Generative-Based Fusion Mechanism for Multi-Modal Tracking

Zhangyong Tang[1], **Tianyang Xu**[1], **Xiaojun Wu**[1*], **Xue-Feng Zhu**[1], **Josef Kittler**[2]

[1]School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, Jiangsu, PR. China
[2]Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK
zhangyong_tang_jnu@163.com; {tianyang.xu;wu_xiaojun}@jiangnan.edu.cn; j.kittler@surrey.ac.uk

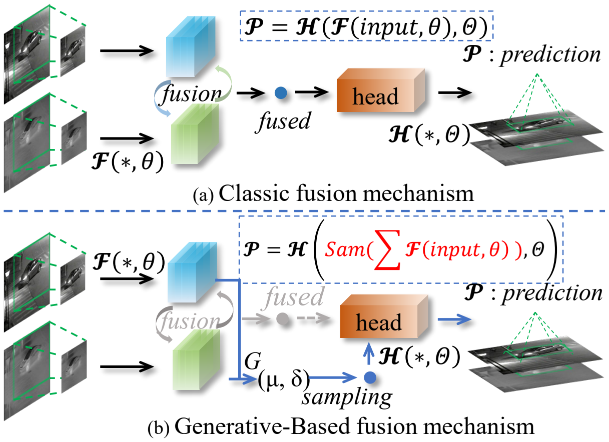
Abstract

Generative models (GMs) have received increasing research interest for their remarkable capacity to achieve comprehensive understanding. However, their potential application in the domain of multi-modal tracking has remained unexplored. In this context, we seek to uncover the potential of harnessing generative techniques to address the critical challenge, information fusion, in multi-modal tracking. In this paper, we delve into two prominent GM techniques, namely, Conditional Generative Adversarial Networks (CGANs) and Diffusion Models (DMs). Different from the standard fusion process where the features from each modality are directly fed into the fusion block, we combine these multi-modal features with random noise in the GM framework, effectively transforming the original training samples into harder instances. This design excels at extracting discriminative clues from the features, enhancing the ultimate tracking performance. Based on this, we conduct extensive experiments across two multi-modal tracking tasks, three baseline methods, and four challenging benchmarks. The experimental results demonstrate that the proposed generative-based fusion mechanism achieves state-of-the-art performance by setting new records on GTOT, LasHeR and RGBD1K. Code will be available at https://github.com/Zhangyong-Tang/GMMT.

Introduction

Due to the strict demand for the robustness of tracking systems in real-world applications, such as surveillance (Lu et al. 2023) and unmanned driving (Zhang et al. 2023a), visual object tracking with an auxiliary modality, named as multi-modal tracking, draws growing attention recently. For example, the thermal infrared (TIR) modality provides more stable scene perception in the nighttime (Tang et al. 2023), and the depth (D) modality provides 3-D perception against occlusions (Zhu et al. 2023b). In other words, the use of auxiliary modalities can complement the visible image in challenging scenarios.

Regarding this, a series of fusion strategies have been explored to aggregate the multi-modal information. These strategies fall into two main categories based on the output of their fusion block. The first category involves adaptive weighting strategies (Xu et al. 2021; Zhang et al. 2021b, 2022b), where the fusion block produces weights (scalars, vectors, or tensors) multiplied to features from each modality. In contrast, the second category focuses on embedded fusion blocks (Zhang et al. 2019; Zhang et al. 2021a; Zhu et al. 2021), which generate fused features using dedicated modules. While these methods differ in the way of producing fused results, their training processes are quite similar. They are trained offline using multi-modal datasets like RGBT234 (Li et al. 2019) and LasHeR (Li et al. 2022), and, from a discriminative perspective, their tracking performance hinges on how well they *match* the training data, rather than *understanding*.


$\boldsymbol{\mathcal{P}} = \boldsymbol{\mathcal{H}}(\boldsymbol{\mathcal{F}}(input, \theta), \Theta)$
$\boldsymbol{\mathcal{P}}$: prediction
fusion
fused
head
$\boldsymbol{\mathcal{H}}(*, \Theta)$
$\boldsymbol{\mathcal{F}}(*, \theta)$

(a) Classic fusion mechanism

$\boldsymbol{\mathcal{F}}(*, \theta)$
$\boldsymbol{\mathcal{P}} = \boldsymbol{\mathcal{H}}\left(Sam(\sum \boldsymbol{\mathcal{F}}(input, \theta)), \Theta \right)$
fused
fusion
head
$\boldsymbol{\mathcal{P}}$: prediction
G
(μ, δ)
sampling
$\boldsymbol{\mathcal{H}}(*, \Theta)$

(b) Generative-Based fusion mechanism

Figure 1: Comparison between the original fusion mechanism and our generative-based fusion mechanism. GM is denoted as G, and we use $\mathcal{F}$ and $\mathcal{H}$ to refer to the feature extractor and task head, respectively. The parameters associated with these two components are symbolised by θ and Θ.

On the contrary, GMs have achieved great success due to their superiority in comprehensive understanding. Accordingly, lots of downstream tasks have achieved promising performance, like image to image translation (Isola et al. 2017) and multi-modal image fusion (Rao, Xu, and Wu 2023). For example, in (Rao, Xu, and Wu 2023), a generator is employed to adaptively extract the salient clues among multiple inputs, and the discriminator further constrains the out-

[*]Corresponding author

put to be high quality by reinforcing the textures globally. However, extending its success on other multi-modal tasks to multi-modal tracking has not been sufficiently discussed yet.

Motivated by the aforementioned observations, the potential of applying GMs to address the multi-modal information fusion is discussed in this paper, with a novel generative-based fusion mechanism being proposed for multi-modal tracking (GMMT), shown in Fig. 1. In order to learn the external projection between the input and output, as well as the internal data distributions, GMs require a longer training time and a bigger size of training data. However, despite the emergence of several multi-modal datasets in recent years (Zhu et al. 2023b; Li et al. 2022; Zhang et al. 2022a), containing around 1000 videos captured across less than 1000 scenarios, there remains a significant diversity gap compared to some widely-used datasets for generation tasks, such as CelebA(Liu et al. 2015) used in face generation. Therefore, the multi-modal feature pair grouped with a random factor is formulated as the input of the GMMT to enlarge the data size and avoid over-fitting. Besides, to facilitate adaptive fusion of the certain image pair when testing, the original information from both modalities are retained as conditions. Based on the above considerations, CGAN (Mirza and Osindero 2014) and DM (Rombach et al. 2022) are implemented in this paper. The generative-based fusion mechanism endows the fusion model with a better awareness of the noise, and thus the fused features are more clean, as shown in Fig. 4(c) and (d), which boosts the tracker to be a more accurate one. To validate the effectiveness of the proposed fusion mechanism, it is implemented on several RGB-T baseline trackers. Consistent improvements can be obtained on all the evaluation metrics. Furthermore, extended experiments on the largest RGB-D benchmark (Zhu et al. 2023b) are also conducted to demonstrate the generalisation of GMMT. In conclusion, our contributions can be summarised as follows:

- We explore the potential of addressing the information fusion part of multi-modal tracking in a generative approach. To achieve this, a novel generative-based fusion mechanism is proposed, which boosts the fused features to be more discriminative.
- A general fusion mechanism is proposed, with its generalisation demonstrated on multiple baseline methods, benchmarks, and two multi-modal tracking tasks.
- Extensive experimental results demonstrate the proposed method as a state-of-the-art one on both RGB-T and RGB-D tracking tasks.

Related Work

RGB-T Trackers

Before the access of extensive RGB-T datasets, including GTOT (Li et al. 2016), RGBT210 (Li et al. 2017), and RGBT234 (Li et al. 2019), traditional RGB-T methods primarily relies on the sparse representation (Li et al. 2016) or handcrafted weighting strategies (Cvejic et al. 2007) to tackle the information fusion task. But these non-deep approaches suffer significant performance degradation in challenging scenarios. As a result, recent researches have been dominated by deep learning techniques. In recent studies, ranging from the simplest operation, concatenation (Zhang et al. 2019), to the more complicated transformer architecture (Hui et al. 2023; Zhu et al. 2023a), the researchers have tried various fusion strategies with multiple intentions, including learning modality importance (Zhang et al. 2021b; Tang, Xu, and Wu 2022), reducing the multi-modal redundancy (Li et al. 2018; Zhu et al. 2019), propagating the multi-modal patterns(Wang et al. 2020), learning the multi-modal prompts from the auxiliary modality (Zhu et al. 2023a), to name a few. With the increment in network complexity and the availability of larger training sets, tracking results have been gradually improved. This improvement is particularly noteworthy since the release of the LasHeR (Li et al. 2022), which promotes the development in a steep way.

Generative Models

While GMs have been one of the classical learning paradigms (Wang and Wong 2002), they initially receive less attention during the early years of deep era compared to the discriminative models. However, their significance is solidified after the introduction of GAN (Goodfellow et al. 2014). GAN first showcases its prowess in image synthesis and subsequently is found to be a success in a range of tasks, including multi-modal image fusion (Rao, Xu, and Wu 2023), text-to-video generation (Luo et al. 2023), and text-to-audio generation (Ruan et al. 2023). After that, although more variations of GMs, such as variational auto-encoder (Kingma and Welling 2014), and flow-based model (Prenger, Valle, and Catanzaro 2019), also draw increasing attention, the downstream applications are still mainly based on GAN.

Until the proposal of the denoising diffusion model (DM) (Ho, Jain, and Abbeel 2020), the interest in GAN falls gradually, as the DM shows superior performance across multiple domains, notably excelling in visual-language generation (Iqbal and Qureshi 2022). Compared to GAN, the DM exhibits a more stable training procedure, generating items in a more refined way.

Generative Models Meet Tracking

In the RGB tracking task, GMs are mainly introduced with two motivations, *i.e.*, generating more samples to improve the diversity (Wang et al. 2018; Han et al. 2020; Yin et al. 2020), and maintaining the most robust and long-lasting patterns (Song et al. 2018; Zhao et al. 2019; Han et al. 2020). The generator is used for the first purpose while the second category employs the discriminator to discard the less distinguishing patterns.

However, the application of generative models in the field of multi-modal tracking has received limited attention. As far as we know, only BD^2Track (Fan et al. 2023) conducts fusion in this way. However, there are two main issues in this method. Firstly, BD^2Track employs a diffusion model to acquire fused classification features while retaining regression features learned discriminatively. This configuration raises questions, since both classification and regression features within each modality encounter challenges and can potentially benefit from modality complementarity (Xiao et al.

2022). Secondly, it lacks in-depth analysis, leaving space for a comprehensive understanding.

In our method, we address the first issue by generating fused features before they are fed into the classification and regression heads. This approach eliminates the need to handle classification and regression features separately within each modality. Furthermore, to tackle the second issue, we provide an intuitive explanation that verifies the superior performance of GMs. This explanation sheds light on why GMs are advantageous for multi-modal information fusion. Notably, we extend our approach by implementing more than one type of GM. This inclusion enriches our discussion of applying GMs in multi-modal tracking, providing a more comprehensive exploration of this approach. Additionally, we thoroughly evaluate the effectiveness of our proposed generative-based fusion mechanism across multiple baseline methods, benchmarks, and tracking tasks.

Methodology

Multi-modal tracking aims to obtain the prediction with the collaboration among multiple modalities, requiring the model to fuse relevant clues from the multi-modal input $\{\mathcal{X}^1,...,\mathcal{X}^m\}$. After pre-processing, the images are sent into the feature extractor and the fusion block. However, these two blocks are sometimes entangled (Hui et al. 2023), and therefore termed as $\mathcal{F}$ in combination. The fused features $fused$ are then forwarded to the task head $\mathcal{H}$ to extract task-specific information. Later, the final prediction $\mathcal{P}$ can be maintained after post-processing. The mathematical description is presented as follows:

$$\mathcal{P} = \mathcal{H}(\mathcal{F}(input, \theta), \Theta), \tag{1}$$

where θ and Θ denote the learnable parameters of $\mathcal{F}$ and $\mathcal{H}$, respectively. $input$ is the multi-modal image pair after pre-processing.

Generative-Based Fusion Mechanism

To fulfill this objective, we introduce a novel fusion mechanism, termed GMMT, in this section. Given that the fusion process is typically applied at the feature level, our GMMT is also carefully designed and discussed within the embedding feature space. Following the typical design of GMS, the original fused features $fused$, the input of our GMMT, should be obtained beforehand, which aligns with our multi-stage training scheme. Other than $fused$, the features from each modality $(f_1, ..., f_m)$ should also be retained, which provides strong conditions to guide the fusion for the specific frame pair. These analysis constrain the input of the GMs, but attach no limitation to the architecture of GMS. Therefore, two popular GMs, *i.e.*, DM and CGAN, are involved in our method.

The DM-based GMMT is depicted in Fig. 2(a). Following the DDIM (Song, Meng, and Ermon 2020), in the training stage, the original fused feature $fused$ serves as the x_0. In the forward diffusion process, x_0 undergoes diffusion through the random Gaussian noise $\overline{z}_t$, as defined by the following formulation:

$$x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1 - \overline{\alpha}_t}\overline{z}_t, \tag{2}$$

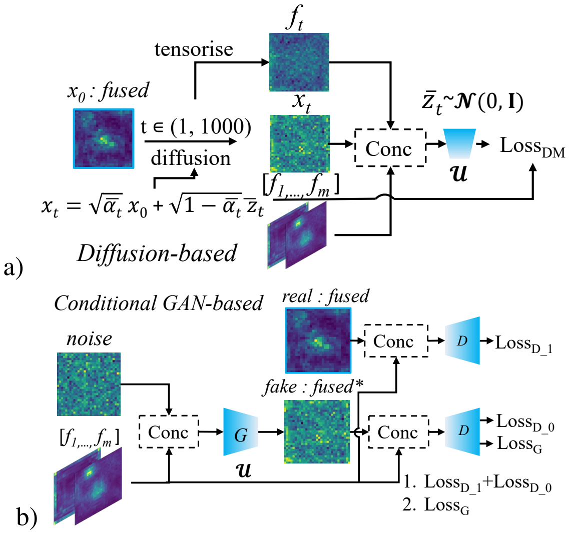


Figure 2: Illustration of the proposed GMMT.

where the subscript t is a random factor chosen from the interval [1, T], which defines how many steps x_0 are performed. $\overline{\alpha}_t$ is a factorial of $\alpha_{1,..,t}$, which are the remainder of $\beta_{1,...,t}$. Here β_t is the predefined diffusion rate and determines how far the t^{th} forward step goes. Once the noisy representation x_t is computed, the reverse diffusion process begins, aiming to recover a clean x_{t-1}. It takes x_t as input, along with $(f_1, ..., f_m)$ as conditions, and the tensorised embedding of t, f_t, as a flag. These elements are concatenated and fed into the U-shaped network $\mathcal{U}$. $\mathcal{U}$ is then optimised by minimizing the $\mathcal{L}2$ loss between the output and noise $\overline{z}_t$, based on which the mean μ_{t-1} and variance σ_{t-1} for the distribution of x_{t-1} can be derived according to Eq. 3.

$$\begin{aligned} \sigma_{t-1} &= \frac{1 - \overline{\alpha}_{t-1}}{1 - \overline{\alpha}_t}\beta_t, \\ \mu_{t-1} &= \frac{\sqrt{\overline{\alpha}_t}(1 - \overline{\alpha}_{t-1})}{1 - \overline{\alpha}_t}x_t + \\ &\frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1 - \overline{\alpha}_t}\frac{1}{\sqrt{\overline{\alpha}_t}}(x_t - \sqrt{1 - \overline{\alpha}_t}\mathcal{U}(x_t, f_1, ..., f_m, f_t)). \end{aligned} \tag{3}$$

Therefore, in the testing phase, the reverse diffusion process is executed iteratively, and in the end, the result can be sampled from the learned distribution of x_0. But at the beginning, x_t is replaced by random noise, and then the time flag t is reversely traversed from T to 1.

In general, with the DM-based GMMT, the typical tracking process described in Eq. 1 develops to Eq.4:

$$\mathcal{P} = \mathcal{H}(Sam(\sum(\mathcal{F}(input, \theta))), \Theta), \tag{4}$$

where Sam is the abbreviation of the sampler, which means sampling data from the generated distribution. Inspired by the total probability formula, $\sum$ is used as a symbol of distribution.

The CGAN-based GMMT is displayed in Fig. 2(b). Following the widely-used CGAN (Mirza and Osindero 2014), the discriminator D and generator G are trained iteratively. To train the D, the synthesised $fused^*$ and the original $fused$ are one-hot labelled, assigning 1 to $fused$ and 0 to $fused^*$. After that, separate losses are computed for $fused^*$ and $fused$, denoted as $Loss_{D_0}$ and $Loss_{D_1}$, respectively. Aiming at distinguishing the real and fake data, D is optimised by minimising $Loss_D = Loss_{D\_0} + Loss_{D\_1}$. After training D, its parameters are frozen, and the learning process of G commences. $fused^*$ is sent into D, with the label change to 1, and the corresponding loss $Loss_G$ is obtained and minimised. Since G is designed to deceive and mislead D, $Loss_G$ is equivalent to $Loss_{D\_1}$. Notably, the loss in this part is calculated by mean square error. To ensure a fair comparison, the architecture of G mirrors that of $\mathcal{U}$ employed in DM-based GMMT. Besides, since only G is employed during inference, the introduction of D is remained in the supplementary material.

In conclusion, the output of the CGAN-based GMMT consists of fake features, signifying that the distribution is not explicitly learned. Consequently, the overall tracking process remains the same to Eq. 1.

Multi-Modal Trackers

The proposed GMMT is implemented on three RGB-T trackers *i.e.*, a self-designed Siamese tracker, the ViPT (Zhu et al. 2023a), and TBSI (Hui et al. 2023), which implies that $m = 2$ during application. During the discussion of GMMT, the fused features $fused$ are assumed pre-defined, indicating that the baseline trackers should be pre-trained beforehand. This necessitates two training stages: one to train the baseline method and another to train the proposed GMMT.

For the three selected baseline trackers, the first training stage consists of two primary steps: training the feature extractor and the fusion block. As to the Siamese tracker, SiamBAN (Chen et al. 2020) with a single region proposal network is trained for each modality. A straightforward convolution-based fusion block is constructed and trained for multi-modal fusion. In this fusion block, the multi-modal features are initially concatenated and then fused through a convolutional block. Regarding ViPT (Zhu et al. 2023a), the feature extractor is pre-trained, but its fusion block is retrained in our implementation. As to TBSI (Hui et al. 2023), we use both the publicly available feature extractor and fusion block. A comprehensive description of the implementation details for these baseline trackers is provided in the supplementary material.

Our GMMT is trained during the second stage of our approach. To provide a stable input to GMMT, the feature extractor and the original fusion block are frozen while training the GMMT. Besides, to harmonise the fusion approach with the tracking task, a learnable tracking head is appended, which means the loss in this stage combines the generative loss and the tracking loss $Loss_{track}$ inherited from the baseline method:

$$Loss = Loss_{track} + \lambda * Loss_{gen}, \tag{5}$$

where λ is a hyper-parameter used to balance the contribution of generative loss.

During the testing phase, the overall tracking process is almost the same. The only change is that the original fusion block is discarded, and the fused feature generated by GMMT serves as the input to the subsequent task head $\mathcal{H}$. Further details are provided in the supplementary material.

Evaluation

Implementation Details

Our experiments are conducted on an NVIDIA RTX3090Ti GPU card. Our GMMT is trained on the training split of LasHeR with the parameters optimised by the SGD optimiser. The learning rate is warmed up from 0.001 to 0.005 in the first 20 epochs and subsequently reduces to 0.00005 for the remaining 80 epochs. We set the value of T to 1000.

Benchmarks and Metrics

The effectiveness of GMMT is verified on GTOT (Li et al. 2016), LasHeR (Li et al. 2022), and RGBD1K (Zhu et al. 2023b) benchmarks. In these benchmarks, precision rate (PR), success rate (SR), normalised precision rate (NPR), recall (RE), and F-score are employed for evaluation, whose detail introductions can be found in the supplementary material.

Ablation Study

In this section, based on the siamese framework, we present the ablation study of our GMMT in Table. 2. We denote GMMT(DM) when the embedded generative model is a diffusion model and GMMT(CGAN) when a conditional GAN is employed within GMMT. Additionally, to demonstrate that the improvement is indeed attributed to the generative-based fusion mechanism rather than the larger fusion block embedded in GMMT, we conduct an experiment where the generative loss is removed. In this situation, the network is trained using an $\mathcal{L}2$ loss between the $fused$ and the network output. This variant is denoted as RAW.For fairness, the network architecture remains consistent for all the competitors. Thus, the primary distinction among these competitors lies in the loss function, which is mathematically defined as follows:

$$Loss_{gen} = \begin{cases} Loss_{RAW} = \mathcal{L}2(fused, output); \\ Loss_{CGAN} = E_{x \sim p_g(x)}[logD(x, f_{rgb}, f_{tir})] \\ \quad + E_{z \sim p_n(z)}[log(1 - D(G(z, f_{rgb}, f_{tir})))]; \\ Loss_{DM} = \mathcal{L}2(noise, output); \end{cases} \tag{6}$$

where $Loss_{gen}$ denotes the loss function for the network $\mathcal{U}$, and it can be switched to multiple choices. $Loss_{CGAN}$ is a widely-used loss in the research of GAN, and its detail can be found in (Mirza and Osindero 2014). $Loss_{DM}$ is activated when the diffusion model is employed (Chen et al. 2023).

As to the quantitative results, on LasHeR, the performance of the Siamese baseline is 39.8 on SR. Replacing the fusion block with a larger network $\mathcal{U}$ (RAW) results in an improvement to 42.5. Later, when using the

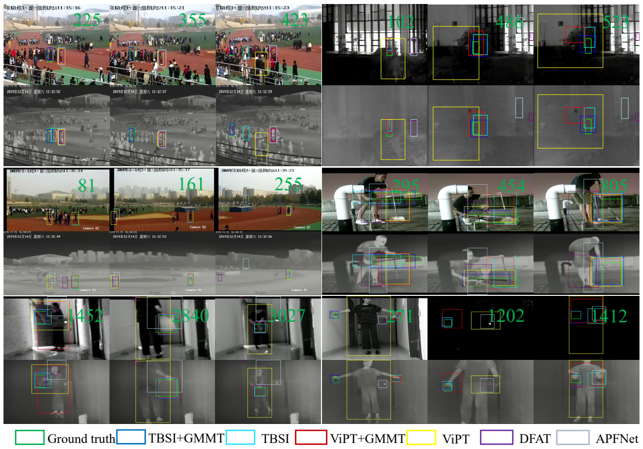


Figure 3: Qualitative visualisation of several advanced RGB-T trackers. The exhibited image pairs are sampled from video 7rightorangegirl, 10runone, ab_bolstershaking, boyinplatform, besom3, ab_rightlowerredcup_quezhen, which are introduced in a top-down and left-right way.

GMMT(CGAN), although a slight degradation of 0.6 appears on SR, a bigger enhancement of 2.3% was maintained on PR, raised from 51.8 to 54.1. Replaced by the DM, consistent improvements are obtained across all the metrics, reaching 57.1, 53.0, and 44.9. Compared to the baseline tracker, significant gains of 6.2%, 5.6%, and 5.1% are observed on PR, NPR, and SR, respectively. It indicates that both the network with deeper architecture and our GMMT contribute to the promising performance. Similar conclusions are drawn from experiments on GTOT. Compared to the baseline method, improvements of 1.7% and 2.3% are displayed in the results of GMMT(DM) on PR and SR. Additionally, since GMMT(DM) performs better than GMMT(CGAN), the rest experiments are conducted based on DM.

Compared with State-of-the-Art Trackers

On LasHeR and RGBT234 benchmarks, several advanced RGB-T trackers are involved, including APFNet (Xiao et al. 2022), ProTrack (Yang et al. 2022), DFAT (Tang et al. 2023), HFRD (Zhang et al. 2023b), ViPT* (Zhu et al. 2023a), TBSI (Hui et al. 2023), and the ViPT* and TBSI modified by GMMT, termed as ViPT*+GMMT and TBSI+GMMT. Here the superscript $*$ represents the results are reproduced by us. As shown in Table. 1, on LasHeR, the best results are obtained by TBSI+GMMT, reaching 70.7, 67.0 and 56.6 on PR, NPR, and SR, respectively. Compared to the original TBSI, GMMT improves the PR, NPR, and SR by 1.5%, 1.3%, and 1.0%. Combining GMMT with ViPT also leads to enhanced performance, with scores rising from 65.0, 61.6, and 52.4 to 66.4, 63.0, and 53.0.

On RGBT234, TBSI+GMMT continuously shows the best performance, reaching 64.7 and 87.9 on the SR and PR metrics, respectively.

On GTOT, we simultaneously display the overall performance and the analysis on each attributes agianst CAT (Li et al. 2020), CMPP (Wang et al. 2020), (Zhang et al. 2022b), JMMAC (Zhang et al. 2021b), ADRNet (Zhang et al. 2021a), MANet++ (Li et al. 2019), HMFT (Zhang et al. 2022a), MacNet (Zhang et al. 2020), APFNet, and TBSI, which are illustrated in the supplymentary material. With the help of the proposed GMMT, TBSI+GMMT shows significantly improvement compared to the competitors. Compared to the baseline method, TBSI, improvements of 2.6% and 2.1% are achieved on SR and PR, boosting the performance from 75.9 and 91.5 to 78.5 and 93.6, respectively, establishing a new state-of-the-art record on this benchmark.

To intuitively show the superiority of GMMT, the tracking results are displayed in Fig. 3, with additional visual comparisons available in the supplementary material. In particular, for ViPT, the enhanced understanding provided by GMMT results in a noticeable improvement, as evident in the comparison between boxes coloured in yellow and red.

RGB-D Extension

To validate the generalisation of GMMT, we also implement it on the RGB-D tracking task, using ViPT-D as the baseline tracker. ViPT-D is an extension of ViPT tailored for RGB-D data.. Initially, we run the official ViPT-D on the RGBD1K dataset, but we notice a performance gap compared to the state-of-the-art, SPT (Zhu et al. 2023b). As RGBD1K videos exhibit a higher diversity with more challenging factors compared to other RGB-D benchmarks, we retrain ViPT-D on the training split of RGBD1K. This retraining effort boosts the F-score from 46.2 to 50.6. The further application of GMMT is based on this retrained variant, which we denote as ViPT-D*.

The quantitative results are displayed in Table. 3, with the competitors being SPT, DDiMP (Bhat et al. 2019), and DeT (Yan et al. 2021). When the UNet is utilised as the embedding network of GMMT, a performance gain of 5.6% is observed. This improvement becomes even more substantial when UNet is replaced by UViT, resulting in an F-score of 57.4. Notably, in addition to the main metric, ViPT-D*+GMMT(V) surpasses ViPT-D* by 6.7% and 7.0% on PR and RE, demonstrating consistent enhancements across all metrics. These results highlight the superiority of GMMT. Besides, although our baseline tracker ViPT-D* falls far behind SPT, which performs the best among all the competitors, a new state-of-the-art is built with the help of our GMMT.

Self-Analysis

Implementation on Multiple Baseline Trackers: To prove our GMMT as a general fusion mechanism, various experiments are conducted on equipping multiple baseline trackers with our GMMT, including a self-designed Siamese tracker, the ViPT and TBSI. On GTOT, when using the UNet (Ho, Jain, and Abbeel 2020) as the embedding network $\mathcal{U}$, the SR of the Siamese baseline is enhanced from 67.0 to 69.3 through the combination of GMMT. On LasHeR, the performance of our Siamese baseline can be significantly boosted from 39.8 to 44.9, with an increment of 5.1%. However, the improvements were relatively modest for ViPT* and TBSI, with gains of around 0.3% and 1.0%, respectively.

	Metrics	APFNet	ProTrack	DFAT	HFRD	ViPT*	TBSI	ViPT*+GMMT	TBSI+GMMT
LasHeR	PR ↑	50.0	53.8	44.6	59.0	65.0	69.2	66.4	70.7
	NPR ↑	43.9	-	40.0	54.5	61.6	65.7	63.0	67.0
	SR ↑	36.2	42.0	33.6	46.4	52.4	55.6	53.0	56.6
RGBT234	PR ↑	82.7	78.6	75.8	82.4	83.5	87.1	84.3	87.9
	SR ↑	57.9	58.7	55.2	58.4	61.7	63.8	61.5	64.7

Table 1: Results on LasHeR and RGBT234 benchmarks.

Dataset	Method	PR↑	NPR↑	SR↑
GTOT	Base	84.0	-	67.0
GTOT	+RAW	81.9	-	67.4
GTOT	+GMMT(CGAN)	81.4	-	67.5
GTOT	+GMMT(DM)	85.7	-	69.3
LasHeR	Base	50.9	47.4	39.8
LasHeR	+RAW	51.8	50.7	42.5
LasHeR	+GMMT(CGAN)	54.1	49.3	41.9
LasHeR	+GMMT(DM)	57.1	53.0	44.9

Table 2: Ablation study on GTOT and LasHeR.

Method	PR ↑	RE ↑	F-score ↑	Δ
DDiMP	55.7	53.4	54.5	
DeT	43.8	41.9	42.8	
SPT	54.5	57.8	56.1	
ViPT-D	45.3	47.2	46.2	
ViPT-D* (Base)	49.2	52.0	50.6	
Base+GMMT(U)	54.7	57.9	56.2	+5.6%
Base+GMMT(V)	55.9	59.0	57.4	+6.8%

Table 3: Results on RGBD1K benchmark.

We attribute this phenomenon to the fusion block in the baseline trackers. In the self-designed Siamese baseline, the fusion block is lightweight. It solely contains a convolutional block, leading to the multi-modal information being insufficiently aggregated. Thus, promoted by GMMT, the performance climbs in a large step. In contrast, ViPT* and TBSI already have well-established fusion processes for multi-modal information, leading to smaller performance increments. In ViPT*, the fusion process occurs in all the 12 self-attention blocks, and it only takes place in the 3^{th}, 6^{th}, and 9^{th} blocks in TBSI. Therefore, the improvement on ViPT* is slightly less than that on TBSI. *In conclusion, the worse the multi-modal information is fused in the baseline tracker, the more it can be boosted by our GMMT.* The overall results are provided in the supplementary material.

The experiments on RGB-D tracking are also in line with this conclusion. Based on the same ViPT baseline, a considerably larger improvement can be found in RGB-D benchmarks. We owe this to the difference in the distinct characteristics of the input data. Although RGB and T data have varying characteristics under various scenarios, they are both imaged based on electromagnetic waves. However, the depth image reflects the distance signal of the surroundings, which has larger heterogeneity to the RGB data. Consequently, the same fusion strategy employed in RGB-T data yields poorer results when applied to RGB-D data. In other words, the multi-modal information in ViPT-D is more inadequately fused than ViPT, which gives reason for the larger enhancement observed in the RGBD1K benchmark.

Learnable Network $\mathcal{U}$ in GMMT: To accomplish the GMMT, an learnable network $\mathcal{U}$ is necessarily introduced in the embedding GM. In our implementation, two renowned networks, UNet (Ho, Jain, and Abbeel 2020) and UViT (Bao et al. 2023), are involved. As their name suggests, both of them follow the U-shaped architecture introduced in supplementary material. However, they differ in two significant aspects: the number of blocks and the detailed architecture of each block. UNet employs a convolution-based block, while UViT constructs its block using the transformer architecture.

With the UNet employed, on the SR metric, the performance of ViPT* reaches 52.7 and that of TBSI is improved from 55.6 to 56.6. Replaced by the UViT, ViPT* performs better but the results of TBSI degrades slightly. But in general, GMMT can boost the baseline methods on all three metrics consistently no matter which inner network is selected. More results are provided in the supplymentary material.

Details of the network architecture and the analysis of the number of blocks, n, are remained to the supplementary material.

Analysis of the Generated Features: To verify the superiority of GMMT intuitively, the generated features are visualised in Fig. 4. Specifically, it is demonstrated globally and locally.

Through the global description, the effectiveness of the generated features is proved. The t-sne tool is used to exhibit thousands of the original and generated features. Fig. 4(a) and Fig. 4(b) are the statistical analysis on GTOT and RGBD1K datasets (red marks denote the generated features and the blues represent the original features). In these two graphs, the original features are marked with blue circles and the generated ones are highlighted by the red pentagram. Apparently, the clusters of the generated and the original fused features are highly overlapped. This overlap indicates that they occupy the same semantic space and share similar properties. Consequently, we believe the generated features are capable of supporting the tracking task, just as the original ones.

After that, the superiority is further demonstrated by the

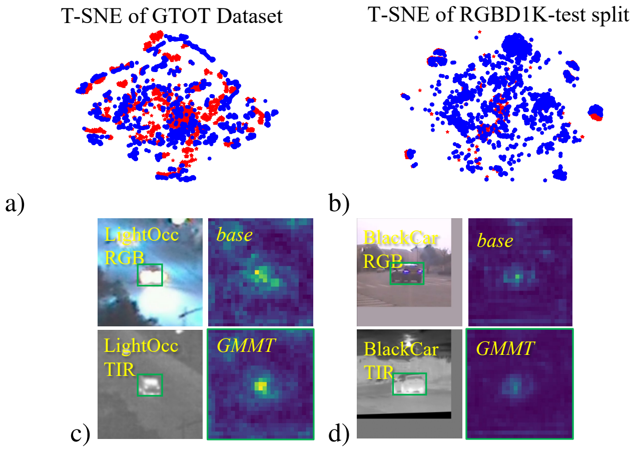


Figure 4: Visualisation of the feature embeddings.

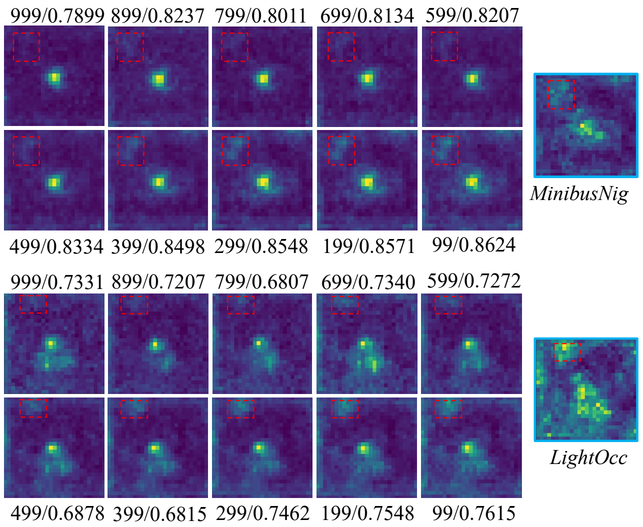


Figure 5: Analysis of diffusion Steps. The content below each feature map is the time step and its structural similarity between the original fused features on the right side.

local description. We provide the visualisation of the feature maps before and after combining GMMT in Fig. 4(c) and Fig. 4(d). Two samples from GTOT are displayed. The left one is from the video $LightOCC$ and another is from $BlackCar$. In these two instances, the targets locate in the centre because they are centre-cropped before sending into the network. Therefore, the ideal feature maps should exhibit a strong response in the central target area while suppressing background regions. In $LightOCC$, the visualisation from GMMT has a higher response in the target area, and the background is clearer because the region with extreme illumination in the RGB modality better discarded. In $BlackCar$, both feature maps are focused on the key position, but the background noise in GMMT is better suppressed. Based on the analysis of these two video samples, we attribute the superiority of GMMT to its ability to generate more discriminative features. Further visualisations and analyses are available in the supplementary material, where consistent conclusions are drawn.

The reason for producing better features is attributed to the generative training paradigm. During training, a random noise together with the RGB and TIR features form the input of the network $\mathcal{U}$. In this setup, $\mathcal{U}$ needs to effectively understand and extract crucial cues from both f_{rgb} and f_{tir} to successfully perform the information fusion task. As a result, compared to a network trained with a purely discriminative paradigm, our approach encourages $\mathcal{U}$ to better extract important information from each modality. This, in turn, leads to fused features with enhanced discrimination, making them more suitable for challenging tracking tasks that involve diverse and complex environmental conditions.

Analysis of Diffusion Steps: Different from the typical fusion blocks, our GMMT(DM) can be recursively executed. Fig. 5 gives the visualisation of $s = 10$ steps (from 999 to 99, in a reverse manner) when T is set to 1000. It can be seen that with s becomes larger, the generated features are more similar to the original fused features $fused$. Unexpectedly, the noise in $fused$ is also better recovered as shown in Fig. 5(a) and (b). The similarity is quantified by the structural similarity between the generated feature map $fused^*$ and $fused$. This indicates the superiority of GMMT is disappearing gradually when s goes larger. The quantitative results and the corresponding analysis are displayed in the supplementary material. Since more steps cost more computational resources and time, s equals 1 in our method for efficiency, reaching 18 frames per second.

Analysis of λ: λ is a crucial factor banding the tracking and generation tasks. Thus, the analysis on it is conducted and exhibited in the supplementary material, with λ valued from (1,2,3,5,10,100). The conclusion is that all the variants perform better than the baseline method, which demonstrate the superiority of GMMT. Additionally, when λ=100, the performance is lightly better than the baseline. This indicates that λ should not be a large value, leading to a small influence of the tracking loss, and, furthermore, the strong supervision of the tracking task is crucial and should not be ignored.

Conclusion

This paper proposes a novel generative-based fusion mechanism for multi-modal tracking, named as GMMT. Its effectiveness has been demonstrated n multiple tracking baselines, multiple challenging benchmarks, as well as two multi-modal tracking tasks. Enhanced by our GMMT, new state-of-the-arts are built on the challenging GTOT, LasHeR, and RGBD1K benchmarks. Furthermore, through the intuitive visualisation, we attribute its superiority to the noisy training paradigm, which forces the model understands and preserves the discriminative clues from each modality to the fused features. Additionally, GMMT tends to yield larger improvements when applied to baseline methods with rough information fusion processes. The supplementary material is available at https://github.com/Zhangyong-Tang/GMMT.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (Grant NO. 62020106012, 62332008, 62106089, U1836218, 62336004), the National Key Research and Development Program of China (2023YFF1105102, 2023YFF1105105), the 111 Project of Ministry of Education of China (Grant No.B12018), and the UK EPSRC (EP/N007743/1, MURI/EPSRC/DSTL, EP/R018456/1, EP/V002856/1).

References

Bao, F.; Nie, S.; Xue, K.; Cao, Y.; Li, C.; Su, H.; and Zhu, J. 2023. All are Worth Words: A ViT Backbone for Diffusion Models. In *CVPR*.

Bhat, G.; Danelljan, M.; Gool, L. V.; and Timofte, R. 2019. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6182–6191.

Chen, S.; Sun, P.; Song, Y.; and Luo, P. 2023. DiffusionDet: Diffusion Model for Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 19830–19843.

Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; and Ji, R. 2020. Siamese Box Adaptive Network for Visual Tracking. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6667–6676.

Cvejic, N.; Nikolov, S. G.; Knowles, H. D.; Loza, A.; Achim, A.; Bull, D. R.; and Canagarajah, C. N. 2007. The Effect of Pixel-Level Fusion on Object Tracking in Multi-Sensor Surveillance Video. In *2007 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–7.

Fan, S.; He, C.; Wei, C.; Zheng, Y.; and Chen, X. 2023. Bayesian Dumbbell Diffusion Model for RGBT Object Tracking With Enriched Priors. *IEEE Signal Processing Letters*, 30: 873–877.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Han, Y.; Zhang, P.; Huang, W.; Zha, Y.; Cooper, G. D.; and Zhang, Y. 2020. Robust Visual Tracking based on Adversarial Unlabeled Instance Generation with Label Smoothing Loss Regularization. *Pattern Recognition*, 97: 107027.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hui, T.; Xun, Z.; Peng, F.; Huang, J.; Wei, X.; Wei, X.; Dai, J.; Han, J.; and Liu, S. 2023. Bridging Search Region Interaction With Template for RGB-T Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13630–13639.

Iqbal, T.; and Qureshi, S. 2022. The survey: Text generation models in deep learning. *Journal of King Saud University-Computer and Information Sciences*, 34(6): 2515–2528.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Kingma, D. P.; and Welling, M. 2014. Stochastic gradient VB and the variational auto-encoder. In *Second international conference on learning representations, ICLR*, volume 19, 121.

Li, C.; Cheng, H.; Hu, S.; Liu, X.; Tang, J.; and Lin, L. 2016. Learning Collaborative Sparse Representation for Grayscale-Thermal Tracking. *IEEE Transactions on Image Processing*, 25(12): 5743–5756.

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

Li, C.; Liu, L.; Lu, A.; Ji, Q.; and Tang, J. 2020. Challenge-Aware RGBT Tracking. In *European Conference on Computer Vision (ECCV)*, 222–237. Springer International Publishing.

Li, C.; Wu, X.; Zhao, N.; Cao, X.; and Tang, J. 2018. Fusing two-stream convolutional neural networks for RGB-T object tracking. *Neurocomputing*, 281: 78–85.

Li, C.; Xue, W.; Jia, Y.; Qu, Z.; Luo, B.; Tang, J.; and Sun, D. 2022. LasHeR: A Large-Scale High-Diversity Benchmark for RGBT Tracking. *IEEE Transactions on Image Processing*, 31: 392–404.

Li, C.; Zhao, N.; Lu, Y.; Zhu, C.; and Tang, J. 2017. Weighted Sparse Representation Regularized Graph Learning for RGB-T Object Tracking. In *Proceedings of the 25th ACM International Conference on Multimedia*, 1856–1864.

Li, C. L.; Lu, A.; Zheng, A. H.; Tu, Z.; and Tang, J. 2019. Multi-Adapter RGBT Tracking. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2262–2270.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Lu, Y.-F.; Gao, J.-W.; Yu, Q.; Li, Y.; Lv, Y.-S.; and Qiao, H. 2023. A Cross-Scale and Illumination Invariance-Based Model for Robust Object Detection in Traffic Surveillance Scenarios. *IEEE Transactions on Intelligent Transportation Systems*, 24(7): 6989–6999.

Luo, Z.; Chen, D.; Zhang, Y.; Huang, Y.; Wang, L.; Shen, Y.; Zhao, D.; Zhou, J.; and Tan, T. 2023. VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10209–10218.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Prenger, R.; Valle, R.; and Catanzaro, B. 2019. Waveglow: A Flow-based Generative Network for Speech Synthesis. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3617–3621.

Rao, D.; Xu, T.; and Wu, X.-J. 2023. TGFuse: An Infrared and Visible Image Fusion Approach Based on Transformer and Generative Adversarial Network. *IEEE Transactions on Image Processing*, 1–1.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruan, L.; Ma, Y.; Yang, H.; He, H.; Liu, B.; Fu, J.; Yuan, N. J.; Jin, Q.; and Guo, B. 2023. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10219–10228.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y.; Ma, C.; Wu, X.; Gong, L.; Bao, L.; Zuo, W.; Shen, C.; Lau, R. W.; and Yang, M.-H. 2018. Vital: Visual tracking via adversarial learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8990–8999.

Tang, Z.; Xu, T.; Li, H.; Wu, X.-J.; Zhu, X.; and Kittler, J. 2023. Exploring fusion strategies for accurate RGBT visual object tracking. *Information Fusion*, 101881.

Tang, Z.; Xu, T.; and Wu, X.-J. 2022. Temporal Aggregation for Adaptive RGBT Tracking. *arXiv preprint arXiv:2201.08949*.

Wang, C.; Xu, C.; Cui, Z.; Zhou, L.; and Yang, J. 2020. Cross-Modal Pattern-Propagation for RGB-T Tracking. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, W.; and Wong, A. K. 2002. Autoregressive Model-Based Gear Fault Diagnosis. *Journal of Vibration and Acoustics*, 124(2): 172–179.

Wang, X.; Li, C.; Luo, B.; and Tang, J. 2018. SINT++: Robust Visual Tracking via Adversarial Positive Instance Generation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4864–4873.

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

Xu, Q.; Mei, Y.; Liu, J.; and Li, C. 2021. Multimodal Cross-Layer Bilinear Pooling for RGBT Tracking. *IEEE Transactions on Multimedia*, 1–1.

Yan, S.; Yang, J.; Kapyla, J.; Zheng, F.; Leonardis, A.; and Kamarainen, J. 2021. DepthTrack: Unveiling the Power of RGBD Tracking. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10705–10713.

Yang, J.; Li, Z.; Zheng, F.; Leonardis, A.; and Song, J. 2022. Prompting for multi-modal tracking. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3492–3500.

Yin, Y.; Xu, D.; Wang, X.; and Zhang, L. 2020. Adversarial Feature Sampling Learning for Efficient Visual Tracking. *IEEE Transactions on Automation Science and Engineering*, 17(2): 847–857.

Zhang, C.; Huang, Z.; Wang, S.; and Hong, Y. 2023a. Decision-making for Overtaking in Specific Unmanned Driving Scenarios based on Deep Reinforcement Learning. In *2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)*, 680–685.

Zhang, H.; Zhang, L.; Zhuo, L.; and Zhang, J. 2020. Object Tracking in RGB-T Videos Using Modal-Aware Attention Network and Competitive Learning. *Sensors*, 20(2): 393.

Zhang, L.; Danelljan, M.; Gonzalez-Garcia, A.; van de Weijer, J.; and Shahbaz Khan, F. 2019. Multi-Modal Fusion for End-to-End RGB-T Tracking. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2252–2261.

Zhang, P.; Wang, D.; Lu, H.; and Yang, X. 2021a. Learning Adaptive Attribute-Driven Representation for Real-Time RGB-T Tracking. *International Journal of Computer Vision*, 129: 2714–2729.

Zhang, P.; Zhao, J.; Bo, C.; Wang, D.; Lu, H.; and Yang, X. 2021b. Jointly Modeling Motion and Appearance Cues for Robust RGB-T Tracking. *IEEE Transactions on Image Processing*, 30: 3335–3347.

Zhang, P.; Zhao, J.; Wang, D.; Lu, H.; and Ruan, X. 2022a. Visible-Thermal UAV Tracking: A Large-Scale Benchmark and New Baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Zhang, T.; Guo, H.; Jiao, Q.; Zhang, Q.; and Han, J. 2023b. Efficient RGB-T Tracking via Cross-Modality Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5404–5413.

Zhang, T.; Liu, X.; Zhang, Q.; and Han, J. 2022b. SiamCDA: Complementarity- and Distractor-Aware RGB-T Tracking Based on Siamese Network. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3): 1403–1417.

Zhao, F.; Wang, J.; Wu, Y.; and Tang, M. 2019. Adversarial Deep Tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(7): 1998–2011.

Zhu, J.; Lai, S.; Chen, X.; Wang, D.; and Lu, H. 2023a. Visual Prompt Multi-Modal Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9516–9526.

Zhu, X.-F.; Xu, T.; Tang, Z.; Wu, Z.; Liu, H.; Yang, X.; Wu, X.-J.; and Kittler, J. 2023b. RGBD1K: A Large-Scale Dataset and Benchmark for RGB-D Object Tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3870–3878.

Zhu, Y.; Li, C.; Luo, B.; Tang, J.; and Wang, X. 2019. Dense Feature Aggregation and Pruning for RGBT Tracking. In *Proceedings of the 27th ACM International Conference on Multimedia*, 465–472.

Zhu, Y.; Li, C.; Tang, J.; Luo, B.; and Wang, L. 2021. RGBT Tracking by Trident Fusion Network. *IEEE Transactions on Circuits and Systems for Video Technology*, 1–1.