Solution for ECCV 2022 BDD100K Challenges

Feng Ni

Abstract. This technical report present an overview of our method for ECCV 2022 BDD100K MOTS (Multiple Object Tracking and Segmentation) Challenges. Our method is based on Unicorn, which can simultaneously solve four tracking problems (SOT, MOT, VOS, MOTS) with a single network using the same model parameters. We achieved 41.86 mHOTA and 34.33 mMOTA on the test set of BDD100K MOTS with single scale, and won the 3rd place.

Keywords: object tracking and segmentation

1 Introduction

Object tracking is one of the fundamental tasks in computer vision, which aims to build pixel-level or instance-level correspondence between frames and to output trajectories typically in the forms of boxes or masks. Over the years, diverse application scenarios and experimental setups divided object tracking into four separate sub-tasks: SOT, MOT, VOS, and MOTS. Consequently, tracking approaches tend to over-specialize on the characteristics of specific sub-tasks, lacking in generalization.

1.1 MOT

Multi-object tracking (MOT) aims at estimating bounding boxes and identities of objects in videos. Trackers of MOT are required to find and associate all instances of specific classes by themselves. The mainstream methods [10,15,16,7,8] follow the tracking-by-detection paradigm. Specifically, an MOT system typically has two main components, an object detector and a certain association strategy.

1.2 MOTS

MOTS is highly related to MOT by changing the form of boxes to fine-grained representation of masks. MOTS benchmarks [9,14] are typically from the same scenarios as those of MOT [6,14]. Besides, many MOTS methods are developed upon MOT trackers.

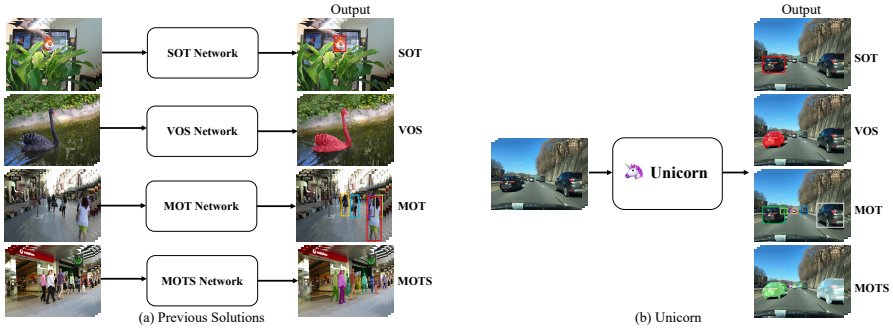

Fig. 1: Comparison between previous solutions and Unicorn.

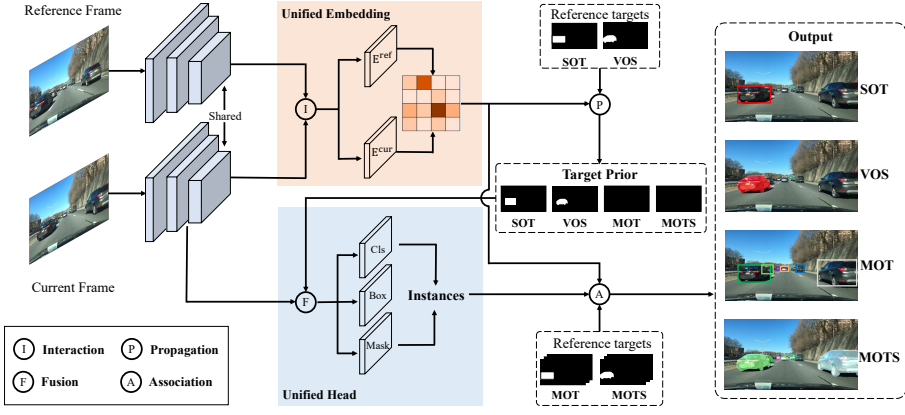

Fig. 2: Unicorn consists of three main components: (1) Unified inputs and backbone (2) Unified embedding (3) Unified head.

1.3 Unicorn

Unicorn [12] solves four tracking tasks with one unified network with the same parameters. Besides, Unicorn can learn powerful tracking representation from a large amount of labeled tracking data. Figure 1 shows the comparison between task-specific methods and the proposed Unicorn.

Unicorn consists of three main components: unified inputs and backbone; unified embedding and unified head. The framework of Unicorn is demonstrated in Figure 2. Given the reference frame $\mathbf{I}_{\text{ref}}$, the current frame $\mathbf{I}_{\text{cur}}$, and the reference targets, Unicorn aims at predicting the states of the tracked targets on the current frame for four tasks with a unified network.

2 Experiments

2.1 Implementation Details

We choose ConvNeXt-Large [4] as the backbone. The input image size is 800 × 1280 and the shortest side ranges from 736 to 864 during multi-scale training.

Table 1: State-of-the-art comparison on the BDD100K MOTS validation set.

Method	Online	mMOTSA↑	mMOTSP↑	mIDF1↑	ID Sw.↓	mAP↑
SortIoU	✓	10.3	59.9	21.8	15951	22.2
MaskTrackRCNN [13]	✓	12.3	59.9	26.2	9116	22.0
STEm-Seg [1]	✗	12.2	58.2	25.4	8732	21.8
QDTrack-mots [7]	✓	22.5	59.6	40.8	1340	22.4
QDTrack-mots-fix [7]	✓	23.5	66.3	44.5	973	25.5
PCAN [3]	✓	27.4	66.7	**45.1**	**876**	26.6
Unicorn	✓	**29.6**	**67.7**	44.2	1731	**32.1**

The model is trained on 16 NVIDIA Tesla A100 GPU with a global batch size of 32. To avoid inaccurate statistics estimation, we replace all Batch Normalization [2] with Group Normalization [11]. Two training stages randomly sample data from SOT&MOT datasets and VOS&MOTS datasets, respectively. Each training stage consists of 15 epochs with 200,000 pairs of frames in every epoch. The optimizer is Adam-W [5] with weight decay of $5e^{-4}$ and momentum of 0.9. The initial learning rate is $2.5e^{-4}$ with 1 epoch warm-up and the cosine annealing schedule. More details can be found in the supplementary materials.Unicorn in four tasks uses the same model parameters.

2.2 Training and Inference

The whole training process divides into two stages: SOT-MOT joint training and VOS-MOTS joint training. In the first stage, the network is end-to-end optimized with the correspondence loss and the detection loss using data from SOT&MOT. In the second stage, a mask branch is added and optimized with the mask loss using data from VOS&MOTS with other parameters fixed.

For MOT&MOTS inference, Unicorn detects all objects of the given categories and simultaneously outputs corresponding instance embeddings.

2.3 Evaluations on BDD100K MOTS

Finally, we evaluate the Unicorn on BDD100K MOTS [14]. The main evaluation metrics are mHOTA and mMOTA. BDD100K MOTS Challenge includes 37 sequences in the validation set. Tab. 1 demonstrates that Unicorn outperforms the previous best method PCAN [3] by a large margin. Tab. 2 shows the final results of BDD100K MOTS test set. we got 41.86 mHOTA and 34.33 mMOTA on the test set of BDD100K MOTS with single scale, and won the 3rd place.

3 Conclusions

In the BDD100K MOTS Challenge of ECCV 2022, we use an enhanced MOTS model Unicorn and achieve significant improvement. Thanks all the previous

Table 2: Final results on the BDD100K MOTS test set.

Method	Online	mHOTA↑	mMOTA↑	mIDF1↑	ID Sw.↓	mAP↑
PCAN [3]	✓	39.17	31.91	**50.42**	-	33.77
Our team	✓	**41.86**	**34.33**	52.93	-	**36.91**

great works PCAN [3], QDTrack [7] and Unicorn [12]. We did it based on the model of Unicorn paper, and made some attempts but no gain, so we still directly use the original model as our best model. In the future, we will further explore higher quality MOTS methods in all aspects of object detection, object tracking and object segmentation.

References

1. Athar, A., Mahadevan, S., Osep, A., Leal-Taixé, L., Leibe, B.: STEm-Seg: Spatio-temporal embeddings for instance segmentation in videos. In: ECCV (2020) 3
2. Ioffe, S., Szegedy, C.: Batch Normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015) 3
3. Ke, L., Li, X., Danelljan, M., Tai, Y.W., Tang, C.K., Yu, F.: Prototypical cross-attention networks for multiple object tracking and segmentation. NeurIPS (2021) 3, 4
4. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. arXiv preprint arXiv:2201.03545 (2022) 2
5. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 3
6. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: MOT16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016) 1
7. Pang, J., Qiu, L., Li, X., Chen, H., Li, Q., Darrell, T., Yu, F.: Quasi-dense similarity learning for multiple object tracking. In: CVPR (2021) 1, 3, 4
8. Sun, P., Jiang, Y., Zhang, R., Xie, E., Cao, J., Hu, X., Kong, T., Yuan, Z., Wang, C., Luo, P.: TransTrack: Multiple-object tracking with transformer. arXiv preprint arXiv:2012.15460 (2020) 1
9. Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B.B.G., Geiger, A., Leibe, B.: MOTS: Multi-object tracking and segmentation. In: CVPR (2019) 1
10. Wang, Z., Zheng, L., Liu, Y., Li, Y., Wang, S.: Towards real-time multi-object tracking. In: ECCV (2020) 1
11. Wu, Y., He, K.: Group Normalization. In: ECCV (2018) 3
12. Yan, B., Jiang, Y., Sun, P., Wang, D., Yuan, Z., Luo, P., Lu, H.: Towards grand unification of object tracking. In: ECCV (2022) 2, 4
13. Yang, L., Fan, Y., Xu, N.: Video instance segmentation. In: ICCV (2019) 3
14. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: BDD100K: A diverse driving dataset for heterogeneous multitask learning. In: CVPR (2020) 1, 3
15. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: FairMOT: On the fairness of detection and re-identification in multiple object tracking. IJCV (2021) 1
16. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. In: ECCV (2020) 1