3.2 Missingness-aware prompting for RGBT tracking

In this paper, we focus on the heavy modality-missing problem in real-world RGBT tracking where random consecutive frame missing occurs for infrared (TIR) and visible (RGB) modalities. We propose a lightweight yet effective missingness-aware prompting (MAP) framework to alleviate such missing issues. The MAP efficiently compensates essential missing information for RGBT tracking via two-stage parameter-efficient prompting instead of recovering all details in the missing frames. We will introduce the details of the proposed MAP framework in the following.

3.2.1 Overview pipeline

The overview pipeline of the proposed missingness-aware prompting framework for RGBT tracking is shown in Fig. 2a. It consists of a patch embedding module, a series of missingness-aware Prompting (MAP) blocks, and a box head for the bounding box regression. If one or several consecutive frame-missing occurs in any modality (e.g. the RGB modality in Fig. 2a), we first pad all missing frames with the nearest non-missing frame in the same modality. Then, the padded multi-modal inputs will be fed into the proposed missingness-aware Prompting (MAP) blocks for feature interaction and missing compensation. Specifically,

the Prototypical missingness-aware prompts (pMAP) consist of learnable tokens that approximate the modality-specific but instance-agnostic missing information. The contextual missingness-aware prompts (cMAP) generated by a Context Network compensate for instance-specific missing information.

The MAP is built upon a series of frozen Multi-modal Transformer blocks and also incorporates the proposed pMAP and cMAP tokens into model tuning, thus generating missing-compensated multi-modal representation for subsequent layers. We incorporate the prompts into each transformer layer via the formula in the following:

$$\begin{aligned} &Attention(\boldsymbol{q}_{\boldsymbol{X}}^{t-n}, [\boldsymbol{k}^{t-n}, \boldsymbol{p}_{MAP_k}], [\boldsymbol{v}^{t-n}, \boldsymbol{p}_{MAP_v}]) \\ &= Softmax(\frac{\boldsymbol{q}_{\boldsymbol{X}}^{t-n}[\boldsymbol{k}^{t-n}, \boldsymbol{p}_{MAP_k}]^T}{\sqrt{d}})[\boldsymbol{v}^{t-n}, \boldsymbol{p}_{MAP_v}] \end{aligned} \tag{6}$$

where [ ] denotes concatenation operation, d is the dimension number of tokens, $\boldsymbol{p}_{MAP_k}$ and $\boldsymbol{p}_{MAP_v}$ denote the prompt embeddings to concatenate with the image key and value, respectively. After N iterations (layers) of interacting and compensating, the modality-missing issue will be progressively and adaptively compensated by learnable tokens in prompts. Eventually, the box head of RGBT trackers exploits the final compensated representation to predict the bounding box of the target. The overall MAP framework is end-to-end tuned with the common RGBT tracking losses (Luo et al. 2023; Hui et al. 2023),

$$\mathcal{L} = \mathcal{L}_{Tracker}(\boldsymbol{X}_{RGB}, \boldsymbol{X}_{TIR}, \theta_p) \tag{7}$$

where θ_p denotes all learnable parameters of the proposed MAP framework, and $\mathcal{L}_{Tracker}$ consisting of the L_1 loss and generalized IoU loss (Rezatofighi et al. 2019).

3.2.2 Prototypical Missingness-aware Prompts (pMAP)

The pMAP is designed to approximate the modality-specific but instance-agnostic missing information and alleviate the instability caused by random modality-missing as well. The key motivation for the pMAP is to progressively learn a continuous prompting vector and then adaptively compensate for the missing information with the instance-agnostic modality prototypes (prompting vector). As shown in Fig. 3b, we insert in total three types of prototypical prompts in each transformer layer in the form of Multi-head Attention, including the RGB-missing prompts compensating for RGB modality, the TIR-missing prompts compensating for TIR modality, and the complete prompts adapting for no-missing inputs. Given the notations for the Query, Key, and Value of the Multi-head Attention in i-th layer as follows:

$$\boldsymbol{q}^i = \boldsymbol{H}^i \boldsymbol{W}_{\boldsymbol{Q}}^i; \boldsymbol{k}^i = \boldsymbol{H}^i \boldsymbol{W}_{\boldsymbol{K}}^i; \boldsymbol{v}^i = \boldsymbol{H}^i \boldsymbol{W}_{\boldsymbol{V}}^i \tag{8}$$

where $\boldsymbol{W}_{\boldsymbol{Q}}^i, \boldsymbol{W}_{\boldsymbol{K}}^i, \boldsymbol{W}_{\boldsymbol{V}}^i \in \boldsymbol{R}^{d \times d}$ are the projection weights of the Multi-head Attention in the i-th layer. We split the learnable prompt tokens $\boldsymbol{p}_{pMAP}^i$ into $\boldsymbol{p}_{pMAP_k}^i$ and $\boldsymbol{p}_{pMAP_v}^i$, then we respectively prefix them to the key and value of each attention module to implement the proposed pMAP, i.e.,

$$f_{attn}^i(\boldsymbol{p}_{pMAP}^i, \boldsymbol{h}^i) = Attention^i(\boldsymbol{p}_{pMAP}^i, \boldsymbol{h}^i) \tag{9}$$

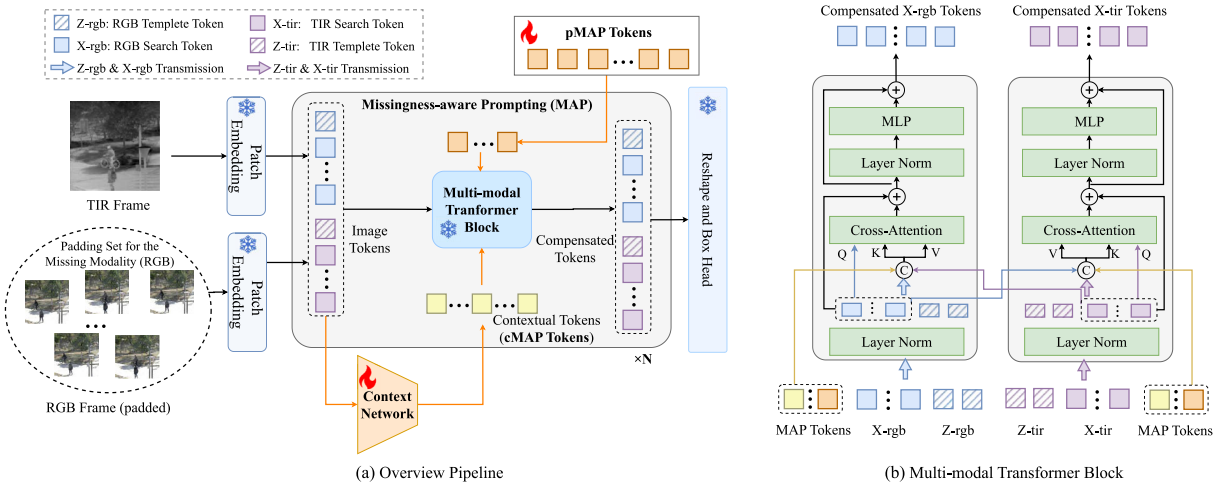

Fig. 2 (a) The overview pipeline of the proposed missingness-aware Prompting (MAP) framework for RGBT tracking. (b) The inner details of the Multi-modal Transformer Block in a typical RGBT tracker backbone

$$Attention(\boldsymbol{q}^i, [\boldsymbol{k}^i, \boldsymbol{p}^i_{pMAP_k}], [\boldsymbol{v}^i, \boldsymbol{p}^i_{pMAP_v}]) = Softmax(\frac{\boldsymbol{q}^i[\boldsymbol{k}^i, \boldsymbol{p}^i_{pMAP_k}]^T}{\sqrt{d}})[\boldsymbol{v}^i, \boldsymbol{p}^i_{pMAP_v}] \quad (10)$$

where f^i_{attn} represents the attention operation of the i-th layer, $\boldsymbol{h}_i$ represents the image features of the i-th layer. As shown in Fig. 3c, since all prompts are prefixed only for the Key and Value but not for the Query, the dimensionality for output tokens remains the same as that of the input sequence. Therefore, no additional post-processing operation is required and we could directly feed the compensated representation from the last transformer layer to the box head for bounding box prediction.

3.2.3 Contextual Missingness-aware Prompts (cMAP)

Although the above pMAP moves an effective step to alleviate the performance degradation caused by random modality-missing, solely relying on prototypical modality information cannot compensate for instance-specific missing information. Therefore, we propose contextual missingness-aware prompts (cMAP) to generate contextual prompt tokens to further compensate for instance-specific information. As shown in Fig. 3a, the cMAP tokens are generated from current and historical video frames which contain abundant contextual information about the missing instances.

Specifically, the contextual prompting tokens are generated by a lightweight Context Network (Fig. 3a). Firstly, We reshape the feature from 1D token sequences back into 2D feature maps. Then, a missingness-aware convolutional layer that chooses learnable convolutional branches for different scenarios (including the RGB missing, TIR Missing, and Complete), where each convolutional branch with kernels of 3 × 3 and stride of 2. This is followed by a two-layer bottleneck structure (Linear-LayerNorm-ReLU-Linear) that generates the final contextual prompting tokens. Eventually, the generated instance-specific cMAP tokens and the above instance-agnostic pMAP tokens are concatenated to collaboratively compensate for missing information that is essential for tracking. Formally, we denote the Context Network comprises of the missingness-aware convolutional layer and two fully connected layers as $h_\theta(\cdot)$, in which the θ denotes its parameter set. The contextual tokens $\boldsymbol{p}_m$ for each transformer layer are obtained via a recursive manner $\boldsymbol{t}_{m+1} = Concat(\boldsymbol{t}_m, \boldsymbol{p}_m)$, $\boldsymbol{p}_m = h_\theta(\boldsymbol{t}_m)$, $\boldsymbol{t}_m \in \{\boldsymbol{k}_m, \boldsymbol{v}_m\}$, $m \in \{1, 2, 3, ..., n\}$, and $\boldsymbol{k}_m, \boldsymbol{v}_m$ are computed via (8) with the representations of the template and search regions for each modality in the m-th layer. The implementation details of prompts incorporation are shown in Figs. 3b and c.

3.2.4 Multi-modal transformer block

As shown in Fig. 2b, the Multi-modal Transformer block is the basic unit for feature interaction and missingness-aware prompting. The interaction between template and search features has been introduced in (4) and (5). The modality interaction in RGBT tracking could be achieved similarly, i.e. further adding the key and value from the other modality to the concatenation operation in (4). Moreover, since this paper tackles the more complicated RGBT tracking scenario that involves modality-missing, we also need to incorporate two types of MAP tokens into multi-modal attention operations. Finally, taking the RGB branch as an example, the interaction and compensation are implemented by (11) and (12) as follows:

$$\begin{aligned} \boldsymbol{k} &= Concat(\boldsymbol{k}_{Z_{rgb}}, \boldsymbol{k}_{X_{rgb}}, \boldsymbol{k}_{X_{tir}}, \boldsymbol{p}_{MAP_k}), \\ \boldsymbol{v} &= Concat(\boldsymbol{v}_{Z_{rgb}}, \boldsymbol{v}_{X_{rgb}}, \boldsymbol{v}_{X_{tir}}, \boldsymbol{p}_{MAP_v}) \end{aligned} \quad (11)$$

$$Attention(\boldsymbol{q}_{X_{rgb}}, \boldsymbol{k}, \boldsymbol{v}) = Softmax(\frac{\boldsymbol{q}_{X_{rgb}} \times \boldsymbol{k}^T}{\sqrt{d}})\boldsymbol{v} \quad (12)$$

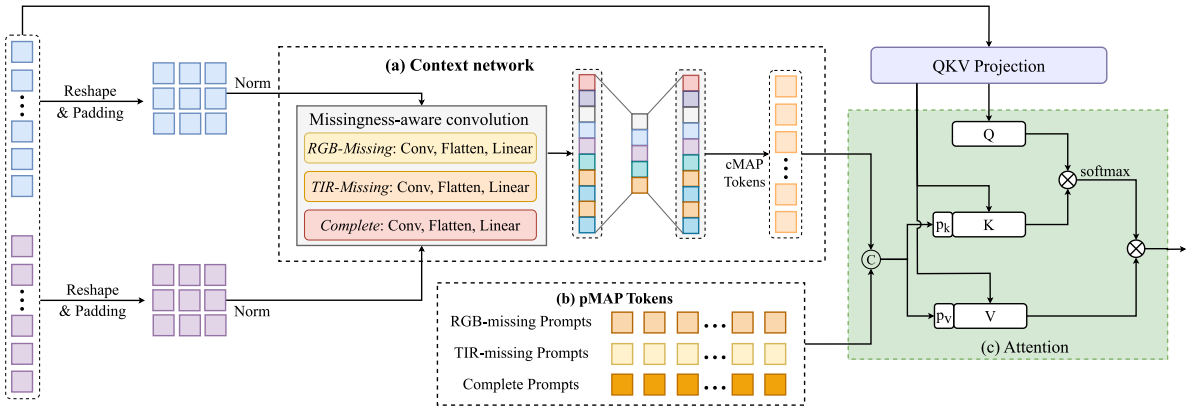

Fig. 3 The implementation details of the MAP in each transformer layer. (a) The implementation details of contextual missingness-aware prompts (cMAP). (b) The implementation details of prototypical missingness-aware prompts (pMAP). (c) The implementation details of prompts incorporation

Algorithm 1 Overall pipeline of the MAP.

Input: Images ($\mathcal{Z}_{rgb}$, $\mathcal{X}_{rgb}$, $\mathcal{Z}_{tir}$, $\mathcal{X}_{tir}$), Number of blocks N
Output: Predicted box (Pred_box)
▷ E denotes Embedding operation
▷ $\bigoplus$ denotes Concatenation operation
▷ The $flag \in \{rgb, tir, cpl\}$ denote missing rgb modality, tir modality and complete modality, respectively
▷ The function MAP is the missing compensation method proposed in this paper
function $p_{MAPk}, p_{MAPv} \leftarrow \text{MAP}(X, i, flag)$
  $p_{pMAPk}, p_{pMAPv} \leftarrow \text{pMAP}(i, flag)$
  $p_{cMAPk}, p_{cMAPv} \leftarrow \text{cMAP}(i, X)$
  $p_{MAPk} \leftarrow p_{pMAPk} \bigoplus p_{cMAPk}$
  $p_{MAPv} \leftarrow p_{pMAPv} \bigoplus p_{cMAPv}$
end function
$Z^1_{rgb}, X^1_{rgb}, Z^1_{tir}, X^1_{tir} = \text{E}(\mathcal{Z}_{rgb}, \mathcal{X}_{rgb}, \mathcal{Z}_{tir}, \mathcal{X}_{tir})$
for $i \leftarrow 1$ to N **do**
  Calculate q, k, v of $Z^i_{rgb}, X^i_{rgb}, Z^i_{tir}, X^i_{tir}$ according to (8)
  ▷ The process of RGB branch
  if RGB is missing **then**
    $p_{MAPk}, p_{MAPv} \leftarrow \text{MAP}(X^i_{rgb}, i, rgb)$
  else
    $p_{MAPk}, p_{MAPv} \leftarrow \text{MAP}(X^i_{rgb}, i, cpl)$
  end if
  Calculate k, v according to (11)
  $X^{i+1}_{rgb} \leftarrow Attn(q_{X^i_{rgb}}, k, v)$
  ▷ The process of TIR branch
  if TIR is missing **then**
    $p_{MAPk}, p_{MAPv} \leftarrow \text{MAP}(X^i_{tir}, i, tir)$
  else
    $p_{MAPk}, p_{MAPv} \leftarrow \text{MAP}(X^i_{tir}, i, cpl)$
  end if
  Calculate k, v according to (11)
  $X^{i+1}_{tir} \leftarrow Attn(q_{X^i_{tir}}, k, v)$
end for
Pred_box $\leftarrow$ Tracking_Head(X_{rgb}, X_{tir})

where the subscript Z and X represent the template and search, while the subscript MAP represents the token concatenation of the pMAP and cMAP.

Eventually, taking the Mixformer (Cui et al. 2022) model as an example, we illustrate implementation of the proposed MAP in Algorithm 1. It contains the the generation and integration process of pMAP and cMAP prompt vectors to collaboratively compensate for modality-missing information, and the implementation details of the Multi-modal Transformer block.

4 Experiments

4.1 Datasets

RGBT210 The original RGBT210 dataset (Li et al. 2017) is an RGBT tracking dataset that comprises 210 video sequences with varying lengths. It contains about 210K frames in total and the lengthiest video sequence extends up to 8K frames. The visible light and infrared video pairs are highly registered and have detailed attribute annotations.

RGBT234 The original RGBT234 dataset (Li et al. 2019) is an extension of above RGBT210 with more accurate annotations and richer scenes. It contains 234 highly aligned video sequences consisting of 234k images in total and the lengthiest video sequence also extends up to 8K frames.

LasHeR The original LasHeR dataset (Li et al. 2021) is a large-scale RGBT dataset that includes 979 training sequences and 245 testing sequences, offering diverse scene variations and dense bounding box annotations. The LasHeR encompasses over 730K frames of RGBT pairs and each pair is temporally and spatially aligned.

Missing Variants (Ours): To evaluate the tracking performance of our missingness-aware prompting framework, we simulate corresponding missing variants for the RGBT210, RGBT234, and LasHeR datasets, respectively. Specifically, we define the missing rate as the proportion of missing duration w.r.t the entire sequence duration. Unless otherwise specified, the overall missing rate is set as 60% consisting of 30% evenly for the RGB and TIR modalities, respectively. In detail, we focus on the challenging missing scenario where random consecutive frame missing occurs within a period for arbitrary modality.

4.2 Evaluation metrics

The Precision Rate (PR) and Success Rate (SR) are the most common two metrics for assessing the performance of RGBT trackers (Li et al. 2016, 2017, 2019). Specifically, the Precision Rate (PR) is the percentage of video frames whose bounding box center distance between the prediction and the groundtruth is less than a predefined distance threshold. For fair comparisons with existing RGBT trackers, we set the distance threshold as 20 pixels following the protocol in Li et al. (2017, 2019, 2021) for the RGBT210 and RGBT234 datasets. Since the PR metric is sensitive to the image resolution and bounding box size. A Normalized Precision Rate (NPR) is also introduced in Muller et al. (2018) for the LasHeR dataset to enhance the evaluation robustness, thus we also report both PR and NPR on the LasHeR dataset. In addition, the Success Rate (SR) is defined as the percentage of video frames in which the bounding box overlap between the prediction and the groundtruth exceeds an overlapping threshold. When varying the overlapping threshold, a success rate curve can be obtained. The final SR score is defined as the area under this success rate curve following Li et al. (2016, 2017, 2019, 2021).

4.3 Implementation details

The image sizes of search images and templates are set as 320 × 320 pixels and 128 × 128 pixels, respectively. Unless otherwise specified, the missing rate is empirically set as 60% for all dataset variants with missing modalities in our experiments. If one or several consecutive frame-missing occurs, we utilize the nearest non-missing frame in the same modality as a padding value for the missing input. The random modality-missing occurs only in search images. Regarding the tracking backbone, we utilize multiple representative off-the-shelf RGBT trackers based on the popular Vision Transformer to validate our missingness-aware Prompting framework, including both the RGBT extension of single-modal RGB tracking models (Mixformer (Cui et al. 2022) and OSTrack (Ye et al. 2022)) and native multi-modal tracking models (TBSI (Hui et al. 2023) and BAT (Cao et al. 2024)).

As for the training details, we use the Adam optimizer with a learning rate value of 0.0001. Following the widely-used protocols in RGBT tracking (Hui et al. 2023; Li et al. 2021; Luo et al. 2023), we utilize the training subset in the large-scale LasHeR dataset as training data and prompt our MAP framework for 35 epochs. Then, we directly evaluate the performances on RGB210, RGBT234, and the testing subset of the LasHeR dataset without any further tuning. For the single-modal RGB tracking backbones, we first extend them to multi-modal RGBT trackers by pre-training them on large-scale multi-modal RGBT dataset LasHeR training subset. We then freeze all tracker parameters and treat it as an off-the-shelf multi-modal RGBT tracker, and further parameter-efficiently prompt it with the proposed MAP framework. Additionally, we construct a padding set using the preceding N frames at each missing timestep. In each training iteration, we randomly select one frame from this padding set and jointly train it with the frames corresponding to non-missing modality, thus simulating and augmenting the real-world modality missing scenarios.

4.4 Experimental results

To validate the effectiveness of the proposed missingness-aware prompting framework (MAP), we conduct comparison experiments on three large-scale datasets (including RGBT210 (Li et al. 2017), RGBT234 (Li et al. 2019), and LasHeR (Li et al. 2021)) and their missing variants for RGBT tracking. We report the Precision Rate (PR) and Success Rate (SR) for these three datasets in Table 1. We compare four representative state-of-the-art RGBT tracking methods under the missingness-free (Fig. 1a), missingness-unaware (Fig. 1b), and missingness-aware (Fig. 1c) paradigms, respectively. First, we directly apply four representative state-of-the-art missingness-free RGBT tracking methods to the missingness-unaware scenario where they undergo severe performance decline (see Table 1, indicating the necessity for exploring missingness-aware approaches. Then, we plug the proposed MAP prompting framework into these RGBT tracking approaches for missing compensation. The results in Table 1 show that our MAP framework significantly improves the performance of RGBT trackers in situations with random missing modalities. Specifically, our MAP boosts the Mixformer (Cui et al. 2022) with large performance gains of 4.8 (PR) and 2.6 (SP) on the RGBT210 dataset. Our MAP brings significant performance gains of 7.8, 9.1, and 3.9 for the TBSI (Hui et al. 2023) on the large-scale LasHeR dataset. Our MAP framework boosts the BAT (Cao et al. 2024) with large performance gains of 3.8 (PR) and 3.0 (SR) on the RGBT210 dataset.

5 Ablation studies

5.1 Contribution examination of the cMAP and pMAP

To analyze the effectiveness of every primary component in the missingness-aware framework (including pMAP and cMAP), extensive ablation experiments have been conducted on the RGBT234 (Li et al. 2019) and LasHeR (Li et al. 2021) datasets (Table 2).

The results from pMAP or cMAP prompting improve the missingness-unaware baseline method (Mixformer (Cui et al. 2022) *) across different datasets, which indicates the individual effectiveness of the proposed two categories of prompts. Specifically, the pMAP exploits prototypical information as compensation for missing modalities enhancing the overall robustness of RGBT trackers. In contrast, the cMAP compensates for missing modalities using instance-specific prompts adapting better to each missing instance. In Table 2, the performance of prototypical missingness-aware prompts (pMAP) surpasses that of contextual missingness-aware prompts (cMAP) on the RGBT234 dataset because the RGBT234 dataset contains relatively simple scenes that the pMAP could be compensated well. However, the cMAP

Table 2 Ablation experiments for assessing the individual contribution of the pMAP and cMAP components in our MAP framework on the RGBT234 and LasHeR datasets

Method	pMAP	cMAP	RGBT234		LasHeR	
			PR↑	SR↑	PR↑	SR↑
Baseline			79.7	57.0	56.0	44.8
1	✓		82.4	58.9	58.8	46.4
2		✓	81.3	58.7	59.4	47.0
3	✓	✓	**83.6**	**60.3**	**60.8**	**48.1**

significantly outperforms the pMAP on the LasHeR dataset since LasHeR contains more complicated scenes, thus the cMAP could provide richer instance-specific context information for miss compensation. Eventually, the full MAP framework that contains both the cMAP and pMAP prompts achieves the best for missingness-aware RGBT tracking by exploiting the advantages of these two prompts.

5.2 Impact of missing rates

To analyze the effectiveness and robustness of our MAP towards different missing rates, we conduct experiments on dataset variants of RGBT234 and LasHeR with various missing rates including 30%, 60%, and 90%, the results are reported in Fig. 4 and Table 3. We observe that the tracking performance of the missingness-unaware baselines Mixformer (Cui et al. 2022)* and TBSI (Hui et al. 2023) decline significantly with the increasing of missing rate, e.g., they drop over 10% regarding the PR and SR metrics on the LasHeR dataset under a missing rate of 90%. At the relatively low missing rate (e.g. 30%), our MAP effectively compensates for modality-missing and reaches a roughly comparable performance with the original missingness-free (complete) method. When the missing rate increased, the missing impact on RGBT tracking became more pronounced. The performance gain obtained from applying our missingness-aware promoting framework became more significant, which indicates the effectiveness and robustness of the proposed MAP framework towards different missing rates. For example, at the missing rate of 60%, our MAP improves 4.8 % on LasHeR datasets regarding PR metrics. While at a 90% missing rate, the performances improve by more than four points regarding 4.2 metrics on the RGBT234 dataset.

5.3 Comparison of different implementations of the cMAP

As shown in Fig. 3a, the contextual missingness-aware prompts (cMAP) are generated via a lightweight network consisting of missingness-aware convolution layers (Conv) and a series of fully-connected layers (FC). The Conv layers operate on the 2D image features for essential representation extraction and dimensionality reduction. The FC layers model the channel-wised global information, followed by a LayerNorm and ReLU for feature recalibration. The experimental results in Table 4 show that individual FC or Conv layers are not enough to generate satisfied missingness-aware contextual tokens, because missing-compensating requires both contextual 2D spatial information and 1D channel-wise global contextual information.

5.4 Comparison of different model tuning methods

In this section, we compare our prompt choice with alternative model tuning methods (including finetuning and a input-level prompting strategy Zhu et al. 2023) on the RGBT234 and LasHeR datasets. Specifically, the fully finetuning method tunes the tracking model on the modality-missing RGBT dataset without prompt (Table 5). The results reveal large performance degradation since the parameter corruption issue on modality-missing datasets, proving the necessity of our prompt learning strategy. In addition, our attention-level prompting operation inserts prompt vectors into the Key (K) and Value (V) in each attention block. To further validate the rationality of this design, we compare it with an input-level prompting variant (Zhu et al. 2023). Although the input-level prompting method also achieves reasonable compensation for missing information, its performance is inferior to our approach (attention-level prompting). Furthermore, the input-level method also requires 1.5 times learnable parameters than our method, which is less parameter-efficient.

5.5 Influence of the size of the padding set

As shown in Fig. 2, the padding frames (for missing modality) are randomly selected from a padding set during training. The padding set consists of N contextual frames from N nearest frame in front of the missing frame at the same modality

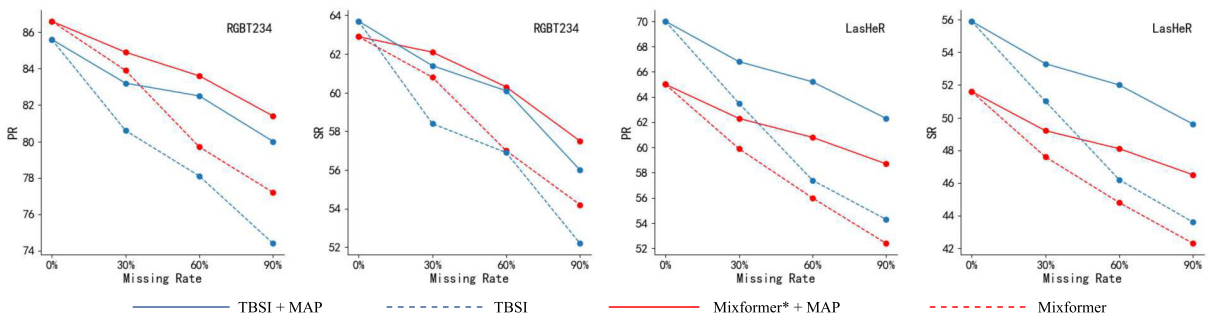

Fig. 4 Impact of different modality-missing rates on various trackers under the proposed MAP framework

Table 3 Effectiveness and robustness validation regarding missing rates on the RGBT234 and LasHeR dataset

Dataset	Method	Complete		Missing					
		0%		30%		60%		90%	
		PR	SR	PR	SR	PR	SR	PR	SR
RGBT234	Mixformer (Cui et al. 2022)* (baseline)	86.6	62.9	83.9	60.8	79.7	57.0	77.2	54.2
	Mixformer (Cui et al. 2022)*+MAP(ours)			$84.9_{(+1.0)}$	$62.1_{(+1.3)}$	$83.6_{(+3.9)}$	$60.3_{(+3.3)}$	$81.4_{(+4.2)}$	$57.5_{(+3.3)}$
	TBSI (Hui et al. 2023) (baseline)	85.6	63.7	80.6	58.4	78.1	56.9	74.4	52.2
	TBSI (Hui et al. 2023)+MAP(ours)			$83.2_{(+2.6)}$	$61.4_{(+3.0)}$	$82.5_{(+4.4)}$	$60.1_{(+3.2)}$	$80.0_{(+5.6)}$	$56.0_{(+3.8)}$
LasHeR	Mixformer (Cui et al. 2022)* (baseline)	65.0	51.6	59.9	47.6	56.0	44.8	52.4	42.3
	Mixformer (Cui et al. 2022)*+MAP(ours)			$62.3_{(+2.4)}$	$49.2_{(+1.6)}$	$60.8_{(+4.8)}$	$48.1_{(+3.3)}$	$58.7_{(+6.3)}$	$46.5_{(+4.2)}$
	TBSI (Hui et al. 2023) (baseline)	70.0	55.9	63.5	51.0	57.4	46.2	54.3	43.6
	TBSI (Hui et al. 2023)+MAP(ours)			$66.8_{(+3.3)}$	$53.3_{(+2.3)}$	$65.2_{(+7.8)}$	$52.0_{(+5.8)}$	$62.3_{(+8.0)}$	$49.6_{(+6.0)}$

Table 4 Implementation rationality of the cMAP

Method	cMAP	RGBT234		LasHeR	
		PR↑	SR↑	PR↑	SR↑
Baseline+pMAP	–	82.4	58.9	58.8	46.4
1	FC	83.3	59.8	60.1	47.4
2	Conv	82.0	58.9	59.5	47.0
3	FC+Conv	**83.6**	**60.3**	**60.8**	**48.1**

during the training process. Since the test subsets of missing datasets are set to randomly and continuously miss 10 to 15 frames meeting a certain level of the missing rate, we vary the size number N in a range of [10,15] on the LasHeR (Li et al. 2021) dataset to examine the choosing padding set size, the results are shown in Fig. 5. We observe that the performance is poor with too large or too small choices of padding set, and achieving the best at a middle padding set of $N = 12$. Therefore, we empirically set the padding set size N as 12 except specially clarified.

In addition, we conduct experimental comparison of different padding methods in our MAP, including nearest padding and random padding strategies (Table 6). For the nearest padding strategy, we utilize the N most recent frames for compensation. For the random padding strategy, we respectively test varying historical ranges of 50, 100, and 200 frames, in which the padding images are randomly selected from historical ranges. The results demonstrate that our nearest padding strategy achieving optimal performance. Moreover, relative larger historical ranges will degrade tracking performance since the context information is less identical with original missing frames. These results effectively validate the soundness of the 12 frames of nearest padding strategy.

5.6 Analysis of computational efficiency

In this section, We analysis the computational efficiency in the metrics of the number of model parameters (Params) and Multiply–Accumulate Operations (MACs). As shown in Table 7, our MAP effectively compensates for modality missing in RGBT tracking with slightly parameter increasing, i.e., only accounting for approximately 10% of total parameters. Notably, as the parameter scale of non-missing baseline models expands, the parameter increasing proportion substantially decreases which is more parameter-efficient. Furthermore, our MAP only introduces a modest computational overhead, typically within 15% of the original non-missing baselines, thus maintaining efficiency of original trackers. This high efficiency stems from our innovative attention-level prompt insertion mechanism, where prompt tokens are inserted into the key and value branches for attention computation, which effectively eliminates redundant operations. Moreover, this innovative insertion strategy will not increase feature size during the attention process, enabling direct bounding box prediction without additional post-processing steps. Additionally, we conduct inference speed comparisons on a same server for Mixformer (Cui et al. 2022)*, OSTrack (Ye et al. 2022)* and TBSI (Hui et al. 2023). As shown in Table 7, the baseline method Mixformer*, OSTrack*, and TBSI respectively achieves an inference speed of 8.8 FPS, 29.8 FPS and 23.2 FPS. In contrast, the enhanced implementation that incorporating our MAP respectively achieves 7.2 FPS, 23.3 FPS and 18.5 FPS, which effectively compensates for modality missing in RGBT tracking with limited speed reduction (i.e., less than 20% compared to corresponding baselines). The results effectively demonstrate the effectiveness and efficiency of the proposed MAP framework in modality-missing RGBT tracking scenarios.

Table 5 Comparison of different model tuning methods

Method	Params	RGBT234		LasHeR	
		PR↑	SR↑	PR↑	SR↑
Baseline	–	79.7	57.0	56.0	44.8
Finetuning	45.99M	78.7	54.8	57.7	46.1
Input-level Prompting	5.27M	82.6	57.4	60.4	47.5
Ours	3.51M	**83.6**	**60.3**	**60.8**	**48.1**

Table 6 Comparison of different padding methods

Padding Method	RGBT234		LasHeR	
	PR↑	SR↑	PR↑	SR↑
Missingness-unaware	79.7	57.0	56.0	44.8
Random padding (50 frames)	82.9	58.2	58.8	46.5
Random padding (100 frames)	82.8	58.1	58.6	46.2
Random padding (200 frames)	81.4	56.8	57.1	45.3
Nearest frames (ours)	**83.6**	**60.3**	**60.8**	**48.1**

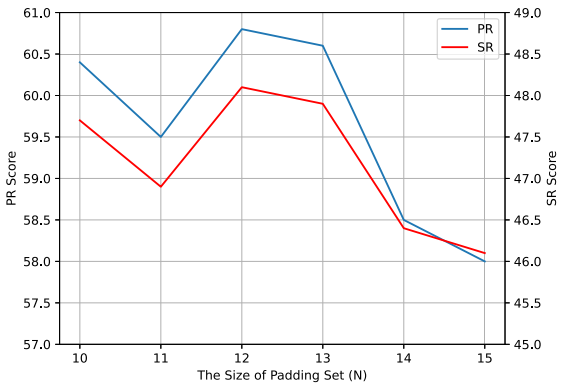

Fig. 5 The influence of the size of the padding set N towards the performance of the proposed MAP framework regarding PR and SR metrics on the LasHeR dataset

Table 7 Comparisons of the computational complexity, the number of model parameters and the inference speed

Method	MACs (G)	Params (M)	FPS
Mixformer (Cui et al. 2022)*	40.51	45.99	8.8
Mixformer (Cui et al. 2022)*+MAP	46.23	49.50	7.2
OSTrack (Ye et al. 2022)*	57.80	97.83	29.8
OSTrack (Ye et al. 2022)*+MAP	66.95	108.93	23.3
TBSI (Hui et al. 2023)	82.52	202.37	23.2
TBSI (Hui et al. 2023)+MAP	91.62	213.48	18.5

5.7 Visualization analysis

To conduct intuitive comparisons, we further visualize the RGBT tracking results from three representative approaches (including the Mixformer (Cui et al. 2022) *, OSTrack (Ye et al. 2022) *, and TBSI (Hui et al. 2023)) under all tracking paradigms, including the missingness-free paradigm (Fig. 1a), missingness-unaware paradigm (Fig. 1b) and missingness-aware paradigm (Fig. 1c). The results of 4 tracking sequences with random missing modalities are shown in Fig. 6, which contains 2 sequences with RGB-missing and 2 sequences with TIR-missing, respectively. The blue boxes represent the tracking results from the original RGBT trackers (missingness-free paradigm), the yellow boxes represent the tracking results from corresponding missingness-unaware RGBT trackers (missingness-unaware paradigm), and the red boxes represent the tracking results corresponding to our missingness-aware prompting (MAP) aided RGBT trackers (missingness-aware paradigm). The results in Fig. 6 indicate that classical RGBT trackers undergo heavy performance degradation when applied to real-world scenarios having inevitable modality-missing. Fortunately, plugging the proposed concise missingness-aware prompting (MAP) framework into classical missingness-unaware RGBT trackers could significantly alleviate the robustness degradation and achieve satisfactory tracking results in real-world scenarios with random missing modalities.

5.8 Discussion

Our MAP framework compensates for random modality missingness with the prototypical missingness-aware prompts (pMAP) and the contextual missingness-aware prompts (cMAP). It is a general prompting framework and

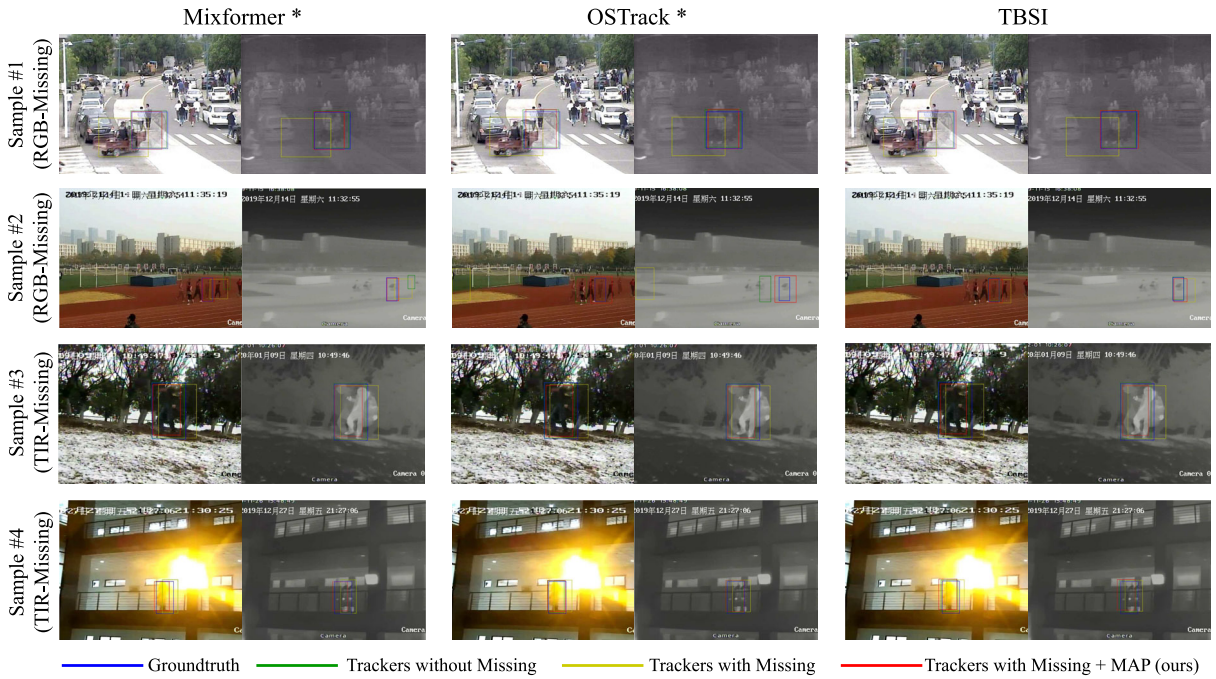

Fig. 6 Visualization comparisons with three representative approaches under three RGBT tracking paradigms (including the Mixfomer (Cui et al. 2022) *, OSTrack (Ye et al. 2022) *, and TBSI (Hui et al. 2023)) on the LasHeR dataset. Each row is a tracking sequence with random missing in the RGB or TIR modality

Table 8 Validation of modality generalization on the RGBD tracking task on the DepthTrack dataset

Paradigm	Method	F_1 ↑	R↑	P↑
Missingness-free	Mixformer (Cui et al. 2022)*	58.5	58.1	59.0
	OSTrack (Ye et al. 2022)*	57.7	57.1	58.3
Missingness-unaware	Mixformer (Cui et al. 2022)*	51.2	50.3	52.1
	OSTrack (Ye et al. 2022)*	50.8	49.7	52.0
Missingness-aware (ours)	Mixformer (Cui et al. 2022)*+MAP	54.2	53.5	55.0
	OSTrack (Ye et al. 2022)*+MAP	53.5	52.6	54.5

don't specify to modality type and tracking task, so it is principally applicable to other visual multimodal modality-missing scenarios, such as the mentioned RGBD (visible and depth images), and RGBE (visible and event images). For example, we extend the proposed MAP framework from RGBT tracking to RGBD tracking tasks on the DepthTrack (Yan et al. 2021) dataset. The results measured by F_1-score, Recall (R) and Precision (P) in Table 8 clearly indicates that our MAP effectively compensates for modality-missing in RGBD tracking scenarios, further demonstrating the effectiveness and generalization for tackling modality-missing under various modalities. However, when applying it to image-text multimodal scenarios, the MAP framework requires additional modification since the text modality cannot be aligned with video modality at frame level in the current MAP, so we leave it for future research.

Besides, we also include a failure case analysis of our MAP framework (Mixformer (Cui et al. 2022) *+MAP) in Fig. 7 to discuss the limitation of our MAP framework. The

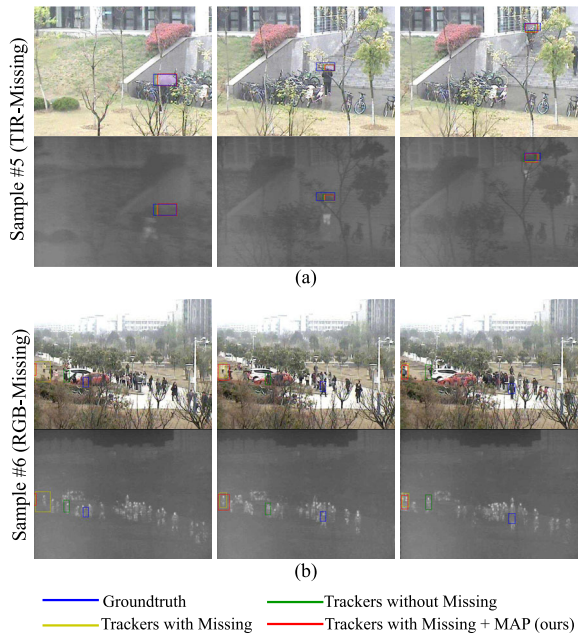

Fig. 7 Failure cases visualization on the RGBT234 dataset. Each row is a tracking sequence with random modality-missing

failure cases of our MAP method primarily stem from two aspects. On one hand, when there is a significant quality gap between the two video modalities and the lower-quality modality is missing (Fig. 7a), our MAP has only limited compensation effect since the missing modality itself has limited impact on the baseline method. On the other hand, when the baseline method itself fails to track the target in some challenging scenarios (Fig. 7b), incorporating our MAP framework still struggles to track the target effectively.

6 Conclusion

In this work, we propose a concise and effective missingness-aware Prompting (MAP) framework for RGBT tracking to alleviate the heavy modality-missing issue in real-world scenarios. The MAP efficiently compensates essential missing information for RGBT tracking via two-stage parameter-efficient prompting instead of recovering all details in the missing frames. Prototypical missingness-aware prompts (pMAP) approximate the modality-specific but instance-agnostic missing information that reduces the information gap between padding values and modality prototypes. Contextual missingness-aware prompts (cMAP) further compensate for instance-specific missing information that reduces the information gap between modality prototypes and missing instances. Eventually, our MAP achieves state-of-the-art performance on three datasets demonstrating its effectiveness in tackling real-world RGBT tracking with random missing modalities.

Author Contributions *Guyue Hu*: Conceptualization, Formal analysis, Methodology, Writing - original draft. *Zhanghuan Wang*: Formal analysis, Methodology, Writing - original draft. *Chenglong Li*: Supervision, Methodology, Writing - review & editing. *Duzhi Yuan*: Software, Visualization. *Bin He*: Validation, Resources. *Jin Tang*: Supervision, Project administration.

Funding This work was supported in part by the National Natural Science Foundation of China (No. 62376004), the Anhui Provincial Natural Science Foundation (No. 2408085QF201), the Natural Science Foundation of Anhui Higher Education Institution (No. 2022AH040014), the Open Project of Anhui Provincial Key Laboratory of Security Artificial Intelligence (No. SAI2024003), and the Open Project of Anhui

Provincial Key Laboratory of Intelligent Detection and Diagnosis for Traffic Infrastructure (No. KY-2025-03).

Availability of data and materials All datasets used in this paper (including RGBT210, RGBT234, and LasHeR) are publicly available at https://github.com/mmic-lcl/Datasets-and-benchmark-code. The code will be made publicly available after acceptance.

Declarations

Competing Interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

References

Cao B, Guo J, Zhu P, Hu Q (2024) Bi-directional adapter for multimodal tracking. In: Proceedings of the AAAI conference on artificial intelligence, vol 38, pp 927–935

Chen S, Ge C, Tong Z, Wang J, Song Y, Wang J, Luo P (2022) Adaptformer: adapting vision transformers for scalable visual recognition. Adv Neural Inf Process Syst 35:16664–16678

Chen L, Sun L, Yang T, Fan L, Huang K, Xuanyuan Z (2017) Rgb-t slam: a flexible slam framework by combining appearance and thermal information. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp 5682–5687

Cui B, Hu G, Yu S (2021) Deepcollaboration: collaborative generative and discriminative models for class incremental learning. In: Proceedings of the AAAI conference on artificial intelligence, vol 35, pp 1175–1183

Cui Y, Jiang C, Wang L, Wu G (2022) Mixformer: end-to-end tracking with iterative mixed attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13608–13618

Dai X, Yuan X, Wei X (2021) Tirnet: object detection in thermal infrared images for autonomous driving. Appl Intell 51(3):1244–1261

Dordevic D, Bozic V, Thommes J, Coppola D, Singh SP (2024) Rethinking attention: exploring shallow feed-forward neural networks as an alternative to attention layers in transformers. In: Proceedings of the AAAI conference on artificial intelligence, vol 38, pp 23477–23479

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S (2021) An image is worth 16x16 words: transformers for image recognition at scale. In: International conference on learning representations

Feng M, Su J (2024) Rgbt image fusion tracking via sparse trifurcate transformer aggregation network. IEEE Trans Instrum Meas 73:1–10

Feng M, Su J (2024) Rgbt tracking: a comprehensive review. Inf Fusion 110:102492

Gao T, Xu C-Z, Zhang L, Kong H (2024) Gsb: group superposition binarization for vision transformer with limited training samples. Neural Netw 172:106133

Gao Y, Li C, Zhu Y, Tang J, He T, Wang F (2019) Deep adaptive fusion network for high performance rgbt tracking. In: Proceedings of the IEEE/CVF international conference on computer vision workshops

Hu G, He B, Zhang H (2023) Compositional prompting video-language models to understand procedure in instructional videos. Mach Intell Res 20(2):249–262

Hui T, Xun Z, Peng F, Huang J, Wei X, Wei X, Dai J, Han J, 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, pp 13630–13639

Jia M, Tang L, Chen B-C, Cardie C, Belongie S, Hariharan B, Lim S-N (2022) Visual prompt tuning. In: European conference on computer vision. Springer, pp 709–727

Karimi Mahabadi R, Henderson J, Ruder S (2021) Compacter: efficient low-rank hypercomplex adapter layers. Adv Neural Inf Process Syst 34:1022–1035

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

Li C, Liang X, Lu Y, Zhao N, Tang J (2019) Rgb-t object tracking: benchmark and baseline. Pattern Recogn 96:106977

Li C, Xue W, Jia Y, Qu Z, Luo B, Tang J, Sun D (2021) Lasher: a large-scale high-diversity benchmark for rgbt tracking. IEEE Trans Image Process 31:392–404

Lialin V, Deshpande V, Rumshisky A (2023) Scaling down to scale up: a guide to parameter-efficient fine-tuning. arXiv:2303.15647

Lian Z, Chen L, Sun L, Liu B, Tao J (2023) Gcnet: graph completion network for incomplete multimodal learning in conversation. IEEE Trans Pattern Anal Mach Intell

Li C, Liu L, Lu A, Ji Q, Tang J (2020) Challenge-aware rgbt tracking. In: European conference on computer vision. Springer, pp 222–237

Liu M, Gao J, Hu G, Hao G, Jiang T, Zhang C, Yu S (2022) Monkeytrail: a scalable video-based method for tracking macaque movement trajectory in daily living cages. Zool Res 43(3):343

Liu T, Hu Y, Gao J, Wang J, Sun Y, Yin B (2024) Multi-modal long document classification based on hierarchical prompt and multi-modal transformer. Neural Netw 176:106322

Li C, Zhao N, Lu Y, Zhu C, Tang J (2017) Weighted sparse representation regularized graph learning for rgb-t object tracking. In: Proceedings of the 25th ACM international conference on multimedia, pp 1856–1864

Long Li C, Lu A, Hua Zheng A, Tu Z, Tang J (2019) Multi-adapter rgbt tracking. In: Proceedings of the IEEE/CVF international conference on computer vision workshops

Lu A, Li C, Yan Y, Tang J, Luo B (2021) Rgbt tracking via multi-adapter network with hierarchical divergence loss. IEEE Trans Image Process 30:5613–5625

Luo Y, Guo X, Feng H, Ao L (2023) Rgb-t tracking via multi-modal mutual prompt learning. arXiv:2308.16386

Ma M, Ren J, Zhao L, Testuggine D, Peng X (2022) Are multimodal transformers robust to missing modality? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 18177–18186

Ma M, Ren J, Zhao L, Tulyakov S, Wu C, Peng X (2021) Smil: multimodal learning with severely missing modality. In: Proceedings of

the AAAI conference on artificial intelligence, vol 35, pp 2302–2310

Marvasti-Zadeh SM, Cheng L, Ghanei-Yakhdan H, Kasaei S (2021) Deep learning for visual tracking: a comprehensive survey. IEEE Trans Intell Transp Syst 23(5):3943–3968

Muller M, Bibi A, Giancola S, Alsubaihi S, Ghanem B (2018) Trackingnet: a large-scale dataset and benchmark for object tracking in the wild. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 300–317

Peng J, Zhao H, Hu Z, Zhuang Y, Wang B (2023) Siamese infrared and visible light fusion network for rgb-t tracking. Int J Mach Learn Cybern 14(9):3281–3293

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR, pp 8748–8763

Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, 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, pp 658–666

Su Y, Wang X, Qin Y, Chan C-M, Lin Y, Wang H, Wen K, Liu Z, Li P, Li J et al (2022) On transferability of prompt tuning for natural language processing. Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 3949–3969

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30:5998–6008

Wang H, Liu X, Li Y, Sun M, Yuan D, Liu J (2024) Temporal adaptive rgbt tracking with modality prompt. In: Proceedings of the AAAI conference on artificial intelligence, vol 38, pp 5436–5444

Wei S, Luo C, Luo Y (2023) Mmanet: margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 20039–20049

Wu H, Xiao B, Codella N, Liu M, Dai X, Yuan L, Zhang L (2021) Cvt: introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 22–31

Xiao Y, Yang M, Li C, Liu L, Tang J (2022) Attribute-based progressive fusion network for rgbt tracking. In: Proceedings of the AAAI conference on artificial intelligence, vol 36, pp 2831–2838

Yang J, Li Z, Zheng F, Leonardis A, Song J (2022) Prompting for multimodal tracking. In: Proceedings of the 30th ACM international conference on multimedia, pp 3492–3500

Yan S, Yang J, Käpylä J, Zheng F, Leonardis A, Kämäräinen J-K (2021) Depthtrack: unveiling the power of rgbd tracking. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10725–10733

Ye B, Chang H, Ma B, Shan S, Chen X (2022) Joint feature learning and relation modeling for tracking: a one-stream framework. In: European conference on computer vision. Springer, pp 341–357

Zhang X, Ye P, Peng S, Liu J, Xiao G (2020) Dsiammft: an rgb-t fusion tracking method via dynamic siamese networks using multi-layer feature fusion. Signal Process: Image Commun 84:115756

Zhang C, Cui Y, Han Z, Zhou JT, Fu H, Hu Q (2020) Deep partial multi-view learning. IEEE Trans Pattern Anal Mach Intell 44(5):2402–2415

Zhang L, Danelljan M, Gonzalez-Garcia A, Van De Weijer J, 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

Zhao J, Li R, Jin Q (2021) Missing modality imagination network for emotion recognition with uncertain missing modalities. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing, pp 2608–2618

Zhou K, Yang J, Loy CC, Liu Z (2022) Learning to prompt for vision-language models. Int J Comput Vision 130(9):2337–2348

Zhou K, Yang J, Loy CC, Liu Z (2022) Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 16816–16825

Zhu Y, Li C, Tang J, Luo B (2020) Quality-aware feature aggregation network for robust rgbt tracking. IEEE Trans Intell Vehicles 6(1):121–130

Zhu J, Lai S, Chen X, Wang D, Lu H (2023) Visual prompt multimodal tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9516–9526

Zhu Y, Li C, Luo B, Tang J, Wang X (2019) Dense feature aggregation and pruning for rgbt tracking. In: Proceedings of the 27th ACM international conference on multimedia, pp 465–472

Authors and Affiliations

Guyue Hu[1,2,3,5,6] · Zhanghuan Wang[1,2,4,7] · Chenglong Li[1,2,3,6] · Duzhi Yuan[6] · Bin He[8] · Jin Tang[1,2,4,7]

✉ Chenglong Li
lcl1314@foxmail.com

Guyue Hu
guyue.hu@ahu.edu.cn

Zhanghuan Wang
zhwang12@stu.ahu.edu.cn

Duzhi Yuan
wa2314009@stu.ahu.edu.cn

Bin He
binhe.cas@foxmail.com

Jin Tang
tangjin@ahu.edu.cn

[1] State Key Laboratory of Opto-Electronic Information Acquisition and Protection Technology, Anhui University, Hefei 230601, China

[2] Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Anhui University, Hefei 230601, China

[3] Anhui Provincial Key Laboratory of Security Artificial Intelligence, Anhui University, Hefei 230601, China

[4] Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University, Hefei 230601, China

[5] Anhui Provincial Key Laboratory of Intelligent Detection and Diagnosis for Traffic Infrastructure, Anhui Jiaojian Traffic Development & Research Center Co., LTD, Hefei 230051, China

[6] School of Artificial Intelligence, Anhui University, Hefei 230601, China

[7] School of Computer Science and Technology, Anhui University, Hefei 230601, China

[8] The 15th Research Institute of China Electronics Technology Group Corporation, Beijing 100083, China