This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Strong Detector with Simple Tracker

Zongheng Tang[(a)*], Yulu Gao[(a)*], Zizheng Xun[(a)*], Fengguang Peng[(a)*],
Yifan Sun[(b)], Si Liu[(a)], Bo Li[(a)]
Beihang University[(a)], Baidu Inc[(b)]
tzhhhh123, gyl97, xunzz, pengfg, liusi, boli @buaa.edu.cn sunyf15 @tsinghua.org.cn

Abstract

Unmanned aerial vehicle (UAV) tracking is a research direction with practical application value and has received sufficient attention in recent years. Challenges such as complex backgrounds, small targets, and motion blur in UAV tracking make it difficult to directly apply existing tracking or detection methods. For example, some state-of-the-art (SOTA) single-object tracking methods such as Ostrack perform poorly when encountering target disappearance or camera offset. Existing detection methods are also difficult to apply directly to this task. This paper proposes a detection-based method with cascading post-processing modules to solve this task. Our entire process includes generating detection candidate boxes, adjusting candidate box scores through video classification, connecting candidate boxes between different frames through a simple tracker, and determining moving targets in the video through background modeling, followed by single-object tracking as post-processing to adjust the results. We finally achieved first place in the 3rd Anti-UAV challenge track1 and top three in track2.

1. Introduction

Unmanned aerial vehicles (UAVs) [6,13] have gained increasing popularity in various fields such as agriculture, geological survey, and aerial photography due to their rapid development in recent years. However, their misuse or illegal operation may threaten social security and stability. This highlights the importance of developing techniques for detecting and tracking UAVs. In situations where obtaining informative appearance information of UAVs is challenging, such as at night, infrared images have unique advantages in capturing targets. Hence, infrared image-based UAV tracking has gained significant attention.

Single object tracking (SOT) is a critical research area in computer vision that involves tracking a single target in the first frame of a video or frame sequence and continuously tracking the target position in subsequent frames.

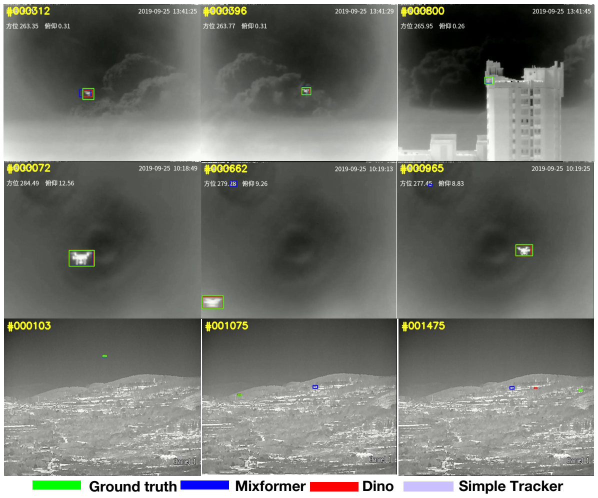

Figure 1. Qualitative comparison of Simple tracker with single object tracker and detector on three challenging sequences. Traditional trackers tend to deviate from the target under the influence of similar distractors. In contrast, our Simple tracker can track accurately due to the novel motion detection and linking mechanism, thus showing strong robustness in a variety of different difficult tracking scenarios.

Single UAV tracking is a natural SOT problem that can be solved using some single-object tracking networks such as OSTrack and Mixformer. However, in real-world scenarios, the movement of UAVs can be complex leading to difficulties in SOT. To address these challenges, we have adopted the tracking-by-detection strategy to replace the original single tracker. This strategy comprises two main modules, namely, the Strong Detector and the Simple Tracker. We have chosen several types of detectors and optimized each one for the unique characteristics of the UAV object resulting in our Strong Detector. The Simple Tracker uses cascading rules that link the results of the Strong Detector to achieve the final tracking results.

However, in infrared images, noise blocks in the background can be similar to the foreground target, resulting in a high probability of detection or tracking failure when relying solely on pure detection methods. To further improve the accuracy of the model, we have utilized temporal information and designed two modules: Video Checker and Motion Model. The Video Checker is a video classifier based on detection results that enlarges the object in the image and crops a local video segment from current and past frames. The segment is then input into the Video Checker for classification, resulting in a new score for the current detection result. The Motion Model is based on background modeling using the frame difference method, which is effective in detecting moving targets with small pixels and can complement detection tasks when dealing with small targets and multiple background clutter similar to foreground targets.

2. Related Work

2.1. Object Detection

YOLOv8 [5] is the latest version of the popular YOLO (You Only Look Once) model for object detection and image segmentation, released by Ultralytics in January 2023. It introduces several innovations, including a new backbone network, a new anchor-free detection head, and a new loss function. YOLOv8 is designed to be fast, accurate, and user-friendly, and can be used as a framework to support all previous versions of YOLO. DINO [12] is an end-to-end object detector that improves upon previous DETR-like models. It uses a contrastive way for denoising training, a mixed query selection method for anchor initialization, and a look-forward-twice scheme for box prediction. DINO performs well on several object detection benchmarks, including COCO and Open Images. EVA [4] is a visual representation learning method proposed by Baidu AI Vision. It is based on a vanilla Vision Transformer (ViT) [3] that is pretrained to reconstruct masked-out image-text aligned vision features (i.e., CLIP features) conditioned on visible image patches. This way, EVA can leverage both visual and semantic information from large-scale unlabeled data without relying on external annotations or labels.

2.2. Object Tracking

There has been significant progress in research on object tracking [2,8–11] in recent years. MixFormer [2] is a single-object tracking (SOT) method that uses a transformer-based backbone with mixed attention to perform end-to-end tracking. It does not use any post-processing, multi-scale feature fusion, or online update strategies that are common in traditional tracking methods. OSTrack [11] is a one-stream tracking framework that unifies feature learning and relation modeling for tracking based on self-attention operators. OSTrack extracts the template and search region features jointly and performs relation modeling between them, enhancing the target awareness and target-background discriminability of the features.

3. Proposed Method

The proposed method is a simple tracker based on strong detector. In this section, we introduce our strong detection and simple tracker. Next, we present our video checker and motion model. Finally, we describe our ensemble strategy and the applications of mainstream single-object trackers in our framework.

Strong Detector: To achieve effective object tracking, a tracking-by-detection approach is employed, where accurate object detection is crucial for determining the overall performance of the tracking system. A range of state-of-the-art detection models are carefully selected, including single-stage, multi-stage, and transformer-based models, such as YOLOv8, EVA, and DINO, respectively. The largest open-source models for each of the aforementioned models, YOLOv8-YOLOv8x6, EVA-SwinL, and DINO-SwinL, are chosen. To account for the uneven distribution and small size of targets, a large input size and small object oversampling strategy is adopted during detector training. Additionally, data augmentation techniques, such as sample strategies for small targets, are employed to enhance the robustness of the detectors. Finally, an ensemble fusion approach is applied to combine the output of multiple detection models for improved accuracy.

Video Checker: Although image-level detection can provide good results, the lack of temporal information between videos can lead to the inability to judge some difficult samples. To improve the accuracy of the test results, a video checker is utilized to differentiate true and false positives. This method effectively utilizes video information and maintains the online property of the model by utilizing only current and past frames. Specifically, we first train DINO to predict detections on both the training and testing sets to detect potential objects. For each image, we select the top 20 boxes, construct positive and negative samples based on the position of the ground truth box, and enlarge all samples 5 times. Additionally, for each sample, we crop the current frame and the thirty previous frames using the enlarged bounding box. Then, we use all samples to train the video checker. During the testing phase, we submit the samples to the video checker to generate a new score, which aids in making a final judgment for the tracker. Currently only a simple score correction has been applied to the top 1 bounding box of a single detector result during testing, and it has not yet been applied to all boxes.

Motion Model: In order to track small targets in video sequences, we introduced the background modeling method vibe [1] for proposal generating in infrared UAV tracking. For each sequence, we first initialize vibe in the first frame and update it in each subsequent frame to get the foreground

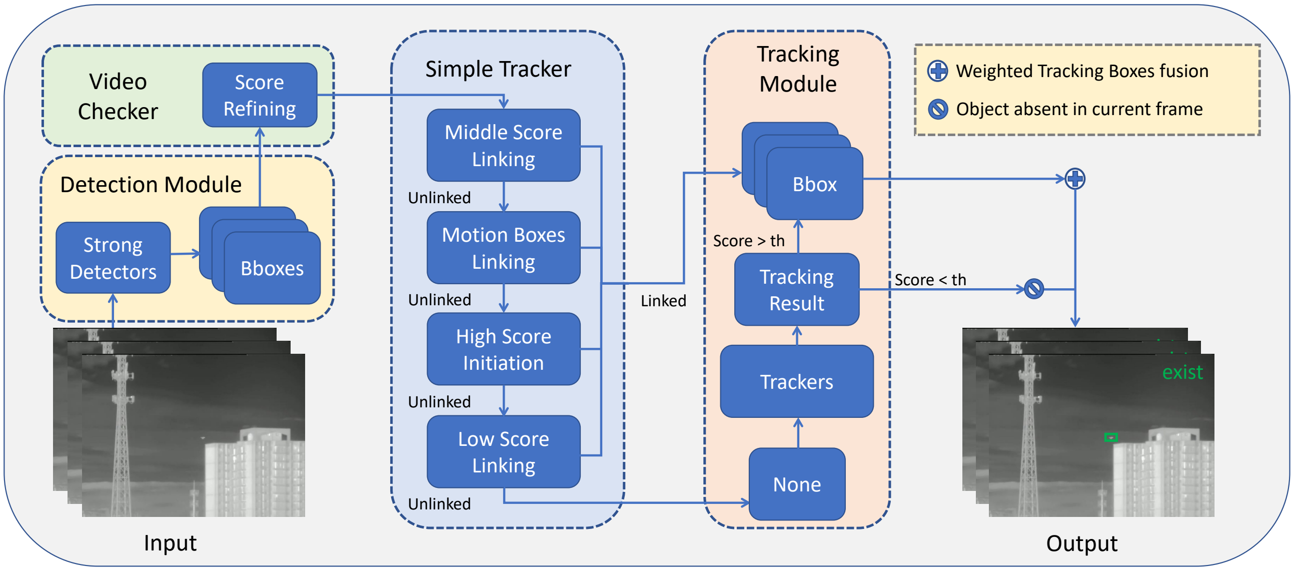

Figure 2. **Overall architecture of our method.** a) The Detection Module comprises several detectors to generate detection results. b) The Simple Tracker links detection results in a cascaded fashion, involving four stages: Middle Score Linking, Motion Boxes Linking, High Score Initiation, and Low Score Linking. c) Tracking Module integrates popular single-target detectors into our framework and tracks unlinked results. Additionally, we utilized Weighted Tracking Boxes Fusion to combine the results.

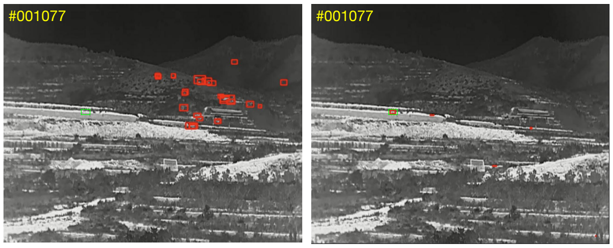

Figure 3. Visualization results of vibe and detection results. The left picture represents the detection result, and the right picture is the vibe result. Among them, the green box represents GT, and the red box represents the generated candidate boxes. In the #001077 frame, the background is still, even if the target is small and has little semantic information, but the motion features are relatively obvious, vibe can successfully perceive the target but the detector cannot.

mask, F_M. Then, open and close operations are performed on F_M, and the boxes are obtained by the $findContours$ and $boundingRect$ functions in OpenCV. At the same time, a pixel summation S_p is performed on F_M, and if S_p is greater than the set threshold, we will reinitialize the vibe in the current frame. Figure 3 shows a visual comparison of vibe and detection results.

Simple Tracker: The anti-UAV Challenge is challenging due to complex and diverse scenarios and small target sizes. Obtaining accurate target locations based on single-image detection is difficult, so we must consider temporal information in videos. One solution is to imitate current MOT work, such as using the post-processing module of ByteTrack or StrongTrack. However, we implemented an online post-processing module called SimpleTrack for the anti-UAV task, which involves only a single type of UAV target.

The idea behind SimpleTrack is intuitive. To maintain consistency in tracking results, we first attempt to match the current frame's detection results with past prediction results. If successful, we update the result. If not, and a sufficiently high confidence detection box exists in the current frame, we use that detection box as the new output result. We measured the similarity between boxes using IOU, although we also tried using the Euclidean distance between box centers divided by box length and width. We set an expiration time for current tracked prediction results and clear past results if they are not linked for several frames. We performed three matches, with medium confidence matching taking priority, followed by motion detection result matching and low confidence matching. The detection boxes generated by the motion result lack confidence, so we placed them in the medium-low confidence interval and increased their IOU threshold during matching.

Ensemble: To improve the tracking performance, we utilize a modified version of the weighted boxes fusion (WBF) [7] in the ensemble phase. While conventional WBF directly

Rank	User Name	Tracking Accuracy
1	**tzhhhh123(Ours)**	**0.700**
2	Undefined	0.688
3	SIA_Ryu	0.680
4	zsl	0.678
5	soro	0.677
6	shan666	0.671
7	stephenx24	0.671
8	Silverfall	0.670
9	shubonlpr	0.667
10	MinkiSong	0.667

Table 1. Main results of Track1 with 200 videos

Rank	User Name	Tracking Accuracy
1	ZY	0.611
2	ryanhe312	0.591
3	**tzhhhh123(Ours)**	**0.570**
4	shan666	0.562
5	stephenx24	0.562
6	HIT_HH	0.550
7	shubonlpr	0.540
8	KKKKKK	0.538
9	QJY0310	0.538
10	Carl_Huang	0.536

Table 2. Main results of Track2 with 200 videos

fuses box results, it may alter the score distribution and thus hinder tracker parameter adjustment. To overcome this issue, we propose a two-step approach called weighted tracking boxes fusion (WTBF). In step 1, the results of each detector are fed into the tracker individually to obtain tracking results. In step 2, the tracking boxes generated by each tracker are combined using ensemble techniques. This design avoids conflicts between tracking and ensemble and simplifies parameter adjustment.

Single Object Tracking: The detection model may detect the whole picture and fail to consider object motion trajectories, so we refine the detection results by connecting single object trackers to the SimpleTrack. Mixformer [2] and OSTrack [11] are chosen as our trackers. For each frame where the target is not detected, we select the nearest previous frame where the target exists as the template. The tracking results are then combined using WTBF. Finally, we determine the target's existence in each re-tracked frame based on the tracker's score.

4. Experiments

DINO [12] is used as the baseline detector throughout the ablation experiments.

	Model	Tracking Accuracy
1	Baseline	0.532
2	+Threshold	0.557
3	+Simple Track	0.599
4	+Motion Model	0.607
5	+SOT Model	0.609

Table 3. Ablation studies on the validation split containing 50 videos.

4.1. Experimental Setup

Dataset: We utilized the training subset of the 3rd Anti-UAV dataset, which comprises 200 videos. This dataset contains challenging video sequences that include dynamic backgrounds, complex movements, and tiny-scale targets, covering a broad range of scenarios with multi-scale UAVs. The training subset offers comprehensive annotation files containing detailed information about the target's existence, location, and various challenges. We followed the official dataset partitioning method and handpicked 50 sequences from the training set to create an evaluation set. Unless explicitly mentioned otherwise, all experimental results presented below were obtained by training on 150 videos and testing on the 50 evaluation videos.

Metric: The metric is illustrated below:

$$acc = \sum_{t=1}^{T} \frac{IoU_t \times \delta\left(v_t > 0\right) + p_t \times \left(1 - \delta\left(v_t > 0\right)\right)}{T} \\ -0.2 \times \left(\sum_{t=1}^{T^*} \frac{p_t \times \delta\left(\nu_t > 0\right)}{T^*}\right)^{0.3} \tag{1}$$

IoU_t is the intersection over the union between the predicted tracking box and its corresponding ground-truth box for each frame t. The predicted visibility flag, p_t, equals 1 when the predicted box is empty and 0 otherwise. The target's ground-truth visibility flag, v_t, is represented by the indicator function $\delta(v_t > 0)$, which equals 1 if $v_t > 0$ and 0 otherwise. The accuracy is averaged over all frames in a sequence. T denotes the total number of frames, and T* denotes the number of frames in which the target is present in the ground truth.

4.2. Main Results

We respectively show the results of our method in the two tracks of the Anti-UAV competition in the Table 1 and Table 2. It is worth noting that our main target challenge is track1, where we have made many attempts and used the complete pipeline process, including video classifiers, tracking post-processing correction, and multiple detector results ensemble. For track2, we only used the detection results of a single detector as output through a simple tracker.

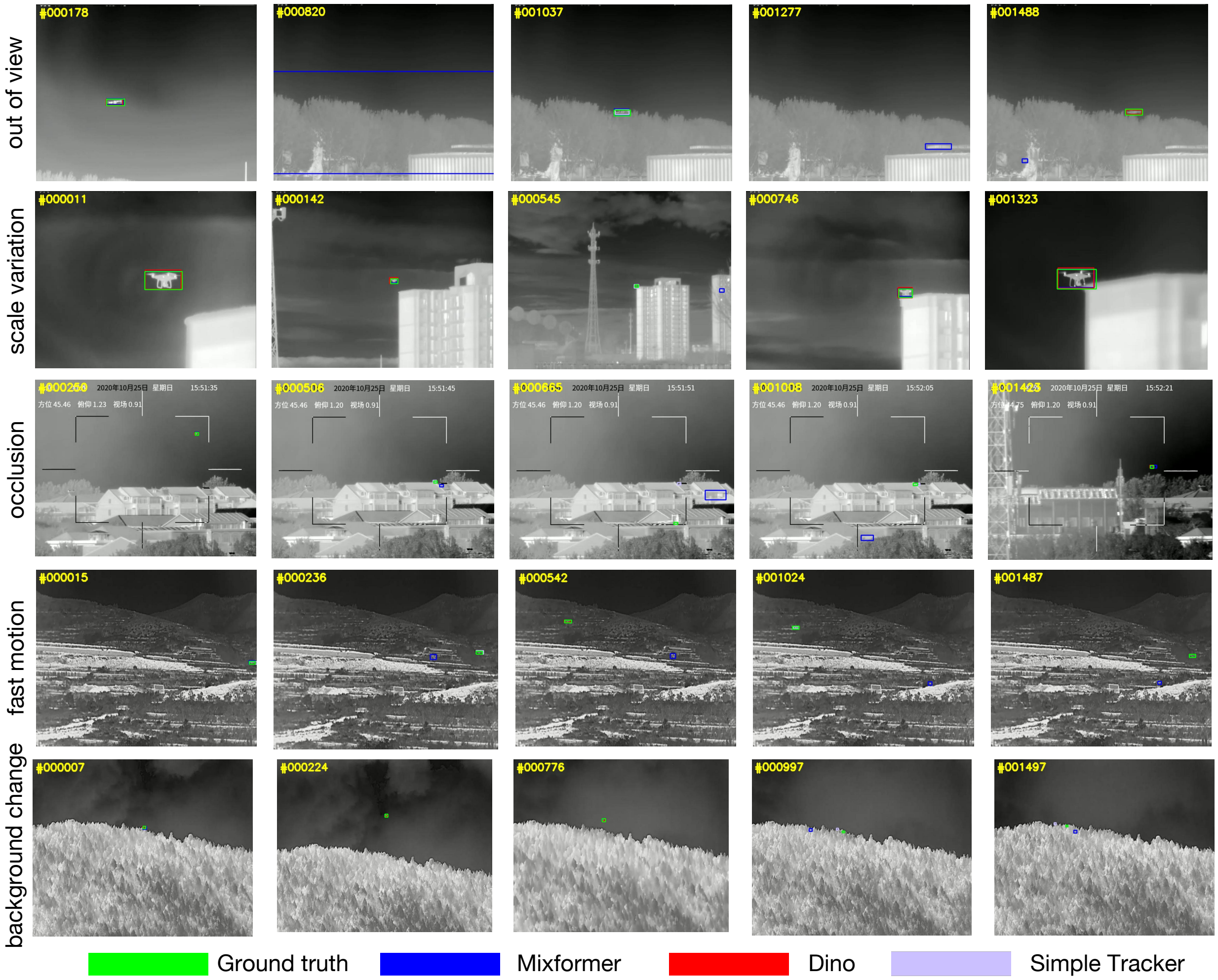

Figure 4. We generated corresponding visualization diagrams for various typical challenges in Antiuav, and compared the results of GT, using SOT method or detection method alone, and our method.

In general, we achieved the championship in track1 and the third place in track2.

4.3. Ablation study

Table 3 investigates the major components through ablation. Baseline refers to using the top1 prediction in each frame of the DINO bounding box as the result for that frame. The threshold represents a simple approach, where if the confidence score of the top 1 prediction is less than the threshold, the current frame is predicted as None. We draw some important observations below.

Simple tracker module brings significant improvements. Comparing Lines *(3)* against the baseline model in Line *(1)*, we observe that simple tracker alone brings noticeable improvement(e.g., 0.532 to 0.599 on validation). The reason why SimpleTrack can bring such a significant improvement is mainly that it utilizes the temporal information in the video and continuously judges the information of multiple frames through the link operation. In challenging scenarios such as Anti-UAV, utilizing the temporal information in the video is an extremely important aspect. In fact, we also found during visualization that if only single frames are used for judgment, even humans would find it difficult to identify many complex scenarios, which must be discovered by combining the video.

Motion Model can further solve difficult scenarios. As shown in Figure 3, we have observed that motion models often detect the location of challenging targets that are difficult for detectors to find. However, this can also introduce a significant amount of noise. By combining the results of the motion model with the overall detection and tracking results, we can effectively suppress this noise and identify useful targets, resulting in further improvement.

Single Object Tracking refines the results. Using single object tracking tracking as a post-processing module can further improve the performance to some extent, but we found that using it together with motion and simple tracker will have some marginal effects. Using it alone with simple tracker can bring relatively considerable improvement, but the improvement after adding motion is not obvious, indicating that the difficult scenarios solved by motion and SOT have a certain degree of intersection.

Other Missing Modules. We only used video classification and ensembled the results of multiple detectors on 200 test videos in track 1. In the current implementation, we only cropped the original video using only the bounding box with the highest confidence per frame and sent it to the video classifier. If the classification result was negative, we halved the confidence of the current bounding box so that it would not become a new initial box in tracking. We used the simple tracker module to obtain the prediction results of each frame for each detector and then ensembled these prediction results. The ensemble module and the video classification module each provided an additional improvement of around 0.3-0.5 on the final track1 test set.

4.4. Qualitative Evaluation

Figure 4 shows qualitative comparisons between our tracker and other state-of-the-art trackers and detectors. The Simple Tracker is shown to significantly outperform other trackers in handling challenging tracking situations such as out-of-view, scale changes, occlusions, fast movements, and background transformations. Due to the proposed box proposal linking mechanism, our Simple Tracker can track tiny objects with complex backgrounds well and quickly recapture objects when lost. Furthermore, the single object trackers in Simple Tracker can obtain favorable motion features based on the detection results, thus confidently tracking tiny objects.

5. Conclusions

This paper proposes a tracking-by-detection strategy to address challenges faced when tracking UAVs using single object tracking (SOT) networks such as OSTrack and Mixformer. The movement of UAVs can be complex leading to difficulties in SOT. The strategy comprises two main modules: Strong Detector and Simple Tracker. We have chosen several types of detectors and optimized each one for the unique characteristics of the UAV object resulting in our Strong Detector. The Simple Tracker uses cascading rules that link the results of the Strong Detector to achieve final tracking results. Infrared images can pose challenges when relying solely on pure detection methods due to noise blocks in the background being similar to foreground targets. To further improve accuracy, we have utilized temporal information and designed two modules: Video Checker and Motion Model.

References

[1] Olivier Barnich and Marc Van Droogenbroeck. Vibe: a powerful random technique to estimate the background in video sequences. In *2009 IEEE international conference on acoustics, speech and signal processing*, pages 945–948. IEEE, 2009. 2

[2] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13608–13618, 2022. 2, 4

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[4] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022. 2

[5] Jocher Glenn. Ultralytics YOLOv8. 2023. 2

[6] Nan Jiang, Kuiran Wang, Xiaoke Peng, Xuehui Yu, Qiang Wang, Junliang Xing, Guorong Li, Qixiang Ye, Jianbin Jiao, Zhenjun Han, et al. Anti-uav: a large-scale benchmark for vision-based uav tracking. *IEEE Transactions on Multimedia*, 2021. 1

[7] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107:104117, 2021. 3

[8] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8238–8249, 2021. 2

[9] Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI*, pages 733–751. Springer, 2022. 2

[10] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10448–10457, October 2021. 2

[11] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 341–357. Springer, 2022. 2, 4

[12] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2022. 2, 4

[13] Jian Zhao, Gang Wang, Jianan Li, Lei Jin, Nana Fan, Min Wang, Xiaojuan Wang, Ting Yong, Yafeng Deng, Yandong Guo, et al. The 2nd anti-uav workshop & challenge: methods and results. *arXiv preprint arXiv:2108.09909*, 2021. 1