Global-Local and Occlusion Awareness Network for Object Tracking in UAVs

Lukui Shi, Qingrui Zhang, Bin Pan, *Member, IEEE*, Jun Zhang, and Yuanchao Su, *Senior Member, IEEE*

***Abstract*—Multiobject tracking in unmanned aerial vehicle (UAV) scenes is a crucial task with numerous applications across various domains. The goal of this task is to track multiple objects such as people and vehicles over time as they move through the scene, often captured by cameras mounted on UAV. However, tracking objects in UAV scenes can be challenging due to several factors, including the scale variance of objects as they move through the scene and the frequent occlusions caused by complex scenes. To address these issues, we propose a global-local and occlusion awareness (GLOA) tracking network for UAVs. It comprises two main components: global-local awareness detector (GLA-D) and the occlusion awareness data association (OADA). The GLA-D uses our specially designed global-local awareness block to extract scale variance feature information from the input frames. It then outputs more discriminative identity information by adding identity embedding branches to the prediction head. The GLA-D is designed to better handle scale variance issues and improve object tracking accuracy. The OADA method used different metrics for high- and low-scoring detection frames was to alleviate occlusion problems in tracking scenarios. By combining these two components, the GLOA provides a more robust and effective solution for multiobject tracking in UAV scenes. Experiments on two public datasets may indicate the effectiveness of the proposed method.**

***Index Terms*—Multiobject tracking, occlusion awareness, unmanned aerial vehicle (UAV).**

I. Introduction

IN RECENT years, unmanned aerial vehicles have been widely applied in object tracking tasks in the field of remote sensing due to their advantages of small size and flexibility. They have proven to be effective in various applications including video surveillance and emergency rescue [1], [2], [3], [4], [5]. However, object tracking in unmanned aerial vehicle (UAV) scenes is still a challenging task due to uncertain flight altitudes, changes in scale variance feature information, and frequent occlusions.

Depending on the task and application scenario the task of tracking objects can be categorized into two types: single-object tracking and multiobject tracking. Single object tracking is an essential research branch in the field of UAV remote sensing. It enables the continuous and real-time monitoring and tracking of a single object, such as a person and vehicle [6], [7], [8], [9], [10]. Although single object tracking is a useful research, it has limitations when it comes to tracking multiple objects in a scene. Therefore, multiobject tracking has emerged as a significant area of research in the field of UAV remote sensing. This article focuses on developing and studying multiobject tracking to meet the needs of tracking multiple objects simultaneously.

There are two main categories of multiobject tracking networks: separate detection and embedding (SDE) and joint learning of detection and embedding (JDE) frameworks [11]. SDE-based multiobject tracking algorithms use separate networks for object detection and reidentification feature extraction [12], [13], [14], [15], [16], [17]. They improve tracking performance in remote sensing by using historical trajectory information and separate network extraction of identity embedding, as shown in [18], [19], [20], [21], [22], [23], and [24]. However, these algorithms are complex and not suitable for real-time tracking. JDE-based multiobject tracking algorithms have higher tracking speed and accuracy because the detectors and embeddings are learned jointly [25], [26]. These algorithms usually simpler and more efficient in tracking objects, making it popular in UAV applications [27], [28], [29], [30], [31]. These network use convolutional neural network to extract features and enhances tracking performance by extracting more discriminative identity embedding. Convolutional neural network (CNN)-based multiobject tracking algorithm have limitations due to their reliance on local perception, difficulty in modeling long-term dependencies and capturing global features, and the negative impact of spatial invariance and pooling operations on tracking accuracy. Transformer has shown great success in various vision tasks due to its ability to capture global features and long-term modeling [32], [33], [34], [35], [36]. Therefore, using only CNN or transformer as the detector of the tracking network cannot adequately capture global and local features [51], [52],

Manuscript received 13 April 2023; revised 15 July 2023; accepted 19 August 2023. Date of publication 23 August 2023; date of current version 29 September 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62001252, in part by the National Natural Science Foundation of China under Grant 42001319, in part of the National Key Research and Development Program of China under Grant 2022YFA1003803, in part by the Beijing-Tianjin Hebei Basic Research Cooperation Project under Grant F2021203109, in part by the Scientific and Technological Research Project of Hebei Province Universities and Colleges under Grant ZD2021311, in part by the Scientific Research Program of the Education Department of Shaanxi Province under Grant 21JK0762, and in part by the University-Industry Collaborative Education Program of Ministry of Education of China under Grant 220802313200859. *(Corresponding author: Bin Pan.)*

Lukui Shi, Qingrui Zhang, and Jun Zhang are with the School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China, and also with the Hebei Province Key Laboratory of Big Data Calculation, Tianjin 300401, China (e-mail: shilukui@scse.hebut.edu.cn; zhangqingrui1@hotmail.com; zhangjun@scse.hebut.edu.cn).

Bin Pan is with the School of Statistics and Data Science, KLMDASR, LEBPS, and LPMC, Nankai University, Tianjin 300071, China (e-mail: panbin@nankai.edu.cn).

Yuanchao Su is with the Department of Remote Sensing, College of Geomatics, Xi'an University of Science and Technology, Xi'an 710054, China (e-mail: suych3@xust.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3308042

especially in the UAV scenario. This leads to a degradation of detection performance, which in turn affects tracking performance. However, using only CNN or transformer can lead to a degradation in tracking performance, especially for UAV video processing, as they cannot capture global and local features well.

Data association is critical in multiobject tracking, involving matching objects between frames, handling new and disappearing objects, and matching objects across frames. Traditional methods for data association use operations research techniques [37], [38], [39], [40]. Bewley et al. [12] used the Kalman filter for prediction, while the authors in [41] and [42] used a greedy matching algorithm for scenes with good detection results. Chen et al. [43] introduced a trajectory scoring mechanism based on length to improve reliability. Zhang et al. [44] addressed low-confidence detections during occlusion and separates high- and low-score frames to identify real objects. Xu et al. [45] applied [44] to UAV tracking, but different treatments are needed for high- and low-scoring frames in different scenarios.

The current UAV multiobject tracking algorithms have the following two bottlenecks.

1) The current feature extraction methods do not fully consider the complexity of the UAV scenes. As a result, they may fail to extract salient features of tracked objects, particularly in special scenes with scale variance feature information.
2) The existing data association methods do not fully consider the special characteristics of UAV scenes, which affects the continuity of tracking trajectories and the accuracy of occlusion objects recognition.

To solve the problem of scale variance during UAV flight, we are inspired by [46] designed global-local awareness block (GLA-block), it use high-order spatial interaction convolution extract global features and model long-term interaction. To further enhance the global and local features, we perform convolutional aggregation and self-attention aggregation on the extracted features, respectively. At the same time, we incorporate GLA-block into the feature extraction phase to form the global-local awareness detector (GLA-D). To alleviate the occlusion problem, we designed occlusion awareness data association (OADA) inspired by [44]. To make it better for drone scenarios processes, the high- and low-score detection frames separately and uses noise-adaptive (NSA) Kalman filter for data association

In this article, we introduce a novel network for global-local awareness occlusion awareness tracking of unmanned aerial vehicles, called GLOA. The network is composed of two main components: GLA-D and OADA. GLA-D employs our novel GLA-block to extract scale variance feature information and output more discriminative identity information by incorporating identity embedding output branches into the prediction head. In addition, we introduce the OADA method, which is specifically designed to address the issue of occlusions in tracking scenarios. Overall, our contributions can be summarized as follows.

1) We propose a global-local and occlusion awareness tracking network, for multiobject tracking, which handles scale variations and occlusions in unmanned aerial vehicles remote sensing scenes.
2) We designed a GLA-D, which enhances the scale variance feature extraction capability by incorporating GLA-block into the feature extraction stage, and output more discriminative identity information by adding identity embedding branch in the prediction head.
3) We developed an OADA, it used different metric for high- and low-scoring frames, respectively, and NSA Kalman filter for motion modeling, which ensures the continuity of tracking trajectories and improves the recognition accuracy of occlusion objects.

II. Method

A. Framework Overview

The overall process of our GLOA, as shown in the Fig. 1. In the feature extraction stage, our GLA-block is used to fully extract global and local features to cope with the scale variance brought about by the complex scene of the UAV. In the prediction head part, we add a branch to extract identity embedding to output more discriminative identity information. In the data association stage, we use the OADA to alleviate occlusion problem and ensure the integrity of the trajectory.

B. Global-Local Awareness Detector

We have inserted the designed GLA-block into the feature extraction stage of GLA-D to enable better extraction of scale variation features. In addition, we have added an identity embedding output branch to the prediction head for discriminative identity information. GLA-D was composed of two main parts.

1) Global-Local Awareness Block: The problem of scale change arises during the flight of UAVs. At higher altitudes, the drone can get a wider global view, but local details may become blurred or invisible. And at lower altitudes, the UAV can capture more local details, but the overall scene may be partially obscured or missing. If global features are missing, it is difficult to accurately analyze and understand the components and interrelationships of the entire scene. If local features are missing, it may be difficult to accurately identify objects in the scene, especially objects that occur to be occluded. But it is difficult to capture this feature information adequately by methods based solely on CNN or transformer. Therefore, we design a feature extraction module GLA-block that combines the advantages of CNN and transformer.

The network structure of GLA-block is shown in the Fig. 2. GLA-block consists of a feature extraction stage composed of recursive gated convolutions (g^nConv), and the stage of feature aggregation for self-attention and Conv, respectively. The feature map undergoes a gated convolution and a feedforward neural network with GELU activation functions to generate the processed feature map. This processed feature map is then convolved with three 1×1 filters to produce three separate feature maps: q, k, and v. The attention weights are calculated by correlating q, k, and location encoding. The weighted summation of v using the attention weights results in integrated features. q, k, and v are also convolved, and the convolved features are

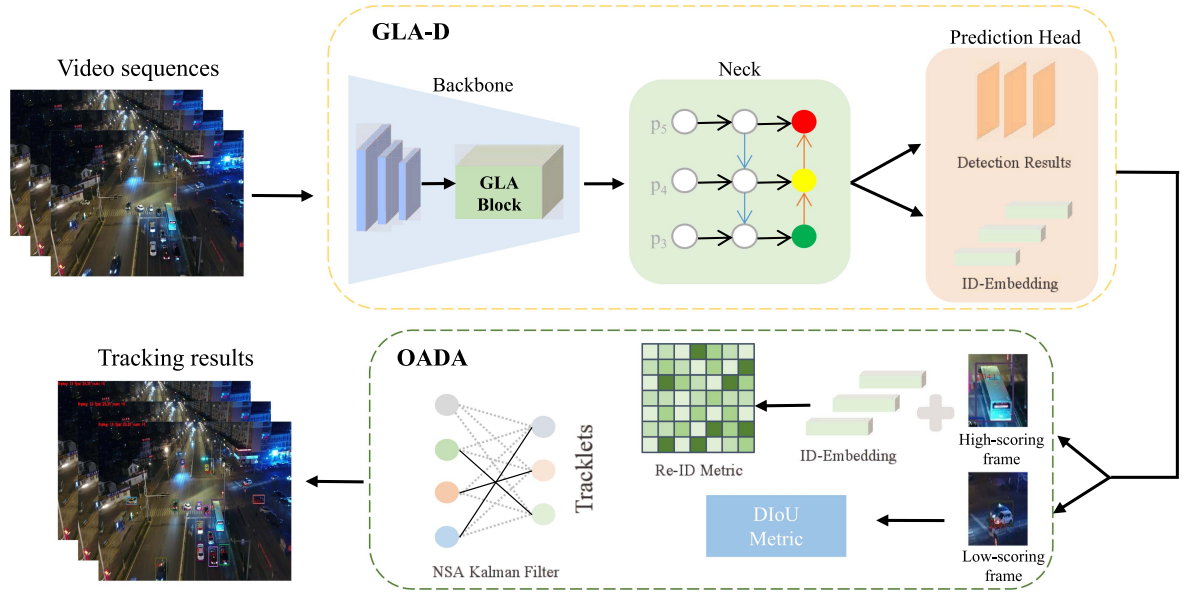


Fig. 1. Overview of the proposed GLOA. Input is a video sequence captured by a drone. The detection results and identity embedding can be output at the same time during a tracking process, and then sent to the OADA method to generate tracking results.

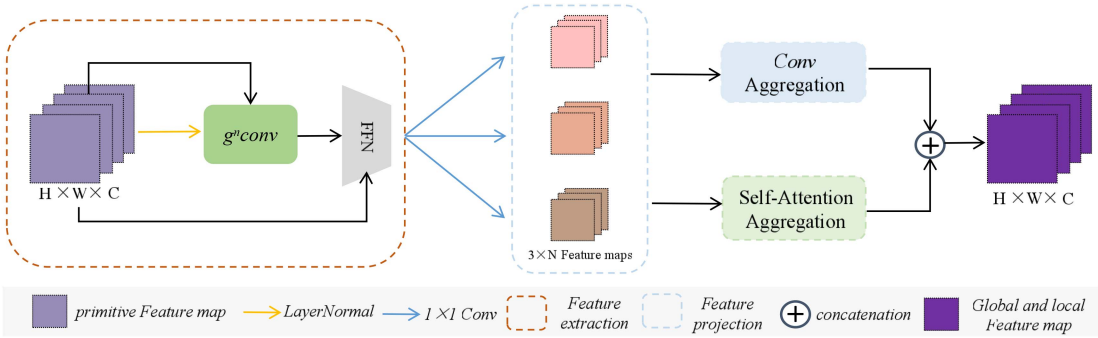


Fig. 2. GLA-block. It is composed of feature extraction and feature aggregation. The global and local feature can be extracted better by GLA-block.

concatenated together. Finally, the concatenated features are further integrated.

The feature extraction stage contains a spatial hybrid layer, consisting of g^nConv and LayerNorm and a feedforward neural network, and the core of this part is g^nConv, as shown in Fig. 3. Since the standard convolution does not consider the high-order spatial interaction information, this results in the loss of global information and long-term dependencies.

So we use high-order gated convolution is to preserve the translation invariance and local relevance of ordinary convolution while better capturing global contextual information and establishing spatial interactions of different orders. g^nConv is constructed from standard convolution, linear projection, and elemental multiplication. The output of the g^nConv can be written as

$$\left[\mathbf{p}_0^{\mathrm{HW}\times C}, \mathbf{q}_0^{\mathrm{HW}\times C}\right] = \phi_{\mathrm{in}}(\mathbf{x}),\ \mathbf{x} \in \mathbb{R}^{\mathrm{HW}\times 2\,C} \tag{1}$$

$$\mathbf{p}_1 = f(\mathbf{q}_0) \odot \mathbf{p}_0,\ \mathbf{p}_1 \in \mathbb{R}^{\mathrm{HW}\times C} \tag{2}$$

$$\mathbf{y} = \phi_{\mathrm{out}}(\mathbf{p}_1),\ \mathbf{y} \in \mathbb{R}^{\mathrm{HW}\times C} \tag{3}$$

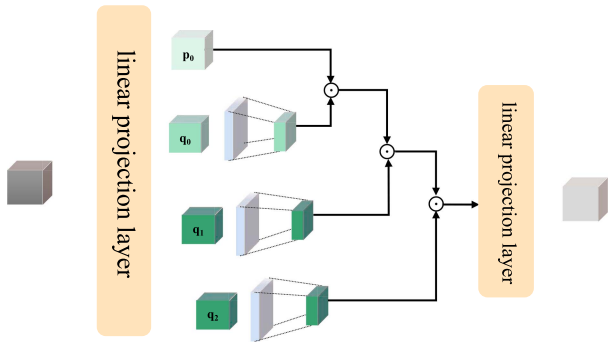


Fig. 3. Overview of g^3Conv. The feature $\mathbf{x}$ passed through a linear projection layer ϕ_{in} to obtain a set of projected features p_0, q_0, q_1, and q_2. Then, gate convolution is recursively performed on these features. Finally, the output of the last recursive step is input into ϕ_{out} to obtain the result of g^3Conv.

where ϕ_{in}, ϕ_{out} are linear projection layers for channel mixing and f is a depth-wise convolution to reduce the repetition rate. The input feature $\mathbf{x}$ first passes through a linear projection ϕ_{in} to

obtain p_0 and q_0. q_0 passes through a depth-wise convolution and is then dotted with p_0 to obtain p_1. At this point, the first-order spatial interaction is extracted.

In our model, we use third-order spatial interactions g^3Conv to extract features. So $\mathbf{x}$ needs to undergo a higher order linear projection to obtain p_0 and q_k (k=0, 1, 2)

$$\mathbf{p}_{k+1} = f_k\left(\mathbf{q}_k\right) \odot g_k\left(\mathbf{p}_k\right)/\alpha, \quad k = 0, 1, 2. \tag{4}$$

Then, gated convolution is performed cyclically, and p_{k+1} can be obtained in turn, where f is the depth-wise convolution, and g is used for dimension alignment at different levels. The depth-wise convolution uses a 7×7 convolution kernel and a global filter to enable it to have a global receptive field to better extract global features and model long-term interactions.

Through the abovementioned operations, the order is increased by 1 every time it is cycled, so g^3Conv can model third-order spatial interactions

$$g_k = \begin{cases} \text{Identity}\,, & k = 0 \\ \text{Linear}\,\left(C_{k-1}, C_k\right), & 1 \le k \le 2. \end{cases} \tag{5}$$

In UAV remote sensing, the integrity of the image is often more important than just focusing on the foreground objects. Therefore, global features are often more suitable to describe the entire image instead of local features that only capture information from small regions. Global features have the disadvantage of not being able to distinguish between foreground and background. This is especially problematic when the objects of interest are partially or fully occluded, as the global features may become distorted. As there are many interferences in the remote sensing scene of UAV, occlusion will occur frequently. To solve this problem, we perform convolution feature aggregation and self-attention feature aggregation on the abovementioned extracted features to make full use of the advantages of Conv and self-attention to better extract global and local features.

Specifically, feature aggregation mainly includes two stages. In the first stage, we enhance the utilization of intermediate features by processing the previously generated feature maps using three 1×1 convolutions. And we convolve these convolved feature maps divide into N groups in the depth direction. This process generates a diverse set of intermediate features that includes $3 \times N$ feature maps. The second stage operates separately for self-attention feature aggregation and convolutional feature aggregation. The self-attention feature aggregation path involves grouping the intermediate features into N groups. Each group is composed of three features obtained from 1×1 convolutions. Similar to the traditional multihead self-attention module, the three corresponding feature maps are utilized as the query, key, and value during the self-attention process. The convolutional feature aggregation path utilizes a lightweight fully connected layer to produce k^2 feature maps, where k is the kernel size. The resulting features are then generated via aggregation and shifting. The features output by the self-attention feature aggregation path has global correlation and long-term dependency, while the output of the features by the convolution feature aggregation path has local correlation and translation invariance. By combining the outputs of the two feature aggregation paths through splicing, we obtain features with both global and local correlations. This approach is more effective in handling the complex scenarios of UAV remote sensing.

2) Prediction Head: We designed GLA-D with the target of obtaining both detection results and identity embedding in a single forward output process. The first target requires us to accurately detect the object. The second target requires us to better extracted to identity embedding. For the first target, it is necessary to select the appropriate value for the double threshold assigned to the foreground and background to differentiate in order to improve the detection effect. Through testing and calculation, it has been determined that an IoU threshold of greater than 0.5 roughly ensures that a bounding box corresponds to a foreground object, which is in line with common settings in object detection. However, for boxes with IoU$<$0.4, it is more appropriate to consider them as background instead of 0.3 in general. This is because preliminary experiments have shown that this threshold is effective in suppressing false positives, which often occur in the presence of severe occlusions. The prediction branch of GLOA has two loss functions: foreground and background classification loss L_α and bounding box regression loss L_β. L_α is formulated as the cross-entropy loss, and L_β is formulated as the smooth-$L1$ loss.

The second target is to obtain identity embedding. We need instances with the same track ID to be close to each other, while instances with different track IDs are far apart. Ideally, the same instance should have the same track ID in different frames, while different instances should have different track IDs. The only information we can obtain is the track ID of the object. Therefore, in the process of network training, we need to convert the acquired identity embedding into sufficiently strong semantic information. The connection layer converts identity embedding information into track ID classification information to enhance semantic information.

Extracting identity embedding belongs to the metric learning task, and the classic loss function is the triplet loss function. The triplet loss function is based on the concept of a set triplet, which consists of an anchor, a positive, and a negative sample. Anchor and positive are different samples of the same type, anchor and negative are heterogeneous samples. The goal of the triplet loss is to learn a feature space where samples of the same category have a smaller distance in comparison to samples of different categories. Specifically, the reference sample (anchor) should be closer to the positive sample (positive), while the distance between the anchor and negative samples (negative) from different categories should be larger. However, this ordinary triplet loss function only samples one negative sample at a time and does not consider other negative samples, which makes it difficult to converge. Therefore, we adopt the modified triplet loss function used in [11]

$$L_{\mathrm{CE}} = -\log \frac{\exp\left(fg^+\right)}{\exp\left(fg^+\right) + \sum_i \exp\left(fg_i^-\right)}. \tag{6}$$

The triplet loss function composed of f, f^-, and f^+ three elements, as shown in the following equation:

$$L_{\text{triplet}} = \sum_i \max\left(0, ff_i^- - ff^+\right). \tag{7}$$

f is the benchmark sample, f^+ means the ID is the same as f, and f^- means the ID is different from f. So f^+f represents the cosine value between samples with the same ID, and f^-f represents the cosine value between samples with different IDs. In the improved triplet loss function L_{CE}, g^+ is the learnable parameter corresponding to f^+, and g^- is the learnable parameter corresponding to f^-. The dot product of the two results in a real number, which is the score. g^+, g^- are the weights of the fully connected layer.

C. Occlusion Awareness Data Association

Data association is an essential part of multiobject tracking. A reliable data association method can solve the problem of object identity switching and occlusion and improve the stability of multiobject tracking. Since the scenes from the drone view are more complex than those from the static camera, a reliable data association method is needed. In the process of multiobject tracking, data association is to associate the detection frames of the detection output, so there will be detection frames with higher discrimination scores and lower scores. In drone scenes, usually, low-score detection frames often indicate the detection of objects. Simply discarding low-scoring detection boxes can result in a significant number of missed detections and trajectory interruptions, degrading the overall tracking performance. We designed a data association method OADA applied in UAV remote sensing scenarios. The OADA we designed uses Re-ID feature metric to calculate appearance similarity to process high-scoring detection frames, uses the DIoU metric to process the low-scoring detection frames, and uses the NSA Kalman filter for motion modeling.

We first divide the detection frame into a high-scoring frame D_{high} and a low-scoring frame D_{low} according to the detection frame score. In the first data association process, we use the high-scoring detection frame D_{high} to match the trajectory generated by NSA Kalman filter. Since the high-scoring detection frame means more obvious features, we use the Re-ID feature metric to calculate the detection frame, and T the similarity between the prediction frames is used to better identify the identity of the tracking object. Finally, the Hungarian algorithm is used to complete the similarity matching and retain the unmatched detection frames $D_{\text{unconfirmed}}$ and tracks $T_{\text{unconfirmed}}$, respectively. The second data association is mainly for low-scoring detection frames D_{low} and unmatched trajectories after the first data association $T_{\text{unconfirmed}}$. We retain unmatched trajectories after this association $T_{\text{reunconfirmed}}$ and delete all unmatched low-scoring detection frames for better to distinguish the foreground and background, we use DIoU to calculate the similarity. Compared with IoU, DIoU not only considers the overlapping area but also takes into account the center distance between the two detection frames. As low-scoring detection frames are often associated with occlusion, incorporating the DIoU metric can enhance the recognition of overlapping occluded objects. A new tracking track will be generated for a detection frame that fails to match the tracking track score after the second data association and has a high enough score. This approach ensures that new objects are detected and tracked, even if they were not associated with any existing track during the initial data association step. For the tracking track that does not match the detection frame, keep 30 frames, and then match it when it appears again.

Algorithm 1: Pseudo-Code of OADA.

Input: UAV video sequences V; detection score threshold τ; object detector GLA − D, NSA Kalman Filter NSA-KF
Output: Tracks T of UAV video
Initialization: $T \Leftarrow \emptyset, D_{\text{high}} \Leftarrow \emptyset, D_{\text{low}} \Leftarrow \emptyset$
for frame f_k in V **do**
  /* predict detection boxes & scores */
  $D_k \leftarrow$ GLA-D(f_k)
  for d in D_k **do**
    if d.score $> \tau$ **then**
      $D_{\text{high}} \leftarrow D_{\text{high}} \cup \{d\}$
    else
      $D_{\text{low}} \leftarrow D_{\text{low}} \cup \{d\}$
    end if
  end for
  /*predict new locations of tracks*/
  for t in T **do**
    $t \leftarrow$ N S A KalmanFilter(t)
  end for
  /* first association */
  Associate T and D_{high} using Re-ID feature metric
  $D_{\text{unconfirmed}} \leftarrow$ unconfirmed object boxes from D_{high}
  $T_{\text{unconfirmed}} \leftarrow$ unconfirmed tracks from T
  /* second association */
  Associate $T_{\text{unconfirmed}}$ and D_{low} using DIoU metric
  $T_{\text{re-unconfirmed}} \leftarrow$ unconfirmed tracks from $T_{\text{unconfirmed}}$
  Delete unconfirmed tracks $T_{\text{re-unconfirmed}}$
  /* initialize new tracks */
  for d in $D_{unconfirmed}$ **do**
    $T \leftarrow T \cup \{d\}$
  end for
end for
return T

The Kalman filter is used to estimate the state of a dynamic system based on a series of measurements with noise and is used in conjunction with a data association algorithm. In the prediction phase, the Kalman filter uses the current state of the object and the motion model to make predictions and obtain the predicted value of the state of the object in the next frame, and in the data association phase the observations are associated to the best matching prediction by calculating the similarity or distance between the observations and the prediction. Finally, the observation information is combined with the prediction information to obtain a more accurate estimate of the object state. To address the limitations of the conventional Kalman filter, which applies a uniform measurement noise scale to all objects regardless of their detection quality, we employ the NSA Kalman filter to achieve more precise motion states. The NSA Kalman filter adjusts the measurement noise scale in an adaptive manner based on the quality of object detection, leading to improved

stability for UAV scenarios with uncertain detection quality. This approach is expected to provide better accuracy and reliability in motion state estimation for UAVs

$$\tilde{R}_k = (1 - c_k) R_k. \tag{8}$$

At the same time, we use DIoU instead of IoU when calculating the similarity of low-scoring detection frames. IoU does not provide any moving gradient when the candidate frame and the real frame do not overlap. DIoU introduces a penalty item based on IoU. b, b^{gt}, respectively, represent the center point of the anchor box and the object box p represents the calculation of the Euclidean distance between the two center points. c represents the diagonal distance of the smallest rectangle that can cover the anchor and the object box at the same time

$$\mathcal{L}_{\mathrm{DIoU}} = 1 - \mathrm{IoU} + \frac{\rho^2(\mathbf{b}, \mathbf{b}^{gt})}{c^2}. \tag{9}$$

One of the key advantages of using the DIoU metric for object detection is its ability to provide accurate moving direction for the bounding box, even when there is no overlap with the object frame. This is achieved by taking into account the distance between the two center points of the anchor frame and the object frame. In addition, the DIoU metric has shown to perform better than other IoU-based metrics in situations of occlusion, where part of the object may be hidden or obscured. By considering both the overlap and distance information, the DIoU metric is able to provide a more comprehensive and reliable measure of object detection accuracy, making it a promising approach for improving the performance of computer vision systems.

III. Experiment

To demonstrate the effectiveness of the GLOA, we chose the appropriate parameters and compare the GLOA with other state-of-the-art trackers. Then, we visualize the qualitative results of the GLOA. Finally, we construct ablation experiments to verify the effectiveness of the GLA-D and the OADA.

A. Experimental Setting

1) Datasets: The GLOA is trained and evaluated on two challenging large-scale datasets captured by drones, namely, VisDrone2019 [47] and UAVDT [48]. The VisDrone2019 dataset was created by the Machine Learning and Data Mining Laboratory of Tianjin University and is designed for tracking and detection in UAV views. The dataset comprises a training set (56 sequences), a validation set (7 sequences), and a test set (17 sequences) that are used in the multiobject tracking task. Each object in every frame is annotated with a bounding box, category label, and tracking ID. The dataset includes ten object categories, namely, pedestrian, person, car, van, bus, truck, motor, bicycle, awning-tricycle, and tricycle, which are considered during the multiobject tracking evaluation.

The UAVDT dataset was introduced by ICCV2018 and includes 50 video sequences with over 80 000 frames, which can be utilized for both object tracking and detection tasks. This dataset is specifically designed for complex scenes and comprises three categories: cars, trucks, and buses. The MOT task is split into a training set (30 sequences) and a test set (20 sequences), focusing only on a single category, cars. The video frames have a resolution of 1080 $\times$ 540 pixels and capture a range of common scenarios, such as squares, arterial streets, and toll stations.

2) Evaluation Metrics: To comprehensively compare the GLOA with other state-of-the-art methods, we applied multiple metrics to measure the tracking performance of our tracker. One of the most commonly used metrics for evaluating object tracking performance is multiple object tracking accuracy (MOTA). MOTA takes into account various errors that may occur during the tracking process and is defined as

$$\mathrm{MOTA} = 1 - \frac{\mathrm{FN} + \mathrm{FP} + \mathrm{IDS}}{\mathrm{GT}} \tag{10}$$

where FN refers to the number of missed detections and FP refers to the number of false alarms. The calculation of FN is based on a comparison between the labeling of the object and the output of the tracker. The calculation of FP is based on the comparison between the output result of the tracker and the labeled objects and IDS refers to the number of identity switches. The number of IDs is affected by the detection performance, the type and number of tracks in the actual tracking process, if the detection performance is better, the more types and numbers of tracks are detected then the more objects are detected, which leads to more ID switches in the same situation. These metrics are computed with respect to the total number of ground truth bounding boxes (GT). Another important metric for tracking robustness is the identification F1 score (IDF1), which is defined as follows:

$$\mathrm{IDF1} = \frac{2\mathrm{IDTP}}{2\mathrm{IDTP} + \mathrm{IDFP} + \mathrm{IDFN}}. \tag{11}$$

IDTP refers to the number of correctly identified trajectories, IDFP refers to the number of falsely identified trajectories, and IDFN refers to the number of missed trajectories.

3) Implementation Details: To improve the robustness of GLOA under complex scenarios, we incorporated YOLOv5's data augmentations and utilized the YOLOv5l backbone as the feature extraction network. The GLOA was trained for 20 epochs on the VisDrone2019 and UAVDT training sets using the SGD optimizer with a batch size of 12. We set the initial learning rate to 1×10^{-2} and resized each input image to a resolution of 1088 $\times$ 608. Our experimental setup comprised a computer equipped with Tesla V100 and used python 3.6 and PyTorch 1.7.0 as the compilation environment.

B. Parameter Setting Experiment

Different training weights are available for YOLOv5, so we tested the different weights on two datasets. Finally, taking into account both the number of parameters and tracking accuracy, we have selected YOLOv5l as our training weights. The experimental results are shown in Table III

At the same time, we set the confidence threshold to 0.3 to 0.5, and observed the impact of different parameters. Experimentally, we found that when the confidence threshold is 0.4, we can have the best tracking effect. Table IV displays the results of the comparison.

TABLE I
COMPARISON OF DIFFERENT METHODS ON VISDRONE2019

Method	MOTA↑(%)	MOTP↑(%)	IDF1↑(%)	MT↑	ML↓	FP↓	FN↓	IDs↓	FM↓
SORT	26.4	73.7	36.4	427	1224	18740	204623	3264	6640
IOUT	28.1	74.7	38.9	467	**670**	36158	**126549**	2393	**3829**
DeepSORT	12.1	72.4	26.9	219	1258	48467	218194	3860	17024
ByteTrack	25.1	72.4	40.8	446	1099	34044	194984	1590	9677
BoT-SORT	23	73.9	41.4	155	1111	27156	208980	**1014**	34440
UAVMOT	25.0	73.4	40.5	443	1105	33603	195522	1644	9702
GLOA (ours)	**39.1**	**76.1**	**46.2**	**581**	824	**18715**	158043	4426	5892

The bold entities is the best experimental results.

TABLE II
COMPARISON OF DIFFERENT METHODS ON UAVDT

Method	MOTA↑(%)	MOTP↑(%)	IDF1↑(%)	MT↑	ML↓	FP↓	FN↓	IDs↓	FM↓
IOUT	36.6	72.1	23.7	534	357	42245	163881	9938	10463
CMOT [49]	36.9	74.7	57.5	**664**	351	69109	144760	1111	3656
SORT	39	74.3	43.7	484	400	**33037**	172628	2350	5787
DeepSORT	40.7	73.2	58.2	595	338	44868	155290	2061	6432
MDP [50]	43	73.5	61.5	647	324	46151	147735	541	4299
UAVMOT	46.4	72.7	67.3	624	221	66352	115940	456	5590
GLOA (ours)	**49.6**	**79.8**	**68.9**	626	**220**	55822	**115567**	**433**	**3589**

The bold entities is the best experimental results.

C. Comparison With State-of-the-Arts

We compare the GLOA with preceding state-of-the-art trackers on VisDrone2019 and UAVDT benchmarks. The experimental results are reported in Tables I and II, the experimental results show that the GLOA has a more significant advantage.

VisDrone2019 dataset: The performance of our method on the VisDrone2019 test dev set is highly promising, as demonstrated by the results presented in Section I. Specifically, our approach achieves an impressive MOTA score of 39.1% and IDF1 score of 46.2%, both of which outperform the existing methods. These results suggest that our method has the potential to significantly improve the accuracy and effectiveness of drone-based visual tracking applications.

UAVDT dataset: We also compare our method with other methods on the UAVDT test set. We train the GLOA using the UAVDT training set and evaluate GLOA on the UAVDT test set. We list a series of indicators, such as MOTA, MOTP, and IDF1, to compare the performance of our method with other methods. As shown in Table II, our method achieves 49.6% on MOTA and 68.9% on IDF1 and gets significantly better results against existing methods.

To verify the real-time performance of our model, we compared the tracking speed with the current representative multiobject tracking algorithms on the VisDrone2019 dataset, including the algorithms DeepSORT and BoT-SORT, etc., based on the SDE framework, and the algorithm UAVMOT based on the JDE framework. The comparison results are shown in the Fig. 6, and we can see that our model has a better real-time performance while ensuring accuracy.

At the same time, we further demonstrate the ability of GLA-D and OADA to solve the scale variation and occlusion problems through Figs. 7 and 8.

D. Visualization

To demonstrate the effectiveness of our method clearly, we have presented the tracking results on the VisDrone2019 test dev set and UAVDT test set. As shown in Fig. 4 and Fig. 5, our GLOA method can adapt well to the dynamic UAV environment, particularly with its ability to capture scale variation features and handle occlusion problems. The visualizations of the results clearly show that the GLOA performs well in completing the MOT task on UAV videos.

E. Ablation Study

In this section, we conduct a series of ablation experiments on the VisDrone2019 test dev set to verify each module of GLOA. In ablation experiments, we use JDE as the baseline model. We consider the GLOA with the GLA-D and the OADA removed as the baseline tracker. The experimental results are summarized in Table V. The GLA-D aims to improve the ability to capture global and local features. When the baseline tracker is equipped with the GLA-D, MOTA, and IDF1

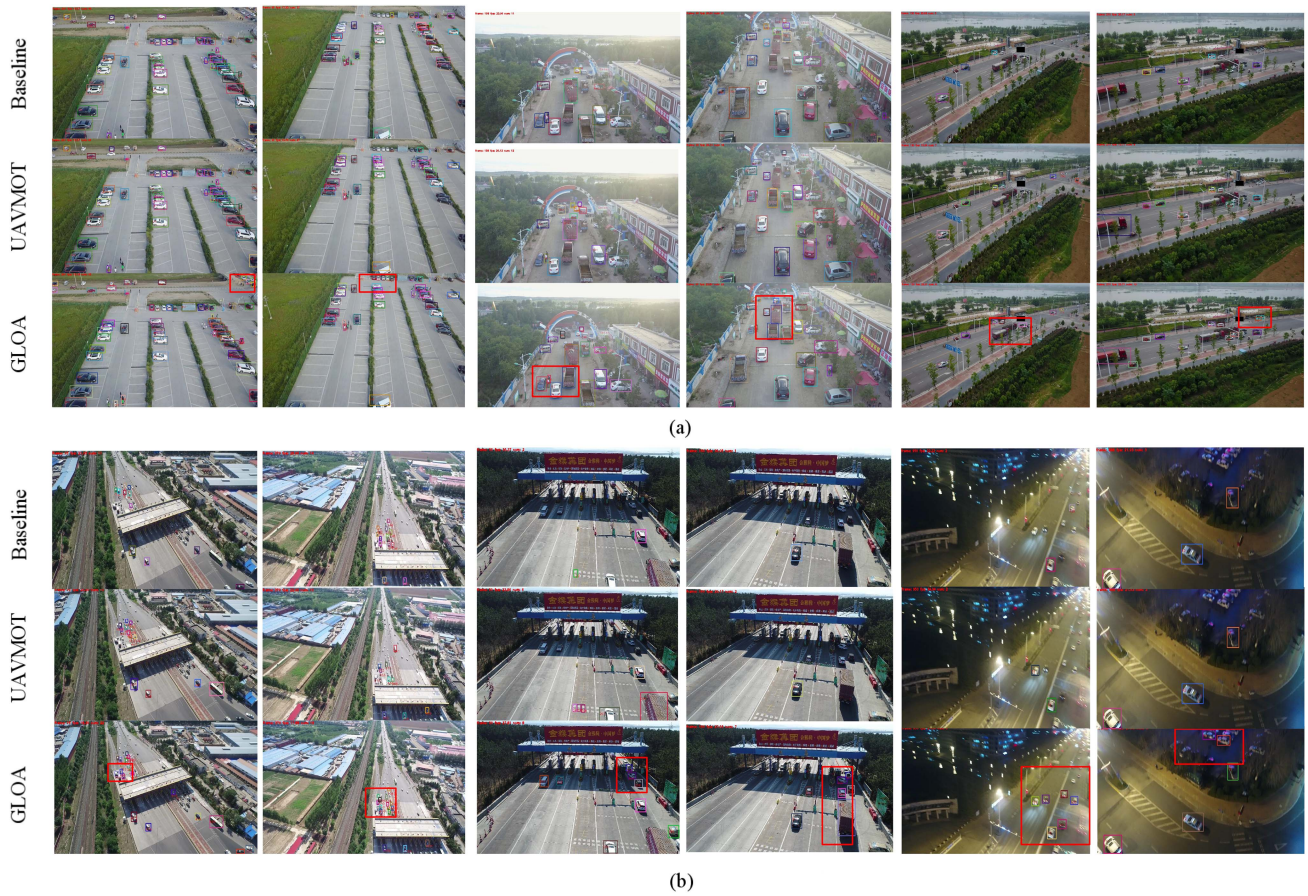


Fig. 4. Visualization results on two datasets (a) and (b). The effectiveness of GLOA has been revealed by visualizing and comparing it with baseline and UAVMOT. To present the visualization results more clearly the key areas have been marked in the figure with red borders. (a) Visualization results on Visdrone2019. (b)Visualization results on UAVDT.

Fig. 5. Visualization of global feature information extraction ability. We extracted three frames from the video for analysis. Taking the tracking object shown by the yellow arrow as an example, our model can accurately locate the object in all three frames as the size and object size change, and the ID of the tracking object is always 1357 without switching. (a) Frame 150. (b) Frame 350. (c) Frame 450.

TABLE III
EFFECT OF WEIGHT SELECTION ON RESULTS

Weight	Dataset	Params$\times(10^6)$	AP↑(%)	MOTA↑(%)
YOLOv5s	VisDrone2019	42.0	29.2	30.5
	UAVDT		59.7	44.2
YOLOv5m	VisDrone2019	48.6	32.4	32.2
	UAVDT		66.8	45.8
YOLOv5l	VisDrone2019	55.0	**38.2**	**39.1**
	UAVDT		**70.2**	**49.6**

The bold entities is the best experimental results.

TABLE IV
EFFECT OF CONFIDENCE THRESHOLD ON RESULTS

Confidence Threshold	MOTA↑(%)	IDF1↑(%)
0.3	37.1	**47.7**
0.4	**39.1**	46.2
0.5	33.6	41.3

The bold entities is the best experimental results.

increase by 12.5% (24.3%–36.8%) and 10.7% (33.1%–43.8%), respectively.OADA is designed to improve occlusion problems and tracking stability. The OADA brings gains of 1.2% (24.3%–25.5%) on MOTA and 2.4% (33.1%–35.5%) on IDF1 for the baseline tracker. Meanwhile, IDS drops from 5392 to 4382. To verify that DIoU can have better results than IoU in the OADA, we conducted experiments on it as shown in Table VI . When we combine both the GLA-D and the OADA into the baseline tracker, GLOA outperforms the baseline tracker by 14.8% (24.3%–39.1%) on MOTA and 13.1% (33.1%–46.2%) on IDF1. The ablation experiments demonstrate the effectiveness of the GLA-D and the OADA.

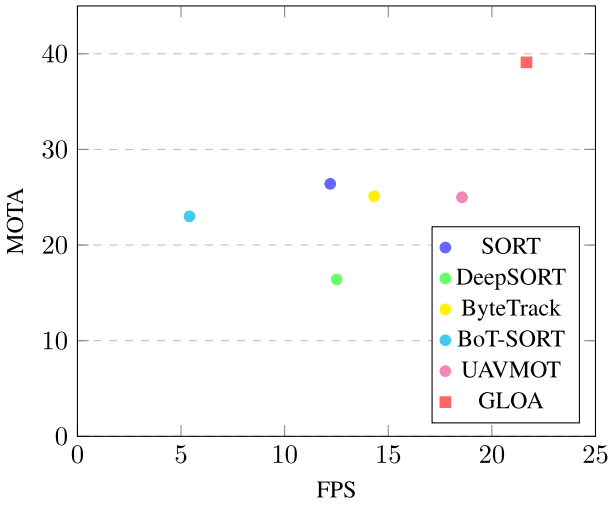


Fig. 6. Speed comparison with other multiobject tracking algorithms.

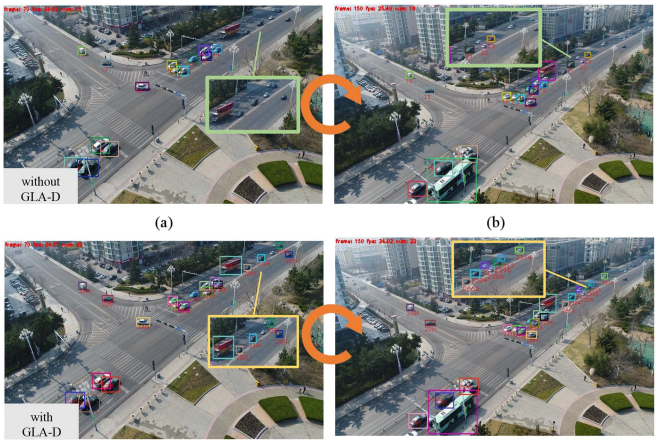


Fig. 7. Scale change robustness. By using our design GLA-D can better cope with scale change scenarios.

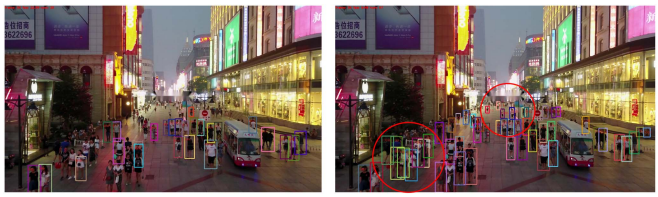


Fig. 8. Ability of occlusion awareness. Occlusion problems in dense scenes can be better mitigated by using OADA. The red circle marked as the key dense areas. (a) Without OADA. (b) With OADA.

TABLE V
RESULTS OF THE ABLATION STUDY

BaseLine	GLA-D	OADA	MOTA↑(%)	IDF1↑(%)	IDs↓
✓			24.3	33.1	5392
✓	✓		36.8	43.8	5756
✓		✓	25.5	35.5	**4382**
✓	✓	✓	**39.1**	**46.2**	4426

The bold entities is the best experimental results.

TABLE VI
EFFECT OF DIOU IN OADA

	MOTA↑(%)	IDF1↑(%)	IDs
OADA+IoU	37.5	44.5	4616
OADA+DIoU	**39.1**	**46.2**	**4426**

The bold entities is the best experimental results.

IV. CONCLUSION

In this article, we proposed a novel GLOA network for multiple object tracking in UAV remote sensing videos. To capture scale variance feature information of the remote sensing scenes better, we fully combine the advantages of CNN and transformer to design GLA-block to extract global and local features and added an identity embedding output branch to the prediction head of GLA-D to extract more discriminatory identity information. In addition, we designed a data association method called OADA, in order to alleviate the occlusion problem during UAV tracking. We conduct a series of experiments on VisDrone2019 and UAVDT datasets and compare GLOA with other methods. The results demonstrate that our method has advantages compared with other existing multiobject tracking methods in the UAV scenes. The impact of extreme weather, such as foggy weather, rain, and snow, on tracking performance is not fully considered in our proposed GLOA. We will improve and address this issue in our future work.

REFERENCES

[1] P. Zhu et al., "Detection and tracking meet drones challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7380–7399, Nov. 2022.
[2] H. Menouar, I. Guvenc, K. Akkaya, A. S. Uluagac, A. Kadri, and A. Tuncer, "UAV-enabled intelligent transportation systems for the smart city: Applications and challenges," *IEEE Commun. Mag.*, vol. 55, no. 3, pp. 22–28, Mar. 2017.
[3] X. Li and J. Wu, "Extracting high-precision vehicle motion data from unmanned aerial vehicle video captured under various weather conditions," *Remote Sens.*, vol. 14, no. 21, 2022, Art. no. 5513.
[4] H. Shen, D. Lin, and T. Song, "Object detection deployed on UAVs for oblique images by fusing IMU information," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jan. 7, 2022, Art. no. 6505305.
[5] Z. Zhang, Y. Liu, T. Liu, Z. Lin, and S. Wang, "DAGN: A real-time UAV remote sensing image vehicle detection framework," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 11, pp. 1884–1888, Nov. 2020.
[6] C. Fu, J. Ye, J. Xu, Y. He, and F. Lin, "Disruptor-aware interval-based response inconsistency for correlation filters in real-time aerial tracking," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6301–6313, Aug. 2021.
[7] 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, Jul. 8, 2021, Art. no. 5606913.
[8] 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.
[9] K. Zhang, W. Wang, and J. Wang, "Robust correlation tracking in unmanned aerial vehicle videos via deep target-specific rectification networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, May 30, 2022, Art. no. 6510605.
[10] H. Zuo, C. Fu, S. Li, J. Ye, and G. Zheng, "DeconNet: End-to-end decontaminated network for vision-based aerial tracking," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 16, 2022, Art. no. 5635712.
[11] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, "Towards real-time multi-object tracking," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., 2020, pp. 107–122.

[12] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 3464–3468.
[13] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 3645–3649.
[14] Y. Du et al., "Strongsort: Make deepsort great again," *IEEE Trans. Multimedia*, to be published, doi: 10.1109/TMM.2023.3240881.
[15] J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. Kitani, "Observation-centric sort: Rethinking sort for robust multi-object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 9686–9696.
[16] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "Bot-sort: Robust associations multi-pedestrian tracking," Jun. 29, 2022, *arXiv:2206.14651*.
[17] G. Maggiolino, A. Ahmad, J. Cao, and K. Kitani, "Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification," Feb. 23, 2023, *arXiv:2302.11813*.
[18] T. Chen, A. Pennisi, Z. Li, Y. Zhang, and H. Sahli, "A hierarchical association framework for multi-object tracking in airborne videos," *Remote Sens.*, vol. 10, no. 9, 2018, Art. no. 1347.
[19] J. Zhang, X. Zhang, Z. Huang, X. Cheng, J. Feng, and L. Jiao, "Bi-directional multiple object tracking based on trajectory criteria in satellite videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jan. 12, 2023, Art. no. 5603714.
[20] Y. Du, J. Wan, Y. Zhao, B. Zhang, Z. Tong, and J. Dong, "Giaotracker: A comprehensive framework for mcmot with global information and optimizing strategies in visdrone 2021," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2809–2819.
[21] W. Li, J. Mu, and G. Liu, "Multiple object tracking with motion and appearance cues," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019.
[22] J. Wu et al., "Multiple ship tracking in remote sensing images using deep learning," *Remote Sens.*, vol. 13, no. 18, 2021, Art. no. 3601.
[23] D. Avola et al., "MS-faster R-CNN: Multi-stream backbone for improved faster R-CNN object detection and aerial tracking from UAV images," *Remote Sens.*, vol. 13, no. 9, 2021, Art. no. 1670.
[24] S. Pan, Z. Tong, Y. Zhao, Z. Zhao, F. Su, and B. Zhuang, "Multi-object tracking hierarchically in visual data taken from drones," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019.
[25] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," *Int. J. Comput. Vis.*, vol. 129, pp. 3069–3087, 2021.
[26] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 474–490.
[27] Y. Lin, M. Wang, W. Chen, W. Gao, L. Li, and Y. Liu, "Multiple object tracking of drone videos by a temporal-association network with separated-tasks structure," *Remote Sens.*, vol. 14, no. 16, 2022, Art. no. 3862.
[28] H. Wu, J. Nie, Z. He, Z. Zhu, and M. Gao, "One-shot multiple object tracking in UAV videos using task-specific fine-grained features," *Remote Sens.*, vol. 14, no. 16, 2022, Art. no. 3853.
[29] S. Liu, X. Li, H. Lu, and Y. He, "Multi-object tracking meets moving UAV," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8876–8885.
[30] H. Wu, Z. He, and M. Gao, "GCEVT: Learning global context embedding for vehicle tracking in unmanned aerial vehicle videos," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Dec. 12, 2022, Art. no. 6000705.
[31] Q. He, X. Sun, Z. Yan, B. Li, and K. Fu, "Multi-object tracking in satellite videos with graph-based multitask modeling," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 6, 2022, Art. no. 5619513.
[32] Y. Hong, D. Li, S. Luo, X. Chen, Y. Yang, and M. Wang, "An improved end-to-end multi-target tracking method based on transformer self-attention," *Remote Sens.*, vol. 14, no. 24, 2022, Art. no. 6354.
[33] P. Sun et al., "Transtrack: Multiple object tracking with transformer," 2020, *arXiv:2012.15460*.
[34] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K. 2020, pp. 213–229.
[35] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei, "MOTR: End-to-end multiple-object tracking with transformer," in *Proc. 17th Eur. Conf. Comput. Vis.*, Tel Aviv, Israel, 2022, pp. 659–675.
[36] P. Chu, J. Wang, Q. You, H. Ling, and Z. Liu, "TransMOT: Spatial-temporal graph transformer for multiple object tracking," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 4870–4880.
[37] T. Fortmann, Y. Bar-Shalom, and M. Scheffe, "Sonar tracking of multiple targets using joint probabilistic data association," *IEEE J. Ocean. Eng.*, vol. OE-8, no. 3, pp. 173–184, Jul. 1983.
[38] R. L. Streit and T. E. Luginbuhl, "Probabilistic multi-hypothesis tracking," Naval Underwater Systems Center, Newport, RI, USA, Tech. Rep. 58, 1995.
[39] G. A. Mills-Tettey, A. Stentz, and M. B. Dias, "The dynamic Hungarian algorithm for the assignment problem with changing costs," Robot. Inst., Pittsburgh, PA, Tech. Rep. CMU-RI-TR-07-27, 2007.
[40] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, vol. 82, pp. 35–45, 1960.
[41] E. Bochinski, V. Eiselein, and T. Sikora, "High-speed tracking-by-detection without using image information," in *Proc. IEEE 14th Int. Conf. Adv. Video Signal Based Surveill.*, 2017, pp. 1–6.
[42] E. Bochinski, T. Senst, and T. Sikora, "Extending IOU based multi-object tracking by visual information," in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, 2018, pp. 1–6.
[43] L. Chen, H. Ai, Z. Zhuang, and C. Shang, "Real-time multiple people tracking with deeply learned candidate selection and person re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2018, pp. 1–6.
[44] Y. Zhang et al., "Bytetrack: Multi-object tracking by associating every detection box," in *Proc. 17th Eur. Conf. Comput. Vis.*, Tel Aviv, Israel, 2022, pp. 1–21.
[45] X. Xu et al., "STN-Track: Multiobject tracking of unmanned aerial vehicles by swin transformer neck and new data association method," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8734–8743, Oct. 10, 2022.
[46] Y. Rao, W. Zhao, Y. Tang, J. Zhou, S.-N. Lim, and J. Lu, "HOrNet: Efficient high-order spatial interactions with recursive gated convolutions," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 6, 2022, vol. 35, pp. 10353–10366.
[47] L. Wen et al., "VisDrone-MOT2019: The vision meets drone multiple object tracking challenge results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019.
[48] D. Du et al., "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 370–386.
[49] S.-H. Bae and K.-J. Yoon, "Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1218–1225.
[50] Y. Xiang, A. Alahi, and S. Savarese, "Learning to track: Online multi-object tracking by decision making," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4705–4713.
[51] H. Zhang, W. Hu, and X. Wang, "Parc-Net: Position aware circular convolution with merits from ConvNets and transformer," in *Proc. 17th Eur. Conf. Comput. Vis.*, Tel Aviv, Israel, 2022, pp. 613–630.
[52] Q. Hou, C.-Z. Lu, M.-M. Cheng, and J. Feng, "Conv2former: A simple transformer-style ConvNet for visual recognition," Nov. 22, 2022, *arXiv:2211.11943*.

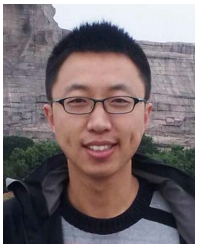
Lukui Shi received the B.S. degree in computer and application and the M.S. degree in computer application technology from the Hebei University of Technology, Tianjin, China, in 1996 and 1999, respectively, and the Ph.D. degree in computer application technology from Tianjin University, Tianjin, in 2006.

Since 2014, he has been a Professor with the School of Artificial Intelligence, Hebei University of Technology. He has authored two books and more than 20 articles. His research interests include machine learning, lung sound recognition, and data digging.

Dr. Shi was a Member of the Discrete Intelligent Computing Professional Committee of the Chinese Association for Artificial Intelligence, and the Visual Big Data Professional Committee of China Society of Image and Graphics.

Qingrui Zhang received the B.S. degree in computer science and technology from Hebei Agricultural University, Hebei, China, in 2019. He is currently working toward the M.S. degree in computer technology with the School of Artificial Intelligence, Hebei University of Technology, Tianjin, China.

His research interests include machine learning and intelligent computing.



Bin Pan (Member, IEEE) received the B.S. and Ph.D. degrees in engineering from the School of Astronautics, Beihang University, Beijing, China, in 2013 and 2019, respectively.

Since 2019, he has been an Associate Professor with the School of Statistics and Data Science, Nankai University, Tianjin, China. His research interests include machine learning, remote sensing image processing, and multiobjective optimization.

Jun Zhang received the B.S. and Ph.D. degrees in engineering from the Hebei University of Technology, Tianjin, China, in 1999 and 2011, respectively.

He is currently an Associate Professor with the School of Artificial Intelligence, Hebei University of Technology. His research interests include machine learning and intelligent computing.

Yuanchao Su (Senior Member, IEEE) received the B.S. and M.Sc. degrees from the Xi'an University of Science and Technology, Xi'an, China, in 2012 and 2015, respectively, and the Ph.D. degree from Sun Yat-Sen University, Guangzhou, China, in 2019, all in engineering.

From 2018 to 2019, he was a Visiting Researcher with Advanced Imaging and Collaborative Information Processing Group, Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN, USA. In 2019, he joined the Department of Remote Sensing, College of Geomatics, Xian University of Science and Technology, where he is an Assistant Professor and a Lecturer. His research interests include hyperspectral unmixing, target detection, neural network, and deep learning.