Multi-Object Tracking by Self-supervised Learning Appearance Model

Kaer Huang
Lenovo
huangke1@lenovo.com

Kanokphan Lertniphonphan
Lenovo
klertniphonp@lenovo.com

Feng Chen
Lenovo
chenfeng13@lenovo.com

Jian Li
Lenovo
lijian30@lenovo.com

Zhepeng Wang
Lenovo
wangzpb@lenovo.com

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 can not 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 the detection and appearance model is rapidly improved. These trends made us consider whether we can achieve SOTA based on only high-performance detection and appearance model. Our paper mainly focuses on exploring this direction based on CBNetV2 with Swin-B as a detection model and MoCo-v2 as a self-supervised appearance model. Motion information and IoU mapping were removed during the association. Our method achieves SOTA results on 2 mainstream MOT datasets and 1 MOTS dataset which is BDD100K MOT, WAYMO 2D Tracking, BDD100K MOTS. Our method yielded a significant improvement of +10.7% and +33.7%, respectively on BDD 100K MOT and MOTS benchmark. The proposed method won first place in BDD100K Multiple Object Tracking (MOT) challenges at CVPR 2022 Workshop on Autonomous Driving. Our method also won first place in BDD100K Multiple Object Tracking (MOT) and Multiple Object Tracking and Segmentation (MOTS) challenges at ECCV 2022 Self-supervised Learning for Next-Generation Industry-level Autonomous Driving (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/MOT_MOTS_without_motion.

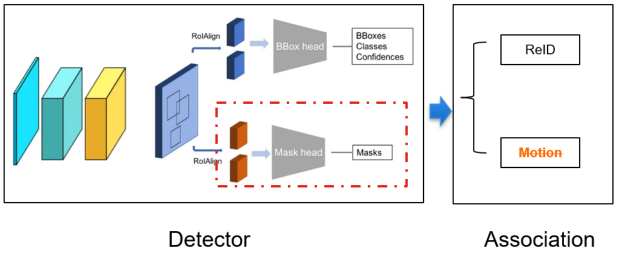

Figure 1. MOT/MOTS overall architecture

1. Introduction

Multiple Object Tracking (MOT) 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 [1]. MOT task aims to simultaneously process detecting and tracking object instances in a given video [2]. The only difference between MOT and MOTS is the latter adds a segmentation branch. Therefore, this paper regards MOT and MOTS as one technical direction. It can be used in video surveillance, autonomous driving, video understanding, etc.

Current mainstream MOT/MOTS methods follow the tracking-by-detection paradigm [3–6]. Until recent years, Transformer-based E2E solutions brought new ideas to MOT and MOTS research areas [7–10], but their performance could not reach SOTA 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 [11, 12]. As deep learning develops, appearance and detection models get rapid improvement in performance. At the same time, the unique diffi-

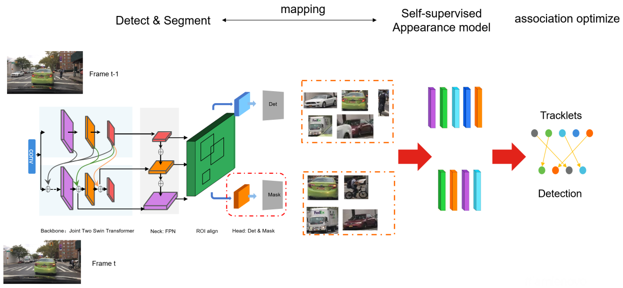

Figure 2. Detail MOT/MOTS framework

culty of the autonomous vehicle dataset includes low video frame rate, fast movement, and large displacement make 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 in the autonomous driving dataset, made us consider whether we can achieve SOTA only based on high-performance detection and appearance model. Our paper tried to explore this direction. We use CBNetV2 Swin-B [13] as the detection model and self-supervised learning MoCo-v2 [14] as a high-quality appearance model. We removed all motion information, including the Kalman filter and IoU mapping. We also introduce ByteTrack [15] innovation to associate the low-score detection boxes and the high-score ones.

Our method remains simple, online, and significantly improves robustness over fast movement and large displacement. Our contributions are summarized as the following:

(1) We propose multi-object tracking and segmentation framework based on high-quality detection plus appearance features which can be applied in the supervised and self-supervised MOT/MOTS dataset;

(2) We perform an ablation test on an extension of the proposed framework on different detectors, appearance models, and association methods;

(3) we add a weighted score to tracklet in order to keep the tracklet representation more smoothly since the detection score tends to get lower when the occluded part gets bigger;

(4) We acquire the state-of-the-art MOT and MOTS in BBD100K dataset on both supervised and self-supervised tasks. At the same time, we also extended our methods to other datasets (WAYMO 2D Tracking), and also achieved good results. We won first place in BDD100K Multiple Object Tracking (MOT) challenges at CVPR 2022 Workshop on Autonomous Driving. we also won first place in BDD100K Multiple Object Tracking (MOT) and Multiple Object Tracking and Segmentation (MOTS) challenges at ECCV 2022 Self-supervised Learning for Next-Generation Industry-level Autonomous Driving (SSLAD) Workshop.

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 [3–6]. With the development of deep learning in recent years, the performance of the detection model is improved rapidly. Currently, most of the latest public work relies on YOLOX [1, 15–17]. Our method selected a stronger performance network CBNetV2 [13] 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) [15, 18, 19], appearance-based (ReID embedding) [1, 20], transformer-based [5, 8, 16, 21, 22], or the combination of them [11, 12, 19]. Our methods remove all motion information and use only a high-performance appearance model which is got by supervised or self-supervised learning.

Multi Object Tracking and Segmentation (MOTS) is highly related to MOT by changing the form of boxes to fine-grained mask representation [1]. The only metrics difference between MOT and MOTS lies in the computation of distance matrices. In MOT, it is computed using box IoU, while for MOTS the mask IoU is used. Many MOTS methods are developed upon MOT trackers [6, 21, 23–25]. Our ideas are similar to theirs. 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 the self-supervised learning methods such as MoCo [14], SimCLR [26], BYOL [27], SwAV [28], etc, has amazing performance which is getting closer to results of supervised learning methods in ImageNet dataset. We leveraged Momentum Contrastive Learning (MoCo-v2) [14] to train a new appearance embedding model without using tracking annotations. The technique not only meets the requirements of self-supervised tracking but also improves the performance of the appearance model.

3. Method

In this section, we present detail of the multi-object tracking framework including detection and segmentation (Sec. 3.2), appearance model (Sec. 3.3), and data association (Sec. 3.4).

3.1. Overall Architecture

As shown in Figure 3, the proposed method is quite simple and it mainly contains three parts: detection and segmentation, appearance model(ReID model), and data as-

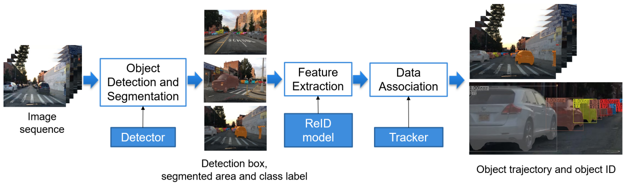
Image sequence
Object Detection and Segmentation
Detector
Detection box, segmented area and class label
Feature Extraction
ReID model
Data Association
Tracker
Object trajectory and object ID

Figure 3. The overall architecture of MOT/MOTS

sociation(Tracker). The detection and segmentation part is mainly responsible for providing high-quality instance bbox and mask information. The appearance model part is mainly responsible for providing high-quality embedding features. Data association(Tracker) part leverage detection output and appearance model to output stable trajectories.

3.2. Detection and Segmentation

Due to the high performance of the transformer, we adopt swin based transformer [29] backbone with Composite Backbone Network V2 (CBNetV2) [13] architecture to predict object bounding box. The CBNetV2 integrates high and low-level features of multiple backbones which connected in parallel. The Feature Pyramid Network (FPN) [30] neck and Hybrid Task Cascade (HTC) [31] detector are attached and trained in each backbone as a main branch and an assistant branch. Only the main branch is used in the inference process.

We use more weight to bound box regression than classification in Loss Function for a more compact detection bound box which will benefit to appearance model performance.

For segmentation, a mask header was added to the detection network. Since a tracking dataset with the segmented mask is rare and has some data distribution shift with MOT dataset, we utilize a model trained on the tracking dataset with bounding box and fine-tuned the mask head with the dataset with the mask.

3.3. Appearance Model

Our appearance model for this framework is MoCo-v2 [14] with ResNet50 backbone. The model extracts feature representations from detected boxes. MoCo-v2 model training by imagenet 1K dataset and then fine-tuning on BDD100K MOT dataset. We also compare with model training by other contrastive learning methods (SimCLR [26], SimCLRv2 [26], MoCo-v2 [14], etc). We also make a comparison between supervised learning and self-supervised learning. Finally, we draw the conclusion that MoCo-v2 [14] has better generalization capacity in the automatic driving dataset.

3.4. Data Association

We adopt Bytetrack [15] concept which is a simple but strong method for matching object id across frames. The detected boxes in each frame are grouped based on their detection score into the high score and low score. Firstly, the method finds the association between the high score box and the tracklet. Then, the rest of the high score and low score boxes are used to find the association from the remained tracklet. The association method can be different in each association step.

Since autonomous dataset usually includes low frame rate and large displacement from object and camera motion, IoU and motion-based tracking are not effective, especially in complex scene with a lot of occlusions. Our method uses only the appearance feature to associate both high and low score boxes with tracklet. In addition, we add a weighted score to tracklet in order to keep the tracklet representation from the higher detection score since the detection score tends to get lower when the occluded part gets bigger.

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 is weighted by the detection score s_j from the previous τ frames.

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

$\hat{e_j}$ is further used for finding the matched box in the data association. We apply the same association method with [20]. A ReId similarity matrix between tracklet and detection box is computed and used to find matching pairs by the Hungarian algorithm [32].

4. Experiments

4.1. Dataset and Metrics

BDD100K [33] is a large-scale autonomous driving video dataset with 100K driving videos. The dataset includes multiple object tracking (MOT) and segmentation (MOTS) datasets. The BDD100K MOT and MOTS datasets provide diverse driving scenarios with high-quality instance segmentation masks under complicated occlusions and reappearing patterns, which serves as a great testbed for the reliability of the developed tracking and segmentation algorithms in real scenes. MOT dataset contains 1400 videos for training, 200 videos for validation, and 400 videos for testing. MOTS dataset contains 154 videos for training, 32 videos for validation, and 37 videos for testing. In addition, the BDD100K also provides object bounding boxes and masks from the detection and instance segmentation sets but does not have tracking annotations for self-supervised learning tracks.

The dataset contains 8 types of objects: pedestrians, riders, cars, buses, trucks, trains, motorcycles, and bicycles. the evaluation metrics employ mean Multiple Object Tracking Accuracy (mMOTA) of MOTA of the 8 categories as a primary evaluation metric for ranking. It also employs the mean ID F1 score (mIDF1) to highlight the performance of tracking consistency which is crucial for object tracking.

ECCV2022 BDD100K MOT and MOTS competitions employ mean Higher Order Tracking Accuracy (HOTA, mean of HOTA of the 8 categories) as a primary evaluation metric for ranking. It also employs mean Multiple Object Tracking Accuracy (mMOTA) and mean ID F1 score (mIDF1), which are previously used as the main metrics. For MOTS, it uses the same metrics set as MOT. The only difference lies in the computation of distance matrices. In MOT, it is computed using box IoU, while for MOTS the mask IoU is used.

Waymo Open dataset [34] has scenes selected from both suburban and urban areas, at different times of the day. In addition to the urban/suburban and time-of-day diversity, scenes in the dataset are selected from many different parts of the cities. the dataset covers an area of 40km2 in Phoenix, and 36km2 combined in San Francisco and Mountain View. The dataset has around 12M labeled 3D LiDAR objects, around 113k unique LiDAR tracking IDs, around 12M labeled 2D image objects, and around 254k unique image tracking IDs.

It contains images from 5 cameras associated with 5 different directions: front, front left, front right, side left, and side right. There are 3,990 videos (790k images) for training, 1,010 videos (200k images) for validation, and 750 videos (148k images) for testing. It annotates 3 classes for evaluation. The videos are annotated at 10 FPS.

It use the multiple object tracking (MOT) metric [34,35]. This metric aims to consolidate detect, localize, and track the identities of objects over time into a single metric to aid in a direct comparison of method quality: MOTA and MOTP.

It computes a MOTA for each difficulty level (L1 and L2). We pick the highest MOTA among all the score cutoffs as the final metric.

4.2. Implementation Details

Detector. The Swin-B backbone [29] was initiated by a model pre-trained on ImageNet-22K [36]. CBNetV2 [13] was trained on both BDD100K object detection and MOT dataset. 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. AdamW optimizer was set 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. During inference, we resize an image to 2880x1920 to better detect the small objects. We applied 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 classes, respectively. For the detection task, we use a combination of classification Cross-Entropy loss and the generalized IoU regression loss [37]. Loss weights λ_1 and λ_2 are set to 1.0 and 10.0 by default, which drives the model output more compact bound box.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{box} \tag{2}$$

Segmentation Head. The backbone, neck, and detection head were initiated by MOT detector. Then, we fine-tuned the MOTS detector with BDD100K instance segmentation and MOTS dataset. The AdamW optimizer 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.

Appearance Model. The backbone of the appearance model is pre-trained on ImageNet-1K. Then, we fine-tuned the backbone by using MoCo-v2 [14] on BDD100K dataset. The training dataset contains cropped object images according to bounding box labels from MOT dataset. The optimizer is SGD with a weight decay of 1e-4, momentum factor of 0.9, and an initial learning rate of 0.12. We trained the model on 4 A100 GPUs with 256 images per GPU.

Tracker. Our method is generally similar to ByteTrack [15], but we used ReID to match high and low detection boxes. We set the high detection score threshold to 0.84 and the low detection score threshold to 0.3.

SSMOT and SSMOTS. We do not rely on the tracking annotations when training our system, thus our method can be applied to SSMOT and SSMOTS task in BBD100K.

4.3. State of the art Comparison on BDD100K

We evaluated the performance of our method on BDD100K MOT & MOTS validation set and test set. We achieve 44.4 mMOTA and 39.6 mMOTSA in BDD100K MOT test set and MOTS validation set which outperform the next place by 10.7% and 33.7%, respectively, as shown in Table 2 and 3. Since we do not use the tracking annotations when training the detector and appearance model, our method can be applied to SSMOT & SSMOTS tasks and achieve the same results as shown in Table 2 and Table 3.

4.4. State of the art Comparison on Waymo 2D tracking

We summit our result to Waymo Open dataset 2D Tracking testset benchmark. We achieve 52.37 MOTA/L1 and 46.57 MOTA/L2 which outperform the second place by 1.1%, as shown in Table 4.

Table 1. Results on BDD100K MOT validation set

Method	mMOTA↑	mIDF1↑
Yu et al. [38]	25.9	44.5
TETer [39]	39.1	53.3
QDTrack [12]	36.6	50.8
Unicorn [1]	41.2	54.0
MOTR [17]	32.3	44.8
MOTRv2 [16]	43.6	56.5
ByteTrack [15]	45.5	54.8
Our	**45.9**	**60.5**

Table 2. Results on BDD100K MOT test set

Method	mMOTA↑	mIDF1↑
Yu et al. [38]	26.3	44.7
DeepBlueAI	31.6	38.7
QDTrack [12]	35.5	52.3
Unicorn [1]	39.5	55.4
ByteTrack [15]	40.1	55.8
Our	**44.4**	**57.4**

Table 3. Results on BDD100K MOTS val set

Method	mMOTSA↑	mIDF1↑
SortIoU	10.3	21.8
MaskTrackRCNN [40]	12.3	26.2
STEm-Seg [25]	12.2	25.4
QDTrack-mots [12]	22.5	40.8
QDTrack-mots-fix [12]	23.5	44.5
PCAN [23]	27.4	45.1
Unicorn [1]	29.6	44.2
Our	**39.6**	**46.2**

Table 4. Results on Waymo 2D Tracking test set

Method	MOTA/L1↑	MOTA/L2↑
Trackor	34.8	28.29
CascadeRCNN-SORTv2	50.22	44.15
HorizonMOT	51.01	45.13
QDTrack [12]	51.18	45.09
LeapMotor-Track	51.79	46.45
Our	**52.37**	**46.57**

4.5. 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 used the original ByteTrack [15] codebase as our strong baseline. Because the original ByteTrack codebase does not public the ReID model code, So we re-implement the ReID model as Unitrack [20]. The framework contains CBNetv2 [13] with Swin-B backbone [29] and a ReID model from Unitrack [20]. The baseline achieves 48.8 mHOTA and 44.5 mMOTA with a high detection rate. However, there are a lot of id switching and lost tracking especially when occlusion occurs. We added weighted on ReID features to give priority to the high detection score features during the occlusion and got 0.4 higher scores on mHOTA and mMOTA. In addition, we trained the ReID model with Resnet-50 backbone on BDD100K by using momentum contrastive learning method [14] on cropped images by class. The fine-tuned model extracts appearance features that can differentiate each object from the same class and improve data association by 0.8 mHOTA and 0.5 mMOTA. Finally, we fine-tuned matching thresholds in ByteTrack and achieved 50.0 mHOTA and 45.9 mMOTA.

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

Method	mHOTA	mMOTA
ByteTrack(IOU)	-	39.4
ByteTrack(ReID)	-	45.0
YOLOX to CBNetV2	48.8	44.5
+ Temporally Weighted smooth	49.2 (+0.4)	45.3 (+0.4)
+ Contrastive Learning Model	50.0 (+0.8)	45.8 (+0.5)
+ Parameters Fine Tuning	50.0	45.9 (+0.1)

Table 6. Ablation study of on BDD100K MOT test set

Method	mHOTA	mMOTA
ByteTrack(ReID)	-	40.1
Our	49	44.4

5. Conclusions

In this paper, we propose a simple yet effective tracking-by-detection framework and achieve state-of-the-art results in BDD100K MOT and MOTS dataset. We discard the motion information and only use the appearance embedding 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 the validation set and 44.0 mMOTA on the test set. We also achieve first place in ECCV2022 SSLAD all 4 BDD100K challenges of MOT, MOTS, SSMOT, and SSMOTS. We hope the simplicity and effectiveness of our method can benefit future research on MOT and MOTS.

6. Limitations

Although mHOTA is more reasonable than mMOTA for MOT Metrics, historical MOT papers do not publish data about HOTA, so we can't compare more fairly under mHOTA. Another limitation is the computational overhead introduced by the CBnetV2 detector, which may hinder our solution for the deployment on edge devices.

7. ACKNOWLEDGMENTS

This work is supported by the STI 2030-Major Projects under Grant 2021ZD0201404.

References

[1] Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking. *arXiv preprint arXiv:2207.07078*, 2022. 1, 2, 5

[2] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. *arXiv preprint arXiv:2207.10661*, 2022. 1

[3] Chao Liang, Zhipeng Zhang, Xue Zhou, Bing Li, Shuyuan Zhu, and Weiming Hu. Rethinking the competition between detection and reid in multiobject tracking. *IEEE Transactions on Image Processing*, 31:3182–3196, 2022. 1, 2

[4] Zhichao Lu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Retinatrack: Online single stage joint detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14668–14678, 2020. 1, 2

[5] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020. 1, 2

[6] Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. Track to detect and segment: An online multi-object tracker. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12352–12361, 2021. 1, 2

[7] 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. 1

[8] Fangao Zeng, Bin Dong, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. *arXiv preprint arXiv:2105.03247*, 2021. 1, 2

[9] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 1

[10] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 1

[11] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129(11):3069–3087, 2021. 1, 2

[12] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 164–173, 2021. 1, 2, 5

[13] Tingting Liang, Xiaojie Chu, Yudong Liu, Yongtao Wang, Zhi Tang, Wei Chu, Jingdong Chen, and Haibin Ling. Cbnetv2: A composite backbone network architecture for object detection. *arXiv preprint arXiv:2107.00420*, 2021. 2, 3, 4, 5

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2, 3, 4, 5

[15] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. *arXiv preprint arXiv:2110.06864*, 2021. 2, 3, 4, 5

[16] Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. *arXiv preprint arXiv:2211.09791*, 2022. 2, 5

[17] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 659–675. Springer, 2022. 2, 5

[18] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016. 2

[19] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 2

[20] Zhongdao Wang, Hengshuang Zhao, Ya-Li Li, Shengjin Wang, Philip Torr, and Luca Bertinetto. Do different tracking tasks require different appearance models? *Advances in Neural Information Processing Systems*, 34:726–738, 2021. 2, 3, 5

[21] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8844–8854, 2022. 2

[22] Zelin Zhao, Ze Wu, Yueqing Zhuang, Boxun Li, and Jiaya Jia. Tracking objects as pixel-wise distributions. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 76–94. Springer, 2022. 2

[23] Lei Ke, Xia Li, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Prototypical cross-attention networks for multiple object tracking and segmentation. *Advances in Neural Information Processing Systems*, 34:1192–1203, 2021. 2, 5

[24] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 7942–7951, 2019. 2

[25] Ali Athar, Sabarinath Mahadevan, Aljosa Osep, Laura Leal-Taixé, and Bastian Leibe. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 158–177. Springer, 2020. 2, 5

[26] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 3

[27] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2

[28] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 2

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. 3, 4, 5

[30] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017. 3

[31] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4969–4978, 2019. 3

[32] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52, 1955. 3

[33] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2633–2642, 2018. 3

[34] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott M. Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2443–2451, 2019. 4

[35] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 4

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2014. 4

[37] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 4

[38] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 5

[39] Siyuan Li, Martin Danelljan, Henghui Ding, Thomas E Huang, and Fisher Yu. Tracking every thing in the wild. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 498–515. Springer, 2022. 5

[40] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5188–5197, 2019. 5