Towards Grand Unification of Object Tracking
—— Supplementary Material ——

Bin Yan[1⋆], Yi Jiang[2,†], Peize Sun[3], Dong Wang[1,†],
Zehuan Yuan[2], Ping Luo[3], and Huchuan Lu[1,4]

[1] School of Information and Communication Engineering, Dalian University of Technology, China [2] ByteDance [3] The University of Hong Kong
[4] Peng Cheng Laboratory

1 Appendix

In this appendix, we provide more details about the unified head architecture and the training process of the Unicorn.

1.1 Unified Head Architecture

The detailed head architecture is shown in Fig. 1. The unified head takes the original FPN feature $\mathbf{F} \in \mathbb{R}^{h\times w\times c}$ and the target prior $\mathbf{P} \in \mathbb{R}^{h\times w\times 1}$ as the inputs. The two inputs are first fused by broadcast sum, getting the fused feature $\mathbf{F}^{'} \in \mathbb{R}^{h\times w\times c}$. Then the fused feature is passed to the detection head [3] and the instance segmentation head [10], predicting final boxes or masks. The head network is fully-convolutional, without any RoI operation such as RoI Align.

1.2 Training Details

Since accurate mask annotations are quite expensive while bounding boxes are relatively cheap, available training data of SOT&MOT is usually dozens of times that of VOS&MOTS. Directly mixing training data from four tasks will lead to a serious data-imbalance problem. To alleviate this problem, we divide the whole training process into two stages. Specifically, in the first stage, we randomly sampled training data from SOT datasets (COCO [5], LaSOT [2], GOT-10K [4] and TrackingNet [7]) and MOT datasets (For evaluating on MOT Challenge, we use Crowdhuman [9], ETHZ [1], CityPerson [14], MOT17 [6]. For evaluating on BDD100K, we use the training set of BDD100K [13]) with a 1:1 sampling ratio to train the whole network without the mask head. In this stage, the network is optimized with the sum of the correspondence loss and the detection loss. $\mathbf{L}_{\text{stage1}} = \mathbf{L}_{\text{corr}} + \mathbf{L}_{\text{det}}$. More details about $\mathbf{L}_{\text{det}}$ can be found in the YOLOX paper. Then in the second stage, to prevent the model from overfitting on the VOS&MOTS and negatively impacting the performance of SOT&MOT, we only train the mask head with the data from VOS (COCO [5], DAVIS [8], Youtube-VOS [12]) and MOTS (MOTS [11], BDD100K [13]), leaving other parameters fixed. $\mathbf{L}_{\text{stage2}} = \mathbf{L}_{\text{mask}}$

⋆ This work was performed while Bin Yan worked as an intern at ByteDance.

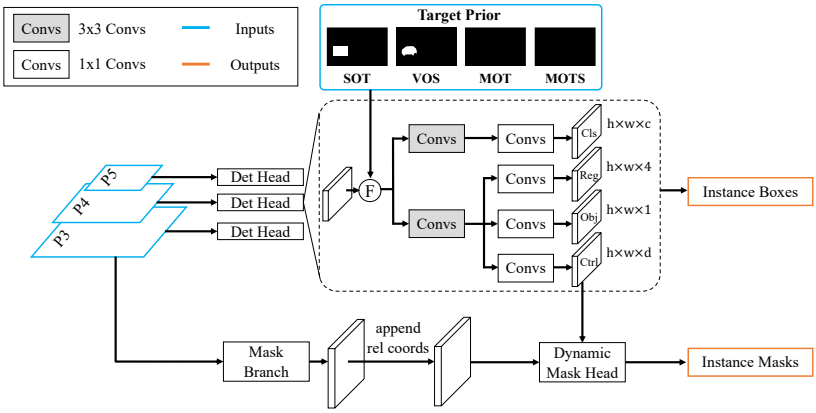

Fig. 1. Unified head architecture of the Unicorn.

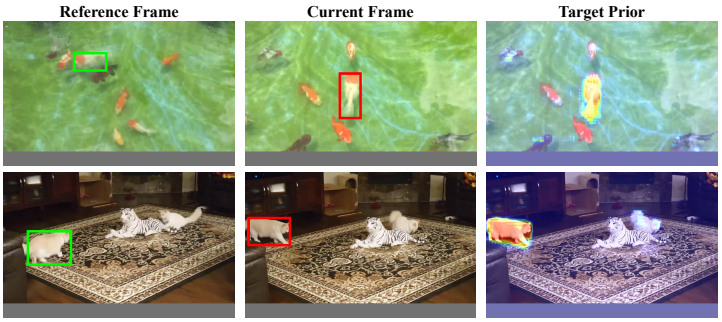

Fig. 2. Visualization of the target prior.

1.3 Visualization

In Figure 2, given the tracked target (highlighted with a green box) on the reference frame, we visualize the predicted target prior on the current frame. It can be seen that Unicorn can predict accurate correspondence in challenging scenarios even though there are many similar distractors.

References

1. Ess, A., Leibe, B., Schindler, K., Van Gool, L.: A mobile vision system for robust multi-person tracking. In: CVPR (2008) 1
2. Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H.: LaSOT: A high-quality benchmark for large-scale single object tracking. In: CVPR (2019) 1
3. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: YOLOX: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021) 1
4. Huang, L., Zhao, X., Huang, K.: GOT-10k: A large high-diversity benchmark for generic object tracking in the wild. TPAMI (2019) 1
5. Lin, T.Y., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014) 1
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. Muller, M., Bibi, A., Giancola, S., Alsubaihi, S., Ghanem, B.: Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In: ECCV (2018) 1
8. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017) 1
9. Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., Sun, J.: Crowdhuman: A benchmark for detecting human in a crowd. arXiv preprint arXiv:1805.00123 (2018) 1
10. Tian, Z., Shen, C., Chen, H.: Conditional convolutions for instance segmentation. In: ECCV (2020) 1
11. 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
12. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: YouTube-VOS: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018) 1
13. 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
14. Zhang, S., Benenson, R., Schiele, B.: Citypersons: A diverse dataset for pedestrian detection. In: CVPR (2017) 1