RoMATer: An end-to-end robust multiaircraft tracker with transformer

Xujie He[+], Jing Jin[+,*],Duo Chen[+], Cangtian Zhou[+], Jiale Jiang[+] and Yuhan Chen[+]
[+]*School of Astronautics, Harbin Institute of Technology, Harbin 150001, China*
hexujie@stu.hit.edu.cn, jinjinghit@hit.edu.cn

2024 International Joint Conference on Neural Networks (IJCNN) | 979-8-3503-5931-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IJCNN60899.2024.10650101

Abstract*—Multiple aircraft tracking (MAT) plays a critical role in military and civil aerial surveillance systems. Many studies have focused on tracking multiple pedestrians and automobiles, leaving a gap in related research on MAT because of the peculiar tracking properties of multiple aircraft, such as small/tiny object formation properties, identical shapes and appearances, severe camera jitter and trail cloud occlusion. In this paper, to fill this gap, building on top of a tracker named TransTrack, we present a robust MAT method to address multiple aircraft tracking, referred to as RoMATer. The improvements are threefold: a receptive field enlarging (RFE) module is first integrated into the backbone of a feature extractor to assist in feature extraction of full-scale aircraft, and a context-aware encoder layer (CaEL) is proposed to introduce global and local contextual embeddings to provide supplementary discriminative tracking information; moreover, a motion and appearance association (MA-A) module is proposed to overcome the tracking challenges of aircrafts possessing highly identical shapes and appearances. Extensive experiments on our established HIT-MATD (MAT Dataset) dataset (the first multiaircraft tracking dataset) verify the SOTA performance of RoMATer on multiaircraft tracking, with an increase of ~10% in terms of the MOTAL and an increase of ~5% with respect to the recall compared to that of TransTrack. Experiments on public RarePlanes dataset verify the effectiveness of the proposed modules in detecting multiple aircraft at full scales. Moreover, RoMATer can also run at a high frame rate (~11FPS, Nvidia-A100).

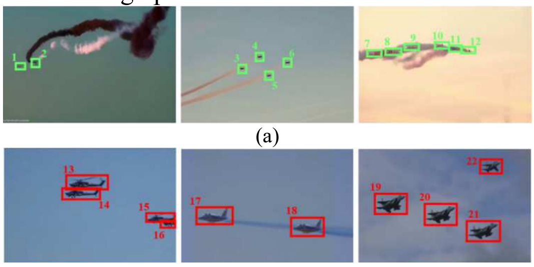
***Keywords*—Multiple aircraft tracking, aerial surveillance systems, receptive field enlarging module, context-aware encoder layer, motion and appearance association module**

I. Introduction

Multiple aircraft tracking (MAT) technology plays a pivotal role not only in military target identification but also in civil air transportation, leading to an important part of the situational awareness-and-avoidance systems of unmanned aerial vehicles (UAVs). Current ATC (air traffic control) systems mostly use surveillance radar and multipoint positioning to track targets in the airspace, but this type of surveillance has shortcomings, such as low image resolution, poor anti-jamming ability and high equipment cost. With the advantages of low cost, small fade area and direct display, visual image-based multiple object tracking technology has become a better alternative to aircraft tracking [1,2] and is gradually being adopted by an increasing number of systems, such as situational awareness systems used in unmanned scout and bionic aircraft.

Current research on aircraft tracking (AT) is relatively scarce compared to that on multiple object tracking (MOT; tracking of both pedestrians and automobiles), and the overwhelming majority of ATs are designed for single aircraft tracking (SAT). To our knowledge, the current methods applied for SAT can be categorized into two groups based on image sources: visual image-based SAT (vi-SAT) and infrared image-based SAT (ii-SAT). Widely used methods for coping with ii-SAT include edge information [3], correlation [4], centroid information [5], mean shift [6], Kalman filtering [7] and neural networks [4,6]. Since infrared images have shortcomings, such as low resolution, low SNR (signal-noise ratio), limited textural features and high expense, visual image-based SATs have radiated new vitality because of their high resolution of images and adequate textural features. Methods for accessing vi-STAs include morphology [8], optical flow [9], Kalman filtering [10], particle filtering [11], clustering [12], mean shift [13], neural networks [6,10,16] and other theories, such as multiple instance learning [14]. However, when regarding MAT, research [15] is not as common as SAT is, and because of the lack of datasets, research on MAT is still in an early stage of its study in the literature and its applications remain largely unexplored. Therefore, addressing MATs remains significant and challenging.

Compared to the multiple object tracking (MOT) methods applied to pedestrians and automobiles, in addition to the difficulties that normally occur in MOT systems, such as occlusion and illumination variation, MAT introduces several other tough problems to address:



(a)

(b)

Fig. 1. Illustration of the challenges that MAT introduces. All the aircraft in (a) are small and even tiny; however, the aircraft in (b) are not as small as those in (a) but have exactly the same shape and appearance, introducing new challenges when tracking. Best view in color.

1) *Small/tiny object formation property*: The vast airspace results in a long distance between aircraft and cameras. Therefore, the aircraft captured in images usually tend to be small or even tiny, as shown in Fig. 1(a), limiting and decreasing the discriminative ability of the features.

2) *Identical shape and appearance property*: The diversity of pedestrians and automobiles contributes to people and automobiles possessing unique appearance features. However, aircrafts lack this good trait, as shown in Fig. 1(b). All of the aircraft appearing in the same field of view may have exactly identical shapes and appearances.

3) *Trail cloud occlusion property*: As shown in Fig. 1(a), the coexistence of small aircraft and trail clouds makes occlusion under multiaircraft tracking more challenging to solve.

In this paper, to address the problems above, we present

[*] Corresponding author.

This work was supported by Heilongjiang Province's "Million and Ten Million" Major Project in Science and Technology of China under Grant No. 2021ZX10A01.

a robust multiaircraft tracker with transformer, referred to as RoMATer. Our method starts with the baseline TransTrack,and the improvements are made accordingly: we first propose a receptive field enlarging (RFE) module. The main idea of the RFE module is to construct feature extraction blocks with larger kernels to enhance the receptive capability of the feature extractor at full scale, including small and large aircraft, and to inset this information into the backbone of the feature extractor. Given that global features have the advantage of favorable perceiving integrity properties of objects and that local features have the advantage of being more resistant to interfeature interference, we propose the context-aware encoder layer (CaEL), which integrates global and local information within a single layer; moreover, to overcome the problems of tracking homogeneous aircraft, we propose a motion and appearance association (MA-A) module, which reinforces the association ability of the tracker by employing bounding-box changing traits along with color traits of the aircraft. To verify the effectiveness of the methods for addressing MAT, we collected and annotated the first MAT dataset, named HIT-MATD, which contains data from challenging scenarios involving tiny aircrafts, homogeneous appearances, heavy trail cloud occlusions and bad weather conditions. Experiments on HIT-MATD demonstrate the SOTA performance of RoMATer for tracking multiple aircrafts, achieving an increase of ~10% in terms of MOTAL and an increase of ~5% in terms of recall compared to TransTarck, and experiments on a public aircraft dataset termed RarePlanes further demonstrate the effectiveness of the proposed modules for detecting multiple aircrafts.

To summarize, the contributions of our work include the following:

1) A robust multiaircraft tracker with a transformer referred to as RoMATer is proposed.

2) To assist feature extractors in better perceiving aircraft at full scale, a receptive field enlarging (RFE) module is proposed; to provide more stable and discriminative information, a context-aware encoder layer (CaEL) is proposed; and to track aircrafts possessing identical shapes and appearances, a motion and appearance association (MA-A) module is adopted.

3) A new multiple-aircraft tracking dataset named HIT-MATD was collected and annotated. To our knowledge, this is the first visual-image-based dataset used for multiaircraft tracking. The HIT-MATD will be made publicly available to encourage further research.

II. Related Work

Research on MATs is scarce since attention has been given primarily to aircraft tracking in SAT. However, MAT has significant applications in both military and civilian domains, such as military reconnaissance and border surveillance. Researching MAT can assist in improving the coordination of multiple aircraft in formation flying, convoy operations, and air traffic management systems, enhancing the collaborative capabilities and overall efficiency of aircraft. Considering that MAT has the same paradigm as MOT applied for pedestrians and automobiles, and given that neural network-based MOT methods have currently been researched in full swing and that their performance is substantially better than that of tradition theory-based methods [16], we therefore refer to related methods and execute some improvements according to the challenging properties of MAT to enhance its performance for multiple aircraft tracking. Current prevailing paradigms for MOT can be categorized into three types: detection-free tracking (DFT), detection-based tracking (DBT) and detection joint tracking (DJT).

Detection-free tracking (DFT): DFT first initializes a certain number of objects in the first frame and subsequently consecutively tracks them in the following frames. For example, [17] segmented trajectories into foreground and background based on saliency and incorporated object connectivity constraints into the foreground topology-based trajectory weight matrix to track multiple interacting people. [18] utilized the logarithmic Euclidean appearance model based on subspace learning to capture object appearance distribution information and mitigate the occlusion problem. [19] proposed a multiobject model-free tracker to enhance the ability of similar objects to model appearance by introducing spatial constraints into an SVM. However, this type of tracking paradigm is the only one among three that separates from detection; moreover, it has the shortcoming of not being able to track new objects appearing in the view, resulting limited applications in real scenarios.

Detection-based tracking (DBT): DBT usually first detects all the objects in the frame and subsequently applies an association module to link the current objects to the objects in previous frames. Therefore, under DBT, designing a robust association metric is important. For example, [20] proposed the use of a Kalman filter and the IOU metric to accomplish this association upon detection. [21] further integrated appearance information into [20] to enhance the similarity embeddings, and [28] further improved the [20] by fixing the noise in the Kalman filter. Given that the similarity metrics above are all handcrafted, [22] proposed an affinity network to output the similarity matrix, including follow-up trackers [23]. [24] reused detections lower than the given threshold to accomplish the association and increase the continuity of trajectories. To strengthen the ability of appearance modeling methods, information from other data sources, such as point clouds [25] and multisensor information [26], is also widely adopted. Currently, the transformer mechanism is also introduced in DBT, such as [27], which encodes the detections from multiple consecutive frames and groups them into different trajectories by track queries. Apparently, this type of paradigm additionally requires an individual detection module integrated in, and then accomplishing tracking in a cascaded manner, which is time- and memory-inefficient and overly relies on the performance of the detector used at times.

Detection joint tracking (DJT): DJT follows the same paradigm as DBT but accomplishes detection and tracking within a single model and is more structure- and time-efficient. Recently, the field of DJT has developed rapidly, and its performance has consequently substantially improved. For example, [29] added a tracking head to the improved R-FCN [30] to simultaneously accomplish detection and tracking and transformed the tracking into a regression task by predicting the relative offset of each target position in two consecutive frames. Likewise, to increase the reusability of features, [31] also introduces a branch of appearance feature extraction head to the original framework to accomplish detection and tracking and elaborate on training multiclass networks; the same goes for [32], which introduces a tracking head to MaskRCNN [33]. [34] proposed increasing the input dimension to two frames, and the observation heatmap [34] can output the object center positions, sizes and relative offsets of the two frames. Given that the ID switches of [31] are mostly caused by inferior ReID features, [35] used deep layer aggregation and deformable convolution to enhance feature embeddings. Based on the real-time detector [36], [37] integrates a simple and effective post-FPN (feature pyramid network) prediction subnetwork into [36] to solve the problem of inadaptability when embedding every instance at different scales. TransTrack [38] is the first MOT tracker with a transformer; it uses the learned object queries and the object feature queries of the previous frame as inputs. The former is decoupled as the detection of the current frame after decoding, and the latter is decoupled as the tracking results after decoding. TransTrack completes the data association in the current frame, which allows it to have a simple association metric, IOU, to accomplish the linking. Later, [39] proposed the use of seamless data association with a new tracking-by-attention paradigm, and [40] applied dense pixel-level multiscale queries together with powerful global representation via a transformer to globally and adequately detect and track the centers of objects. However, because of the peculiarities of MAT, as mentioned in *Section 1*, the above MOT methods are not completely applicable for MAT. In this paper, due to the high efficiency of TransTrack and the ability of TransTrack to incorporate motion information to conduct tracking with transformers, we adopt TransTrack as our baseline to address MAT and improvements were made accordingly. The advancement of our method is that RoMATer is the first MAT tracker and is able to track multiple aircraft robustly.

III. RoMATer

A. Overview

RoMATer, for the first time, follows the paradigm of detection joint tracking, so that it can address multiaircraft detection and tracking in a single model. The proposed RoMATer is composed of four components: an improved backbone with the receptive field enlarging (RFE) module, three context-aware encoder layers (CaEL), double decoders with three layers and a motion and appearance association (MA-A) module, as shown in Fig. 1. Specifically, two consecutive frames, $\boldsymbol{I}_{t\text{-}1}$ and $\boldsymbol{I}_t\in\mathbb{R}^{\mathrm{H}\times\mathrm{W}\times3,}$ along with feature embeddings $\mathcal{F}_{t\text{-}1}\in\mathbb{R}^{500\times256}$ of detections and tracked aircraft locations $\mathcal{B}_{t\text{-}1\ \text{and}\ id}\in\mathbb{R}^{n\times4}$ of the last frame, are required as inputs (the first frame is only used for detection). After feature extraction, $\boldsymbol{I}_t\in\mathbb{R}^{\mathrm{H}\times\mathrm{W}\times3}$ is converted into a preliminary feature embedding $\mathcal{F}_b\in\mathbb{R}^{\mathrm{M}\times256}$, where M = H/8×W/8+ ••• + H/64×W/64. $\mathcal{F}_b$ is then fed into the following context-aware encoder layers to encode features of two consecutive frames and obtain the encoded features $\mathcal{F}_{CaEL}\in\mathbb{R}^{\mathrm{M}\times256}$. We retain the operations of TransTrack, which transmit $\mathcal{F}_{CaEL}$ into two separate (dual) groups of decoder layers. The first group of decoder layers is exploited to detect all aircrafts $\mathcal{B}_{t\mathrm{D}}\in\mathbb{R}^{n\times4}$ of the current frame and outputs its corresponding feature embeddings $\mathcal{F}_t\in\mathbb{R}^{n\times256}$. The second group of decoder layers is exploited to propagate all aircrafts of the last frame to the current frame $\mathcal{B}_{t\mathrm{T}}\in\mathbb{R}^{m\times4}$. Finally, we utilize a motion and appearance association module (MA-A) to match the detected and propagated aircraft to conduct interframe aircraft linking.

In the following section, we elaborate on the receptive field enlarging (RFE) module, context-aware encoder layer (CaEL) and motion and appearance association (MA-A) module, which reside only in RoMATer.

B. Receptive field enlarging module

First, because aircraft tracking has the aforementioned small/tiny object formation property, it is necessary to consider detection for small/tiny aircraft. Previous research [41] has also reported that single-stage detectors cut poor figures when detecting small objects; second, small/tiny object detection for anchor-free detectors, such as the detection head of TransTrack, should be attributed mainly to the receptive field (RF) problem rather than to the registration problem of anchors to objects. In other words, the output under this situation is more influenced by the pixels within the perceptive field zones. At this point, small/tiny aircraft need a larger RF to be perceived by the models; otherwise, the recall rate will be low.

TABLE I

DETAILS OF THE RFE MODULE. NOTE THAT 1-3 IN THE KERNEL COLUMN INDICATE TWO CONVOLUTIONS WITH KERNEL SIZES OF 1×1 AND 3×3, RESPECTIVELY. B, R, AND C INDICATE GROUP NORMALIZATION, RELU AND CONVOLUTION, RESPECTIVELY.

Layer name	Operation	Kernel	Out channels
E0	GN-R-C-C	1-3	2048
E1	C-GN	3	512
E2	C-GN	7	512
SqL	C-GN-R	1	2048

However, for a well-designed neural network, an RF in the same layer can cover an image at a smaller scale but may fail to cover a larger image; for the different objects of a certain image, if the RF has been able to cover the whole image, then a lower layer can cover an exact small object, and the higher layer may cover multiple objects. Therefore, we construct a receptive field enlarging (RFE) module, as shown in Fig. 1. The details of the RFE are listed in Table I. We can obtain the value of the receptive field (*rf*) of each layer by:

$$rf_i(s,k)=\begin{cases} rf_{i-1}+(k-1)\times\prod_{j=1}^{i-1}s_j, & i\geq 1\\ 1, & i=0 \end{cases} \tag{1}$$,

where s and k indicate the adjusted stride size and kernel size of the current layer, respectively, and i is the layer index. As shown in [42], larger kernels still have the potential to address small feature maps; we eventually utilize as 7×7 kernels to provide larger receptive fields.

C. Context-aware encoder layer

The transformer structure has the ability to perceive all the input features due to the self-attention mechanism and feed-forward networks used. Nonetheless, its computational complexity is squared with respect to the input feature size, resulting in inefficiency compared to CNNs, especially for vision tasks such as multiaircraft tracking. Moreover, the global features extracted by the Transformer contain a large amount of redundant information about noninteresting objects. Considering these problems and given that local features use

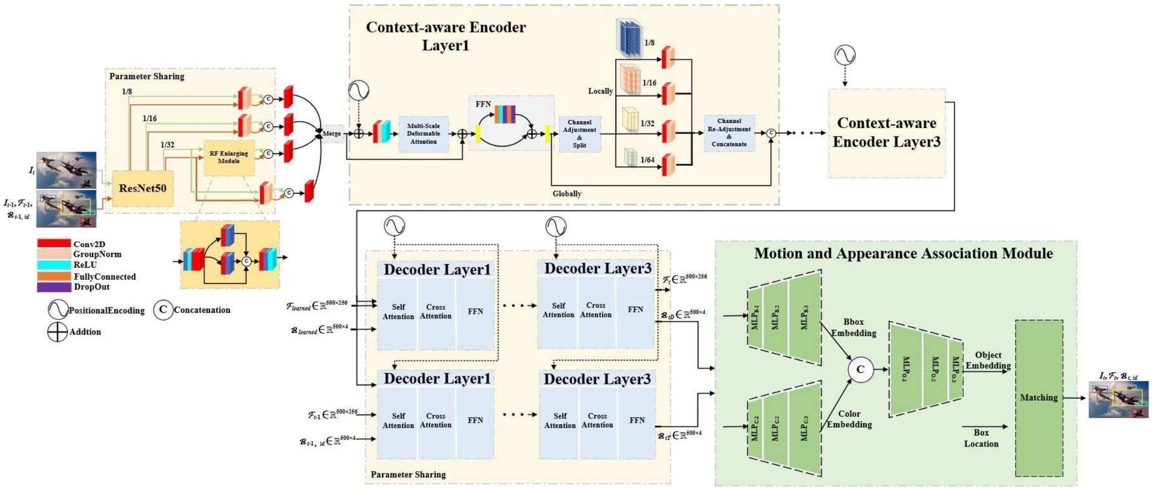


Fig. 2. Deployment of RoMATer. RoMATer takes two consecutive frames, $\boldsymbol{I}_t$ and $\boldsymbol{I}_{t-1}$, along with feature embeddings of detections F_{t-1} and tracked aircrafts $\mathsf{B}_{t-1,id}$ of the last frame as inputs (the first frame conducts detection only so that the latter two information is not needed.). $\boldsymbol{I}_t$ and $\boldsymbol{I}_{t-1}$ are passed into the backbone (ResNet50), which is enhanced by the receptive field enlarging (RFE) module to obtain preliminary features, the preliminary features of $\boldsymbol{I}_t$ and $\boldsymbol{I}_{t-1}$ are merged and are passed into the following three context-aware encoder layers (CaEL) to encode the features that contain interframe information. The encoded features are then transmitted into two groups of decoders along with F_{t-1} and $\mathsf{B}_{t-1,id}$ to decode the encoded features into detections F_t and extended aircrafts B_{tT}. Finally, the motion and appearance association (MA-A) module is applied to associate F_t and B_{tT} of the current frame to accomplish interframe aircraft association. Best view in color.

TABLE II

Details of the CaEL Module. Note that GN, LN, R, and C Indicate Group Normalization, Layer Normalization, ReLU and Convolution, Respectively. C-AS and C-rAS Are Channel Adjustments (readjustments) and Splits, Respectively.

Layer Name	Operation	Kernel	In→Out dimensionality
Conv1	C-GN-R	1	256→128
MSDA			128
FFN	LN-F-R-D-F-R-LN		128
CAS			128
Projector1	C-GN	3	128
Projector2	C-GN	3	128
Projector3	C-GN	3	128
Projector4	C-GN	3	128
C-rAS			128

the most representative, discriminative and stable features to characterize objects within regions of interest, local features are still able to effectively restore the overall information when objects are fuzzy or partial occlusion occurs. At this point, we further introduce the local features of multiple scales in multiaircraft tracking and propose the context-aware encoder layer (CaEL), as shown in Fig. 1.

Specifically, the input frame is converted into feature embeddings $\mathcal{F}_b$ with a size of M×256 after being passed into the backbone, where M=H/8×W/8+…+H/64×W/64. It can be noted that the image information has been flattened into vector embeddings as each row of $\mathcal{F}_b$ indicates a feature location from feature maps of different downsampling rates. Then, $\mathcal{F}_b$ is passed into a multiscale deformable attention module followed by a feed forward network (FFN). It is understandable why traditional encoders of transformers are rich in global information, especially when using an attention mechanism and fully connected layers that take all of the features as input. Therefore, after obtaining the encoded feature embeddings of the FFN ($\mathcal{F}_{FFN}$), we duplicate the $\mathcal{F}_{FFN}$ and utilize it to obtain local feature embeddings, as shown in context-aware encoder layer 1 of Fig. 1 (the upper branch after the FFN). Specifically, we first split the duplicated $\mathcal{F}_{FFN}$ into four segments relying on the four downsampling rates and then reshape the split four segments back into the shape of the image domain (channel adjustment & split/C-AS). The inner features that possess four different scales of the image domain of encoders are consequently obtained. Four projectors are applied to locally encode the obtained features. Finally, we aggregate the local feature embeddings and the global feature embeddings. The whole structure of the CaEL can be formalized as follows:

$$\begin{aligned}\boldsymbol{F}_{\mathrm{CaEL}} &= \mathrm{Cat}[\boldsymbol{F}_{g,p}, \boldsymbol{F}_{l,p}] \\ &= \mathrm{Cat}[\boldsymbol{F}_{g,p}, \mathrm{MSP}(\boldsymbol{F}_{g,p})]\end{aligned} \quad (2),$$

where p is the index of the encoder layer, l and g indicate local and global feature embeddings, respectively, and Cat is the concatenation function. $\boldsymbol{F}_{g,p}$ is given by:

$$\begin{aligned}\boldsymbol{F}_{g,p} &= \mathrm{FFN}\circ\mathrm{MSDA}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) \\ &= \mathrm{FFN}\circ\square\{(\boldsymbol{V}\otimes\boldsymbol{W}^{V})\langle\sum_{h=1}^{8}\sum_{n=1}^{4}(\boldsymbol{Q}\times\boldsymbol{W}^{K})\langle:,:,h,:,n,:\rangle+\boldsymbol{K}\rangle\}\otimes(\boldsymbol{Q}\otimes\boldsymbol{W}^{Q})\square \\ &= \boldsymbol{W}^{FFN}\otimes\square(\boldsymbol{V}\otimes\boldsymbol{W}^{V})\langle\sum_{h=1}^{8}\sum_{n=1}^{4}(\boldsymbol{Q}\times\boldsymbol{W}^{K})\langle:,:,h,:,n,:\rangle+\boldsymbol{K}\rangle\otimes(\boldsymbol{Q}\times\boldsymbol{W}^{Q})\square\end{aligned} \quad (3),$$

where FFN and MSDA are the feed-forward network and multiscale deformable attention, respectively. $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are queries, keys and values used in the MSDA, and $\boldsymbol{W}^{\boldsymbol{Q}}$, $\boldsymbol{W}^{\boldsymbol{K}}$, and $\boldsymbol{W}^{\boldsymbol{V}}$ are learnable weights. MSP($\boldsymbol{F}_{g,p}$) is given by:

$$\mathrm{MSP}(\boldsymbol{F}_{g,p}) = \boldsymbol{F}_{g,p}\times\boldsymbol{W}^{\mathrm{M}} \quad (4),$$

where MSP is the multiscale projector used, which includes projectors 1, 2, 3 and 4. The structural details of the CaEL are listed in Table II. Notably, the dimensionality of the feature embeddings after the CaEL is first reduced to 128. After the corresponding operations of C-rAS, we aggregate the feature embeddings from C-rAS that are extracted locally and feature embeddings from the FFN that are extracted globally to formalize local-global feature embeddings.

D. Motion and appearance association module

Since MAT possesses the challenge of tracking homogeneous shapes and appearance properties, as previously mentioned (Fig. 1), it is insufficient to merely utilize appearance feature embeddings to perform tracking. To address this issue, motion information should be consequently considered. TransTrack integrates motion information by extending the objects of the last frame to the current frame using decoders (which is also one of the aspects in which we adopt TransTrack for MAT). However, the tracking accuracies of MOTAL and IDF1 are still not promising when tracking multiple aircraft (refer to *Section IV*), which indicates that the motion information of TransTrack is still not robust for addressing MAT.

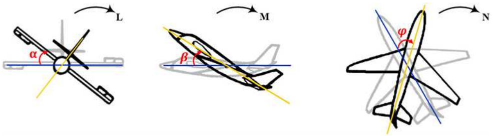


Fig. 3. Illustration of aircraft poses, including roll, pitch and yaw. Best view in color.

Different from previous work regarding motion information that was addressed by modeling trajectories, we propose integrating motion information with changes in the bounding boxes of aircraft. The motivations are derived from the facts that, different from pedestrian and automobile tracking, i) more maneuvering poses, such as pitch, roll and yaw, as shown in Fig. 3, and ii) the high-speed maneuvering property of aircrafts make interframe variations in aircraft more severe, and this severe motion variation can be effectively captured from changes in the bounding boxes. As shown in Fig. 4, we plot the width and height of the annotated boxes of the six aircraft into different figures. To better investigate the maneuvering property of aircraft, we also plot the width and height of annotated boxes of six representative pedestrians in the MOT17 dataset [43]. We note that the majority of the slopes of the curves plotted using aircrafts are greater than those plotted using pedestrians. In addition, due to the poses of aircraft, such as pitch, yaw and roll, the height and width of aircraft change simultaneously over time, unlike pedestrians, who change in only one plane dimension. Based on these aspects, we propose the motion and appearance association (MA-A) module, as shown in Fig. 2.

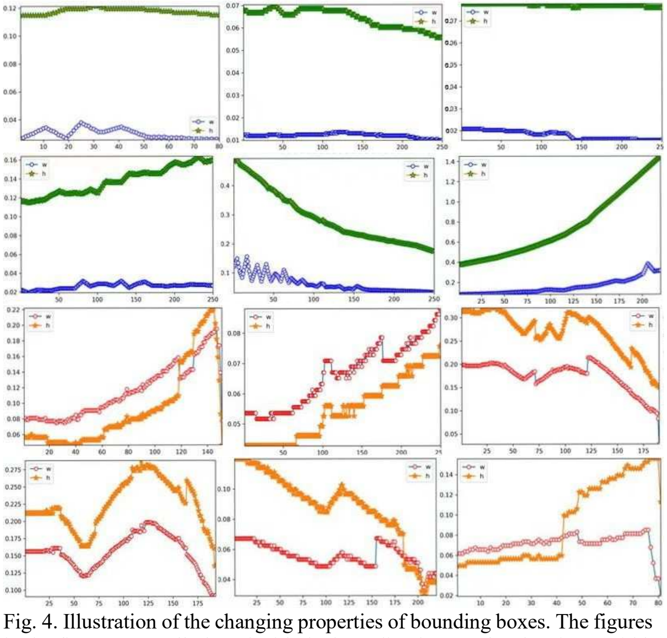


Fig. 4. Illustration of the changing properties of bounding boxes. The figures in the first row are all plotted with the bounding boxes of pedestrians (width and height are reported in green and blue), and the figures in the second row are all plotted with the bounding boxes of aircraft (width and height are reported in red and orange). Best view in color.

TABLE III
DETAILS OF MA-A. NOTE THAT C AND B REPRESENT CONVOLUTION AND BATCH NORMALIZATION, RESPECTIVELY. THE NUMBERS IN THE IN AND OUT COLUMNS INDICATE THE CHANNEL DIMENSIONALITY.

Motion			**Appearance**		
Operation	In	Out	Operation	In	Out
$\mathrm{MLP_{B1}}$(C-B)	2	64	$\mathrm{MLP_{C1}}$(C-B)	3	64
$\mathrm{MLP_{B2}}$(C-B)	64	128	$\mathrm{MLP_{C2}}$(C-B)	64	28
$\mathrm{MLP_{B3}}$(C-B)	128	256	$\mathrm{MLP_{C3}}$(C-B)	128	256

Fusion		
Operation	In	Out
$\mathrm{MLP_{O1}}$(C-B)	256×2	256
$\mathrm{MLP_{O2}}$(C-B)	256	128
$\mathrm{MLP_{O3}}$(C-B)	128	64

Details of the MA-A are listed in Table III. We utilized three groups of MLPs, all with a squashing rate of 1 to encode motion and appearance features. After encoding with the MLPs, aircraft embeddings with a size of k×64 can be obtained, where k represents the number of aircraft shown in the view by the current time stamp. Specifically, we employ the weight, height and appearance of the centroids of bounding boxes to formalize motion and appearance information according to the equation given by:

$$\begin{aligned}\mathrm{OE}(\boldsymbol{B}) &= \mathrm{MLP_O}\cdot[\mathrm{MLP_C}(\boldsymbol{B})+\mathrm{MLP_B}(\boldsymbol{B})]\\ &=\mathrm{MLP_O}\cdot[\boldsymbol{B}(\textstyle\prod_{i=1}^{3}(\boldsymbol{W}^{C,i}+\boldsymbol{W}^{B,i}))]\\ &=\boldsymbol{B}\cdot[\textstyle\prod_{i=1}^{3}(\boldsymbol{W}^{C,i}+\boldsymbol{W}^{B,i})]\times\prod_{j=1}^{3}\boldsymbol{W}^{O,j}\end{aligned} \quad (5),$$

where $\mathsf{B} = \{B_t^D, B_t^T\}$; MLP-C, B, and O are the MLPs applied on color, boxes and final aircraft objects, respectively. $\boldsymbol{W}^{\mathrm{C}}$, $\boldsymbol{W}^{\mathrm{B}}$ and $\boldsymbol{W}^{\mathrm{O}}$ are the learnable weights. Finally, we likewise conduct the interframe association using the aircraft embeddings above together with the box location by simply using the IOU metric and hungarian matching.

IV. EXPERIMENTS

A. *Comparison experiments*

Experimental design: We first conducted a group of comparison experiments. The methods adopted for comparison all follow the paradigm of detection joint tracking and are all SOTA methods and representative of tracking multiple objects, including FairMOT [35], TransTrack [38], TransCenter [40], CenterTrack [34] and our proposed RoMATer.

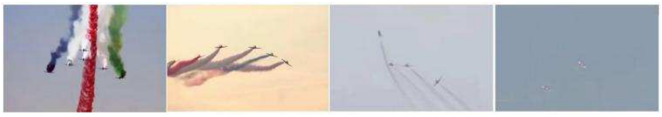


Fig. 5. Frames captured from HIT-MATD. As shown in the figure, aircrafts are tiny, possess homogeneous shapes and appearances, and can easily be occluded by trail clouds. Best view in color.

TABLE IV
DETAILS OF HIT-MATD.

Type	**Videos**	**Length**	**GTBoxes**	**Trajectories**
Train	19	7282	21087	82
Test	13	9223	23876	62

Dataset: There are several datasets that can be used for single aircraft tracking, such as MAV-VID and anti-UAV datasets and sequences containing aircraft, including VOT2015.

However, regarding MAT, there is currently no dedicated dataset available for use in the literature due to commercial and military applications. We therefore collected and meticulously annotated a MAT dataset, which is referred to as the Harbin Institute Technology MAT dataset (HIT-MATD). To our knowledge, this is the first visual-image-based MAT dataset in the literature. To encourage further research on MAT, we plan to make it publicly available to the community. The details of the HIT-MATD are as follows:

The HIT-MATD dataset is the first dataset that can be adopted for multiple aircraft tracking. There are 32 video sequences composed of visual frames of real scenes in total; ~69% of the videos are filmed by cameras mounted on the ground, and the remaining ~31% of the videos are filmed by airborne cameras. The height and width of the frames of the HIT-MATD ranged from 354 to 640 and 190 to 342, respectively. HIT-MATD includes numerous complex and frequently occurring tracking scenarios, such as small/tiny aircraft, identical objects, trail cloud occlusion, considerable scale variation, clutter scenes, heavy fog, frequent interactions, camera jitter and frame blur. Therefore, conducting experiments on HIT-MATD is quite challenging. Details of the HIT-MATD dataset are listed in Table IV and Fig. 5 visualizes four frame samples captured in HIT-MATD.

Evaluation metrics: We utilize widely used metrics to evaluate the tracking results; these metrics include MOTAL, IDF1, ML, MOTP, FP, FN, ID and recall. All the methods adopted were trained with 50 epochs. The initial learning rate (lr_0) was set to 2e-4 and decayed at 35 epochs in the form of $lr_0 \times 0.1$. The image size and batch size were set to 768 and 3, respectively, with GPU memory.

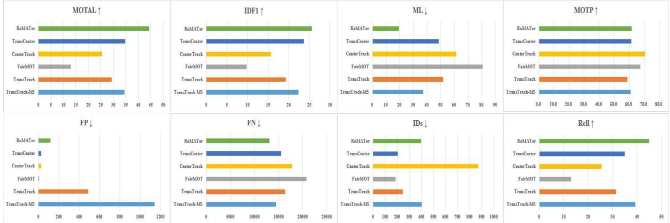


Fig. 6. Bar charts of the comparison results conducted on the HIT-MATD dataset. Best view in color.

Results: The results of the comparisons are listed in Table V. For clarity, we also plot all the results in bar charts, as shown in Fig. 6. We note that our tracker, RoMATer, yields the best results in 5 out of 8 metrics, accounting for 62.5%. Specifically, i) RoMATer yields the best results in terms of a compound tracking performance MOTAL of 44.4, the best results in terms of ID preservation IDF1 of 25.6 among the trackers involved in comparison; ii) RoMATer yields the highest recall of 44.9; iii) RoMATer substantially outperforms the baseline TransTrack in MOTAL, with an increase of ~10% compared to TransTrack with multiscale training and an increase of ~15% compared to TransTrack without multiscale training; additionally, iv) CenterTrack yields the best results in the MOTP of 70.6%, which focuses more on box precision.

B. Ablation study

To further investigate the effectiveness of each module, we conducted another group of ablation experiments on the architectures. The methods involved in ablation were TransTrack with RFE (TransTrack-R), TransTrack with CaEL (TransTrack-C), TransTrack with MA-A (TransTrack-M), TransTrack together with RFE and CaEL (TransTrack-RC), and RoMATer (RFE, CaEL and MA-A). The training protocols and evaluation metrics are consistent with those used in the comparison experiments.

Quantitative results: The ablation results are listed in Table VI, and we also plot the results in bar charts, as shown in Fig. 7, for better observation. From the table and figure, we

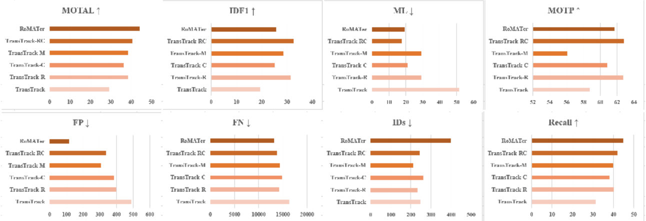
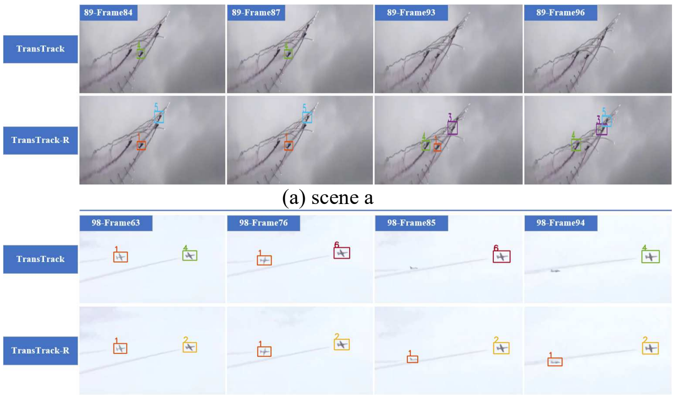


Fig. 7. Bar charts of ablation results conducted on HIT-MATD. Best view in color.



(a) scene a

(b) scene b

Fig. 8. Tracking results under TransTrack and TransTrack-R (TransTrack with REF module). The caption in the top right corner indicates the name of the video and its corresponding frame index. Best view in color.

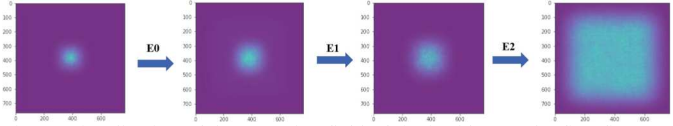


Fig. 9. Visualization of the receptive field of the backbone. The first figure represents the receptive field of the original backbone (ResNet50), and the following three figures represent the receptive field of the backbone after introducing E0, E1 and E2 of the RFE module. Best view in color.

TABLE V

COMPARISON RESULTS ON HIT-MATD. ↑ INDICATES THAT HIGHER VALUES ARE EXPECTED, AND ↓ INDICATES THE OPPOSITE. THE BEST RESULTS ARE REPORTED IN BOLD, RED FONT.

Metrics Tracker	MOTAL ↑	IDF1 ↑	ML ↓	MOTP ↑	FP ↓	FN ↓	IDs ↓	Recall ↑
TransTrack-MS	34.5	22.4	37.1	61.0	1142	14490	404	39.3
TransTrack	29.4	19.3	51.6	58.8	488	16373	248	31.4
FairMOT	13.0	9.8	80.7	67.5	**8**	20750	**188**	13.1
CenterTrack	25.4	15.7	61.3	**70.6**	25	17776	872	25.5
TransCenter	34.8	23.7	48.4	61.3	26	15530	205	35.0
RoMATer	**44.4**	**25.6**	**19.4**	61.7	118	**13149**	399	**44.9**

TABLE VI
ABLATION EXPERIMENTAL RESULTS CONDUCTED ON THE HIT-MATD DATASET. THE BEST RESULTS ARE REPORTED IN BOLD, RED FONT.

Metrics / Tracker	MOTAL ↑	IDF1 ↑	ML ↓	MOTP ↑	FP ↓	FN ↓	IDs ↓	Recall ↑
TransTrack	29.4	19.3	51.6	58.8	488	16373	248	31.4
TransTrack-R	38.6	31.4	29.0	62.7	397	14258	234	40.3
TransTrack-C	36.4	25.2	21.0	60.8	384	14810	264	38.0
TransTrack-M	38.4	28.7	29.0	56.1	307	14401	**212**	39.7
TransTrack-RC	40.6	**32.6**	**17.7**	**62.8**	338	13832	245	42.1
RoMATer	**44.4**	25.6	19.4	61.7	**118**	**13149**	399	**44.9**

note that, i) RoMATer yields 4 (MOTAL, Recall, FP, and FN) out of the 8 best results and 1 competitive result (ML), accounting for 75% of the total; ii) when a single module is added to TransTrack, which are TransTrack-R, TransTrack-C and TransTrack-M, the majority of the evaluation metrics are considerably improved, except for MOTP; and iii) when REF is further introduced together with CaEL, the tracking results are further improved, especially recall, ML, MOTP and MOTAL. FP and IDs at this point slightly increase but are still competitive with those of TransTrack; iii) when the three modules are used together, the results continue to increase, and FP increases when RFE and CaEL decrease again, achieving a final 44.4% MOTAL.

Qualitative results: Fig. 8 represents two groups of tracking results under TransTrack and TransTrack-R (with REF). We note that, i) in scene a, when using TransTrack to track multiple aircraft, only one aircraft is detected and tracked before frame 87; after frame 87, the only detected and tracked aircraft with **ID 4** is lost and keeps being lost until frame 96. However, when using TransTrack-R, the overwhelming majority of the aircraft shown in the view are correctly detected; moreover, the IDs of the detected aircraft all remain through frame 96. ii) In scene b, when using TransTrack, it is noteworthy that the aircraft according to **ID 1** is lost after frame 85, and the aircraft according to **ID 4** frequently switches within **ID 4** and **ID 6** in subsequent frames; when using TransTrack-R, all the aircraft showing in the view are detected, and no switching occurs through frame 94. Additionally, we visualize the receptive field of the backbone by using [42] before and after introducing the RFE module, as shown in Fig. 9. We note that the receptive field becomes increasingly larger after gradually introducing E0, E1 and E2 of the RFE module, as expected.

C. Experiments on the detection head

RoMATer follows the paradigm of detection joint tracking. Therefore, we use RoMATer for detecting aircraft in this section to investigate its performance in aircraft detection. Because no aircraft association is performed, we use only TransTrack-R, TransTrack-C and TransTrack-RC to conduct aircraft detection experiments in this section.

Dataset: The dataset used for aircraft detection is currently the largest publicly available ultrahigh-resolution aircraft detection dataset, named RarePlanes [44], which is composed real scenario data and synthetic data. The real scenario data we used in our experiments span 112 locations and carry 14700 manually annotated aircraft. All the images were classified according to 5 features, 10 attributes and 33 subattributes.

Results: The quantitative results are listed in Table VII. The metric we adopted for detection evaluation is the mAP, which indicates the average precision of a detector; note that the mAP reflects the recall performance. We note that TransTrack-R, TransTrack-C and TransTrack-RC all show substantial improvements in aircraft detection compared with TransTrack; in particular, TransTrack-R yields 4 out of the 5 best metrics, accounting for 80%.

TABLE VII
DETECTION RESULTS ON RARAPLANES. THESUBSCRIPTS 50 AND 75 REPRESENT THE APS OBTAINED BY IOU THRESHOLDS OF 50 AND 75, RESPECTIVELY. S, M AND L INDICATE THAT THE OBJECT SIZES ARE SMALLER THAN 32^2 PIXELS, LARGER THAN 32^2 BUT SMALLER THAN 96^2 PIXELS AND LARGER THAN 96^2 PIXELS, RESPECTIVELY.

Metrics / Tracker	mAP_{50}	mAP_{75}	mAP_S	mAP_M	mAP_L
TransTrack	14.6	1.7	3.3	4.7	10.5
TransTrack-R	**55.3**	**26.4**	24.1	**29.1**	**37.5**
TransTrack-C	51.5	25.1	**24.2**	26.0	34.1
TransTrack-RC	44.5	19.4	21.7	22.1	31.5

V. CONCLUSION

In this paper, we first present a multiaircraft tracker referred to as RoMATer by integrating the receptive field enlarging (RFE), context-aware encoder layer (CaEL) and motion and appearance association (MA-A) modules into a single pipeline. Moreover, we collected and meticulously annotated the first visual image-based MAT dataset, referred to as HIT-MATD, which covers challenging scenarios that often occur during multiaircraft tracking, and HIT-MATD will be made publicly available to the scientific community to encourage further research on MAT. Extensive experiments on HIT-MATD demonstrate the superior performance of RoMATer for multiaircraft tracking, quantitatively yielding an increase of 10%~15% with respect to MOTAL compared to TransTrack. In addition, experiments on a public Rareplanes dataset demonstrate the SOTA performance of the proposed modules in detecting multiple aircraft at full scale. Future work will further consider improving the metric MOTP, which indicates more regarding box precision and metric IDs.

REFERENCES

[1] M. Farhadmanesh, N. Marković, and A. Rashidi, “Automated Video-Based Air Traffic Surveillance System for Counting General Aviation Aircraft Operations at Non-Towered Airports,” Transp. Res. Rec. J. Transp. Res. Board, p. 036119812211150, 2022, doi: 10.1177/03611981221115087.

[2] M. Ahmed, A. Maher, and X. Bai, “Aircraft tracking in aerial videos based on fused RetinaNet and low-score detection classification,” no. August, pp. 687–708, 2023, doi: 10.1049/ipr2.12665.

[3] B. Wang, L. L. Chen, and Z. Y. Zhang, “A novel method on the edge detection of infrared image,” Optik (Stuttg)., vol. 180, no. November 2018, pp. 610–614, 2019, doi: 10.1016/j.ijleo.2018.11.113.

[4] S. Wu, K. Zhang, S. Li, and J. Yan, “Joint feature embedding learning and correlation filters for aircraft tracking with infrared imagery,” Neurocomputing, vol. 450, pp. 104–118, 2021, doi: 10.1016/j.neucom.2021.04.018.

[5] G. C. B and Z. Haibei, “Realization of Target Tracking Technology for Generated Infrared Images”, Springer, Singapore, 2021, doi: 10.1007/978-981-15-8411-4.

[6] Y. Xu, X. Luo, and F. Luo, “Low Slow Small Aircraft Surveillance System Based on Computer Vision,” Proc. - 2018 5th Int. Conf. Inf. Sci. Control Eng. ICISCE 2018, pp. 312–315, 2019, doi: 10.1109/ICISCE.2018.00072.

[7] F. S. Leira, H. H. Helgesen, T. A. Johansen, and T. I. Fossen, “Object detection, recognition, and tracking from UAVs using a thermal camera,” J. F. Robot., vol. 38, no. 2, pp. 242–267, 2021, doi: 10.1002/rob.21985.

[8] G. Fasano, D. Accardo, A. E. Tirri, A. Moccia, and E. De Lellis, “Morphological filtering and target tracking for vision-based UAS sense and avoid,” 2014 Int. Conf. Unmanned Aircr. Syst. ICUAS 2014 - Conf. Proc., pp. 430–440, 2014, doi: 10.1109/ICUAS.2014.6842283.

[9] O. D. M. Granillo and Z. Z. Beltran, “Real-Time Drone (UAV) Trajectory Generation and Tracking by Optical Flow,” Proc. - 2018 Int. Conf. Mechatronics, Electron. Automot. Eng. ICMEAE 2018, pp. 38–43, 2018, doi: 10.1109/ICMEAE.2018.00014.

[10] J. Yang et al., “Aircraft tracking based on fully conventional network and Kalman filter,” IET Image Process., vol. 13, no. 8, pp. 1259–1265, 2019, doi: 10.1049/iet-ipr.2018.5022.

[11] M. K. K. Navya and M. Wilscy, “Object ranging and tracking for aircraft landing system,” Int. Conf. Signal Process. Image Process. Pattern Recognit. 2013, ICSIPR 2013, vol. 1, pp. 278–282, 2013, doi: 10.1109/ICSIPR.2013.6497939.

[12] R. Ming et al., “Optical Tracking System for Multi-UAV Clustering,” IEEE Sens. J., vol. 21, no. 17, pp. 19382–19394, 2021, doi: 10.1109/JSEN.2021.3091280.

[13] H. Zheng, X. Mao, L. Chen, and X. Liang, “Adaptive edge-based mean shift for drastic change gray target tracking,” Optik (Stuttg)., vol. 126, no. 23, pp. 3859–3867, 2015, doi: 10.1016/j.ijleo.2015.07.160.

[14] C. Fu, A. Carrio, M. A. Olivares-Mendez, R. Suarez-Fernandez, and P. Campoy, “Robust real-time vision-based aircraft Tracking from Unmanned Aerial Vehicles,” Proc. - IEEE Int. Conf. Robot. Autom., pp. 5441–5446, 2014, doi: 10.1109/ICRA.2014.6907659.

[15] A. Maher, H. Taha, and B. Zhang, “Realtime multiaircraft tracking in aerial scene with deep orientation network,” J. Real-Time Image Process., vol. 15, no. 3, pp. 495–507, 2018, doi: 10.1007/s11554-018-0780-1.

[16] M. K. B, M. Danelljan, R. Pflugfelder, O. Drbohlav, and L. He, VOT2020 Challenge Results, vol. 1. 2020. doi: 10.1007/978-3-030-68238-5.

[17] K. Fragkiadaki and J. Shi, “Detection free tracking: Exploiting motion and topology for segmenting and tracking under entanglement,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 2073–2080, 2011, doi: 10.1109/CVPR.2011.5995366.

[18] W. Hu, X. Li, W. Luo, X. Zhang, S. Maybank, and Z. Zhang, “Single and multiple object tracking using log-euclidean riemannian subspace and block-division appearance model,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 12, pp. 2420–2440, 2012, doi: 10.1109/TPAMI.2012.42.

[19] L. Zhang and L. Van Der Maaten, “Preserving structure in model-free tracking,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 4, pp. 756–769, 2014, doi: 10.1109/TPAMI.2013.221.

[20] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, “Simple online and realtime tracking,” Proc. - Int. Conf. Image Process. ICIP, vol. 2016-Augus, pp. 3464–3468, 2016, doi: 10.1109/ICIP.2016.7533003.

[21] N. Wojke, A. Bewley, and D. Paulus, “Simple online and realtime tracking with a deep association metric,” Proc. - Int. Conf. Image Process. ICIP, vol. 2017-September, pp. 3645–3649, 2017, doi: 10.1109/ICIP.2017.8296962.

[22] S. Sun, N. Akhtar, H. Song, A. Mian, and M. Shah, “Deep Affinity Network for Multiple Object Tracking,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 1, pp. 104–119, 2021, doi: 10.1109/TPAMI.2019.2929520.

[23] J. Pang et al., “Quasi-Dense Similarity Learning for Multiple Object Tracking,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 164–173, 2021, doi: 10.1109/CVPR46437.2021.00023.

[24] Y. Zhang et al., “ByteTrack: Multi-Object Tracking by Associating Every Detection Box”, vol. 1. Springer Nature Switzerland, 2021. doi: 10.1007/978-3-031-20047-2.

[25] T.-X. Xu, Y.-C. Guo, Y.-K. Lai, and S.-H. Zhang, “CXTrack: Improving 3D Point Cloud Tracking with Contextual Information,” pp. 1084–1093, 2023, doi: 10.1109/cvpr52729.2023.00111.

[26] P. Karle, F. Fent, S. Huch, F. Sauerbeck, and M. Lienkamp, “Multi-Modal Sensor Fusion and Object Tracking for Autonomous Racing,” IEEE Trans. Intell. Veh., vol. 8, no. 7, pp. 3871–3883, 2023, doi: 10.1109/TIV.2023.3271624.

[27] X. Zhou, T. Yin, V. Koltun, and P. Krahenbuhl, “Global Tracking Transformers,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2022-June, pp. 8761–8770, 2022, doi: 10.1109/CVPR52688.2022.00857.

[28] J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. Kitani, “Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking,” pp. 9686–9696, 2023, doi: 10.1109/cvpr52729.2023.00934.

[29] C. Feichtenhofer, A. Pinz, and A. Zisserman, “Detect to Track and Track to Detect,” Proc. IEEE Int. Conf. Comput. Vis., vol. 2017-Octob, pp. 3057–3065, 2017, doi: 10.1109/ICCV.2017.330.

[30] Y. Li, K. He, J. Sun, and others, “R-fcn: Object detection via region-based fully convolutional networks,” Adv. Neural Inf. Process. Syst., no. Nips, pp. 379–387, 2016, [Online]. Available: http://papers.nips.cc/paper/6465-r-fcn-object-detection-via-region-based-fully convolutional-networks.pdf

[31] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, “Toward Real-Time Multi-Object Tracking,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12356 LNCS, pp. 107–122, 2020, doi: 10.1007/978-3-030-58621-8_7.

[32] P. Voigtlaender et al., “Mots: Multiobject tracking and segmentation,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2019-June, pp. 7934–7943, 2019, doi: 10.1109/CVPR.2019.00813.

[33] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

[34] X. Zhou, V. Koltun, and P. Krähenbühl, “Tracking Objects as Points,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12349 LNCS, pp. 474–490, 2020, doi: 10.1007/978-3-030-58548-8_28.

[35] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, “FairMOT: On the Fairness of Detection and Reidentification in Multiple Object Tracking,” Int. J. Comput. Vis., vol. 129, no. 11, pp. 3069–3087, 2021, doi: 10.1007/s11263-021-01513-4.

[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[37] Z. Lu, V. Rathod, R. Votel, and J. Huang, “RetinaTrack: Online single stage joint detection and tracking,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 14656–14666, 2020, doi: 10.1109/CVPR42600.2020.01468.

[38] P. Sun et al., “TransTrack: Multiple Object Tracking with Transformer,” 2020, [Online]. Available: http://arxiv.org/abs/2012.15460

[39] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer, “TrackFormer: Multi-Object Tracking with Transformers,” pp. 8834–8844, 2022, doi: 10.1109/cvpr52688.2022.00864.

[40] Y. Xu, Y. Ban, G. Delorme, C. Gan, D. Rus, and X. Alameda-Pineda, “TransCenter: Transformers With Dense Representations for Multiple-Object Tracking,” IEEE Trans. Pattern Anal. Mach. Intell., pp. 1–17, 2022, doi: 10.1109/TPAMI.2022.3225078.

[41] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection,” 2020, [Online]. Available: http://arxiv.org/abs/2004.10934

[42] X. Ding, X. Zhang, J. Han, and G. Ding, “Scaling Up Your Kernels to 31×31: Revisiting Large Kernel Design in CNNs,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2022-June, pp. 11953–11965, 2022, doi: 10.1109/CVPR52688.2022.01166.

[43] I. Reid, S. Roth, K. Schindler, A. Milan, and L. Leal-taix, “MOT16 : A Benchmark for Multi-Object Tracking,” 2016, pp. 1–12.

[44] J. Shermeyer, T. Hossler, A. Van Etten, D. Hogan, R. Lewis, and D. Kim, “RarePlanes: Synthetic data takes flight,” Proc. - 2021 IEEE Winter Conf. Appl. Comput. Vision, WACV 2021, pp. 207–217, 2021, doi: 10.1109/WACV48630.2021.00025.