Siamese Transformer Network for Real-Time Aerial Object Tracking

HAIJUN WANG AND SHENGYAN ZHANG
Aviation Information Technology Research and Development, Binzhou University, Binzhou 256603, China

Corresponding author: Haijun Wang (whjkyx@163.com)

This work was supported in part by the Shandong Provincial Natural Science Foundation, China, under Grant ZR2020MF142 and Grant ZR2019PF021; in part by the Doctoral Foundation of Binzhou University under Grant 2021Y04; in part by the Dual Service Projects of Binzhou University under Grant BZXYSFW201805; and in part by the Research Fund Project of Binzhou University under Grant 2019ZD03 and Grant BZXYL1803.

ABSTRACT Recently, deep learning (DL) based trackers have attracted tremendous interest for their high performance. Despite the remarkable success, most trackers utilizing deep convolution features commonly neglect tracking speed, which is crucial for aerial tracking on mobile devices. In this paper, we propose an efficient and effective transformer based aerial tracker in the framework of Siamese, which inherits the merits from both transformer and Siamese architectures. Specifically, the outputs from multiple convolution layers are fed into transformer to construct robust features of template patch and search patch, respectively. Consequently, the interdependencies between low-level information and semantic information are interactively fused to improve the ability of encoding target appearance. Finally, traditional depth-wise cross correlation is introduced to generate a similarity map for object location and bounding box regression. Extensive experimental results on three popular benchmarks (DTB70, UAV123@10fps, and UAV20L) have demonstrated that our proposed tracker outperforms other 12 state-of-the-art trackers and achieves a real-time tracking speed of 71.3 frames per second (FPS) on GPU, which can be applied in mobile platform.

INDEX TERMS Aerial object tracking, transformer, self-attention, Siamese network, deep neural network.

I. INTRODUCTION Aerial object tracking [1], [2], [3] has gained considerable attention in recent years because of its application in motion object analysis, aerial cinematography, military monitoring and maneuvering targets tracking. Given the initial state of target, the objective of unmanned aerial vehicle (UAV) tracking is to predict the position and size of target in the subsequent frames captured by onboard cameras. Although great progress has been made in the past decades [4], [5], [6], [7], designing an efficient and effective aerial object tracker remains a challenging task due to harsh calculation resources and factors such as full occlusion, abrupt target motion, low resolution and scale variation, etc. Hence, an efficient and effective aerial object tracker is desired to cope with the aforementioned problems.

Existing aerial object tracking approaches mainly concentrate on discriminative correlation filters (DCF)-based trackers [8] and DL-based trackers [9]. DCF-based trackers have attracted widespread attention due to their high computational efficiency and promising tracking results. These methods learn a DCF from a limited set of training samples, which derive from cyclically shifted samples. The circulant structure achieves highly efficient training and tracking by the use of Fast Fourier Transform (FFT). However, this circular shift technique also introduces unwanted boundary effects which degrade the tracking performance. Furthermore, most DCF-based trackers may suffer from model drift problem, which will result in losing targets in complex scenes. Thus, despite the incredible tracking speed, DCF-based trackers have inferior performance. These drawbacks make DCF-based trackers impractical to UAV platforms for aerial tracking.

The recent ten years have witnessed considerable progress in DL-based trackers. Through deeper networks and complex
structures, these methods can promote tracking robustness greatly. In spite of the excellent tracking performance, most of DL-based trackers [10], [11] using deep neural network require high complexity for features extraction and suffer from low efficiency, which limit their real-world applications on mobile devices. In order to find a good balance between performance and speed, Siamese-based trackers are developed for real-time tracking while lowering redundancy. The pioneering Siamese-based tracker is the fully convolutional Siamese network (SiamFC) [12], which calculates the similarity between template patch and search patch. The great majority of subsequent Siamese-based trackers [13], [14] are developed from SiamFC. In order to meet the needs of real-time tracking requirements, a series of lightweight models are adopted in Siamese-based trackers for fast features extraction. However, it is almost impossible for lightweight model to extract robust features, which are crucial for tracking targets precisely from mobile platform.

Recently, transformer has achieved impressive success in visual tracking task. Most transformer-based tracking algorithms utilize ResNet50 [15] for features extraction, which constrains their application in the field of aerial tracking. In this paper, motivated by the idea of transformer, we alleviate the above problems by designing a
hierarchical features fusion network and propose a novel Siamese transformer-based tracker for real-time aerial tracking (named STAT). We use AlexNet as the lightweight backbone to extract hierarchical features, which deals with the loss of efficiency caused by traditional deep network. The feature fusion mechanism through transformer with an encoder-decoder structure effectively integrates features from multiple layers to correlation for response generation. FIGURE 1 gives some representative tracking results, demonstrating that our STAT method outperforms better than other competing trackers.

To be specific, in order to obtain good generalization performance of target feature, we feed hierarchical features from the last three convolutional layers into transformer. Therefore, the features from the shallow layer are used to enhance the discriminability and the semantic features from the deeper layer are utilized to deal with low resolution objects. Finally, a robust target feature is achieved by transformer fusion module, which is able to cope with complex scenes with high efficiency and effectiveness. The main contributions of our STAT tracker in this work are summarized as follows,

- We propose a novel transformer-based tracker in the framework of Siamese, consisting of multi-layer features extraction, hierarchical features fusion and head prediction modules.
- We simultaneously take account of hierarchical features from a lightweight network to better explore potential relationships amongst multi-level features. The higher spatial resolution information from the earlier layer and the semantic information from the latter layer are integrated to one whole, which can obtain a robust fused feature information.
- Extensive experiments on three famous UAV benchmarks prove that our proposed STAT tracker outperforms favorably other state-of-the-art trackers and achieves an average speed of 71.3 FPS on GPU, which is suitable for real-time aerial tracking.

The remainder of this paper is organized as follows. We review the related survey of aerial tracking in Section II. In Section III, we introduce the proposed aerial tracking algorithm in details. Section IV demonstrates the experimental results and discussion. We finally conclude this paper in Section V.

II. RELATED WORK

In this section, we give a few relevant state-of-the-art trackers which are closely related to our aerial tracking method. For comprehensive review, we refer the interested readers to [8], [9].

A. CORRELATION FILTER BASED TRACKING

Recently, DCF-based trackers have been widely utilized in aerial tracking due to their high computational efficiency. With the use of fast Fourier transform, DCF-based trackers are able to achieve fast training and detection. Bolme et al. [16] first introduced correlation filter into the field of visual tracking through minimum output sum of squared error (MOSSE). This method can achieve a speed of more than 600 FPS. Henriques et al. [17] proposed a circulant structure with kernels (CSK) to improve the performance of MOSSE tracker. Danelljan et al. [18] presented the discriminative scale-space tracker (DSST), which can estimate the scale of target by training another single dimension filter. Li et al. [19] built an augmented memory for correlation filter (AMCF), which utilizes the historical views and the current view to adapt the rapid change of appearance. Lin et al. [4] leveraged temporal information to develop response reasoning-based correlation filter (ReCF), which is robust to intricate UAV tracking scenarios. In spite of high efficiency, the accuracy and robustness of DCF-based trackers only using hand-crafted features can hardly meet the need of UAV tracking in challenging scenarios. The recent 10 years have witnessed the growing interest in deep neural networks on visual tracking problems. Fu et al. [20] combined convolutional features and hand-crafted features to enhance expressions for UAV target and presented multi-kernelized correlators (MKCT). Li et al. [21] adopted both the hand-crafted and deep features to select keyframes imposed on context learning intermittently (KAOT). Although DCF-based trackers using deep features improved tracking performance greatly, these methods ran at low speed and can not meet the requirements of UAV tracking.

B. SIAMESE BASED TRACKING

In order to find a good balance between speed and accuracy, Siamese network based trackers flourished in recent years and attracted widespread attention both at home and abroad. Bertinetto et al. [12] first developed a two-branch fully-convolutional Siamese network for tracking (SiamFC), which compares the similarity between target template and search region, and finds the tracking results by a score map. Liu et al. [22] proposed a multi-level similarity network in the framework of Siamese network for thermal infrared visual tracking (MLSSNet). In order to solve the scale variation of target effectively, Li et al. [11] introduced region proposal network (RPN) into SiamFC tracking framework (SiamRPN) and trained classification branch and regression branch simultaneously. Li et al. [10] applied deeper network as the backbone and further promoted the tracking robustness (SiamRPN++). Yang et al. [23] introduced an attention network into SiamRPN++ framework and achieved the state-of-the-art performance (SiamAtt). All the methods mentioned above based on RPN are anchor-based trackers, which need to estimate the offsets of bounding boxes and are easily affected by numbers, sizes, and aspect ratios of predefined anchor boxes. To cope with these problems, anchor-free trackers based on Siamese framework were developed to directly predict the location of objects. Guo et al. [24] built a simple Siamese based classification and regression (SiamCAR) network and demonstrated that anchor-free trackers
can perform better than anchor-based trackers. Xu et al. [25] proposed a novel fully convolutional Siamese tracker++ (SiamFC++) with accurate target estimation guidelines. Fu et al. [5] designed a novel anchor proposal network and reduced hyper-parameters related to anchor boxes hugely (SiamAPN).

Despite superior tracking performance, the robustness of Siamese-based trackers relied on features extracted from deep neural network heavily. Trackers with lightweight network like AlexNet lacked global context and blocked improvement of tracking performance. Trackers with deep neural network like ResNet suffered from huge computation and could not meet the real-time requirements onboard UAV. In this paper, we explore a novel multi-level features fusion scheme through transformer in Siamese framework for effective and efficient aerial tracking.

C. TRANSFORMER BASED TRACKING
Transformer has been widely adopted in many artificial intelligence fields in the past few years, such as natural language processing, computer vision and speech processing. Transformer was originally developed as a sequence-to-sequence model for machine translation. Later works demonstrated that transformer can achieve state-of-the-art performance in visual tracking community. Wang et al. [26] first modified the classic transformer to suit visual tracking task and proposed a neat transformer-assisted tracking framework (TrDiMP). Liu et al. [27] introduced a dual-level deep representation for thermal infrared tracking which consists of a holistic correlation module and a fine-grained aware network like transformer architecture. Chen et al. [28] explained that attention-based feature fusion mechanism based on transformer is better than correlation conducted on original Siamese tracking framework (TransT). Yan et al. [29] built an end-to-end transformer architecture for visual tracking which can explore global spatio-temporal features effectively and obtain promising tracking performance (STARK). Zhao et al. [30] replaced cross-correlation operation with self- and cross-attention to explore global and rich information. Cao et al. [31] combined hierarchical response maps from multiple convolutional layers into transformer for fast UAV tracking (HiFT). Xing et al. [32] built a novel and lightweight Siamese transformer pyramid network (SiamTPN) to inherit the advantages from both transformer and CNN architecture.

Although transformer has achieved great success in visual tracking task, its superiority can not be efficiently extended to tracking on mobile platforms. Up to now, most trackers based on transformer usually extracted features from deep neural network like ResNet, which leads to intractable computations. In this work, we build a novel multi-level features fusion scheme with a lightweight encoder-decoder structure. By virtue of interactive combination of features from shallow and deep layers, the proposed method is able to promote the discriminability greatly and improves tracking performance significantly. Meantime, as using the AlexNet as the backbone, our method is able to afford real-time aerial tracking.

III. PROPOSED METHOD
In this section, the workflow of our proposed method is given in Figure 2. It consists of three modules: one lightweight Siamese network for feature extraction, one transformer based multi-level features fusion network and one prediction head for binary classification and regression.

A. FEATURE EXTRACTION NETWORK
To meet the requirements of mobile platform, we adopt the AlexNet as the backbone for feature extraction. From HCFT [33], we can learn that deep features in the earlier layers retain higher spatial resolution for precise localization with low-level visual information and features in the latter layers capture more semantic information. So we use hierarchical features of the last three convolutional layers to exploit these multiple levels of abstraction to represent target objects. Similar to Siamese-based trackers, the feature extraction network consists of two share-architecture branches: template branch and search branch. The input of template branch is the cropped image $Z$ from the initial frame of a sequence and the input of search branch is the cropped image $X$ from the current frame. For clarity, $Z$ denotes the template image and $X$ stands for the search image. After passing through the two branches, the $k$-th layer output feature of template branch is represented as $ϕ_k(Z) \in \mathbb{R}^{C_k \times H_k \times W_k}$ and the $k$-th layer output feature of search branch is represented as $ϕ_k(X) \in \mathbb{R}^{C_k \times H_k \times W_k}$. Here, $k \in \{3, 4, 5\}$ is the layer number of feature extraction. $(C_k, H_k, W_k)$ and $(C_k, H_k, W_k)$ mean the the channel, height and width of the $k$-th layer output feature from template branch and search branch, respectively. The output feature channel of the 3-th and the 4-th layers are 384. The output feature channel of the 5-th layer is 256. In order to reduce the computation burden, we use a $3 \times 3$ convolution kernel to reduce all the three output feature channels to 192. The parameters (padding of convolution operation) corresponding to the 3-th, 4-th and 5-th layer output features are 0, 1 and 2, respectively. Thus, we get the final output feature $ϕ_k(Z) \in \mathbb{R}^{C_k \times H_k \times W_k}$ and $ϕ_k(X) \in \mathbb{R}^{C_k \times H_k \times W_k}$ from template branch and search branch, respectively.

B. MULTI-LEVEL FEATURES TRANSFORMER FUSION NETWORK
In this section, in order to learn the interdependent relationship between high-resolution features and low-resolution features, we build a novel feature fusion network based on transformer to effectively fuse hierarchical features from template branch and search branch, respectively. Similar to the original transformer, our transformer fusion network can be divided into two parts: transformer encoder with high-resolution features and transformer decoder with low-resolution features.

Transformer encoder. The encoder is composed of a multi-head attention (MHA) component and a feed forward network (FFN). Given queries $Q \in \mathbb{R}^{N \times D_q}$, keys $K \in \mathbb{R}^{M \times D_k}$ and values $V \in \mathbb{R}^{M \times D_v}$, a single attention function can be
where $\Phi$ is generated by a sine function. Then, $F$ is denoted as $H$ and expressed in the scale dot-production way as:

$$H = \text{MultiHeadAttn}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_H)W^Q$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

where $W^Q \in \mathbb{R}^{D_q \times D_k}$, $W_i^Q \in \mathbb{R}^{D_q \times d_{head}}$, $W_i^K \in \mathbb{R}^{D_k \times d_{head}}$ and $W_i^V \in \mathbb{R}^{D_v \times d_{head}}$ are parameter matrices. $d_{head} = \frac{D_v}{H}$ and $H$ is the number of single attention head.

In our method, $(\psi'(Z) + Pos_Z) \in \mathbb{R}^{C_Z \times H_Z \times W_Z}$ and $(\psi'(Z) + Pos_Z) \in \mathbb{R}^{C_Z \times H_Z \times W_Z}$ are utilized as the input of template branch. $Pos_Z$ denotes the positional encoding, which is generated by a sine function. Then, $(\psi'(Z) + Pos_Z) \in \mathbb{R}^{C_Z \times H_Z \times W_Z}$ is reshaped to $(\psi'(Z) + Pos_Z) \in \mathbb{R}^{H_Z W_Z \times C_Z}$. It is denoted as $F_{Z3}$. $(\psi'(Z) + Pos_Z) \in \mathbb{R}^{C_Z \times H_Z \times W_Z}$ is reshaped to $(\psi'(Z) + Pos_Z) \in \mathbb{R}^{H_Z W_Z \times C_Z}$. It is denoted as $F_{Z4}$. Subsequently, we can get the output feature of MHA from template branch:

$$F_{Ze} = \text{MultiHeadAttn}(\text{Norm}(F_{Z3} + F_{Z4}), \text{Norm}(F_{Z3} + F_{Z4}), F_{Z3})$$

Here, $\text{Norm}$ stands for the normalization layer. The remaining mechanism can be summarized as:

$$F'_{Ze} = \text{Norm}(F_{Ze} + F_{Z3})$$

$$F''_{Ze} = \text{Norm}(\text{FFN}(F'_{Ze}) + F'_{Ze})$$

where $F''_{Ze}$ stands for the final output feature from template branch. At the same time, we can obtain the output feature of MHA from search branch,

$$F_{Xe} = \text{MultiHeadAttn}(\text{Norm}(F_{X3} + F_{X4}), \text{Norm}(F_{X3} + F_{X4}), F_{X3})$$

where $F_{X3}$ and $F_{X4}$ represent the reshaped features in the search branch. Then the final output features from search branch can be formulated as

$$F'_{Xe} = \text{Norm}(F_{Xe} + F_{X3})$$

$$F''_{Xe} = \text{Norm}(\text{FFN}(F'_{Xe}) + F'_{Xe})$$

Through the operation of MHA, the connection between features from the third layer and the fourth layer can be exploited effectively. Besides, the abundant high-resolution feature maps $F''_{Ze}$ and $F''_{Xe}$ can be enriched by introducing the global context information.

**Transformer decoder.** The decoder adopts the similar architecture of classical transformer and takes the feature $\psi'_5(Z) \in \mathbb{R}^{C_Z \times H_Z \times W_Z}$ and $\psi'_2(X) \in \mathbb{R}^{C_X \times H_X \times W_X}$ from the fifth layer as its input. Similar to the encoder, we first reshape these features to $F_{Z5} \in \mathbb{R}^{H_Z W_Z \times C_Z}$ and $F_{X5} \in \mathbb{R}^{H_Z W_Z \times C_Z}$. Then $F_{Z5} \in \mathbb{R}^{H_Z W_Z \times C_Z}$ is fed to template branch and $F_{X5} \in \mathbb{R}^{H_Z W_Z \times C_Z}$ is put to search branch. But unlike

![FIGURE 2. Workflow of our proposed STAT tracker.](image)
the encoder, position encoding is not introduced here. Take template branch as an example, the output feature of the first MHA is formulated as

\[ F_{Zd} = \text{MultiHeadAttn}(F_{Z5}, F_{Z5}, F_{Z5}) \]  

(7)

Subsequently, the output feature of the second MHA is expressed as

\[
\begin{align*}
F'_{Zd} & = \text{norm}(F_{Z5} + F_{Zd}) \\
F''_{Zd} & = \text{MultiHeadAttn}(F'_{Zd}, F''_{Zd}, F'_Zd) \\
\end{align*}
\]

(8)

Finally, the output feature of decoder from template branch is elaborated as

\[
\begin{align*}
F''''_{Zd} & = \text{norm}(F'_{Zd} + F''_{Zd}) \\
F''''''_{Zd} & = \text{norm}(\text{FFN}(F''''_{Zd}) + F''''_{Zd}) \\
\end{align*}
\]

(9)

Thus, the output feature of decoder from search branch is elaborated as

\[
\begin{align*}
F''''''_{Xd} & = \text{norm}(\text{FFN}(F''''_{Xd}) + F''''_{Xd}) \\
\end{align*}
\]

(10)

Instead of cross-correlation, we perform a depth-wise correlation on output features from template branch and search branch.

\[
M = F''''_{Zd} \ast F''''_{Xd} 
\]

(11)

Here \( \ast \) stands for the depth-wise correlation. \( M \) means the correlation response map and is adopted as the input to prediction head. The detailed workflow of our proposed STAT tracker is shown in FIGURE 3.

C. CLASSIFICATION AND REGRESSION NETWORK

We use the similar network as what HiFT used to directly classify and regress object bounding box at each location. The first branch intends to determine the anchor with the largest intersection over union (IoU) with the ground truth box. The second classification branch centers on the positive samples determined by euclidean distance between the center of ground truth and the corresponding points. The regression branch concentrates on outputting the center point, height and width of the bounding box.

The overall loss function is

\[
\mathcal{L} = \gamma_1 \mathcal{L}_{cls1} + \gamma_2 \mathcal{L}_{cls2} + \gamma_3 \mathcal{L}_{loc} 
\]

(12)

where \( \mathcal{L}_{cls1}, \mathcal{L}_{cls2} \) denote for the cross-entropy loss and binary cross-entropy loss for classification. \( \mathcal{L}_{loc} \) represents the IoU loss for regression. Constants \( \gamma_1, \gamma_2 \) and \( \gamma_3 \) weight the contributions of each loss.

IV. EXPERIMENTS

In this section, we perform numerous experiments to comprehensively analyze and evaluate our proposed STAT method. First, we introduce some implementation details and evaluation metrics. Then, we evaluate our method on three popular aerial tracking benchmarks including DTB70 [34], UAV123@10fps [35] and UAV20L [35]. 12 state-of-the-art trackers are included in the evaluation, i.e., HiFT [31], SiamAPN [5], ECO [36], MCPF [37], MCCT [38], UDT [39], BiCF [40], DeepSTRCF [41], IBCCF [42], ARCF [43], AutoTrack [44], CoKCF [45]. Finally, we demonstrate the qualitative results on some challenging sequences.

A. IMPLEMENTATION DETAILS

Our tracker is trained on the images extracted from COCO [46], ImageNet VID [47], GOT10K [48] and Youtube-BB [49] and is implemented using PyTorch on a PC with an Intel Xeon(R) Silver 4216 CPU, a 512GB RAM, and a Tesla V100. By following classical SiamFC, we adopt the AlexNet as the backbone, which is initialized with parameters pretrained on ImageNet. The first two convolution layers of AlexNet are frozen and the last three layers are fine-tuned. We train a total of 70 epochs, using a learning rate decreased from 0.01 to 0.0001 in the log space. Besides, the sizes of template patch and search patch are set to \( 3 \times 127 \times 127 \) pixels and \( 3 \times 287 \times 287 \) pixels, respectively.

B. EVALUATION METRICS

We use the precision plots and success plots in one-pass evaluation (OPE) [50] to assess the tracking performance. The precision plots compute the percentage of frames in which the Euclidean distance between the tracked location and the annotated center is smaller than a given threshold. All the trackers in precision plots are ranked based on the distance precision (DP) score at a threshold of 20 pixels. Meanwhile, the success plot calculates the proportion of frames where the overlap rate is larger than a threshold ranging from 0 to 1. The overlap rate is defined as

\[ \text{Score} = \frac{\text{area}(R_E \cap R_G)}{\text{area}(R_E \cup R_G)}, \]

where \( R_E \) means the estimated tracking bounding box and \( R_G \) stands for the ground truth bounding box. \( \cap \) and \( \cup \) denote the intersection and union operators, respectively. All the trackers in success plots are ranked based on the area under curve (AUC) score.

C. EVALUATION ON UAV BENCHMARKS

1) OVERALL PERFORMANCE

FIGURE 4 gives the precision and success plots of STAT and the other state-of-the-art 12 trackers using OPE on three UAV benchmarks. It can be clearly seen that our proposed STAT tracker is able to achieve the outstanding tracking performance compared with the other 12 trackers on almost all the three UAV benchmarks.

DTB70: The DTB70 benchmarks includes 15777 pictures, making up 70 severe sequences captured by low-altitude drone camera. As shown in FIGURE 4(a), STAT achieves the second best DP score (0.757), 2.7% behind the best tracker SiamAPN. Meanwhile, STAT obtains the best AUC score (0.592), followed by SiamAPN (0.586) and HiFT (0.573).

UAV123@10fps: UAV123@10fps consists of 123 image sequences with a frame rate of 10FPS, which is temporarily downsampled from the 30FPS version. The larger frame interval means the more abrupt appearance change between successive frames and brings more challenging difficulties.
to the tracking task. Therefore, UAV123@10fps is chosen to evaluate the robustness of trackers in this work. In FIGURE 4(b), our proposed STAT gets the best DP score (0.758) compared to the second best tracker SiamAPN (0.752) and the third best tracker HiFT (0.751). Moreover, STAT also achieves the best AUC score (0.586), which is 0.8% higher than HiFT and 2.0% higher than SiamAPN. The promising performance demonstrates that STAT could be a better choice in abrupt UAV tracking scenes.

UAV20L: The UAV20L benchmark focuses on practical long-term UAV tracking task with 20 representative sequences (2934 frames per image sequence on average). As illustrated in FIGURE 4 (c), STAT obtains favorably with an improvement of 2.8% on DP score and 3.3% on AUC score comparing to the second-best tracker.

2) ATTRIBUTE-BASED EVALUATION

In this section, we carry out numerous experiments to show the tracking results in terms of each attribute on three UAV benchmarks.

DTB70: The DTB70 benchmark contains 11 challenging factors including scale variation (SV), aspect ratio variation (ARV), occlusion (OCC), deformation (DEF), fast camera motion (FCM), in-plane rotation (IPR), out-of-plane rotation (OPR), out-of-view (OV), background clutter (BC), similar objects around (SOA) and motion blur (MB).

UAV123@10fps: The 12 attributes from the UAV123@10fps benchmark are fast motion (FM), camera motion (CM), low resolution (LR), illumination variation (IV), viewpoint change (VC), scale variation (SV), aspect ratio change (ARC), partial occlusion (POC), similar object (SOB), out-of-view (OV), background clutter (BC), and full occlusion (FOC).
attribute-based evaluation on DTB70 dataset. The value in the bracket of the title denotes the number of videos associated with the respective attribute.

UA V20L: It contains 12 challenging factors including scale variation (SV), aspect ratio change (ARC), low resolution (LR), fast motion (FM), full occlusion (FOC), partial occlusion (POC), out-of-view (OV), background clutter (BC), illumination variation (IV), viewpoint change (VC), camera motion (CM) and similar object (SOB).

To further show the excellence of our method, success plots of different challenging factors on three UAV benchmarks are given in FIGURE 5, FIGURE 6 and FIGURE 7, respectively. On the DTB70 benchmark, STAT provides the best AUC score in almost all the challenging attributes, except for OCC (ranks second), BC (ranks third) and SOA (ranks third). More specifically, STAT improves the AUC by 4.0%, 3.0%, 2.2% and 2.8%, in the environments of SV, DEF, IPR and OV, respectively. On the UAV123@10fps benchmark, STAT also has the leading AUC score in almost all the challenging attributes, except for VC (ranks second). In particular, STAT exceeds the second best tracker HiFT by 2.7%, 2.2% and 2.6% in the attribute of FOC, BC and CM, respectively. On the UAV20L benchmark, our STAT tracker achieves superior performance in all the attributes. The experimental results show that STAT is able to outperform...
favorably than the other state-of-the-art 12 trackers in most attributes.

3) QUALITATIVE EVALUATION
In this section, some qualitative comparisons are demonstrated with STAT and other 5 top trackers involved, i.e., HiFT, SiamAPN, ECO, MCCT and AutoTrack. FIGURE 8 gives some sample tracking results of the evaluated trackers on six challenging sequences, i.e., BMX3 from DTB70, ChasingDrones from DTB70, person19_2 from UAV123@10fps, wakeboard2 from UAV123@10fps, car9 from UAV20L and group3 from UAV20L.

In the BMX3 sequence, the appearance of target changes severely when the player is keep riding. ECO and AutoTrack lose target completely due to the challenging interference.
factors. It can be clearly seen that STAT performs well in the whole sequence and is able to cope with DEF, FCM, SV and BC effectively. In the ChasingDrones sequence, the UAV target keeps moving and undergoes FCM, SV and ARV. From the tracking results, we can observe that most of the trackers lose target (e.g., ♯212). However, only our method and HiFT are able to track target successfully through the entire sequence. In the person19_2 sequence, the person target goes across the square and experiences with OV and SV. In the 230-th sequence image, the target disappears due to change of sight range from UAV. When the man reappears in the scene, our method has the ability to estimate the target position accurately again. HiFT, SiamAPN and ECO lose target and drift away. In the wakeboard2 video, the rider is fastened to a board and towed behind a motorboat at speeds of around 30 miles per hour. Although target undergoes severe FM and SOB, our method locates target more precisely than other trackers. In the car9 sequence, the car is severely occluded by a billboard when it is moving (e.g., ♯780). It is obvious that our method is robust to severe occlusion and can track target steadily. In the group3 video, the person wearing a green coat is occluded by trees twice in a row. Only our STAT method is able to handle occlusion effectively in long-term videos.
FIGURE 8. Qualitative comparison of our proposed STAT tracker with other 5 advanced trackers. From top to bottom, these sequences are BMX3 from DTB70, ChasingDrones from DTB70, person19_2 from UAV123@10fps, wakeboard2 from UAV123@10fps, car9 from UAV20L and group3 from UAV20L.
TABLE 2. Comparison with transformer based trackers.

| Method   | Backbone | Precision | Success rate | FPS |
|----------|----------|-----------|--------------|-----|
| Trans. [28] | ResNet50 | 0.819 | 0.631 | 39.9 |
| TriDiMP [26] | ResNet50 | 0.768 | 0.620 | 15.7 |
| HCAT [51] | ResNet18 | 0.676 | 0.510 | 104.1 |
| SiamTPN [32] | ShuffleNetV2 | 0.793 | 0.607 | 32.9 |
| HiFT [31] | AlexNet | 0.736 | 0.553 | 92.2 |
| STAT      | AlexNet  | 0.764 | 0.586 | 71.3 |

TABLE 3. Ablation studies of transformer fusion module on three UAV benchmarks.

|          | Baseline       | Baseline+transformer(STAT) |
|----------|----------------|---------------------------|
| Precision | 0.554          | 0.592 | 0.784 |
| Success rate | 0.745          | 0.784 |
| UAV123+10fps | 0.560          | 0.586 | 0.758 |
| UAV20L   | 0.500          | 0.586 | 0.764 |
| Average  | 0.538          | 0.588 | 0.768 |

4) COMPARISON OF DIFFERENT METHODS IN SPEED ON UAV20L

Tracking speed is very critical for industrial application of UAV tracking algorithm. It is unrealistic to apply trackers with a slow speed into industrial UAV products. This section shows the comparison of different tracking methods in speed on UAV20L. In order to demonstrate the comparison of different UAV tracking methods validly and fairly, all the nine methods with deep features (including STAT, SiamAPN, HiFT, ECO, MCCT, DeepSTRCF, MCPF, CoKCF and UDT) in Table 1 are conducted on the same PC mentioned in subsection A of section IV. The other four methods with handcrafted features in Table 1 are conducted on the same PC platform without GPU. From Table 1 and FIGURE 4, it can be seen that our tracker is able to get the best tracking results with faster tracking speed, which is more suitable for UAV platform.

5) COMPARISON WITH TRANSFORMER BASED TRACKERS

Recently, transformer has achieved great success in computer vision, especially in the field of visual tracking. In this section, in order to validate the superiority of our proposed method, we compare our STAT method with five state-of-the-art transformer based trackers on UAV20L. The comparison results are demonstrated in Table 2. All these methods are trained using the code provided by the authors without any modification in the same hardware platform as mentioned above. TransNet and TriDiMP use ResNet50 as the backbone to extract deeper features and get better performance than our method. However, the slow tracking speed of these two methods limits their application in UAV platform. HCAT utilizes ResNet18 as the backbone and abandons the self-attention module. HiFT uses the correlation responses from multiple layers as the input of transformer module. Although HCAT and HiFT are able to track the object with fast speed, the tracking performance is inferior to their industrial application. Taking account of both the efficiency and the effectiveness of trackers, our proposed STAT method is more suitable for UAV tracking.

6) ABLATION STUDIES

To demonstrate the effectiveness of the proposed transformer fusion structure, ablation studies are carried out on three UAV authoritative benchmarks. Please note that the baseline represents the pure tracker. The features in the baseline tracker from the last three layers are put into the correlation module directly. From the comparison in Table 3, by virtue of the feature fusion, both the precision and the success rate are improved greatly on three UAV benchmarks. Eventually, the transformer fusion module promotes the tracking performance 5% on the average precision and 4.1% on the average success rate.

V. CONCLUSION

In this work, we present an efficient and high-performance aerial tracking method based on transformer in the framework of Siamese. Features from multiple layers are put into transformer to promote the expression ability of model. Specifically, deep features from the earlier layer with higher spatial resolution and features from the deeper layer with more semantic information are interdependently integrated into one whole. Hierarchical features of target are fully exploited and fused in the framework of transformer, which is robust to large appearance variation and occlusion. Extensive experiments conducted on three UAV benchmarks demonstrate that our proposed STAT algorithm achieves considerably better than the state-of-the-art methods while tracking at a real-time speed.

REFERENCES

[1] C. Fu, J. Xu, F. Lin, F. Guo, T. Liu, and Z. Zhang, “Object saliency-aware dual regularized correlation filter for real-time aerial tracking,” IEEE Trans. Geosci. Remote Sens., vol. 58, no. 12, pp. 8940–8951, Dec. 2020.
[2] C. Fu, F. Ding, Y. Li, J. Jin, and C. Feng, “Learning dynamic regression with automatic distractor regression for real-time aerial tracking,” Eng. Appl. Intell. Intell., vol. 98, Feb. 2021, Art. no. 104116.
[3] C. Fu, J. Jin, F. Ding, Y. Li, and G. Lu, “Spatial reliability enhanced correlation filter: An efficient approach for real-time UAV tracking,” IEEE Trans. Multimedia, early access, Oct. 8, 2021, doi: 10.1109/TMM.2021.3118891.
[4] F. Lin, C. Fu, Y. He, W. Xiong, and F. Li, “ReCF: Exploiting response reasoning for correlation filters in real-time UAV tracking,” IEEE Trans. Intell. Transp. Syst., vol. 23, no. 8, pp. 10469–10480, Aug. 2022.
[5] C. Fu, Z. Cao, Y. Li, J. Ye, and C. Feng, “Onboard real-time aerial tracking with efficient Siamese anchor proposal network,” IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–13, 2022.
[6] X.-F. Zhu, X.-J. Wu, T. Xu, Z.-H. Feng, and J. Kittler, “Robust visual object tracking via adaptive attribute-aware discriminative correlation filters,” IEEE Trans. Multimedia, vol. 24, pp. 301–312, 2022.
[7] X. Li, Q. Liu, N. Fan, Z. He, and H. Wang, “Hierarchical spatial-aware Siamese network for thermal infrared object tracking,” Knowl.-Based Syst., vol. 166, pp. 71–81, Feb. 2019.
[8] C. Fu, B. Li, F. Ding, F. Lin, and G. Lu, “Correlation filters for unmanned aerial vehicle-based aerial tracking: A review and experimental evaluation,” IEEE Geosci. Remote Sens. Mag., vol. 10, no. 1, pp. 125–160, Mar. 2022.
[9] P. Li, D. Wang, L. Wang, and H. Lu, “Deep visual tracking: Review and experimental comparison,” Pattern Recognit., vol. 76, pp. 323–338, Apr. 2018.
[10] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, “SiamRPN++: Evolution of Siamese visual tracking with very deep networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 4277–4286.
D. Bolme, J. Beveridge, B. Draper, and Y. Lui, “Visual object tracking with Siamese region proposal network,” in Proc. CVPR, Jun. 2018, pp. 8971–8980.

L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, “Fully-convolutional Siamese networks for object tracking,” in Proc. Eur. Conf. Comput. Vis. Amsterdam, The Netherlands, Oct. 2016, pp. 850–865.

Z. Chen, B. Zhong, G. Li, S. Zhang, and R. Li, “Siamese box adaptive network for visual tracking,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 6667–6676.

S. Cheng, B. Zhong, G. Li, X. Liu, Z. Tang, X. Li, and J. Wang, “Learning to filter: Siamese relation network for robust tracking,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 4419–4429.

K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Feb. 2016, pp. 770–778.

D. Bolme, J. Beveridge, B. Draper, and Y. Lui, “Visual object tracking using adaptive correlation filters,” in Proc. Int. Conf. Comput. Vis. Pattern Recognit., Sep. 2010, pp. 2544–2550.

J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, “Exploiting the circulant structure of tracking-by-detection with kernels,” in Proc. Eur. Conf. Comput. Vis. Florence, Italy, Oct. 2012, pp. 702–715.

M. Danelljan, G. Häger, S. J. H. Khan, and M. Felsberg, “Discriminative scale space tracking,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 8, pp. 1561–1575, Apr. 2017.

Y. Li, C. Fu, F. Ding, Z. Huang, and J. Pan, “Augmented memory for correlation filters in real-time UAV tracking,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Oct. 2020, pp. 1559–1566.

C. Fu, Y. He, F. Lin, and W. Xiong, “Robust multi-kernelized correlators for UAV tracking with adaptive context analysis and dynamic weighted filters,” Neural Comput. Appl., vol. 32, no. 16, pp. 12591–12607, Aug. 2020.

Y. Li, C. Fu, Z. Huang, Y. Zhang, and J. Pan, “Intermittent contextual learning for keyfilter-aware UAV object tracking using deep convolutional feature,” IEEE Trans. Multimedia, vol. 23, pp. 810–822, 2021.

Q. Liu, D. Yuan, N. Fan, P. Gao, X. Li, and Z. He, “Learning dual-level deep representation for thermal infrared tracking,” IEEE Trans. Multimedia, early access, Jan. 2022, doi: 10.1109/TMM.2022.314029.

K. Yang, Z. He, Z. Zhou, and N. Fan, “SiamAtt: Siamese attention network for visual tracking,” Knowl.-Based. Syst., vol. 203, Jan. 2021, Art. no. 106079.

D. Guo, J. Wang, Y. Cui, Z. Wang, and S. Chen, “SiamCAR: Siamese fully convolutional classification and regression for visual tracking,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 608–6276.

Y. Xu, Z. Wang, Z. Li, Y. Yuan, and G. Yu, “SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines,” in Proc. AAAI Conf. Artif. Intell., vol. 34, no. 7, Apr. 2020, pp. 12549–12556.

N. Wang, W. Zhou, J. Wang, and H. Li, “Transformer meets tracker: Exploiting temporal context for robust visual tracking,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2021, pp. 1571–1580.

Q. Liu, X. Li, Z. He, N. Fan, D. Yuan, and H. Wang, “Learning deep multi-level similarity for thermal infrared object tracking,” IEEE Trans. Multimedia, vol. 23, pp. 2114–2126, 2021.

Y. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, “Transformer tracking,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 8122–8131.

B. Yan, H. Peng, J. Fu, D. Wang, and H. Lu, “Learning spatio-temporal transformer for visual tracking,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 10428–10437.

M. Zhao, K. Okada, and M. Inaba, “TrTr: Visual tracking with transformer,” 2021, arXiv:2105.03817.

Z. Cao, C. Fu, J. Ye, B. Li, and Y. Li, “HiFT: Hierarchical feature transformer for aerial flying,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 15437–15446.

D. Xing, N. Evangelio, A. Tsoukalas, and A. Tzes, “Siamese transformer pyramid networks for real-time UAV tracking,” in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), Jan. 2022, pp. 1890–1907.

C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, “Robust visual tracking via hierarchical convolutional features,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 11, pp. 2709–2723, Nov. 2019.

S. Li and D.-Y. Yeung, “Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models,” in Proc. Conf. Assoc. Adv. Artif. Intell. (AAAI), vol. 31, no. 1, Feb. 2017, pp. 1–7.