Deep Learning with Historical Features and Timewise Similarity for Multiple Objects Tracking

Tuan Manh Tao, Changha Lee, Minsu Jeon and Chan-Hyun Youn*
School of Electrical Engineering
Korea Advanced Institute of Science and Technology
Daejeon, Korea
{ttmanh, changha.lee, msjeon, chyoun}@kaist.ac.kr

***Abstract*—Multiple Object Tracking (MOT) in computer vision is a fundamental task focused on identifying and monitoring the movement of multiple objects within a video sequence. MOT plays a crucial role in various applications, including surveillance, autonomous driving, and human-computer interaction. The primary objective is to consistently and accurately follow the trajectories of individual objects across frames while dealing with challenges such as occlusions, and varying appearances. This research paper presents an approach for tackling the challenging task of multiple categories of object tracking using deep learning techniques, combined with the utilization of enriching contextual features during training. In this study, we address the complexities of tracking objects by using temporal-wise similarity to improve features for consecutive frames. To enhance the performance of our tracking framework, we introduce a training strategy by adding to the original dataset a sub-dataset wherein large input images are divided into sub-patches to reach competitive results regarding tracking accuracy and precision with 72.4% and 81.6% on MOT dataset, respectively.**

***Index Terms*—Object Tracking, Augmented Training Data, Multiple Categories Tracking, Sub-Patches Based Annotation.**

I. Introduction

In recent years, the field of computer vision has witnessed a remarkable evolution, transforming the way we perceive and interact with our visual environment. One of the most challenging and crucial tasks within computer vision is object tracking as the next evolution of object detection [1], [2]. Object tracking is a fundamental component in various applications such as smart traffic, robotics, and augmented reality. Accurate and robust tracking of multiple object categories in complex scenes remains a formidable challenge due to the diverse range of change object appearances, temporary occlusions, and environmental dynamics conditions.

Object tracking, essentially the process of following and predicting the trajectory of objects across frames in a video sequence, demands a comprehensive understanding of object features and their interactions with the surrounding context. Traditional methods often rely on handcrafted features which may struggle to handle the inherent variability and complexity present in real-world scenarios. The advent of deep learning has revolutionized this domain, offering a data-driven approach that learns intricate patterns directly from the data, opening the way for significant advancements in object tracking.

* Corresponding author

One approach involves the integration of convolutional neural networks (CNNs) into the tracking pipeline. Studies [3], [4] have demonstrated the effectiveness of employing deep networks for object detection, feature extraction, and data association, resulting in enhanced tracking performance even in challenging scenarios. Besides, the concept of end-to-end multiple object tracking has gained success, exemplified in research [5] by Wu et al., This paradigm seeks to unify object detection, tracking, and segmentation within a single framework, leveraging on the strengths of each task to achieve more accurate and comprehensive tracking outcomes.

Two prominent and influential challenges in this domain are the Multiple Object Tracking (MOT) Challenge [6] and the VisDrone Challenge [7] with up-to-date respective datasets. The Multiple Object Tracking Challenge (MOT Challenge) has become a standard in the field focusing on the task of tracking multiple objects across frames in video sequences. The challenge datasets encompass a list of real-world scenarios with diverse challenges such as scale variations, and object interactions. As a result, the MOT Challenge stands as a fundamental study for the improvement of object-tracking techniques applying autonomous systems. In parallel, the VisDrone Challenge embodies the difficulties and intricacies of aerial sight, accentuating the examination and comprehension of imagery obtained by drones. The challenge datasets consist of aerial pictures and videos obtained in various circumstances, encompassing varied heights, alterations in lighting, and different weather conditions. The VisDrone Challenge has not only fostered advancements in computer vision for aerial purposes but has also underscored the wider significance of visual cognition in the era of drones with some studies [8], [9].

In this work, we focus capitalizes on the power of deep learning by introducing an approach that leverages augmented training data. Augmented training data refers to the strategic enrichment of the training dataset with artificially generated examples, aimed at enhancing the robustness and generalization capabilities of the model. By splitting consecutive image frames into sub-patches, other objects within the subsequential images are reassigned as new objects with new identifiers. Besides, we need to consider new coordinates as well as new bounding boxes of local objects in the methodology section.

II. Related work

In this section, we briefly introduce research papers related to MOT has been applied to both MOT [6] and Visdrone [7] datasets. Integrating detection and tracking through a unified network have commenced to captivate increased interest. The detection branch combines with re-ID feature extractor [10] in a single network to decrease inference time while keeping competitive tracking accuracy. Also, jointing detection and motion prediction in [11] enhanced occlusion handling and reduced identity-switching issues. The works inside these papers are close to our direction that boosts the accuracy detection as well as object tracking in the Aerial dataset.

A. FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking

The primary issue with framing multi-object tracking as a combination of multi-task learning involving object detection and re-identification within a single network is that the conventional approach of prioritizing detection over re-identification in existing one-shot trackers leads to an unfair learning of the re-ID network. In training, the model becomes inclined towards generating precise object proposals rather than developing superior re-ID features. Consequently, even though the detection outcomes are satisfactory, this imbalance contributes to a significant count of identity switches.

FairMOT [3] tackles the issue of favoring the main detection task by adopting an anchor-free methodology for the detection and re-identification components. This technique involves estimating object center positions [12] and dimensions through position-sensitive measurement maps. By doing so, it rectifies the inequitable treatment of the detection aspect and successfully develops good re-ID characteristics. This leads to a balanced compromise between detection and re-identification performance.

Within the framework of FairMOT, the training process involves a mix of a detection loss, a center-ness loss, and a re-ID loss. This combination incentivizes the model to acquire excellent features for both the detection and re-identification assignments.FairMOT discovered that acquiring lower-dimensional characteristics is more advantageous for one-shot MOT. Conversely, high-dimensional re-ID features detrimentally affect object detection precision, as the two tasks compete with each other, consequently causing adverse effects on the ultimate tracking accuracy.

FairMOT has been trained on mainly MOT and CrowdHuman [13] datasets with unified object sizes. Experimentally, we realized that when applying FairMOT to the Visdrone dataset, tracking small objects is not really good, especially with objects moving close to the edge of the frame. Therefore, dividing into sub-part consecutive images allows the model to isolate and analyze the small object with less interference from the surrounding context like Figure 1.

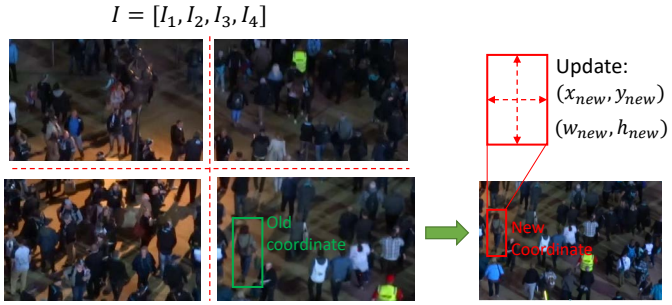

Fig. 1: Adding Augmented training data by splitting the original image into sub-patches: Updating new coordinates for object following new size of sub-images.

B. GIAOTracker: A comprehensive framework for MCMOT with global information and optimizing strategies

The GIAOTracker [9] method is a thorough system for tracking multiple objects in drone videos, covering various classes. It employs worldwide data and optimization techniques. It comprises three phases: real-time tracking, global connection, and post-processing. The initial stage produces dependable tracklets by considering camera and object movements, as well as object appearance. In the second phase, it connects these tracklets into trajectories by utilizing global information. In the end, the trajectories undergo refinement using four post-processing approaches including denoising, interpolation, rescoring and fusion.

By considering these multifaceted factors of camera and motion, the system generates tracklets that exhibit a high degree of reliability and accuracy. This initial step serves as the foundation upon which subsequent phases build, ensuring that the tracking process is not only precise but also adaptable to the dynamic nature of drone video environments.

The global connection phase takes the intermediate stage, where the GIAOTracker method leverages the power of global information integration. This holistic approach to trajectory construction enables the system to capture and represent the intricate interactions and movements of objects within the video, providing a comprehensive overview of object behavior. GIAOTracker has shown competitive performances related to tracking accuracy; however, with multiple stages framework, it is not easy to adapt to real-time applications that required short inference time.

III. Methodology

Our study has developed based on FairMOT [3] as the backbone framework. Our target is to provide flexibility-augmented training data which helps the model generalize better to different scales and viewpoints when adding to the original training data a sub-dataset. Besides, the model employs a local tracklet re-ID that involves calculating a similarity matrix of the current and historical tracklets and using it to enhance the tracklet features. The whole end-to-end unified system is presented in Figure 2.

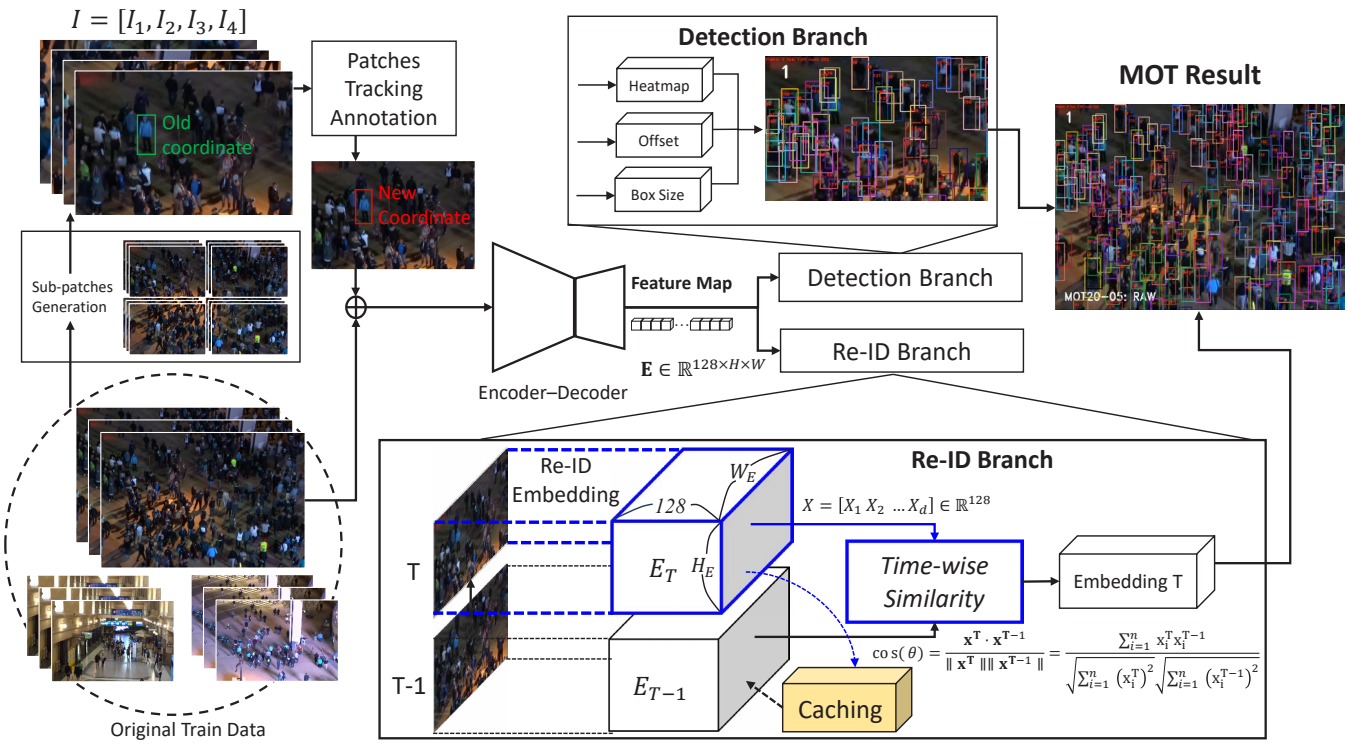

Fig. 2: Multiple Object Tracking approach with Enriching Training Data Strategy and Enhanced Re-ID features with Timewise Similarity: Training data is expanded by adding Sub-patches dataset into the Original data. The embedding feature of frames is enhanced when comparing comprehensively with historical tracklets in previous frames.

A. Flexibility-Augmented Training Data on Aerial Videos

Popular aerial video datasets such as MOT and Visdrone have their own characteristics. The MOT dataset focuses on crowds of people with a variety of backgrounds and lighting conditions; while the Drone dataset witnesses a variety of objects classes and sizes. Failure to recognize small objects is a challenge in object tracking. When we divide bigger images into smaller sections, we generate additional training samples, which may improve the range and inclusiveness of the training dataset. The accurate detection of small objects can be difficult, particularly when they are depicted by a limited number of pixels in a large image. By breaking down the image, the model can concentrate on tinier areas, thereby augmenting the likelihood of correctly identifying the small objects. The amount of training dataset contains the original data and the extended part from sub-patches. Objects in the center positions of larger frames can become new objects located at the edges of sub-patches. This thing makes the model detect well objects at the edge of videos when the model has enough contextual edge information from the features extractor.

While there are benefits to splitting large images for small object detection as well as handling occlusions, it is important to consider some potential drawbacks related to object fragmentation leading to increasing false positives for inference phases. Dividing the large aerial images or videos for efficient training is a common solution in the computer vision field; however, with MOT tasks we can note that our splitting images approach is not a naively normal solution when just breaking down bigger images into sub-patches. Algorithm 1, named Sub-Patches Tracking Annotation, has been proposed to address potential errors. The key considerations are the proportion of original training data used for Algorithm 1 and the need to include the bounding box threshold as a criterion for identifying new objects.

For simplicity, initial large images (H,W,C) will be divided into four sub-patches: top-left, top-right, bottom-left, and bottom-right (h,w,C) where $h = H/2$ and $w = W/2$. After breaking down, the coordinates will be updated following the division ratio and be determined to belong to which sub-patches. We assume that a frame I has been divided into four sub-patches I_1, I_2, I_3 and I_4 corresponding to top-left, top-right, bottom-left and bottom-right. The distribution coordinates for new objects $D(x, y)$ with (x, y) are (top, left) bounding box coordinates as follows:

$$D(x,y) = \begin{cases} I_1, & \text{if } (x,y) < (w,h) \\ I_2, & \text{if } x \geq w \text{ and } y < h \\ I_3, & \text{if } x < w \text{ and } y \geq h \\ I_4, & \text{if } (x,y) \geq (w,h) \end{cases} \quad (1)$$

Especially for object tracking tasks, the identifier of objects is unique to avoid misunderstanding for the model during the

Algorithm 1 Sub-Patches Tracking Annotation

Input: Original Training data and Split Percentage

procedure PATCH BASED ANNOTATE($Data, percentage$)
  Split large images into 4 parts
  $w_{sub} \leftarrow w/2$ ▷ Update new image-patch size
  $h_{sub} \leftarrow h/2$
  for $images$ `in` $video$ **do**
    Read_Annotation_Line() ▷ Obtain ($\bar{x}$,$\bar{y}$,bbox,id)
    Distribute objects to respective patches
    // Check size bounding box with threshold bbox ϵ
    if $w/2 + bbox_{w,h}/2 - (\bar{x}, \bar{y}) < bbox_{w,h} \times \epsilon$ **then**
      Continue ▷ Ignore fragmented box
    else
      Update new coordinates and bbox size
      Assign local identifiers with respective patches
    end if
  end for
  Get_Max_ID() ▷ Get maximum id in original video
  Update_Global_ID() ▷ Id will be unique in datasets
  return ($patches$, $new\_annotations$)
end procedure

Output: Sub-dataset along additional IDs for new objects

training phase. The object identifier in augmented datasets that are generated from the original datasets becomes new objects with non-duplicated identifiers. Algorithm 1 also covers complex cases when objects at the boundary of the split zone move back and forth between the divided regions. New identifiers of objects in sub-patches are defined as:

$$ID = \begin{cases} I_i.index(ID_{obj}) + 1, & \text{if local ID} \\ max + I_i.index(ID_{obj}).i + i, & \text{if global ID} \end{cases} \quad (2)$$

Where i indicates the sub-patches orders and max indicates the maximum number of global objects.

B. Improvement of re-ID features through historical tracklet with TimeWise Similarity

Adding augmented training data will make bias potential issues; thus, improvement of re-ID features is necessary to avoid missing trajectory of objects throughout the whole surveillance videos. We have proposed the TimeWise Similarity to measure the similarity and correlation between tracklets across two adjacent image frames in a video. TimeWise Similarity assesses how consistent the spatial and temporal characteristics of the tracklets are between consecutive frames. When applying backbone extractor [3], we have embedding features $\mathbf{E} \in \mathbb{R}^{128 \times H_E \times W_E}$. The value of each element inside TimeWise Similarity matrix can be obtained by computing cosine similarity between two feature vectors $\mathbf{X_i^T}$ and $\mathbf{X_j^{T-1}}$ $\in \mathbf{R}^{128}$ respective with past and current re-ID frame features when $(i, j) \in (0, wh]$:

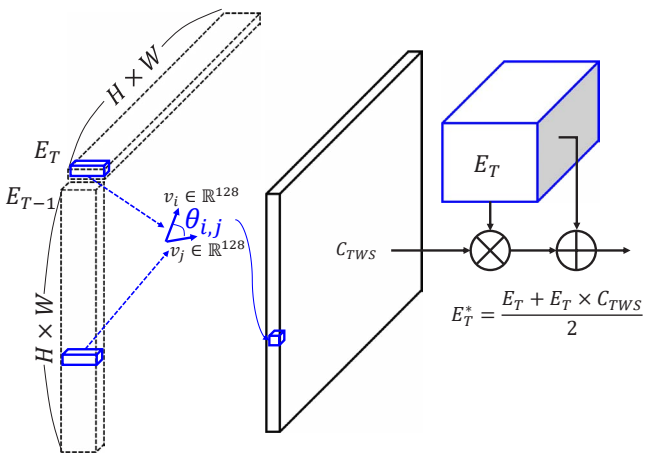

Fig. 3: The proposed Timewise Similarity Matrix and Enhancement Embedding Feature Map

$$TWS_{wh \times wh} = \begin{bmatrix} \cos(\theta_{11}) & \cos(\theta_{12}) & \cdots & \cos(\theta_{1wh}) \\ \cos(\theta_{21}) & \cos(\theta_{22}) & \cdots & \cos(\theta_{2wh}) \\ \vdots & \vdots & \ddots & \vdots \\ \cos(\theta_{wh1}) & \cos(\theta_{wh2}) & \cdots & \cos(\theta_{wh^2}) \end{bmatrix}$$

The size of each embedding feature vector in the re-ID branch is fixed. The value of $\cos\theta_i j$ represents the similarity of two feature vectors X_i^T and X_j^{T-1}:

$$\begin{aligned} \cos(\theta) &= \frac{\mathbf{X^T} \cdot \mathbf{X^{T-1}}}{\|\mathbf{X^T}\|\|\mathbf{X^{T-1}}\|} \\ &= \frac{\sum_{i=1}^{n} X_i^T X_i^{T-1}}{\sqrt{\sum_{i=1}^{n}(X_i^T)^2}\sqrt{\sum_{i=1}^{n}(X_i^{T-1})^2}} \end{aligned} \quad (3)$$

Then we have the position vector of a current frame compared to a past adjacent frame by getting averaging values following the row of Timewise Similarity (TWS) matrix. This position vector as supported information provides the change positions as well as identifies objects in consecutive image frames in videos. After that, we obtain the enhanced re-ID feature map $\mathbf{E}_T^*$ by re-computing feature vectors as:

$$\mathbf{T}_i^* = \frac{\mathbf{T}_i + \mathbf{T}_i.\overline{\cos(\theta_i)}}{2}, with\ i \in (0, wh] \quad (4)$$

In case the object appears in the previous frame and disappears in the current frame, the enhanced re-ID feature $\mathbf{T}^*$ may contain historical position information of these objects. However, the primary tracking condition is still the detection branch along offset as well as the box size of objects. Therefore, the tracking results are less adversely affected in this case.

IV. EXPERIMENTS AND RESULT

Here we provide our experimental results demonstrating the effect of the Augmented Training Data approach related to tracking accuracy and precision of model. Our multiple objects

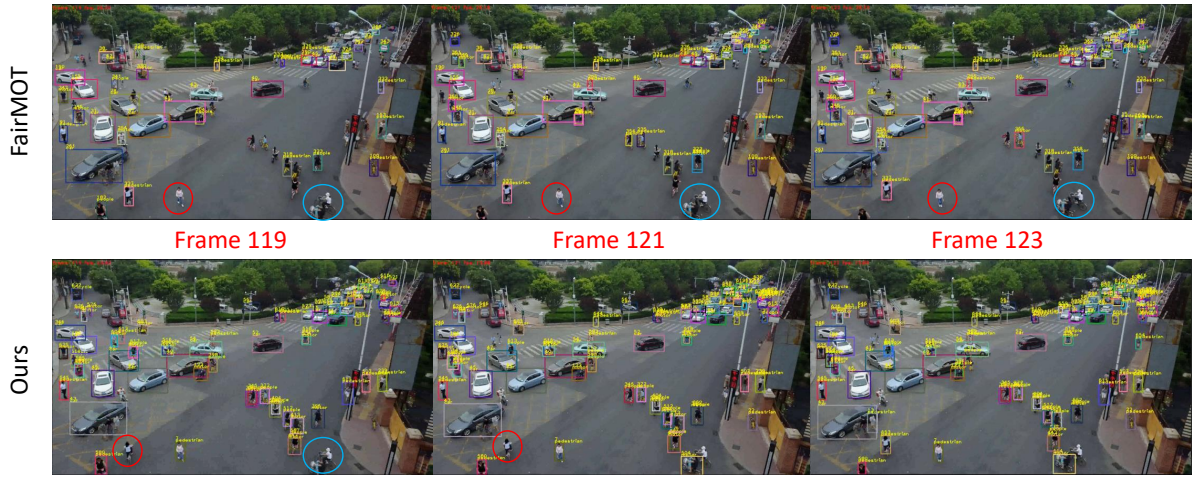

Fig. 4: The visual results on Visdrone 2019; our method showed the improvement related to detecting and tracking small objects. Objects are assigned consistent IDs throughout consecutive image frames. Best viewed in color mode.

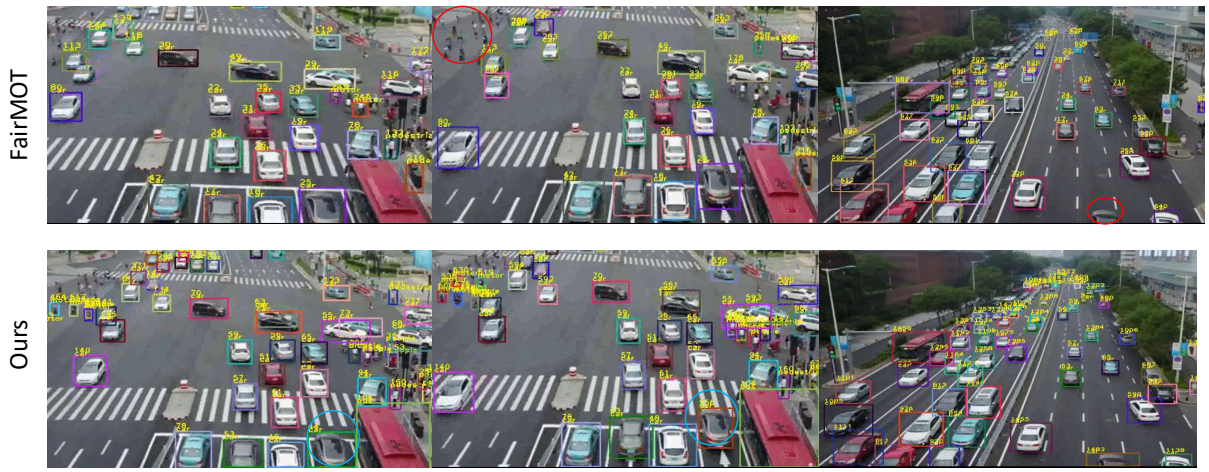

Fig. 5: Comparison of two approaches with changing camera views and speed of objects in video. MOT framework reached better results when detecting small objects with denser bounding boxes. ID switching issues still occur in some cases.

Table I: Comparison of Augmented Training Data with different split percentages for original training data on MOT20*

Methods	MOTA (%)	MOTP (%)	IDF1 (%)	IDs	Recall (%)	Precision (%)
FairMOT	71.8	81.4	76.8	2725	77.4	93.6
Ours_A	**72.4**	**81.8**	**77.4**	2708	**77.6**	93.8
Ours_B	72.3	80.8	76.7	**2665**	76.7	**94.9**
Ours_C	71.9	80.7	76.7	2693	76.4	94.7

Table II: Performances of Multiple Objects Tracking on VisDrone 2019 Dataset

Methods	MOTA (%)	MOTP (%)	IDF1	FP	FN
SORT [15]	18.1	65.1	32.2	104,453	78,467
FairMOT	29.8	73.3	**46.1**	**17,683**	58,657
Ours	**30.5**	73.3	44.9	18,145	**57,832**

tracking experiments have been conducted using MOT20 and Visdrone 2019-MOT dataset.

Our approach has been evaluated on some popular indicators using for measuring objects tracking containing MOTA, MOTP, IDF1, IDs [14]. MOTA measures how well the model can track multiple objects in a scene, and MOTP measures how closely the model can match the shape and size of each object. IDF1 measures how accurately the model can identify each object without confusing it with others, while IDs measure how often the model switches the identity of the same object. Higher values of MOTA, MOTP, and IDF1 indicate better performance, while lower values of IDs indicate more consistent tracking.

Table I gives the comparison of MOT performance between FairMOT as a baseline and our approach with the percentage of data augmentation is 25%, 50% and 75% in A, B, and

C results, respectively. The experiments have been deployed on MOT20 validation data. Overall, our approach has shown efficient performance when increasing the amount of training data; especially, increasing 25% has brought the best results with 72,4% MOTA and 81.8% MOTP. The accuracy of model represents a decrease with higher percentages; extension of sub-patches training data can raise the false positives leading to affect detection and tracking performance. However, our approach with 50% increased training data shows better keep-tracking ability when having the lowest IDs value along with 94.9 % detection precision.

By leveraging the performance of table I, increasing 25% training data is a chosen strategy that applies to Visdrone 2019 datasets. Note that, the performance on table II has been collected when only considering object detection and tracking resulting bounding boxes during the whole video. The visualized demonstrations are still represented with multiple categories of object tracking in Figure 3 and Figure 4. Training dataset has been added sub-patches helped model more focus on small objects in the crowd. This thing leads to a decreasing amount of false negatives from 58.6k in FairMOT to 57.8k; our approach also reaches a better MOTA value of 30.5 %. One drawback is that detecting small objects makes the model become more confused among similar classes leading to increasing false positive values.

Regarding qualitative results, our illustrations have been conducted on Visdrone 2019 Dataset with the different views and speed changes of scenarios. Each row demonstrates the tracking results over time. We consider some important points inside consecutive frames between two approaches: the circle manually bounding boxes expressed the negative points for each frame. Our approach fixed object detection at the edges of the frame when the model has been trained with sub-patches having more edge samples. It is shown in Figure 3 with denser bounding boxes from various direction motions of objects. In Figure 4, the Augmented Training Data approach still shows improvement in tracking more tiny samples as well as objects that just appear and are located on the border of frames. In general, both quantitative and qualitative comparisons have represented competitive results in the evaluation; it shows that our method performs well in MOT tasks.

V. Conclusion

In this paper, we proposed an improvement method for multiple object tracking of different categories using extended training data. We showed how splitting large images into smaller sub-patches and annotating them with constraints can improve the detection and tracking of objects in various scenarios. Besides, using historical tracklets from previous frames to enhance embedding features brought more accurately in tracking objects. We evaluated our method on the different MOT datasets and achieved competitive results in terms of accuracy and precision. Our method can handle challenges such as occlusions, appearance changes, and object interactions. Our work contributes to the advancement of MOT applications in computer vision.

Acknowledgment

This research was supported by Korea Institute of Marine Science & Technology Promotion(KIMST) funded by the Ministry of Oceans and Fisheries(RS-2023-00238652).

References

[1] Redmon, Joseph, et al. "You only look once: unified, real-time object detection (2015)." arXiv preprint arXiv:1506.02640 (2015).

[2] Carion, Nicolas, et al. "End-to-end object detection with transformers." European conference on computer vision. Cham: Springer International Publishing, 2020.

[3] Zhang, Yifu, et al. "Fairmot: On the fairness of detection and re-identification in multiple object tracking." International Journal of Computer Vision 129 (2021): 3069-3087.

[4] Du, Y., et al. "Strongsort: Make deepsort great again. arXiv." arXiv preprint arXiv:2202.13514 (2022).

[5] Wu, Jialian, et al. "Track to detect and segment: An online multi-object tracker." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[6] Dendorfer, Patrick, et al. "Motchallenge: A benchmark for single-camera multiple target tracking." International Journal of Computer Vision 129 (2021): 845-881.

[7] Du, Dawei, et al. "VisDrone-DET2019: The vision meets drone object detection in image challenge results." Proceedings of the IEEE/CVF international conference on computer vision workshops. 2019.

[8] Lin, Yeneng, et al. "Multiple Object Tracking of Drone Videos by a Temporal-Association Network with Separated-Tasks Structure." Remote Sensing 14.16 (2022): 3862.

[9] Du, Yunhao, et al. "Giaotracker: A comprehensive framework for mcmot with global information and optimizing strategies in visdrone 2021." Proceedings of the IEEE/CVF International conference on computer vision. 2021.

[10] Lu Z, Rathod V, Votel R, Huang J (2020) Retinatrack: Online single stage joint detection and tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14668–14678

[11] Sun P, Jiang Y, Zhang R, Xie E, Cao J, Hu X, Kong T, Yuan Z, Wang C, Luo P (2020) Transtrack: Multiple-object tracking with transformer. arXiv preprint arXiv:201215460

[12] Zhou, Xingyi, Jiacheng Zhuo, and Philipp Krahenbuhl. "Bottom-up object detection by grouping extreme and center points." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[13] Shao, Shuai, et al. "Crowdhuman: A benchmark for detecting human in a crowd." arXiv preprint arXiv:1805.00123 (2018).

[14] Ristani, Ergys, et al. "Performance measures and a data set for multi-target, multi-camera tracking." European conference on computer vision. Cham: Springer International Publishing, 2016.

[15] Bewley, Alex, et al. "Simple online and realtime tracking." 2016 IEEE international conference on image processing (ICIP). IEEE, 2016.