Visible-Thermal UAV Tracking: A Large-Scale Benchmark and New Baseline

Pengyu Zhang[1], Jie Zhao[1], Dong Wang[1†], Huchuan Lu[1,2], Xiang Ruan[3]
[1]School of Information and Communication Engineering, Dalian University of Technology, China
[2]Peng Cheng Laboratory [3]Tiwaki Co.Ltd.
{pyzhang, zj982853200}@mail.dlut.edu.cn, {wdice, lhchuan}@dlut.edu.cn, ruanxiang@tiwaki.com

Abstract

With the popularity of multi-modal sensors, visible-thermal (RGB-T) object tracking is to achieve robust performance and wider application scenarios with the guidance of objects' temperature information. However, the lack of paired training samples is the main bottleneck for unlocking the power of RGB-T tracking. Since it is laborious to collect high-quality RGB-T sequences, recent benchmarks only provide test sequences. In this paper, we construct a large-scale benchmark with high diversity for visible-thermal UAV tracking (VTUAV), including 500 sequences with 1.7 million high-resolution ($1920 * 1080$ *pixels) frame pairs. In addition, comprehensive applications (short-term tracking, long-term tracking and segmentation mask prediction) with diverse categories and scenes are considered for exhaustive evaluation. Moreover, we provide a coarse-to-fine attribute annotation, where frame-level attributes are provided to exploit the potential of challenge-specific trackers. In addition, we design a new RGB-T baseline, named Hierarchical Multi-modal Fusion Tracker (HMFT), which fuses RGB-T data in various levels. Numerous experiments on several datasets are conducted to reveal the effectiveness of HMFT and the complement of different fusion types. The project is available at here.*

1. Introduction

Given the initial position of a model-agnostic target, visual object tracking is to capture the target in the subsequent frames [28], where the target may suffer out-of-view, occlusion, illumination variation, and motion blur. Previous algorithms solve those challenges within visible modality, providing limited information when the target is in dark, rainy, foggy and other extreme conditions (the first row in Fig. 1). By contrast, thermal image, as a complementary cue, is insensitive to illumination variation, while it cannot distinguish the target when the target and background are in similar temperature (the second row in Fig. 1). To this end, with the portability and low-price of multi-modality sensors, tracking with visible-thermal (RGB-T) data enlarges the application scope by providing complementary information, which has attached more attentions [23, 47]. Li *et al.* [19] release a gray-scale RGB-T dataset with 50 videos. Later, RGBT210 [25] and RGBT234 [20] are proposed, containing 210 and 234 test videos. In 2019 [15] and 2020 [16], VOT committee holds the VOT-RGBT tracking subchallenge, which selects 60 sequences from RGBT234 to evaluate the accuracy and robustness of competitors. Furthermore, various algorithms are proposed by considering performance and time cost. Li *et al.* [22] propose a Multi-

Figure 1. Sample frames in our dataset. *Scenes - super class (sequence length)* are shown on the top. Sequence-level attributes are shown at bottom, including camera movement (C), deformation (D), extreme illumination (E), partial occlusion (P), full occlusion (F), scale variation (S), thermal clustering (H), fast moving (M), out-of-view (O), and low resolution (L).

† Corresponding author: Dr. Dong Wang, wdice@dlut.edu.cn

Adaptor network to learn modality-shared and modality-specific representations. Zhang *et al.* [48] design a real-time RGB-T tracker that exploits the effectiveness of attribute annotation. Zhang *et al.* [46] extend DiMP [2] to RGB-T tracking, obtaining the best ranking in VOT2019-RGBT.

However, the lack of training data becomes the main bottleneck for RGB-T tracking. Existing datasets (GTOT, RGBT210, RGBT234, and VOT-RGBT) contain 284 unique short-term sequences overall. Trackers have to be trained on another test set [22, 48] or synthetic data generated from visible modality [46, 50], which suffer limited generalization ability and training gap. Moreover, test sequences are captured with monitoring devices, thereby leading to limited viewpoint, frame length, and imaging quality. To fully exploit the potential of the RGB-T tracker, this paper presents a large-scale RGB-T tracking dataset with high diversity. The main contributions are listed as follows:

- We construct a large-scale benchmark with high diversity for visible-thermal UAV tracking (VTUAV). To our best knowledge, VTUAV is the largest multi-modal tracking dataset with the highest resolution. Moreover, we take short-term, long-term tracking and segmentation mask prediction into consideration to achieve a comprehensive evaluation with wider applications. We also provide an exquisite attribute annotation in frame and sequence levels, which can meet the requirement of training a challenge-specific tracker.

- We propose a new baseline for RGB-T tracking, namely HMFT, which unifies various multi-modal fusion strategies (including image fusion, feature fusion and decision fusion) into a hierarchical fusion framework. We implement corresponding versions for short-term and long-term tracking. Furthermore, we provide an in-depth analysis on various fusion types to develop RGB-T trackers. Exhaustive experiments on GTOT, RGBT210, RGBT234 and VTUAV conclude the complement of various fusion types.

2. Related Work

RGB-T tracking benchmarks. The first dataset used for RGB-T tracking is OTCBVS [39], which contains 6 sequences with 7200 frames. In 2012, LITIV [36] was proposed with 9 video clips and 6300 image pairs. These two datasets are out-of-date since they are not particularly designed for RGB-T tracking and with limited data. In 2016, Li *et al.* [19] propose a grayscale-thermal tracking dataset, namely GTOT, which contains 7800 frames. GTOT contains various challenging scenes to evaluate the trackers' robusteness in extreme conditions. The RGBT210 [25] dataset is released, with 210 videos and more than 104K frames. Later, RGBT234 [20], an extended version of RGBT210, enlarges the number of sequence to 234 and provides a modality-independent annotation, which can be used to learn individual models separately. In 2019, VOT committee selects 60 sequences and construct a new dataset VOT-RGBT [15], which utilizes Expected Average Overlap (EAO) to evaluate the accuracy and robustness of trackers. The LSS dataset [50] is a newly-built synthetic dataset, where either visible or thermal images are generated from another modality using image translation or video colorization methods. Recently, LasHeR [24] contains 1224 short-term videos and 730K frames with multiple scenes and viewpoints. In this paper, we propose a unified large-scale RGB-T tracking dataset with high-quality training pairs. Compared with the most recent dataset (LasHeR), three main differences can be summarized. First, we have higher quality images and a wider distribution in frame length. Second, LasHeR focuses on short-term evaluation while our dataset measures trackers' performance from three mainstream perspectives including tracking accuracy, target re-detection, and pixel-level estimation. Third, detailed frame-level attribute annotations are provided, which can meet the requirement of challenge-aware trackers [21, 48].

RGB-T tracking algorithms. Recent RGB-T trackers focus on exploiting the correspondence and discriminability of multi-modal information [45, 51, 54, 55]. Several fusion methods are proposed, which can be categorized as image fusion, feature fusion, and decision fusion. For image fusion, Peng *et al.* [33] utilize a group of layers to learn complementary information by sharing weights for heterogeneous data. Image-fusion-based method can provide a shared representation of multi-modal, while highly depends on image alignment and has not been exploited sufficiently. Most trackers aggregate the representation by fusing features [38], which can be detailed as two types, i.e., modality interaction and direct fusion. The former is to refine uni-modal feature guided by another one, and then the features from both modalities are combined, thereby achieving comprehensive representations [37, 38]. By contrast, using the multi-modal features as input, the latter combines them first and learns a fused representation by direct concatenation [46, 51] or attention technique [33]. Feature fusion has the potential of high flexibility and can be trained with massive unpaired data, which is well-designed to achieve significant promotion. Decision fusion models each modality independently, and the scores are fused to obtain the final candidate. JMMAC [49] adopts a multi-modal fusion network to ensemble the responses by considering modality-level and pixel-level importance. Luo *et al.* [30] utilize independent frameworks to track in RGB-T data, and then the results are combined by adaptive weighting. Decision fusion avoids the heterogeneity of different modalities and is not sensitive to modality registration. In this work, we also design a new baseline for RGB-T tracking using hierar-

chical fusion manner, which derives benefit from all aforementioned three fusion types. Numerous results on three popular RGB-T datasets show that the information fused in various levels can provide comprehensive contributions to obtain a superior result.

3. VTUAV Benchmark

3.1. Benchmark Features and Statistics

- *Large-scale sequences with high diversity*. Recent RGB-T datasets use the multi-sensor surveillance camera with a 2 degree-of-freedom turnable platform. The image quality and flexibility cannot meet the requirements for tracking. Moreover, the target cannot be tracked for a long time with the still camera, leading to a limited frame length. To address these issues, our dataset is captured by a professional UAV (DJI Matrice 300 RTK) with Zenmuse H20T camera, which can achieve stable flight in extreme conditions, such as night, foggy, and windy scenes. The thermal camera captures 8-14μm and we control the flight height from 5-20m for a proper target size. We collect 500 sequences with 1,664,549 RGB-T image pairs. The images are in high quality, which are stored with 1920*1080 resolution[1] with jpg format and sampled in 30 fps. We separate 250 sequences as training set and the other 250 sequences as test set. All sequences are also split into long-term and short-term sets, according to the absence of targets[2]. In the training set, there are 207 short-term sequences and 43 long-term sequences. A total of 74 of 250 test sequences belong to long-term set to evaluate the performance for long-term tracking. Another 176 test sequences are used for short-term tracking evaluation. We also provide mask annotation for 100 sequences selected in short-term subset (50 sequences for training and 50 sequences for test), which can be used for video object segmentation and scale estimation learning. Comparison between existing datasets can be found in Tab. 1 and Fig. 2 (c).
- *Generic object and scene category*. Related datasets are mainly recorded at the road, school for safely monitoring scenes, with limited scenes and object categories. We aim to construct a highly diverse dataset with adequate object types and scenes. As shown in Fig. 2 (a), the tracked target can be divided into 5 super-classes (pedestrian, vehicle, animal, train and ship) and 13 sub-classes, which can cover most of category for real-world applications. Sequences are captured at 15 scenes cross two cities, which include the road, street, bridge, park, sea, beach, court, and school, etc. To emphasize the effectiveness of both modalities, the data acquisition lasts for a whole year for various weather conditions and climates. Specifically, 325 sequences are captured at daytime and 175 sequences are at night with different conditions, such as windy, cloudy, and foggy weather.
- *Hierarchical attributes*. Previous methods [21, 35, 48] aim to exploit the potential of attribute information and achieve satisfying performance in challenging cases. However, existing visual and RGB-T datasets [20, 26, 40] label attributes at the sequence-level, which involves various challenges into single sequence in a coarse manner. In this paper, besides the sequence-wise attributes, we achieve a hierarchical attribute annotation by additionally labeling frame-level attributes for training sequences to fully investigate the attribute-based methods. Instead of separating the whole sequences into several clips [9], we maintain the sequence continuity, which allows the frames to be annotated with multiple labels or none. Challenges are summarized as 13 attributes, including target blur (TB), camera movement (CM), extreme illumination (EI), deformation (DEF), partial occlusion (PO), full occlusion (FO), scale variation (SV), thermal crossover (TC), fast moving (FM), background clustering (BC), out-of-view (OV), low-resolution (LR) and thermal-visible separation (TVS). Fig. 2 (b) lists the number of each attribute from sequence-level and frame-level. *The description of attribute is summarized in supplementary material*.
- *Alignment*. Given that multi-sensor device cannot ensure the photocardic polymerization, thereby view differences occur. Previous RGB-T datasets apply frame-level alignment to calculate homography transformation and unify the view scope frame by frame, incurring immense labor costs and is impracticable in real-world applications. In our dataset, we operate modality alignment in the initial frame for each video and apply it to all frames. We note that most frames are well-aligned. *The comparison of different alignment methods can be found in supplementary material.*

3.2. High-quality Annotation

In our dataset, we provide sufficient expert annotations in three formats, including bounding boxes, segmentation masks, and attribute annotations. Examples are shown in Fig. 1 and Fig. 2 (d).

- *Bounding boxes*. In VTUAV, we carefully annotate bounding boxes for both modalities, individually. We provide sparse annotations in an interval of 10 frames.

[1]The thermal images are captured in 640*480 resolution and we rescale them to achieve registration in alignment process.

[2]We label the sequence as long-term, where the target is out-of-view for more than continuous 20 frames.

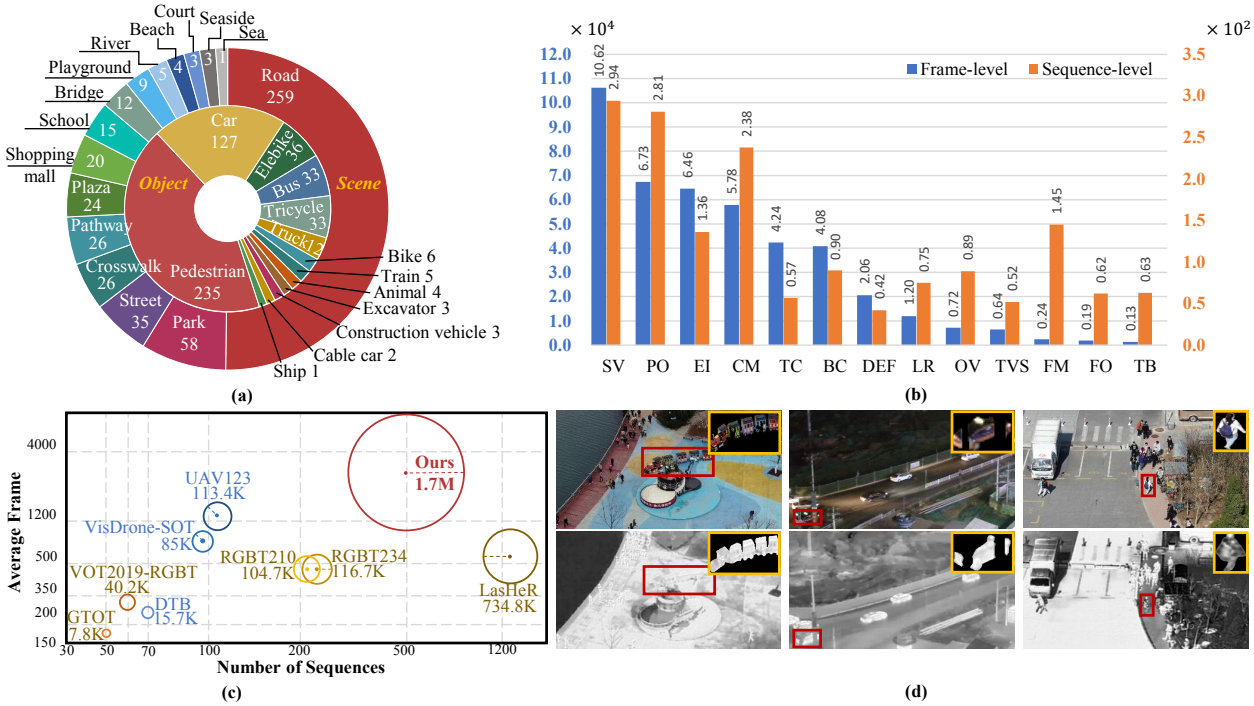

Figure 2. Main features and statistics of the proposed dataset. (a) Distribution of scene (outer) and object category (inner). (b) Statistics of frame-level and sequence-level attributes. (c) Comparison among existing datasets and the proposed dataset. The area of each cycle denotes the number of total frame. (d) Precise annotation with bounding box and segmentation mask. Better viewed in color with zoom in.

Table 1. Statistics comparison among existing multi-modality and UAV tracking datasets.

	Benchmark	Num. Seq.	Avg. Frame	Min. Frame	Max. Frame	Total Frame	Resolution	Train subset	Long-term subset	Num. Seg.	Multi-modal	Year
RGB-T	GTOT [19]	50	157	40	376	7.8K	384 × 288	×	×	×	✓	2016
	RGBT210 [25]	210	498	40	4140	104.7K	630 × 460	×	×	×	✓	2017
	VOT2019-RGBT [15]	60	334	40	1335	40.2K	630 × 460	×	×	×	✓	2019
	RGBT234 [20]	234	498	40	4140	116.7K	630 × 460	×	×	×	✓	2019
	LasHeR [24]	**1224**	600	57	12862	734.8K	630 × 480	✓	×	×	✓	2021
UAV	UAV123 [31]	123	1246	109	5527	113.4K	1280 × 720	×	✓	×	×	2016
	DTB [26]	70	225	68	699	15.7K	1280 × 720	×	×	×	×	2017
	VisDrone-SOT [53]	96	892	92	3135	85K	1360 × 765	✓	×	×	×	2018
	VTUAV	500	**3329**	**196**	**27213**	**1.7M**	**1920 × 1080**	✓	✓	**24.4K**	✓	2021

As described in [32], dense annotation can be achieved with the guidance of state-of-the-art trackers. In this manner, we provide 326,961 high-quality bounding box annotations in total.

- *Segmentation masks*. We annotate the target mask at 1 fps for visible and thermal images. A total of 24,464 masks are generated using Labelme toolkit.

- *Attribute annotations*. In our dataset, we provide frame-level attribute annotations to conduct the detailed attribute-based analysis. Most attributes[3] are labeled by a full-time expert. Therefore, we totally label 301,678 frames with 430,960 attributes and provide 500 * 13 sequence-level annotations.

[3]The attribute of FM, SV, LR and TVS are automatically annotated according to their descriptions.

3.3. Evaluation Metrics

In our experiment, all the trackers are run in one-pass evaluation (OPE) protocol and evaluated by maximum success rate (MSR) and maximum precision rate (MPR), which are widely used in RGB-T tracking [19,20,25]. Overall performance for all sequences and attribute-based performance for attribute-specific sequences are considered. For mask evaluation, we measure the results with Jaccard index ($\mathcal{J}$) and F-score ($\mathcal{F}$) [34].

- *Maximum success rate (MSR).* Success rate (SR) measures the ratio of tracked frames, determined by the Interaction-over-Union (IoU) between tracking result and ground truth. With different overlap thresholds, a success plot (SP) can be obtained, and SR is calculated as the area under curve of SP. Owing to the modality-level displacement, we adopt the maximum overlap in frame level as the final score.

- *Maximum precision rate (MPR).* Similar to precision rate (PR), MPR is to calculate the percentage of frames, where the center distance between prediction and ground truth is smaller than a threshold τ. τ is set to 20 in our experiment.

- *Jaccard index ($\mathcal{J}$).* Jaccard index is defined as the averaging pixel-level IoU between the predicted mask $\mathbf{M}$ and ground truth $\mathbf{G}$ for all N frames. which can be formulated as $\mathcal{J} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{M}^i \cap \mathbf{G}^i}{\mathbf{M}^i \cup \mathbf{G}^i}$.

- *F-score ($\mathcal{F}$).* F-score calculates region-based precision Pr and recall Re based between the closed contours in $\mathbf{M}$ and $\mathbf{G}$, which can be expressed via $\mathcal{F} = \frac{2Pr*Re}{Pr+Re}$.

4. Hierarchical Multi-modal Fusion Tracker

In this section, we will introduce a new baseline for RGB-T tracking to fully exploit various fusion types in a unified framework, shown as Fig. 3. It contains three main modules: Complementary Image Fusion (CIF), Discriminative Feature Fusion (DFF) and Adaptive Decision Fusion (ADF). CIF aims to learn shared pattern between two modalities. DFF introduces a channel-wise combination of heterogeneous representations. Finally, ADF is to provide the final target candidate by considering the responses from discriminative and complementary classifiers.

4.1. Complementary Image Fusion

As shown in the third row in Fig. 1, RGB-T images are captured in the same scenes and the complementary information (such as semantic and contour, etc) exists in both modalities, which can be propagated to each other for a robust feature representation [27]. To this end, we utilize a shared backbone, i.e., ResNet50 [11], to extract the common features. To leverage the consistency between the two modalities, we introduce a divergence loss $\mathcal{L}_{div}$ to constrain the multi-modal feature distribution by measuring their KL-divergence, which can be expressed as follows,

$$\begin{aligned}\mathcal{L}_{div} &= \sum_{i=1}^{L} \mathrm{KL}(\mathbf{P}_v^i || \mathbf{P}_t^i) \\ &= \frac{1}{N}\sum_{i=1}^{L}\sum_{n=1}^{N}(p_{vn}^i \log(p_{vn}^i - p_{tn}^i))\end{aligned} \tag{1}$$

where, $\mathbf{P}_v^i \in \mathbb{R}^{C*H*W}$ and $\mathbf{P}_t^i \in \mathbb{R}^{C*H*W}$ denote the complementary features output by i-th block for visible and thermal modalities, and p_{vn}^i and p_{tn}^i are the n-th items in $\mathbf{P}_v^i$ and $\mathbf{P}_t^i$, respectively. L denotes the number of block in ResNet50. The learned representations are then concatenated to form the overall complementary feature $\mathbf{P}_a \in \mathbb{R}^{2C*H*W}$, where C, H, W denote the channel number, height and width of the feature, respectively. Complementary feature is more robust when all modalities work well and achieves accurate scale estimation.

4.2. Discriminative Feature Fusion

Dual modalities can provide heterogeneous information, where visible images provide detailed context, and thermal images obtain more contour information according to the temperature difference, achieving robustness to illumination changes. To exploit the potential of both modalities, we first use an individual feature extractor to model each modality. Then we propose a Discriminative Feature Fusion (DFF) module to fuse those representations. Considering the information from visible and thermal images, DFF provides a fused feature map by introducing a channel-wise modality weight. In DFF, feature maps from visible and thermal images $\mathbf{D}_v \in \mathbb{R}^{C*H*W}$ and $\mathbf{D}_t \in \mathbb{R}^{C*H*W}$ are summed, and we embed global vector $\mathbf{d}_g$ from both modalities by Global Average Pooling (GAP) and Fully-Connected (FC) layer, which can be expressed as,

$$\mathbf{d}_g = \mathcal{F}_g(\mathrm{GAP}(\mathbf{D}_v + \mathbf{D}_t)), \tag{2}$$

where $\mathcal{F}(\cdot)$ denotes the fully-connected layer. Then, two FC layers are adopted to produce channel-wise weights $\mathbf{w}_v, \mathbf{w}_t \in \mathbb{R}^{C*1*1}$ for each modality, which is followed by a softmax operation, as shown in Eq. (3) and Eq. (4).

$$\mathbf{w}_v = \frac{e^{\mathcal{F}_v(\mathbf{d}_g)}}{e^{\mathcal{F}_v(\mathbf{d}_g)} + e^{\mathcal{F}_t(\mathbf{d}_g)}} \tag{3}$$

$$\mathbf{w}_t = \frac{e^{\mathcal{F}_t(\mathbf{d}_g)}}{e^{\mathcal{F}_v(\mathbf{d}_g)} + e^{\mathcal{F}_t(\mathbf{d}_g)}} \tag{4}$$

Finally, the aggregated features can be obtained by weighted summation among channels via,

$$\mathbf{D}_a^i = w_v^i * \mathbf{D}_v^i + w_t^i * \mathbf{D}_t^i \tag{5}$$

where the superscript i denotes the i-th channel of all the variables. With the proposed DFF, we construct a comprehensive feature, which fuses the latent representations of visible-thermal modalities.

4.3. Adaptive Decision Fusion

The aforementioned CIF and DFF make decisions individually. They model complementary and discriminative cues, respectively, which output as two response maps. It

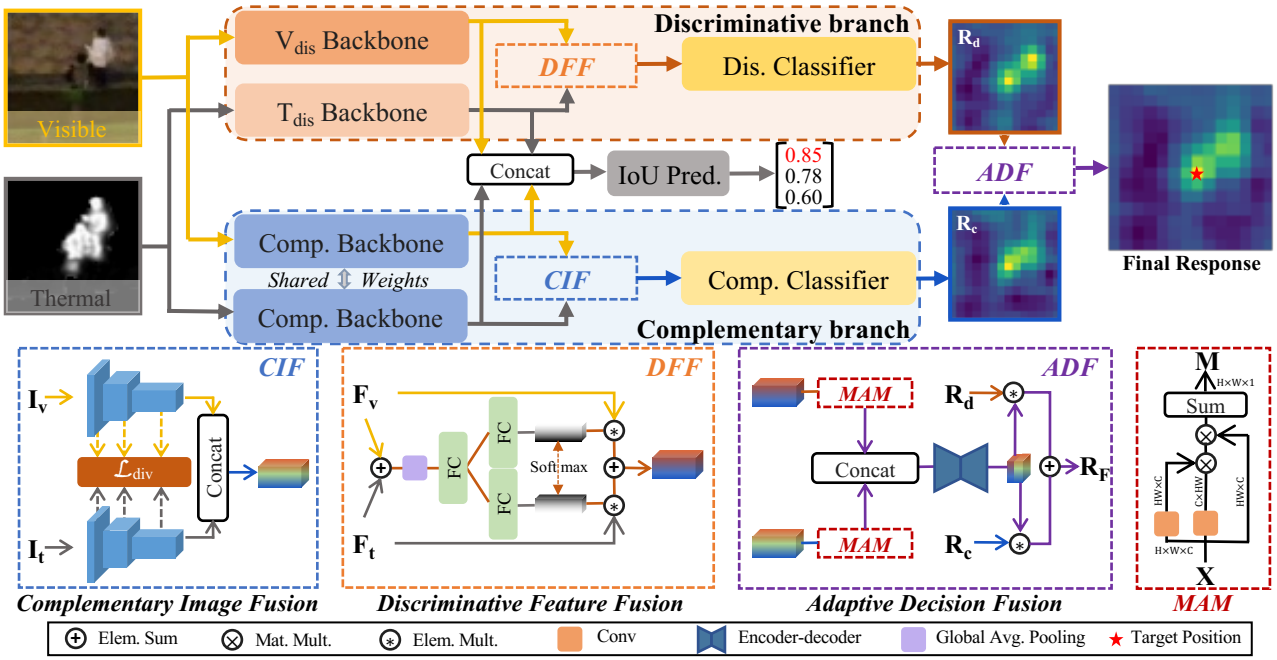

Figure 3. Overview of HMFT. Three fusion types are combined to learn a comprehensive representation and predict accurate results, which consists of Complementary Image Fusion (CIF), Discriminative Feature Fusion (DFF) and Adaptive Decision Fusion (ADF). CIF aims to extract the modality-shared representation, while DFF fuses the individual features to learn the modality-independent map. Both two features are utilized to locate the target, and ADF is to make a final decision by combining the outputs of both branches.

is crucial to determine which cue is reliable for target location. Thus, we introduce Adaptive Decision Fusion (ADF) to fuse these two response maps according to their modality confidences. First, the Modality Aggregation Module (MAM) is designed to obtain the confidence for each modality. MAM is a self-attention network, which produces the modality confidence $\mathbf{M}_d$ and $\mathbf{M}_c$. It mines the modality information in a non-local manner, which takes the whole coordinates into consideration. The MAM process can be formulated as,

$$\mathbf{M} = \mathcal{S}(\mathbf{A} \times \mathbf{X}), \tag{6}$$

$$\mathbf{A} = \mathcal{R}(\phi_{1*1}(\mathbf{X})) \times \mathcal{R}(\varphi_{1*1}(\mathbf{X})), \tag{7}$$

where $\mathbf{X}$ is the input feature, i.e., $\mathbf{P}_a$ or $\mathbf{D}_a$. $\phi_{1*1}(\cdot)$ and $\varphi_{1*1}(\cdot)$ are the learnable $1*1$ Convolution layers. $\mathcal{R}(\cdot)$ and $\mathcal{S}(\cdot)$ are the reshape operation and channel-wise summation, respectively. $\times$ denotes matrix multiplication. When the modality confidences for discriminative and complementary branches $\mathbf{M}_d$ and $\mathbf{M}_c$ are calculated, they are concatenated and sent to a two-layer encoder-decoder network to generate the weight maps $\mathbf{E}_d \in \mathbb{R}^{H*W}$ and $\mathbf{E}_c \in \mathbb{R}^{H*W}$ for the response. Finally, the final response is obtained by $\mathbf{R}_F = \mathbf{R}_d \odot \mathbf{E}_d + \mathbf{R}_c \odot \mathbf{E}_c$, where $\odot$ denotes element-wise production.

4.4. Implementation Details

We take DiMP [2] as our base tracker, with the truncated ResNet50 as the backbone network. We conduct a multi-step training process to fit different purposes in various modules. First, we train both backbones for discriminative and complementary branches. The losses for both branches are expressed as follows,

$$\begin{aligned} \mathcal{L}_d &= \mathcal{L}_{bb} + \beta \mathcal{L}_{cls} \\ \mathcal{L}_c &= \mathcal{L}_{bb} + \beta \mathcal{L}_{cls} + \gamma \mathcal{L}_{div} \end{aligned} \tag{8}$$

where, $\mathcal{L}_{bb}$ and $\mathcal{L}_{cls}$ are the bounding box estimation and target classification losses, respectively, which are detailed in DiMP [2]. β and γ are the weights for classification and divergence loss, which are set to 100. After that, the learned backbones are fixed, and we learn DFF module and the classifiers with Eq. (8). Finally, with all the backbones fixed, we start to learn ADF and IoU prediction modules [13] and fine-tune both classifiers to fit the learned representations. The learning rates for DFF and ADF are $2e^{-5}$ and $2e^{-4}$. We use the same setting in DiMP [46] to train the backbone. We fine-tune the network with multiplying a decreasing factor to the original learning rate, which is set to 0.1. HMFT is implemented on Pytorch platform and run on a single Nvidia RTX Titan GPU with 24G memory.

Table 2. Comparison results for short-term tracking on the proposed dataset and existing RGB-T tracking benchmarks, including GTOT, RGBT210 and RGBT234. The top-three trackers are marked in red, blue and green fonts.

Tracker	VTUAV		GTOT		RGBT210		RGBT234		FPS
	MSR	MPR	MSR	MPR	MSR	MPR	MSR	MPR	
DAFNet [10]	45.8	62.0	71.2	89.1	48.5	72.6	54.4	79.6	21.0
ADRNet [48]	46.6	62.2	73.9	90.4	53.4	77.8	57.1	80.9	25.0
FSRPN [15]	54.4	65.3	69.5	89.0	49.6	68.9	52.5	71.9	30.3
mfDiMP [46]	55.4	67.3	49.0	59.4	52.2	74.9	42.8	64.6	28.0
HMFT (Ours)	62.7	75.8	74.9	91.2	53.5	78.6	56.8	78.8	30.2

5. Experimental Analysis for RGB-T Tracking

5.1. Short-term Evaluation

Overall performance. We select four RGB-T trackers (DAFNet [10], ADRNet [48], FSRPN [15] and mfDiMP [46]). As shown in Tab. 2, HMFT with real-time speed achieves the top performance with 62.7% MSR and 75.8% MPR. The runner-up tracker is mfDiMP, which equips a box regression module to apply scale estimation. Siamese-based tracker (FSRPN) and multi-domain networks (DAFNet and ADRNet) obtain inferior results.

Comparison on existing datasets. We also conduct analysis on three popular RGB-T tracking benchmarks, including GTOT, RGBT210 and RGBT234. To adapt on different benchmarks, we fine-tune HMFT on GTOT and test on RGBT210 and RGBT234. The results are shown in Tab. 2. HMFT obtains satisfying performance in all public benchmarks with real-time speed. Specifically, HMFT achieves the state-of-the-art performance in GTOT and RGBT210, with 74.9% and 53.5% MSR and 91.2% and 78.6% MPR. In RGBT234, our tracker obtains the top-three results against all the competitors with 56.8% and 78.8% in MSR and MPR. All the results show the effectiveness of HMFT, which has great potential for a strong baseline tracker.

5.2. Long-term Evaluation

Overall performance. Given no available long-term RGB-T trackers, following the idea of [6], we implement the HMFT_LT with a global tracker, in which we utilize GlobalTrack [12] as a global detector and RTMDNet [14] as the tracker switcher. When RTMDNet recognizes the absence of target, GlobalTrack is selected to find the target in the whole images. We test all the competitors and the results are shown in Tab. 3. Our long-term variant (HMFT_LT) sets a new baseline for RGB-T long-term tracking, which outperforms short-term verison (HMFT) with 29.8% and 29.4% relative promotion in MSR and MPR, respectively.

5.3. Ablation Study

The ablation analysis of HMFT is shown in Fig. 4. Given that visible images contain more detailed information, which is more capable of recognizing the object, DiMP with the visible image is much more superior to that with the thermal image. Furthermore, both image fusion and feature fusion achieve obvious promotion against trackers with single modality (DiMP-RGB and DiMP-T), where DiMP+CIF and DiMP+DFF gain 3.2% and 4.4% promotion in MSR, respectively. DiMP+CIF+DFF simply averages the responses from complementary and discriminative branches, leading to a slight performance decrease. Our final model (DiMP+CIF+DFF+ADF) achieves 2.6% and 2.7% improvement in MSR and MPR, indicating the adaptation of our decision fusion module.

Table 3. Quantitative comparison against state-of-the-art RGB-T trackers on the long-term subset.

Tracker	MSR	MPR	FPS
ADRNet [48]	17.5	23.5	10.3
DAFNet [10]	18.8	25.3	7.1
mfDiMP [46]	27.2	31.5	25.8
FSRPN [15]	31.4	36.6	36.8
HMFT (Ours)	35.5	41.4	25.1
HMFT_LT (Ours)	46.1	53.6	8.1

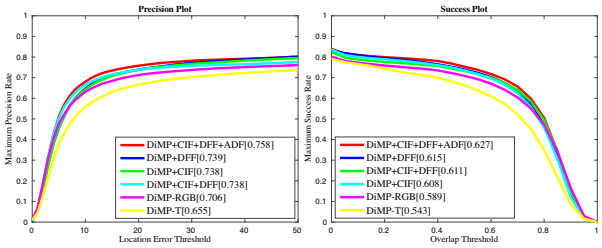

Figure 4. Ablation analysis of HMFT.

5.4. Qualitative Analysis

Fig. 5 provides visualization results between HMFT and the competitors. HMFT shows accurate tracking results on various challenges, such as occlusion, camera movement and scale variation, while other trackers miss the target or cannot estimate the scale properly.

6. Experimental Results on VTUAV-V Subset

We state that VTUAV has great potential for conventional visual tracking task. To unveil the power of RGB tracking, we construct a subset, namely VTUAV-V, which only contains the visible modality for RGB track-

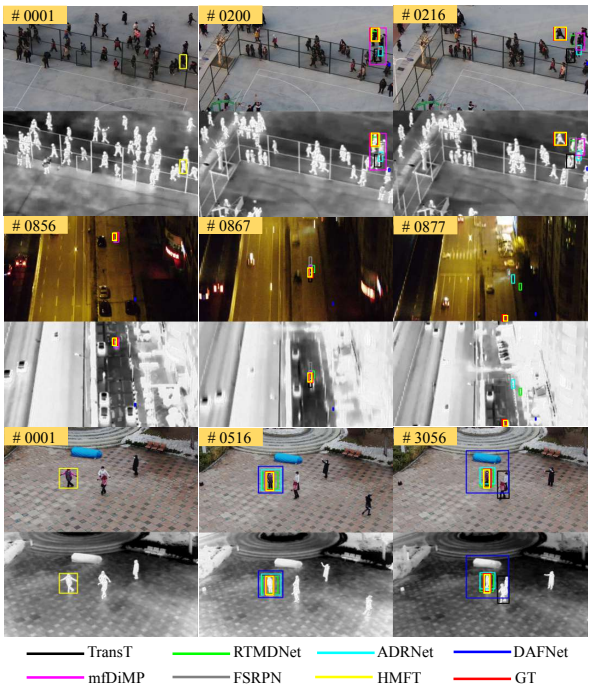

Figure 5. Qualitative comparison of HMFT. Our method shows strong performance on PO, CM and SV.

ing. We evaluate numerous popular short-term and long-term RGB trackers with various frameworks on VTUAV-V dataset, including, LTMU [6], STARK [42], TransT [5], DiMP [2], SiamRPN++ [17], ATOM [8], LightTrack [43], Ocean [52], SiamRPN [18], SiamTPN [41], SiamAPN++ [4], ECO [7], D3S [29], RTMDNet [14], HiFT [3], SPLT [44], SiamFC [1] and GlobalTrack [12]. Since only visible modality is used for tracking. We measure their performance using SR and PR instead of calculating the maximum score between those two modalities. Each tracker is tested without any modification or retraining.

6.1. Short-term Evaluation

As shown in Fig. 6, transformer-based trackers (STARK, TransT) obtain the top performances, which shows the dominative strength in tracking. STARK achieves the top performance with 64.9% and 75.3% in SR and PR, respectively. With the global tracker, LTMU can redetect the target when suffering large camera motion and viewpoint change, leading to a satisfying performance with respect to PR. Trackers with online updating (LTMU, DiMP, ATOM, ECO) obtain the following-up performances and Siamese-based trackers (SiamRPN++, LightTrack, Ocean, SiamRPN, SiamTPN, SiamAPN++, HiFT) achieve inferior results, which indicates the importance of model updating.

6.2. Long-term Evaluation

As shown in Fig. 7, compared with the performances on short-term subset, all the trackers show an decreasing

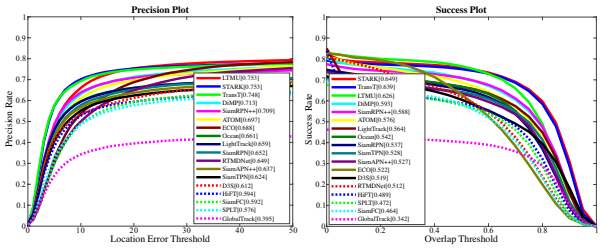

Figure 6. Evaluation results on short-term subset of VTUAV-V.

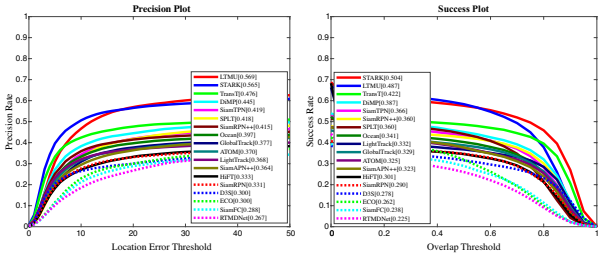

Figure 7. Evaluation results on long-term subset of VTUAV-V.

performance. STARK and LTMU obtain the top performance on SR and PR, respectively. With the global detection mechanism, all long-term trackers (LTMU, GlobalTrack and SPLT) promote their rankings signicantly due to their redetection modules.

7. Conclusion

In this paper, we release a large-scale benchmark for RGB-T tracking. Three main breakthroughs have been achieved. First, we address the issue of few available training data by providing diverse and high-resolution paired RGB-T images captured in various conditions. Second, to the best of our knowledge, this is the first unified RGB-T dataset that considers short-term tracking, long-term tracking and pixel-level prediction to evaluate the trackers comprehensively. Third, we annotate 13 challenges in sequence and frame levels, which can meet the requirement of scene-specific trackers and fully exploit the effectiveness of attributes. Moreover, a new baseline, called HFMT, is designed by combining both image fusion, feature fusion and decision fusion. The remarkable performance on three benchmarks show the complement of those fusion methods, and the importance of adequate training data.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Grant nos. 62022021, 61806037, 61725202, U1903215 and 61829102, and in part by the Science and Technology Innovation Foundation of Dalian under Grant no. 2020JJ26GX036 and Dalian Innovation leader's support Plan under Grant no. 2018RD07.

References

[1] Luca Bertinetto, Jack Valmadre, Joao F. Henriques, Andrea Vedaldi, and Philip H. S. Torr. Fully-convolutional siamese networks for object tracking. In *European Conference on Computer Vision Workshop*, pages 850–865, 2016. 8

[2] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *IEEE International Conference on Computer Vision*, pages 6182–6191, 2019. 2, 6, 8

[3] Ziang Cao, Changhong Fu, Junjie Ye, Bowen Li, and Yiming Li. Hift: Hierarchical feature transformer for aerial tracking. In *IEEE International Conference on Computer Vision*, pages 15457–15466, 2021. 8

[4] Ziang Cao, Changhong Fu, Junjie Ye, Bowen Li, and Yiming Li. Siamapn++: Siamese attentional aggregation network for real-time UAV tracking. In *International Conference on Intelligent Robots and Systems*, pages 3086–3092, 2021. 8

[5] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8126–8136, 2021. 8

[6] Kenan Dai, Yunhua Zhang, Dong Wang, Jianhua Li, Huchuan Lu, and Xiaoyun Yang. High-performance long-term tracking with meta-updater. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6298–6307, 2020. 7, 8

[7] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ECO: Efficient convolution operators for tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6638–6646, 2017. 8

[8] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ATOM: Accurate tracking by overlap maximization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4660–4669, 2019. 8

[9] Heng Fan, Fan Yang, Peng Chu, Yuewei Lin, Lin Yuan, and Haibin Ling. TracKlinic: Diagnosis of challenge factors in visual tracking. In *IEEE Winter Conference on Applications of Computer Vision*, pages 970 – 979, 2021. 3

[10] Yuan Gao, Chenglong Li, Yabin Zhu, Jin Tang, Tao He, and Futian Wang. Deep adaptive fusion network for high performance RGBT tracking. In *IEEE International Conference on Computer Vision Workshop*, pages 1–9, 2019. 7

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[12] Lianghua Huang, Xin Zhao, and Kaiqi Huang. GlobalTrack: A simple and strong baseline for long-term tracking. In *AAAI Conference on Artificial Intelligence*, pages 11037–11044, 2020. 7, 8

[13] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *European Conference on Computer Vision*, pages 784–799, 2018. 6

[14] Ilchae Jung, Jeany Son, Mooyeol Baek, and Bohyung Han. Real-time MDNet. In *European Conference on Computer Vision*, pages 83–98, 2018. 7, 8

[15] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, and et al. The seventh visual object tracking VOT2019 challenge results. In *IEEE International Conference on Computer Vision Workshop*, pages 1–36, 2019. 1, 2, 4, 7

[16] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, and et al. The eighth visual object tracking VOT2020 challenge results. In *European Conference on Computer Vision Workshop*, pages 547–601, 2020. 1

[17] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4282–4291, 2019. 8

[18] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8971–8980, 2018. 8

[19] Chenglong Li, Hui Cheng, Shiyi Hu, Xiaobai Liu, Jin Tang, and Liang Lin. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Transactions on Image Processing*, 25(12):5743–5756, 2016. 1, 2, 4

[20] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. RGB-T object tracking: Benchmark and baseline. *Pattern Recognition*, 96(12):106977, 2019. 1, 2, 3, 4

[21] Chenglong Li, Lei Liu, Andong Lu, Qing Ji, and Jin Tang. Challenge-aware RGBT tracking. In *European Conference on Computer Vision*, pages 222–237, 2020. 2, 3

[22] Chenglong Li, Andong Lu, Aihua Zheng, Zhengzheng Tu, and Jin Tang. Multi-adapter RGBT tracking. In *IEEE International Conference on Computer Vision Workshop*, pages 2262–2270, 2019. 1, 2

[23] Chenglong Li, Xiaohao Wu, Nan Zhao, Xiaochun Cao, and Jin Tang. Fusing two-stream convolutional neural networks for RGB-T object tracking. *Neurocomputing*, 281:78–85, 2018. 1

[24] Chenglong Li, Wanlin Xue, Yaqing Jia, Zhichen Qu, Bin Luo, and Jin Tang. LasHeR: A large-scale high-diversity benchmark for RGBT tracking. *arXiv preprint arXiv:2104.13202*, 2021. 2, 4

[25] Chenglong Li, Nan Zhao, Yijuan Lu, Chengli Zhu, and Jin Tang. Weighted sparse representation regularized graph learning for RGB-T object tracking. In *ACM International Conference on Multimedia*, pages 1856–1864, 2017. 1, 2, 4

[26] Siyi Li and Dit-Yan Yeung. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In *AAAI Conference on Artificial Intelligence*, pages 4140 – 4146, 2017. 3, 4

[27] Yuqi Li, Haitao Zhao, Zhengwei Hu, Qianqian Wang, and Yuru Chen. IVFuseNet: Fusion of infrared and visible light images for depth prediction. *Information Fusion*, 58:1 – 12, 2020. 5

[28] Huchuan Lu and Dong Wang. *Online Visual Tracking*. Springer, 2019. 1

[29] Alan Lukezic, Jiri Matas, and Matej Kristan. D3S-A discriminative single shot segmentation tracker. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7133–7142, 2020. 8

[30] Chengwei Luo, Bin Sun, Ke Yang, Taoran Lu, and Wei-Chang Yeh. Thermal infrared and visible sequences fusion tracking based on a hybrid tracking framework with adaptive weighting scheme. *Infrared Physics and Technology*, 99:265–276, 2019. 2

[31] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for UAV tracking. In *European Conference on Computer Vision*, pages 445–461, 2016. 4

[32] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. TrackingNet: A large-scale dataset and benchmark for object tracking in the wild. In *European Conference on Computer Vision*, pages 300 – 317, 2018. 4

[33] Jingchao Peng, Haitao Zhao, Zhengwei Hu, Yi Zhuang, and Bofan Wang. Siamese infrared and visible light fusion network for RGB-T tracking. *arXiv preprint arXiv:2103.07302*, 2021. 2

[34] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016. 4

[35] Yuankai Qi, Shengping Zhang, Weigang Zhang, Li Su, Qingming Huang, and Ming-Hsuan Yang. Learning attribute-specific representations for visual tracking. In *AAAI Conference on Artificial Intelligence*, pages 8835–8842, 2019. 3

[36] Atousa Torabi, Guillaume Massé, and Guillaume-Alexandre Bilodeau. An iterative integrated framework for thermal-visible image registration, sensor fusion, and people tracking for video surveillance applications. *Computer Vision and Image Understanding*, 116(2):210–221, 2012. 2

[37] Chaoqun Wang, Chunyan Xu, Zhen Cui, Ling Zhou, Tong Zhang, Xiaoya Zhang, and Jian Yang. Cross-modal pattern-propagation for RGB-T tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7064–7073, 2020. 2

[38] Xiao Wang, Xiujun Shu, Shiliang Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. MFGNet: Dynamic modality-aware filter generation for RGB-T tracking. *arXiv preprint arXiv:2107.10433*, 2021. 2

[39] James W.Davis and Vinay Sharma. Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer Vision and Image Understanding*, 106(2-3):162–182, 2007. 2

[40] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1834–1848, 2015. 3

[41] Daitao Xing, Nikolaos Evangeliou, Athanasios Tsoukalas, and Anthony Tzes. Siamese transformer pyramid networks for real-time UAV tracking. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1898–1907, 2022. 8

[42] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *IEEE International Conference on Computer Vision*, pages 10428–10437, 2021. 8

[43] Bin Yan, Houwen Peng, Kan Wu, Dong Wang, Jianlong Fu, and Huchuan Lu. Lighttrack: Finding lightweight neural networks for object tracking via one-shot architecture search. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 15180–15189, 2021. 8

[44] Bin Yan, Haojie Zhao, Dong Wang, Huchuan Lu, and Xiaoyun Yang. 'skimming-perusal' tracking: A framework for real-time and robust long-term tracking. In *IEEE International Conference on Computer Vision*, pages 2385–2393, 2019. 8

[45] Hui Zhang, Lei Zhang, Li Zhuo, and Jing Zhang. Object tracking in RGB-T videos using modal-aware attention network and competitive learning. *Sensors*, 20(2):1–19, 2020. 2

[46] Lichao Zhang, Martin Danelljan, Abel Gonzalez-Garcia1, Joost van de Weijer, and Fahad Shahbaz Khan. Multi-modal fusion for end-to-end RGB-T tracking. In *IEEE International Conference on Computer Vision Workshop*, pages 1–10, 2019. 2, 6, 7

[47] Pengyu Zhang, Dong Wang, and Huchuan Lu. Multi-modal visual tracking: Review and experimental comparison. *arXiv preprint arXiv:2012.04176*, 2020. 1

[48] Pengyu Zhang, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Learning adaptive attribute-driven representation for real-time RGB-T tracking. *International Journal of Computer Vision*, 129:2714–2729, 2021. 2, 3, 7

[49] Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Jointly modeling motion and appearance cues for robust RGB-T tracking. *IEEE Transactions on Image Processing*, 30:3335 – 3347, 2021. 2

[50] Tianlu Zhang, Xueru Liu, Qiang Zhang, and Jungong Han. SiamCDA: Complementarity- and distractor-aware RGB-T tracking based on siamese network. *IEEE Transactions on Circuits and Systems for Video Technology*, 99:1–16, 2021. 2

[51] Xingming Zhang, Xuehan Zhang, Xuedan Du, Xiangming Zhou, and Jun Yin. Learning multi-domain convolutional network for RGB-T visual tracking. In *International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, pages 1–6, 2018. 2

[52] Zhipeng Zhang, Houwen Peng, Jianlong Fu, Bing Li, and Weiming Hu. Ocean: Object-aware anchor-free tracking. In *European Conference on Computer Vision*, pages 771–787, 2020. 8

[53] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Qinghua Hu, and Haibin Ling. Vision meets drones: Past, present and future. *arXiv preprint arXiv:2001.06303*, 2020. 4

[54] Yabin Zhu, Chenglong Li, Yijuan Lu, Liang Lin, Bin Luo, and Jin Tang. FANet: Quality-aware feature aggregation network for RGB-T tracking. *arXiv preprint arXiv:1811.09855*, 2018. 2

[55] Yabin Zhu, Chenglong Li, Bin Luo, Jin Tang, and Xiao Wang. Dense feature aggregation and pruning for RGBT tracking. In *ACM International Conference on Multimedia*, pages 465–472, 2019. 2