Resource-Efficient RGBD Aerial Tracking

Yang, Jinyu; Gao, Shang; Li, Zhe; Zheng, Feng; Leonardis, Ales

Resource-Efficient RGBD Aerial Tracking

Jinyu Yang[1,2,†], Shang Gao[1,†], Zhe Li[1,†], Feng Zheng[1,3*], Aleš Leonardis[2]
[1]Southern University of Science and Technology [2]University of Birmingham [3]Peng Cheng Laboratory
jinyu.yang96@outlook.com gaos2021@mail.sustech.edu.cn zhe.li.cs@outlook.com
f.zheng@ieee.org a.leonardis@cs.bham.ac.uk

Abstract

Aerial robots are now able to fly in complex environments, and drone-captured data gains lots of attention in object tracking. However, current research on aerial perception has mainly focused on limited categories, such as pedestrian or vehicle, and most scenes are captured in urban environments from a birds-eye view. Recently, UAVs equipped with depth cameras have been also deployed for more complex applications, while RGBD aerial tracking is still unexplored. Compared with traditional RGB object tracking, adding depth information can more effectively deal with more challenging scenes such as target and background interference. To this end, in this paper, we explore RGBD aerial tracking in an overhead space, which can greatly enlarge the development of drone-based visual perception. To boost the research, we first propose a large-scale benchmark for RGBD aerial tracking, containing 1,000 drone-captured RGBD videos with dense annotations. Then, as drone-based applications require for real-time processing with limited computational resources, we also propose an efficient RGBD tracker named EMT. Our tracker runs at over 100 fps on GPU, and 25 fps on the edge platform of NVidia Jetson NX Xavier, benefiting from its efficient multimodal fusion and feature matching. Extensive experiments show that our EMT achieves promising tracking performance. All resources are available at https://github.com/yjybuaa/RGBDAerialTracking.

1. Introduction

Aerial robots have been widely used in complex missions. For example, Unmanned Aerial Vehicles (UAVs) equipped with cameras are able to perceive and understand unknown environments and have wide applications on agriculture and surveillance [11, 43]. Specifically, color-based visual tracking with drones has been rapidly developed, thanks to large-scale datasets [27, 43] and dedicated algorithms [2–4, 9, 10, 12, 17, 24, 35]. However, these UAVs merely equipped with color-based sensors generally fail to deal with the challenges in complex environments, such as background clutters and dark scenes, which break the visibility and illumination limitations in color-only domain. For example, current drones have difficulties on tracking a person in dark scenes. While, RGBD tracking is effective to tackle such kinds of tracking failures.

However, for a long time, depth sensors are only incorporated with UAVs to enable aerial autonomy and collision avoidance [14]. Visual perception like RGBD tracking with drones is unexplored due to the multiple limitations. For example, commercial RGBD sensors are strictly limited by application scenarios and depth measurement range. On the other hand, we notice that current UAV tracking datasets record video sequences in the manner of aerial photography [8, 43]. The captured objects mainly focus on pedestrians and vehicles, and the captured scenes are in urban environments from a birds-eye view.

In this work, we explore RGBD aerial tracking from a more practical viewpoint. Different from existing UAV tracking works, we focus on the unexplored overhead space (2 - 5 meters above the ground), aiming to save the ground space greatly with drone-based visual perception. Instead of mainly focusing on people and vehicles, our research can include more generic objects of different categories, such as hands, cups, or balls. Thus, multimodal aerial platforms in this space are very important, as flying robots with short-range perception capabilities can potentially be used in a wider range of scenarios, such as human-robot interaction.

Notably, the new task brings challenges in drone-based visual perception, which can be concluded as follows:

Complex real-world circumstances. The real-world flight comes with complicated and changeable natural environments. On the one hand, the high mobility of drones brings intense pose changes, resulting in huge variations of target scale and considerable motion blurs. Except for the common challenges in visible situations, drone vision also suffers from other problems like low illumination, similar objects and background clutter.

† Equal contribution. ∗ Corresponding author.

Limited onboard computational resources. In practical applications, flying platforms generally require higher efficiency on edge platforms with limited resources, while state-of-the-art trackers can only run on powerful GPUs. Especially for multimodal trackers, the model efficiency is always the least valued in model design.

Real-time practical applications. Real time is a basic requirement in aerial tracking. Moving platforms require real-time responses and real-world applications also require trackers to function in real-time speed. However, most of current state-of-the-art trackers even cannot achieve real-time speed on powerful GPUs, not to mention their real-world applications.

Therefore, to achieve UAV visual tracking with depth, we first build a novel RGBD aerial platform to collect videos. The platform is particularly designed to simulate the environments in real-world applicaitions. The captured videos can comprehensively reflect those challenges to be tackled. Using this aerial platform, a large-scale dataset for **D**rone-based RGB**D** aerial tracking, named **D^{2}Cube**, is built. Some examples in our dataset are given in Fig. 1. In total, 1,000 sequences are provided with dense bounding box annotations. The settings of captured videos cover diverse scenarios in daily life.

Furthermore, we propose an efficient tracker named **EMT** to facilitate the development of on-board RGBD tracking. The proposed EMT can be treated as a strong baseline for on-board multimodal tracking to simultaneously tackle above three issues. Thanks to the efficient multimodal fusion and feature matching, our proposed tracker can successfully balance the tight computational budget and tracking accuracy. We perform extensive experiments in diverse scenarios and various platforms to validate the effectiveness of our EMT. Competitive tracking performance is observed in comparison with state-of-the-art RGB-only and multimodal trackers, in which EMT runs at a high frame rate of over 100 FPS. Practical application tests are given on *NVIDIA Jetson NX Xavier*, where our EMT can run at a frame rate of over 25 FPS. To conclude, our dataset covers complex aerial tracking scenarios and our method shows a promising balance of accuracy, resources and speed.

The contributions are summarised below:

- **New Problem:** We propose a new task of RGBD air tracking for newly defined overhead space (2m - 5m). Unlike previous aerial tracking, this task is more relevant to human life and has wider applications.
- **New Benchmark:** We construct a large-scale high-diversity benchmark for RGBD aerial tracking. The advantage is that much more categories (34 classes) can be considered than existing aerial tracking datasets. As far as we know, this is the first dataset that can test multimodal aerial tracking models.
- **New Baseline:** An efficient tracking baseline is proposed for RGBD aerial tracking, which is the first real-time tracker for efficient on-board multimodal tracking. It performs better than classical UAV trackers and maintains comparable efficiency.

2. Related Work

2.1. Aerial Tracking

In general, aerial tracking, *i.e.*, UAV-based tracking, is to track target objects in consecutive frames with drone-based views. Various drone-based datasets are proposed for color-based object tracking, as shown in Table 1. We notice that existing UAV tracking datasets focus on high-altitude aerial photography with capturing vehicles mainly. For example, the well-known VisDrone [43] and UAVDT [8] both contribute to vehicle tracking in a birds-eye view. The limitations of them are obvious. On the one hand, they are captured at high altitudes, which has a gap with our daily life scenes. On the other hand, the UAVs and cameras can only work under visible conditions. More complex scenarios will lead to flight and data failure. In contrast to the above datasets, our proposed D^{2}Cube contains multimodal information and more diverse scenarios, bringing new challenges to aerial tracking tasks.

At the arithmetic level, UAV-based tracking faces the challenges of both limited computational resources and strict real-time speed requirements, impeding the usage of state-of-the-art trackers. Thus, UAV tracking requires for efficient tracking algorithms. Existing UAV trackers have shown their effectiveness in RGB-based tracking. LightTrack [33] achieves a lightweight tracking framework by using NAS. TCTrack [4] provides a holistic temporal encoding framework to handle temporal contexts in Siamese-based aerial tracking. HCAT [5] achieves high tracking speed thanks to the hierarchical cross-attention transformer. However, unlike color-based UAV tracking, there are much less attentions paid to multimodal tracking efficiency and multimodal tracker's speed on edge devices.

2.2. RGBD Tracking

RGBD tracking gains lots of attention thanks to the development of accessible depth sensors. A series of tracking datasets and baselines are proposed to boost this area [18–22, 36, 38]. As shown in Table 1, there have been some efforts on RGBD tracking datasets. Besides, compared to single-modal trackers, RGBD trackers have to process multimodal information and focus more on cross-modal fusion. Qian *et al.* proposed DAL [28] which designed a depth-aware deep correlation filter. Yan *et al.* proposed DeT [34] which is trainable on RGBD data with two-stream feature extraction backbones. However, RGBD tracking still suffers from bad speed and performance balance. To the best

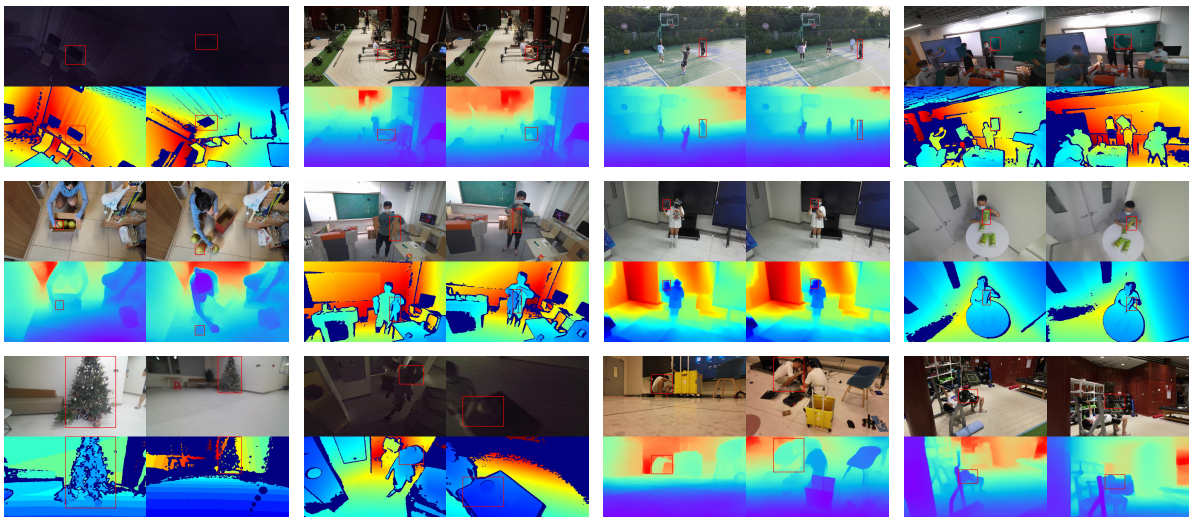

Figure 1. Annotated example video sequences in the proposed dataset. As shown, our D^{2}Cube contains multiple challenges.

Table 1. Comparison of related datasets for aerial tracking and RGBD tracking. T=Thermal, D=Depth, L=Language, A=Audio.

Scope	Dataset	Modality	Object Type	Scenario	Videos	Year
Aerial Tracking	UAV123 [27]	RGB	Generic	Outdoor	123	2016
	VisDrone-SOT [43]	RGB	Human; Vehicle	Outdoor	167	2018
	UAVDT-SOT [8]	RGB	Human; Vehicle	Outdoor	100	2018
	VT-UAV [41]	RGB; T	Generic	Outdoor	500	2021
	WebUAV-3M [40]	RGB; L; A	Generic	Outdoor	4,485	2022
RGBD Tracking	PTB [30]	RGB; D	Generic	Indoor	100	2016
	STC [32]	RGB; D	Generic	Indoor; Outdoor	36	2018
	CDTB [26]	RGB; D	Generic	Indoor; Outdoor	80	2019
	DepthTrack [34]	RGB; D	Generic	Indoor; Outdoor	200	2021
RGBD Aerial Tracking	**D^{2}Cube**	**RGB; D**	**Generic**	**Indoor; Outdoor**	**1,000**	**2022**

of our knowledge, this work is the first one contributing to RGBD tracking efficiency, in which our proposed method can run on edge platforms with real-time speed.

3. Dataset Construction

3.1. Data Collection

Flight platforms. We present our real-world data collection on a handcrafted flight platform, mounted with advanced RGBD cameras, *i.e.*, *Microsoft Azure Kinect DK*, *ZED 2i Stereo Camera*, and *Intel RealSense D455*. The flight platform is to provide the aerial view and the RGBD cameras are to acquire high-quality synchronous color and depth flows. A *Nvidia Jetson NX Xavier* computer running Ubuntu 18.04 is mounted in our UAV for computational support. The overall weight (including LiPo battery and propellers) is about 2.5 kg, with dimensions of $450 \times 450 \times 250$ mm.

RGBD acquisition setups. The following three RGBD acquisition setups are used to increase the dataset diversity in terms of hardware: (i) *Microsoft Azure Kinect DK* is based on Time-of-Flight (ToF) method, measuring depth in a range of 0.5m to 5.46m. It is used for indoor scenarios. (ii) *Intel RealSense D455* uses structure light for depth perception, with an ideal depth measurement range of 0.6m to 6m, designed for both indoor and outdoor scenarios. (iii) *ZED 2i Stereo Camera* uses stereo vision and neural networks to reproduce human vision, enabling depth perception from 0.2m to 20m for outdoor applications. All three devices provide synchronized RGB and depth camera streaming with configurable delay between cameras. RGBD cameras are connected to the drone by pan-tilt, thus the capturing viewpoints can be flexibly changed. All videos are captured under 30 fps, with resolution normalized to 1280×720 pixels.

3.2. Dataset Statistics

Statistics. We provide 1,000 challenging video clips (1,030,097 frames) in total, including 900 sequences for training (929,370 frames), and 100 for testing (100,727 frames), with an average video length of 1030 frames (about 34 seconds). Regarding the training set, we do not provide a further partition and users can split the training and vali-

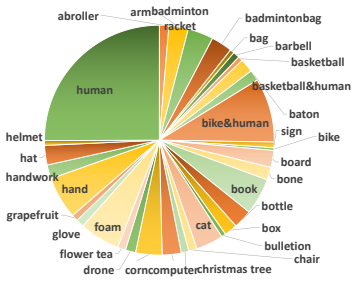

Figure 2. Object classification and distribution in test set.

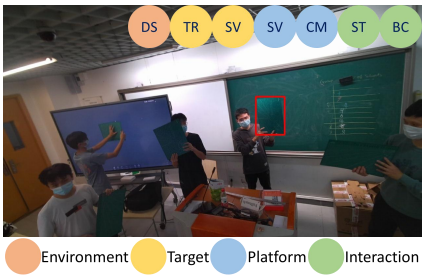

Figure 3. An annotated example with bounding box and attributes.

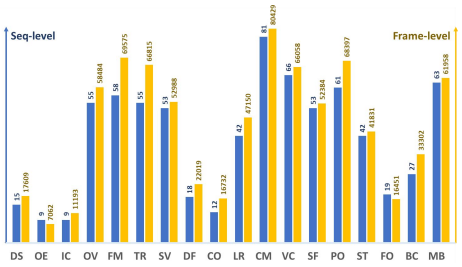

Figure 4. Attribute distribution in our test set.

Table 2. Attributes and corresponding description.

Level	ABB.	Description
Environment	DS	*Dark Scene.* The light is too low to distinguish the target.
	OE	*Overexposure.* The illumination is too high to distinguish the target.
	IC	*Illumination Change.* There are illumination changes during one video sequence.
Target	OV	*Out of View.* Object partially or fully leave the view.
	FM	*Fast Motion.* The average per-frame object motion is larger than 20 pixels.
	TR	*Target Rotation.* Target rotates in plane or out of plane.
	SV	*Scale Variation.* Ratio change of target size between minimum and maximum is more than 50%.
	DF	*Deformation.* The object is deformable.
	CO	*Composite Objects.* The target object is an ensemble of multiple objects (e.g. man with a basketball).
Platform	LR	*Low Resolution.* The ratio of the object area to the image size is lower than 5%.
	CM	*Camera Motion.* The camera moves/shakes.
	MB	*Motion Blur.* The target is blurred due to the motion of itself or the camera.
	VC	*Viewpoint Change.* The viewpoint is not fixed because the capturing angle changes.
	SF	*Sensor Failure.* At least one camera cannot provide useful information.
Interaction	PO	*Partial Occlusion.* The object is partially occluded.
	ST	*Similar Targets.* There are similar objects.
	FO	*Full Occlusion.* The object is fully occluded.
	BC	*Background Clutter.* There are distractors around the target object.

dation sets by themselves.

Objects. Unlike previous drone-based tracking datasets which only contain limited object categories, our D^{2}Cube covers generic objects. Fig. 2 shows the object classes in our test set, in which 34 classes are included. Specifically, we include some classes rarely appeared in the semantic area, *e.g.*, part of an entire object (upper part of a body) or composite object (man holding a basketball). The whole D^{2}Cube includes more than 100 categories and covers diverse objects, it thus is representative for daily scenarios.

Scenarios. In this work, we mainly focus on daily life scenarios. We involve many applicable scenarios for RGBD aerial tracking, *e.g.*, sports, work, service and entertainment, in which aerial robots and depth cameras can both work well. In detail, our recording scenarios cover daily life scenes, including office, bedroom, meeting room, gym, stadium, kitchen, and so on. Both short-term and long-term tracking scenarios are included.

Annotations. As shown in Fig. 1, we provide tight axis-aligned bounding box annotations for the target objects at the frame level. A professional team annotates D^{2}Cube rigorously. The annotation process follows the following rules: (i) if the target appears in the frame, we annotate the visible part of the target by the tightest bounding box. (ii) if the target does not appear in the frame, we will mark this frame with an "target loss" tag.

Attributes. We define 18 tracking challenges in RGBD aerial tracking and classify the attributes in a hierarchical manner. All the attributes are defined in four levels: environment, target, platform and interaction. At different levels, correspondingly there are different challenges. Details of each attribute are given in Table 2. With such a hierarchical classification of different challenges, we can justify what challenges RGBD trackers are indeed suffering from. We also give an annotated example in Fig. 3. The distribution of each attribute in our test set is given in Fig. 4.

4. Efficient Multimodal Tracker

To achieve RGBD tracking on UAV platforms, trackers' ability to run on edge platforms with limited resources is of importance. However, the vast majority of RGBD trackers focus on architectural design with heavy backbones and additional modules. Such complex frameworks cannot satisfy the real-time requirements of aerial tracking. In this section, we propose Efficient Multimodal Tracker (EMT) for RGBD aerial tracking, which discards the heavy backbones and additional modules.

4.1. Multimodal Fusion and Matching Architecture

The proposed EMT contains four main steps, *i.e.*, efficient modality-aware fusion, feature extraction, efficient attention-based feature matching, and target prediction. The overall architecture is shown in Fig. 5.

Efficient modality-aware fusion. Firstly, to speed up the fusion, we design a novel Efficient Modality-Aware Fusion (EMAF) module that can first fuse the raw data from

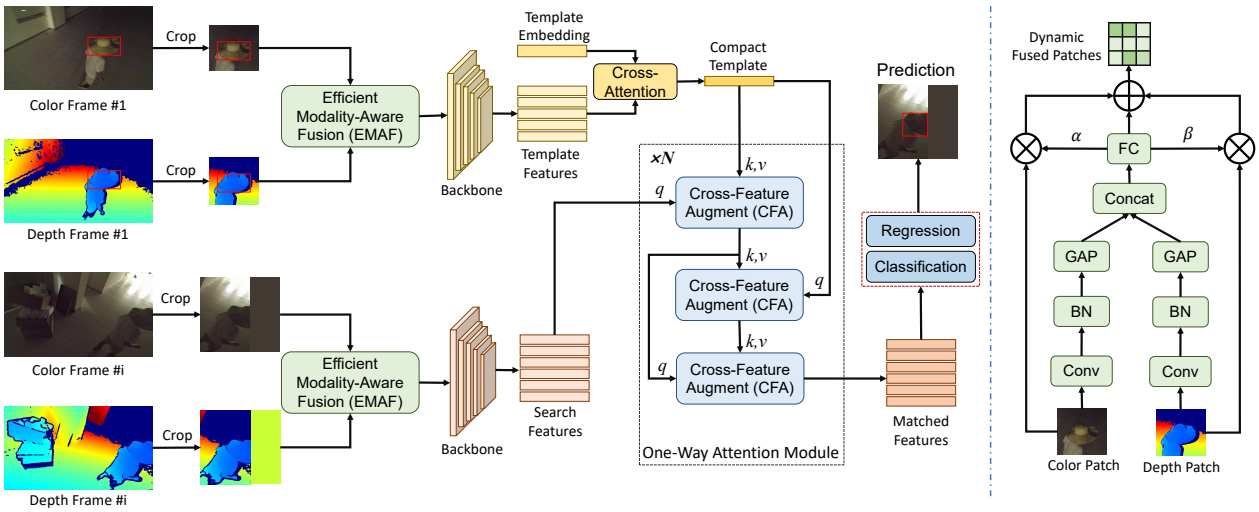

Figure 5. Overview of our proposed Efficient Multimodal Tracker (EMT). Left: Pipeline for EMT. Right: Architecture of Efficient Modality-Aware Fusion (EMAF) module.

multiple modalities at a very early stage. The key to speeding up is that once the high-dimensional features are well integrated, the amount of computation that follows is greatly reduced. In addition, in order to effectively adapt to the scene, we also employ a fusion strategy that can dynamically estimate the importance of features from two modalities. The details of the fusion module to obtain fused image patches $\mathbf{T_{fused}}$ and $\mathbf{X_{fused}}$ are referred to 4.2.

Feature extraction. Secondly, we assume that the fused template patch is $\mathbf{T_{fused}} \in \mathbb{R}_{\mathbf{3\times H_{t0}\times W_{t0}}}$ and the fused search patch is $\mathbf{X_{fused}} \in \mathbb{R}_{\mathbf{3\times H_{x0}\times W_{x0}}}$. Then, we treat them as the input of the parameter-sharing backbone network for feature extraction. A modified ResNet-18 Network [6] is used as backbone network to obtain features maps for templates $f_t \in \mathbb{C} \times \mathbb{H}_t \times \mathbb{W}_t$ and search regions $f_x \in \mathbb{C} \times \mathbb{H}_x \times \mathbb{W}_x$. Here $(H_t, W_t) = (\frac{H_{t0}}{16}, \frac{W_{t0}}{16})$, $(H_x, W_x) = (\frac{H_{x0}}{16}, \frac{W_{x0}}{16})$ and $C = 256$.

Efficient attention-based feature matching. Thirdly, as we obtain the multimodal feature maps for the template and the search region, the next step is to match the corresponding features. To further speed up this procedure, we first use a trainable embedding to reduce the dimension of template features and then design a one-way attention-based fusion module to efficiently fuse the template features and search area features. We give a detailed description of template-to-search matching in Sec. 4.3.

Target prediction. With the template-to-search map, we obtain target predictions by using regression head and classification head. The regression head is to regress the overlap between groundtruth and bounding box candidates. The classification head is to classify the objects and background.

4.2. Efficient Modality-Aware Fusion

Our EMT takes the dual-modal image patches as input, and performs a weighted fusion of modalities online. Unlike the modal independent backbone network design of traditional RGBD trackers, our proposed EMT reduces the model's size through a very early learnable fusion.

Raw patch preparation. Specifically, four image patches will be treated as the input, including the color and depth template patches, and the color and depth patches for search regions. On the one hand, the template image patches $\mathbf{T_{rgb}}$ and $\mathbf{T_{depth}}$ are obtained by expanding the target bounding box of the first frame twice in the video. To effectively enhance discrimination, these patches should include the local surrounding information. And, the perturbation is also added to the target to avoid learning location bias. On the other hand, the patches for search regions $\mathbf{X_{rgb}}$ and $\mathbf{X_{depth}}$ are obtained by expanding the target box in the previous frame by four times instead of the whole original image, which utilizes the temporal context in the video sequence and reduces the computational cost.

Dynamic cross-modal fusion. The aim of this step is to fuse the data and reduce the dimension at a very early stage. The intrinsic reason we can fuse these two types of data at such an early stage is that they are pixel-level aligned in image space. Moreover, it requires to dynamically judge the environment of the tracking target by extracting the global information of the image patches. To this end, we calculate the importance of the two modalities in the current frame based on the context of the two modalities. Based on the importance ratio, the RGB image patch and depth image patch can be fused by the weights for early fusion.

Taking the search branch as an example, the RGB and

depth image patches go through 3×3 convolution layer (Conv) and a batch norm layer (BN) to extract discriminative features. Then, a global pooling layer (GAP) is used to extract the global context of the two modalities. Next, the features are concatenated and as input of the fully connected layers (FC) to output the importance (α and β) of the two modalities in the current frame. Finally, RGB image patches and depth image patches can be fused by importance weights. The process can be formulated as follows:

$$\begin{aligned} \mathbf{X_{rgb}} &= (\mathrm{GAP}(\mathrm{BN}(\mathrm{Conv}(\mathbf{X_{rgb}})), \\ \mathbf{X_{depth}} &= \mathrm{GAP}(\mathrm{BN}(\mathrm{Conv}(\mathbf{X_{depth}}))), \\ \alpha, \beta &= \mathrm{FC}(\mathrm{Cat}(\mathbf{X_{rgb}}, \mathbf{X_{depth}})), \\ \mathbf{X_{fused}} &= \alpha * \mathbf{X_{rgb}} + \beta * \mathbf{X_{depth}}, \end{aligned} \tag{1}$$

where $\mathbf{X_{fused}}$ has the same size of original images.

The immediate benefit of this early fusion is that we avoid extracting features from two modalities by using two separate backbones. Therefore, both the size of memory and the amount of computation have been greatly reduced.

4.3. Efficient Attention-based Feature Matching

For a given template feature f_t, we will use an attention-based module to find corresponding features in f_x for a search region. Naturally, similar to [7], both cross-attention between features and self-attention within features can be directly used. However, to speed computation up, we design a more streamlined network to achieve feature matching.

Compact template representation. To make the template compact, we use a learnable embedding to reduce the dimensions of template vectors f_t. As shown in Fig. 5, through the cross-attention with dimension reduction embedding, we obtain a compact template representation f_t^c.

One-Way Attention (OWA) module. With such a compact template representation f_t^c, we then utilize an efficient matching from template to search area. Here, we merely deploy the cross-attention based Cross-Feature Augment (CFA) module [7] for fusion, in which multi-head cross-attention is used in a residual form. However, CFA is utilized in a one-way manner, due to the fact that we only need the template project on search areas to make a prediction in the search area [5]. OWA can be repeated by several times. With such a one-way cross-attention module, we can achieve feature matching more efficiently.

4.4. Training and Inference

Loss Function. The losses are computed between the outputs of target prediction and groundtruth:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{bbox} + \lambda_3 \mathcal{L}_{GIou}. \tag{2}$$

Here, classification loss $\mathcal{L}_{cls}$ is to dicriminate the object from background. Regression loss consists of $\mathcal{L}_{bbox}$ - Mean Squared Error (MSE) between the predicted bounding boxes and the groundtruth bounding boxes, and $\mathcal{L}_{GIou}$ - generalized IoU loss [29]. We use $\lambda_1 = 0.8344$, $\lambda_2 = 5$, $\lambda_3 = 2$ in our experiments following [7].

Training phase. The training process follows the standard training recipe of current trackers [13, 28, 34]. We use ResNet-18 [16] pretrained on the ImageNet as our backbone and fine-tune the whole tracking framework with the training set of our proposed D^{2}Cube and the training data recipe of [34]. We randomly sample two frames within an interval of 30 frames in sequences as template and search region. Subsequently, templates and search regions are obtained with some data augmentation such as jitter, brightness change, and then resized to 128×128 and 256×256 pixels, respectively. The template embedding size is 16. The model is trained with AdamW [25] optimizer, and the learning rate for the backbone and EMAF module are $1e-5$ and $1e-4$, respectively. Weight decay is $1e-4$. The learning rate decays 10 times every 50 epochs. We sample 128,000 pairs in each epoch and the whole tracker is trained for 100 epochs on a single 32GB Tesla V100 GPU with a batch size of 128.

Tracking phase. During inference, the template and search image are resized to fixed size. Dynamic cross-modal fusion module performs early fusion of the two modalities. After feature extraction and correlation, the prediction head outputs 256 bounding boxes and classification scores. A penalty window is employed to filter distractors.

5. Experiments

5.1. Experimental Settings

Hardware. All comparison experiments, except for the onboard tests, are executed on a single 32GB Tesla V100 GPU. A widely-used UAV onboard processor NVidia Jetson NX Xavier is used for onboard tests.

Evaluation protocols. We follow the evaluation principles in long-term RGBD tracking from VOT challenge [19]. One-Pass Evaluation (OPE) is used to test trackers' performance on our proposed D^{2}Cube. At frame t, θ_t is a prediction confidence score and τ_θ is a classification threshold. If the predicted θ_t is not below τ_θ, $A_t(\tau_\theta)$ is used to denote the corresponding prediction. Otherwise, we set $A_t(\tau_\theta) = \varnothing$. Thus, $\Omega(A_t(\tau_\theta), G_t)$ indicates the intersection-over-union (IoU) between the prediction result $A_t(\tau_\theta)$ and the groundtruth G_t. We here calculate the precision-recall over the whole test set as follows:

$$\begin{aligned} Pr(\tau_\theta) &= \frac{1}{M} \sum_{i=1}^{M} \frac{1}{N_p} \sum_{A_t(\tau_\theta) \neq \varnothing} \Omega(A_t(\tau_\theta), G_t), \\ Re(\tau_\theta) &= \frac{1}{M} \sum_{i=1}^{M} \frac{1}{N_g} \sum_{G_t \neq \varnothing} \Omega(A_t(\tau_\theta), G_t), \end{aligned} \tag{3}$$

Table 3. Performance comparison of state-of-the-art RGB aerial trackers on D^{2}Cube dataset. The top 3 results are shown in red, green, and blue. Speed in FPS (frames per second).

Method	Pr	Re	F-score	Speed
LightTrack [33]	0.500	0.531	0.515	119.5
HiFT [2]	0.404	0.430	0.417	66.9
TCTrack [4]	0.416	0.448	0.432	78.1
SiamAPN [9]	0.413	0.441	0.427	140.2
SiamAPN++ [3]	0.411	0.436	0.423	114.9
DaSiamRPN [44]	0.392	0.415	0.403	200.6
HCAT [5]	0.544	0.578	0.561	148.2
UDT+ [31]	0.387	0.412	0.399	50.4
SiamRPN++ [23]	0.459	0.488	0.473	83.3
UDAT-CAR [39]	0.462	0.492	0.476	33.9
EMT	0.653	0.609	0.630	120.3

where N_p denotes the number of frames in which the target is predicted visible in a video sequence, and N_g denotes the number of frames in which the target is indeed visible. $Pr(\tau_\theta)$ and $Re(\tau_\theta)$ denote the precision and recall metrics for M test videos. F-score is obtained by $F(\tau_\theta) = \frac{2Re(\tau_\theta)Pr(\tau_\theta)}{Re(\tau_\theta)+Pr(\tau_\theta)}$.

5.2. Comparison with Aerial Trackers

To show the superiority of multimodal tracking in aerial tracking area, we compare the performance of our proposed EMT with existing state-of-the-art aerial trackers. Results are given in Table 3. According to both tracking accuracy and speed, EMT has performed favorably against other state-of-the-art deep aerial trackers. As shown, EMT outperform UDAT [39], TCTrack [4] and HCAT [5] on F-score by 15.4%, 19.8% and 6.9%, with maintaining high speed. The huge performance differences between EMT and sota aerial trackers show the effectiveness of depth information, especially when trackers work in complex scenarios.

5.3. Comparison with RGBD Trackers

We also compare our model with state-of-the-art RGBD trackers. Our EMT significantly beats most state-of-the-art RGBD trackers and is on par with ProTrack [37] on tracking accuracy. Specifically, EMT outperforms DeT [34] and DMT [13] by 3.3% and 5.4% on F-score. Besides, we compare the efficiency between EMT and state-of-the-art RGBD trackers. Here, we calculate the MACs, parameters and tracking speed for a fair comparison on efficiency. EMT achieves comparable performance with ProTrack with 15× fewer params, 25× fewer MACs and 20× higher speed. Therefore, our EMT can definitely achieve a balance of accuracy, resources and speed.

Table 4. Performance comparison of state-of-the-art RGBD trackers on D^{2}Cube dataset. The top 3 results are shown in red, green, and blue. Speed in FPS (frames per second).

Method	DAL [28]	TSDM [42]	DeT [34]	DMT [13]	ProTrack [37]	EMT
Pr	0.529	0.521	0.608	0.584	0.669	0.653
Re	0.565	0.492	0.587	0.569	0.644	0.609
F-score	0.547	0.506	0.597	0.576	0.656	0.630
MACs	15.78G	74.08G	30.57G	40.44G	82.58G	3.43G
Params	19.60M	114.59M	34.63M	38.97M	159.61M	10.05M
Speed	21.3	18.2	26.8	25.5	5.4	120.3

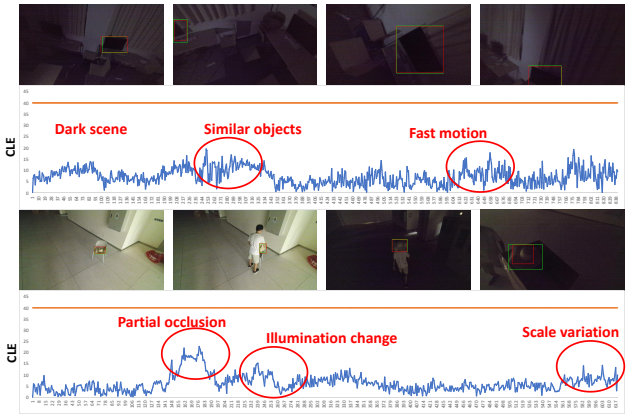

Figure 6. The proposed EMT is tested on the UAV platform with Nvidia NX Xavier. The tracking results and ground truth are marked with red and green box respectively.

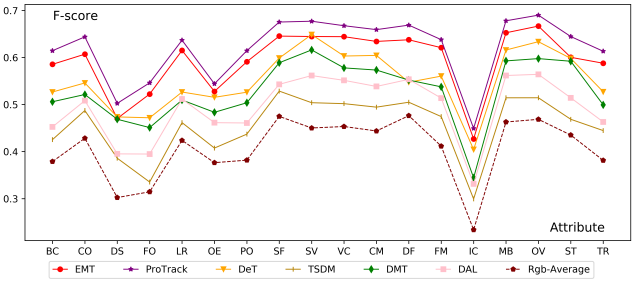

Figure 7. Attribute-based performance comparison on D^{2}Cube.

5.4. On-board Tests

We deploy representative trackers on a commonly-used UAV onboard processor *NVIDIA Jetson NX Xavier* to simulate real-world UAV tracking circumstances. With onboard tests, trackers' real-time capabilities can be evaluated and verified. Fig. 6 shows several tests of our EMT in some challenging real-world tests. As shown, the tests cover multiple challenging scenarios, *e.g.*, dark scenes, fast motion, similar objects and so on. While our EMT can perform successful tracking with an on-board speed of 25 fps. Center Location Error (CLE) refers to the center error between the predictions and groundtruth. We set the error to be within 40 pixels for successful tracking in real-world applications.

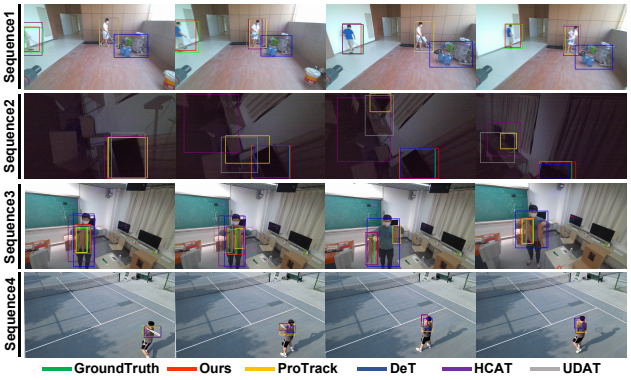

Figure 8. Qualitative results of representative RGB and RGBD trackers on D^{2}Cube dataset.

Table 5. Ablation study on different ways for cross-modal fusion.

Method	Pr	Re	F-score	Speed
Add(RGB,Depth)	0.574	0.415	0.472	147.5
Mean(RGB,Depth)	0.531	0.432	0.476	144.6
Max(RGB,Depth)	0.484	0.396	0.435	140.7
EMAF (Proposed)	0.653	0.609	0.630	120.3

5.5. Attribute-based Performance

We also investigate trackers' performance against different kinds of challenges according to our annotated attributes. As shown in Fig. 7, RGBD trackers outperform RGB-only trackers in all attributes, especially in the case of attributes like *dark scenes* and *illumination change*, with which the target appearance is not such informative. This verifies that the addition of depth information enhances the discriminative ability of trackers in complex environments. Among RGBD tackers, EMT achieves comparable performance with ProTrack, while the model size is 10× smaller and the speed is 20× faster. Besides, EMT far outperforms other sotas in terms of uav-specific challenges, *e.g*, *low resolution*, *camera motion*, *fast motion*, *viewpoint change*, and *sensor failure*. It confirms that EMT can maintain high performance in complex UAV scenarios. We give some result visualization in Fig. 8 to show the qualitative comparison of representative RGB and RGBD trackers against difficulties.

5.6. Ablation Study

Different ways for cross-modal fusion. We investigate the impact of using different methods for cross-modal fusion. As shown in Table 5, common operations, *i.e.*, *add*, *mean* and *max*, show relatively lower performance with F-score degradation of over 10%, compared to the proposed EMAF. This demonstrates that our module can dynamically determine the importance of two modalities in terms of different environments and perform an effective fusion.

Different dimensions of template embedding. As we utilize a compact representation for template, we also investigate the impact of different template dimensions. As reported in Table 6, the 16-dimension embedding gives similar performance to 32-dimension one, while both of them are much more higher than the 4-dimension one, confirming that the modality information is redundant [1, 15] and our compact representation is efficient and effective.

Table 6. Ablation study on the dimension of template features.

Dimension	Pr	Re	F-score	Speed
4	0.467	0.421	0.443	122.5
16 (Default)	0.653	0.609	0.630	120.3
32	0.653	0.604	0.628	114.6

Table 7. Ablation study on the number of OWA modules.

OWA modules	Pr	Re	F-score	Params	Speed
1	0.579	0.543	0.561	7.42M	135.5
2 (Default)	0.653	0.609	0.630	10.05M	120.3
4	0.569	0.506	0.536	15.31M	87.1

Different numbers of One-Way Attention (OWA) modules. In our experiments, we use the one-way attention modules twice. Table 7 gives the performance comparison with different numbers of OWA modules. As reported, two OWA modules perform best, exceeding the one-module approach by 19% with high speed. However, as the number of OWA modules increases to 4, the model performance decreases. It can be explained that too many OWA modules may force the model to aggregate the attention on the invalid information, *e.g.*, the failed value in depth images.

6. Conclusions

In this paper, to explore aerial perception in overhead space, we define a new RGBD aerial tracking task. Compared to the previous research scenario, this new task enables more complex drone-based perception. To validate models for this task, we collect a large-scale dataset covering more scenarios and categories than existing aerial tracking datasets. To facilitate research, a strong baseline has been proposed for the RGBD aerial tracking task, and experimental results on our new dataset clearly demonstrate the efficiency and effectiveness of the proposed model.

Limitations. As this is the first work dedicated to RGBD aerial tracking, the research is still at preliminary stage, we will consider enlarging the dataset and extending the method in serial works to facilitate the whole community.

Acknowledgements

This work was supported by the National Key R&D Program of China (Grant NO. 2022YFF1202903) and the National Natural Science Foundation of China (Grant NO. 62122035 and 61972188).

References

[1] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. *arXiv preprint arXiv:2204.01678*, 2022. 8

[2] Ziang Cao, Changhong Fu, Junjie Ye, Bowen Li, and Yiming Li. Hift: Hierarchical feature transformer for aerial tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15457–15466, 2021. 1, 7

[3] Ziang Cao, Changhong Fu, Junjie Ye, Bowen Li, and Yiming Li. Siamapn++: Siamese attentional aggregation network for real-time uav tracking. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3086–3092. IEEE, 2021. 1, 7

[4] Ziang Cao, Ziyuan Huang, Liang Pan, Shiwei Zhang, Ziwei Liu, and Changhong Fu. Tctrack: Temporal contexts for aerial tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14798–14808, 2022. 1, 2, 7

[5] Xin Chen, Ben Kang, Dong Wang, Dongdong Li, and Huchuan Lu. Efficient visual tracking via hierarchical cross-attention transformer. *arXiv preprint arXiv:2203.13537*, 2022. 2, 6, 7

[6] Xin Chen, Dong Wang, Dongdong Li, and Huchuan Lu. Efficient visual tracking via hierarchical cross-attention transformer. *arXiv preprint arXiv:2203.13537*, 2022. 5

[7] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8126–8135, 2021. 6

[8] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 370–386, 2018. 1, 2, 3

[9] Changhong Fu, Ziang Cao, Yiming Li, Junjie Ye, and Chen Feng. Onboard real-time aerial tracking with efficient siamese anchor proposal network. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2021. 1, 7

[10] Changhong Fu, Adrian Carrio, Miguel A Olivares-Mendez, Ramon Suarez-Fernandez, and Pascual Campoy. Robust real-time vision-based aircraft tracking from unmanned aerial vehicles. In *2014 ieee international conference on robotics and automation (ICRA)*, pages 5441–5446. IEEE, 2014. 1

[11] Changhong Fu, Kunhan Lu, Guangze Zheng, Junjie Ye, Ziang Cao, and Bowen Li. Siamese object tracking for unmanned aerial vehicle: A review and comprehensive analysis. *arXiv preprint arXiv:2205.04281*, 2022. 1

[12] Changhong Fu, Ramon Suarez-Fernandez, Miguel A Olivares-Mendez, and Pascual Campoy. Real-time adaptive multi-classifier multi-resolution visual tracking framework for unmanned aerial vehicles. *IFAC Proceedings Volumes*, 46(30):99–106, 2013. 1

[13] Shang Gao, Jinyu Yang, Zhe Li, Feng Zheng, Aleš Leonardis, and Jingkuan Song. Learning dual-fused modality-aware representations for rgbd tracking. *arXiv preprint arXiv:2211.03055*, 2022. 6, 7

[14] Botao He, Haojia Li, Siyuan Wu, Dong Wang, Zhiwei Zhang, Qianli Dong, Chao Xu, and Fei Gao. Fast-dynamic-vision: Detection and tracking dynamic objects with event and depth sensing. In *IROS*, pages 3071–3078. IEEE, 2021. 1

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 8

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[17] Ziyuan Huang, Changhong Fu, Yiming Li, Fuling Lin, and Peng Lu. Learning aberrance repressed correlation filters for real-time uav tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2891–2900, 2019. 1

[18] Matej Kristan, Aleš Leonardis, Jiří Matas, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Hyung Jin Chang, Martin Danelljan, Luka Čehovin Zajc, Alan Lukežič, et al. The tenth visual object tracking vot2022 challenge results. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 431–460. Springer, 2023. 2

[19] Matej Kristan, Aleš Leonardis, Jiří Matas, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Martin Danelljan, Luka Čehovin Zajc, Alan Lukežič, Ondrej Drbohlav, et al. The eighth visual object tracking vot2020 challenge results. In *European Conference on Computer Vision*, pages 547–601. Springer, 2020. 2, 6

[20] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka ˇCehovin Zajc, Tomas Vojir, Goutam Bhat, Alan Lukezic, Abdelrahman Eldesokey, et al. The sixth visual object tracking vot2018 challenge results. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 2

[21] Matej Kristan, Jiří Matas, Aleš Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Hyung Jin Chang, Martin Danelljan, Luka Cehovin, Alan Lukežič, et al. The ninth visual object tracking vot2021 challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2711–2738, 2021. 2

[22] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kamarainen, Luka ˇCehovin Zajc, Ondrej Drbohlav, Alan Lukezic, Amanda Berg, et al. The seventh visual object tracking vot2019 challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[23] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4282–4291, 2019. 7

[24] Yiming Li, Changhong Fu, Fangqiang Ding, Ziyuan Huang, and Geng Lu. Autotrack: Towards high-performance visual tracking for uav with automatic spatio-temporal regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11923–11932, 2020. 1

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[26] Alan Lukezic, Ugur Kart, Jani Kapyla, Ahmed Durmush, Joni-Kristian Kamarainen, Jiri Matas, and Matej Kristan. Cdtb: A color and depth visual object tracking dataset and benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10013–10022, 2019. 3

[27] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for UAV tracking. In *ECCV (1)*, volume 9905 of *Lecture Notes in Computer Science*, pages 445–461. Springer, 2016. 1, 3

[28] Yanlin Qian, Song Yan, Alan Lukežič, Matej Kristan, Joni-Kristian Kämäräinen, and Jiří Matas. DAL : A deep depth-aware long-term tracker. In *International Conference on Pattern Recognition*, 2020. 2, 6, 7

[29] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 6

[30] Shuran Song and Jianxiong Xiao. Tracking revisited using rgbd camera: Unified benchmark and baselines. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 233–240, 2013. 3

[31] Ning Wang, Yibing Song, Chao Ma, Wengang Zhou, Wei Liu, and Houqiang Li. Unsupervised deep tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1308–1317, 2019. 7

[32] Jingjing Xiao, Rustam Stolkin, Yuqing Gao, and Aleš Leonardis. Robust fusion of color and depth data for rgb-d target tracking using adaptive range-invariant depth models and spatio-temporal consistency constraints. *IEEE Transactions on Cybernetics*, 48(8):2485–2499, 2017. 3

[33] 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 *CVPR 2021*, June 2021. 2, 7

[34] Song Yan, Jinyu Yang, Jani Kapyla, Feng Zheng, Ales Leonardis, and Joni-Kristian Kamarainen. Depthtrack: Unveiling the power of rgbd tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10725–10733, 2021. 2, 3, 6, 7

[35] Jinyu Yang, Wenrui Ding, Chunlei Liu, and Zechen Ha. A saliency-based object tracking method for UAV application. In *PRCV (4)*, volume 11259 of *Lecture Notes in Computer Science*, pages 115–125. Springer, 2018. 1

[36] Jinyu Yang, Zhe Li, Song Yan, Feng Zheng, Aleš Leonardis, Joni-Kristian Kämäräinen, and Ling Shao. Rgbd object tracking: An in-depth review. *arXiv preprint arXiv:2203.14134*, 2022. 2

[37] Jinyu Yang, Zhe Li, Feng Zheng, Ales Leonardis, and Jingkuan Song. Prompting for multi-modal tracking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3492–3500, 2022. 7

[38] Jinyu Yang, Zhongqun Zhang, Zhe Li, Hyung Jin Chang, Ales Leonardis, and Feng Zheng. Towards generic 3d tracking in RGBD videos: Benchmark and baseline. In *ECCV (22)*, volume 13682 of *Lecture Notes in Computer Science*, pages 112–128. Springer, 2022. 2

[39] Junjie Ye, Changhong Fu, Guangze Zheng, Danda Pani Paudel, and Guang Chen. Unsupervised domain adaptation for nighttime aerial tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8896–8905, 2022. 7

[40] Chunhui Zhang, Guanjie Huang, Li Liu, Shan Huang, Yinan Yang, Yuxuan Zhang, Xiang Wan, and Shiming Ge. Webuav-3m: A benchmark unveiling the power of million-scale deep uav tracking. *arXiv preprint arXiv:2201.07425*, 2022. 3

[41] Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale benchmark and new baseline. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8876–8885, 2022. 3

[42] Pengyao Zhao, Quanli Liu, Wei Wang, and Qiang Guo. Tsdm: Tracking by siamrpn++ with a depth-refiner and a mask-generator. In *2020 25th International Conference on Pattern Recognition*, pages 670–676. IEEE, 2021. 7

[43] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 1, 2, 3

[44] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 101–117, 2018. 7