1st Place Solution for ECCV2022 SSLAD BDD100K MOT/MOTS/SSMOT/SSMOTS Challenges

Kaer Huang[1], Kanokphan Lertniphonphan[1], Feng Chen[1], Tao Zhang[2], Jun Xie[1], Huabing Liu[3], Qigang Wang[1], Zhepeng Wang[1]✉

[1]Lenovo Research [2]Tsinghua University
[3]LCFC (Hefei) Electronics Technology Co., Ltd.

Abstract. In recent years, dominant Multi-object tracking (MOT) and segmentation (MOTS) methods mainly follow the tracking-by-detection paradigm. Transformer-based end-to-end (E2E) solutions bring some ideas to MOT and MOTS, but they cannot achieve a new state of the art (SOTA) performance in major MOT and MOTS benchmarks. Detection and association are two main modules of the tracking-by-detection paradigm. Association techniques mainly depend on the combination of motion and appearance information. As deep learning has been recently developed, the performance of detection and appearance model are rapidly improved. These trends made us consider whether we can achieve SOTA based on only high-performance detection and appearance model. Our paper mainly focus on exploring this direction based on CBNetV2 with Swin-B as detection model and MoCo-v2 as self-supervised appearance model. Motion information and IoU mapping were removed during the association. Our method achieves SOTA results on BDD100K MOT and MOTS dataset and win 1st place of all tracks in track 4 challenges, which consist of MOT, MOTS, SSMOT, and SSMOTS, in ECCV2022 SSLAD workshop. We hope our simple and effective method can give some insights to the MOT and MOTS research community. Source code will be released under this git repository https://github.com/CarlHuangNuc.

Keywords: MOT, MOTS, Self-Supervised Learning

1 Introduction

Object tracking is one of the fundamental tasks in computer vision, which used to build instance-level correspondence between frames and output trajectories with boxes or masks [18]. MOT and MOTS tasks aim to simultaneously process detecting, segmenting and tracking object instances in a given video [17]. It can be used in video surveillance, autonomous driving, video understanding, etc.

Current mainstream methods follow the tracking-by-detection paradigm [9, 11, 13, 16]. Until recent years, Transformer-based E2E solutions brought new ideas to MOT and MOTS research areas [3–5, 19], but their performance could not reach SOTA in major MOT and MOTS benchmarks. Detection and association are two main modules of tracking-by-detection paradigm. Association techniques mainly depend on the combination of motion and appearance information [12, 21]. As deep learning developed, appearance and detection models get rapid improvement in performance. At the same time, the difficulty of the autonomous vehicle dataset includes low video frame rate, fast movement, and large displacement. The traditional association methods based on IoU and motion do not perform well in this kind of situations.

The challenge of association based on motion information, made us consider whether we can archive SOTA only based on high-performance detection and appearance model. Our paper tried to explore this direction. We use CBNetV2 Swin-B [10] as detection model and self-supervised learning MoCo-v2 [7] as high-quality appearance model. We removed all motion information, including Kalman filter and IoU mapping, and archived SOTA on BDD100K dataset. Our method win 1st Place in CVPR2022 WAD BDD100K MOT challenge, and 1st Place in ECCV2022 SSLAD track 4 BDD100K challenges, including MOT, MOTS, SSMOT, and SSMOTS tracks. We hope our simple and effective method can give some insight to the MOT and MOTS research community.

2 Related Work

Multi Object Tracking (MOT) is a very general algorithm and has been studied for many years. The mainstream methods follow the tracking-by-detection paradigm [9, 11, 13, 16]. With the development of deep learning in recent years, the performance of the detection model is improved rapidly. Currently, most of the work relies on YOLOX [18, 20]. Our method selected a stronger performance network CBNetV2 [10] which is used to verify the potential of the detector in our hypothesis. Another important component of MOT is an association strategy. Popular association methods include motion-based (IoU matching, Kalman filter) [1], appearance-based (ReID embedding) [15], transformer-based [19], or the combination of them [12, 21]. Our methods remove all motion information, and use only high-performance appearance model.

Multi Object Tracking and Segmentation (MOTS) is highly related to MOT by changing the form of boxes to fine-grained mask representation [18]. Many MOTS methods are developed upon MOT trackers [8, 14]. Our ideas are similar to their. A mask header was added on the basis of MOT network in our MOTS solution.

Self-Supervised Learning has made significant progress in representation learning in recent years. Contrastive learning, one of self-supervised learning methods such as MoCo[7], SimCLR[2], BYOL[6], etc, has performance which is getting closer to results of supervised learning methods in ImageNet dataset. We leveraged Momentum Contrastive Learning (MoCo-v2)[7] to train a new appearance embedding model without using tracking annotations. The technique is not only meets the requirements of SSMOT and SSMOTS, but also improves the performance of appearance model.

3 Method

The overview of our framework is shown in figure 1. The framework is based on tracking-by-detection paradigm. Object bounding boxes are detected in each image by a detector in MOT. In MOTS, a segmentation head is added to the detector to extract binary masks within each detected box. A ReID model extracts features from the bounding boxes. Then, a tracker process the data association to match object ID in the image sequence.

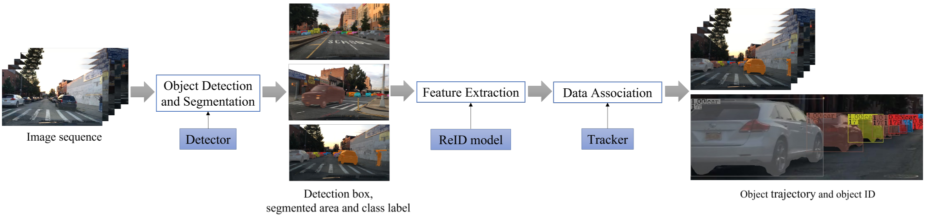

Fig. 1. Our framework

3.1 Detection and Segmentation

We applied CBNetV2 architecture to connect two Swin-B with FPN backbones in parallel. Features from high and low level from the backbones are integrated to improve detector performance. The HTC detection head was used to predict box and binary mask. The mask head is trained with a multi-steps training strategy. Firstly, the model was trained for box detection by using a relatively large number of box labeled data. Then, the whole network with mask branch was fine-tuned based on MOTS labeled dataset. In addition, multi-class NMS threshold is applied to reduce data imbalance problem.

3.2 Re-Identification

We used Unitrack as ReID module for MOT and MOTS. Our appearance model for this framework is MoCo-v2 with ResNet50 backbone. The model extracts feature representations from detected boxes. The tracklet features are weighted by the detection score and combined within τ frames to maintain the object representation during occlusion. The weighted feature $\hat{e_j}$ combined tracklet feature e_j which weighted by the detection score s_j from the previous τ frames.

$$\hat{e_j} = \frac{\sum_{t=1}^{\tau} e_j^t \times s_j^t}{\sum_{t=1}^{\tau} s_j^t} \quad (1)$$

$\hat{e_j}$ is further used for computing ReID distance in the data association.

3.3 Tracking

ByteTrack method, which divides detection boxes into high and low detection score for data association, is used in our framework. Firstly, the high score boxes are used to associate with the tracklet. The remained high score boxes will be kept as tentative boxes, which will become a new tracklet after appearing for 2 consecutive frames. Then, the low score boxes are used to find the matching with the remained tracklet. From our experiments, using ReID distance has the best results in all high and low score box association. Then, the Hungarian algorithm uses the distance to assign the tracking ID in each association step. For the lost and occluded tracklets, they are kept within 10 frames.

4 Experiments

In this section, we introduce the dataset and evaluation metrics. Then, we explain our implementation details for experiments. Finally, we report the main results on ECCV2022 BDD100K Challenges test server and ablation study of major methods.

4.1 Dataset and Evaluation Metrics

We conducted experiments on BDD100K dataset which is a large-scale autonomous driving video dataset with 100K driving videos (40 seconds each). BDD100K provides the multi-task annotations for MOT and MOTS. MOT dataset contains 1400 and 200 videos with annotation for training and validation, respectively, and 400 videos for testing. MOTS dataset contains 154 and 32 videos with annotation for training and validation, respectively, and 37 videos for testing.

Mean Higher Order Tracking Accuracy (mHOTA) is used as a main metric for ranking in ECCV2022 BDD100K challenges. Mean Multiple Object Tracking Accuracy (mMOTA) and mean ID F1 score (mIDF1) are used as secondary metrics to evaluate MOT and MOTS performance. In MOT, box IoU is used to calculate distance matrices, while the mask IoU is used in MOTS. Self-supervised MOT (SSMOT) and self-supervides MOTS (SSMOTS) leverage the same metrics as MOT and MOTS.

4.2 Implementation Details

Detector. CBNetV2 was trained on both BDD100K object detection and MOT dataset. The Swin-B backbone was initiated by a model pretrained on ImageNet-22K. We applied multi-scale augmentation to scale the shortest side of images to between 640 and 1280 pixels and applied random flip augmentation during training. The optimizer is AdamW with an initial learning rate of 1e-6 and weight decay of 0.05. We trained the model on 4 A100 GPUs with 1 image per GPU for 10 epochs. At inference time, we resize the image size to 2880x1920 to better detect the small objects. We applied the multi-class NMS thresholds 0.6, 0.1, 0.5, 0.4, 0.01, 0.01, 0.01, and 0.4 for pedestrian, rider, car, truck, bus, train, motorcycle, and bicycle class, respectively.

Segmentation Head. The backbone, neck, and detection head was initiated by MOT detector. Then, we fine-tuned the MOTS detector with BDD100K instance segmentation and MOTS dataset. The AdamW optimizer was set the initial learning rate of 5e-7 and weight decay of 0.05. We trained the model on 4 A100 GPUs with 1 image per GPU for 20 epochs.

ReID. The backbone of ReID is pretrained on ImageNet-1K. Then, we fine-tuned the backbone by using MoCo-v2 on BDD100K dataset. The training dataset contains cropped object images according to bounding box labels from MOT dataset. The optimizer is SGD with weight decay of 1e-4, momentum factor of 0.9, and initial learning rate of 0.12. We trained the model on 4 A100 GPUs with 256 images per GPU.

We do not rely on the tracking annotations when training the detector, segmentation head, and ReID model, thus our method can be applied to SSMOT and SSMOTS.

Tracker. Our method is generally similar to ByteTrack, but we used ReID to match high and low detection boxes. We set the high detection score threshold to 0.84 and low detection score threshold to 0.3.

4.3 Main Results

We evaluated the performance of our method on BDD100K MOT and MOTS test set. We achieve 49.2 and 44.0 mHOTA in BDD100K MOT and MOTS which outperform the next place by 2.9 and 2.1 mHOTA, respectively, as shown in Table 1 and 2. Since we do not use the tracking annotations when training detector and ReID model, our method can be applied to SSMOT and SSMOTS tasks and achieve the same results as shown in Table 3 and Table 4.

4.4 Ablation Study

We performed ablation experiments to study the effect of each module on BDD100K MOT validation set and reported the results in Table 5. We use ByteTrack as our strong baseline. The framework contains CBNetv2 with Swin-B backbone detector and a ReID model from Unitrack. The baseline achieves 48.8 mHOTA. Then, we added weighted ReID features module and got 0.4 higher

Table 1. Comparison with other methods on **BDD100K MOT** test set. **Bold** represents the best metrics.

Team	mHOTA	mMOTA	mIDF1	mDetA	mAssA	mMOTP
Ours	**49.2**	**43.0**	**59.5**	**43.9**	**56.4**	**81.4**
bbq (v7)	46.3	38.1	55.2	41.0	53.9	81.1
Anonymous	44.4	40.4	53.0	39.9	50.2	72.9
CMSQ	42.4	36.2	53.4	35.5	52.3	77.0
Host_38176_Team (QDtrack)	41.9	35.7	52.4	34.6	52.4	77.8

Table 2. Comparison with other methods on **BDD100K MOTS** test set.

Team	mHOTA	mMOTA	mIDF1	mDetA	mAssA	mMOTP
Ours	**44.0**	**41.1**	**54.9**	**39.3**	**50.8**	**69.7**
Anonymous	41.9	34.4	52.9	36.9	49.1	67.7
vdig	41.9	34.3	52.9	36.9	49.0	67.7
OKC	40.0	32.6	50.3	35.5	46.7	67.4
Host_58935_Team	39.2	31.9	50.4	33.8	46.3	66.5

Table 3. Comparison with other methods on **BDD100K SSMOT** test set.

Team	mHOTA	mMOTA	mIDF1	mDetA	mAssA	mMOTP
Ours	**49.2**	**43.0**	**59.5**	**43.9**	**56.4**	**81.4**
Host_34931_Team	37.8	35.4	46.8	32.0	46.0	71.2

Table 4. Comparison with other methods on **BDD100K SSMOTS** test set.

Team	mHOTA	mMOTA	mIDF1	mDetA	mAssA	mMOTP
Ours	**44.0**	**41.1**	**54.9**	**39.3**	**50.8**	**69.7**
Host_28547_Team	36.8	25.6	46.7	32.1	43.3	65.7

Table 5. Ablation study of each module on **BDD100K MOT validation set**.

Method	mHOTA	mMOTA
Baseline (CBNetv2_Swin-B + ByteTrack + ReID)	48.8	44.5
+ Weighted ReID Features	49.2 (+0.4)	45.3 (+0.4)
+ Contrastive Learning ReID Model	50.0 (+0.8)	45.8 (+0.5)
+ Parameters Fine Tuning	50.0	45.9 (+0.1)

score on mHOTA and mMOTA. Next, we trained the ReID model with Resnet-50 backbone on BDD100K by using momentum contrastive learning method and improved 0.8 mHOTA and 0.5 mMOTA. Finally, we fine-tuned matching thresholds in ByteTrack and achieved 50.0 mHOTA and 45.9 mMOTA in BDD100K MOT validation set.

5 Conclusions

In this paper, we propose a simple yet effective tracking-by-detection framework for multi-object tracking (MOT) and segmentation (MOTS) and achieve the state-of-the-art results in BDD100K MOT and MOTS dataset. We discard the motion information and only use the appearance embeddings to associate the objects. The training of detection and appearance models does not rely on tracking annotations which can be costly to obtain. Our method achieves the first place in CVPR2022 WAD BDD100K MOT Challenge with 45.6 mMOTA on validation set and 44.0 mMOTA on test set. We also achieve the first place in ECCV2022 SSLAD all 4 BDD100 challenges of MOT, MOTS, SSMOT and SSMOTS. We hope the simplicity and effectiveness of our method can benefit future research of MOT and MOTS.

References

[1] Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: 2016 IEEE international conference on image processing (ICIP). pp. 3464–3468. IEEE (2016)

[2] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)

[3] Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1290–1299 (2022)

[4] Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. Advances in Neural Information Processing Systems **34**, 17864–17875 (2021)

[5] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[6] Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)

[7] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)

[8] Ke, L., Li, X., Danelljan, M., Tai, Y.W., Tang, C.K., Yu, F.: Prototypical cross-attention networks for multiple object tracking and segmentation. Advances in Neural Information Processing Systems **34**, 1192–1203 (2021)

[9] Liang, C., Zhang, Z., Zhou, X., Li, B., Zhu, S., Hu, W.: Rethinking the competition between detection and reid in multiobject tracking. IEEE Transactions on Image Processing **31**, 3182–3196 (2022)

[10] Liang, T., Chu, X., Liu, Y., Wang, Y., Tang, Z., Chu, W., Chen, J., Ling, H.: Cbnetv2: A composite backbone network architecture for object detection. arXiv preprint arXiv:2107.00420 (2021)

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

[12] Pang, J., Qiu, L., Li, X., Chen, H., Li, Q., Darrell, T., Yu, F.: Quasi-dense similarity learning for multiple object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 164–173 (2021)

[13] Sun, P., Cao, J., Jiang, Y., Zhang, R., Xie, E., Yuan, Z., Wang, C., Luo, P.: Transtrack: Multiple object tracking with transformer. arXiv preprint arXiv:2012.15460 (2020)

[14] Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B.B.G., Geiger, A., Leibe, B.: Mots: Multi-object tracking and segmentation. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 7942–7951 (2019)

[15] Wang, Z., Zhao, H., Li, Y.L., Wang, S., Torr, P., Bertinetto, L.: Do different tracking tasks require different appearance models? Advances in Neural Information Processing Systems **34**, 726–738 (2021)

[16] Wu, J., Cao, J., Song, L., Wang, Y., Yang, M., Yuan, J.: Track to detect and segment: An online multi-object tracker. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12352–12361 (2021)

[17] Wu, J., Liu, Q., Jiang, Y., Bai, S., Yuille, A., Bai, X.: In defense of online models for video instance segmentation. arXiv preprint arXiv:2207.10661 (2022)

[18] Yan, B., Jiang, Y., Sun, P., Wang, D., Yuan, Z., Luo, P., Lu, H.: Towards grand unification of object tracking. arXiv preprint arXiv:2207.07078 (2022)

[19] Zeng, F., Dong, B., Wang, T., Zhang, X., Wei, Y.: Motr: End-to-end multiple-object tracking with transformer. arXiv preprint arXiv:2105.03247 (2021)

[20] Zhang, Y., Sun, P., Jiang, Y., Yu, D., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. arXiv preprint arXiv:2110.06864 (2021)

[21] Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision **129**(11), 3069–3087 (2021)