The Solutions for ECCV 2022 BDD100K MOT and MOTS Challenges

Cheng Gao, Qinzhen Guo, Bin Sun, and Bin Lu

ByteDance Inc.
{gaocheng.kosung, guoqinzhen, sunbin.824, bin.lu}@bytedance.com

Abstract. This report introduces the solutions for ECCV 2022 Workshop SSLAD Track 4 - BDD100K Multiple Object Tracking (MOT) and Multiple Object Tracking and Segmentation (MOTS) challenges. Based on Unicorn, a unified tracking framework, we win the second place of MOTS challenge and the third place of MOT challenge using a single model with the same model parameters.

Keywords: BDD100K, Multiple Object Tracking, Multiple Object Tracking and Segmentation

1 Dataset

BDD100K[9] MOT and MOTS dataset is a challenging large-scale tracking dataset for autonomous driving scenarios. BDD100K MOT set contains 2,000 fully annotated 40-second 5 FPS sequences containing a total of 160K instances and 4M objects, with 1,400/200/400 videos for train/val/test. The MOTS set uses a subset of MOT videos, with 154/32/37 videos for train/val/test, containing 25K instances and 480K object masks. MOT challenge employs mean Higher Order Tracking Accuracy (mHOTA)[5] as the primary evaluation metric for ranking and also uses mean Multiple Object Tracking Accuracy (mMOTA)[1] and mean ID F1 score (mIDF1)[7]. For MOTS, the same metrics set as MOT is used. 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.

2 Method

2.1 Unicorn

We use Unicorn[8], a unified model for tracking tasks, to address both the MOT and MOTS challenges of BDD100K. With a single model with the same model parameters, Unicorn is able to perform the four tracking tasks: Single Object Tracking (SOT), Video Object Segmentation (VOS), MOT and MOTS simultaneously.

As illustrated in Fig. 1, the overall framework of Unicorn consists of three main components: (1) Unified inputs and backbone are responsible for obtaining powerful visual representation. (2) Unified embedding is is employed to establish precise correspondence. (3) Unified Head is to detect different tracked targets.

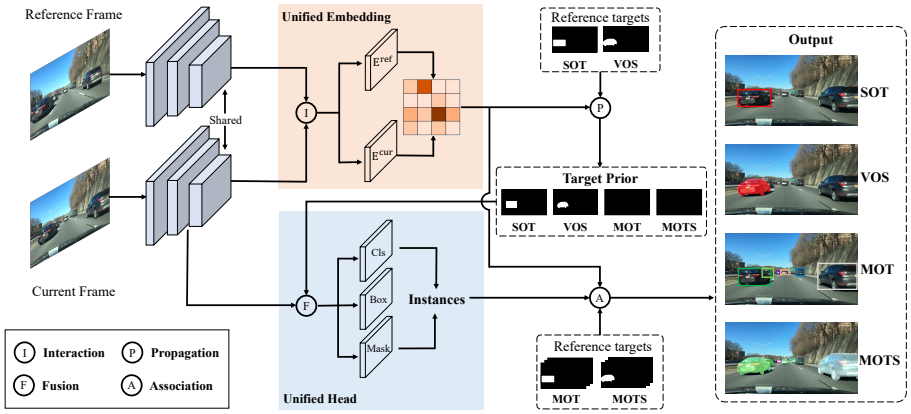

Fig. 1. Unicorn consists of three main components: (1) Unified inputs and backbone (2) Unified embedding (3) Unified head.[8]

2.2 Ablation Studies

As in the official paper[8], we mainly explore the performance of Unicorn with ResNet-50[2] and ConvNeXt-L[4] as the backbone in the BDD100K MOT and MOTS challenges. We train the models on 8 NVIDIA Tesla v100 GPU and all other hyper-parameters in this work follow [8] if not specified.

MOT. The performance of different methods and different backbones on the BDD100K MOT validation set is shown in Table 1, where QDTrack[6] is the baseline provided by the organizer. The mHOTA of Unicorn improves by 0.94% over QDTrack when using ResNet-50 as the backbone. Switching the backbone of Unicorn from ResNet-50 to ConvNeXt-L leads to a significant improvement of 3.43% mHOTA.

Finally, as presented in Table 2, we obtain 44.36% mHOTA on the test set with Unicorn using ConvNeXt-L as the backbone, and win the third place in the BDD100K MOT challenge.

Table 1. Ablation study on the BDD100K MOT validation set.

Method	Backbone	mHOTA	mMOTA	mIDF1	mMOTP	HOTA	MOTA	IDF1
QDTrack	ResNet-50	41.05	36.62	51.62	70.66	60.01	63.93	71.50
Unicorn	ResNet-50	41.99	36.61	50.31	72.90	60.40	64.07	69.30
Unicorn	ConvNeXt-L	**45.42**	**42.10**	**54.06**	**73.71**	**62.62**	**66.77**	**71.52**

MOTS. Table 3 shows the performance of different methods and different backbones in the BDD100K MOTS validation set, where PCAN[3] is the baseline provided by the organizer. It is obvious that Unicorn improves 1.33% over PCAN

Table 2. Leaderboard of BDD100K MOT Challenge.

Rank	Team	mHOTA	mMOTA	mIDF1	mMOTP	HOTA	MOTA	IDF1
1	Lenovo_LR_PCIE	49.15	42.97	59.50	81.35	61.52	68.58	71.28
2	bbq	46.33	38.13	55.21	81.06	62.75	67.42	72.01
3	Our Team	44.36	40.38	52.98	72.89	62.87	67.62	72.24
4	CMSQ	42.38	36.15	53.44	77.02	60.72	65.19	72.98
5	Host Team	41.85	35.67	52.36	77.82	60.51	64.55	72.45
6	OKC	40.69	33.98	50.88	70.08	61.19	65.58	72.92

in terms of mHOTA when ResNet-50 is used as the backbone. Different from MOT, switching the backbone of Unicorn from ResNet-50 to ConvNeXt-L does not improve mHOTA, but on the contrary reduces it by 0.03%.

However, the generalization ability of using ResNet-50 as backbone on the test set is not as strong as ConvNeXt-L. The mHOTA of Unicorn ResNet-50 is only 39.75%, which is 1.95% lower than that of ConvNeXt-L. In the end, our team win the second place with mHOTA of 41.87%.

Table 3. Ablation study on the BDD100K MOTS validation set.

Method	Backbone	mHOTA	mMOTA	mIDF1	mMOTP	HOTA	MOTA	IDF1
PCAN	ResNet-50	35.93	28.11	45.42	66.73	55.86	56.11	65.63
Unicorn	ResNet-50	**37.26**	**30.78**	**47.13**	67.49	56.5	57.83	65.9
Unicorn	ConvNeXt-L	37.23	29.68	44.30	**67.65**	**58.14**	**60.50**	**67.54**

Table 4. Leaderboard of BDD100K MOTS Challenge.

Rank	Team	mHOTA	mMOTA	mIDF1	mMOTP	HOTA	MOTA	IDF1
1	Lenovo_LR_PCIE	44.01	41.09	54.91	69.67	60.55	63.43	70.08
2	Our Team	41.87	34.36	52.94	67.72	59.20	60.64	70.15
3	vdig	41.86	34.33	52.93	67.72	59.10	60.52	69.99
4	OKC	40.00	32.59	50.34	67.41	58.00	59.24	68.56
5	ACT	40.00	32.59	50.34	67.41	58.00	59.24	68.56
6	CMSQ	39.74	30.68	50.26	67.59	56.97	57.73	67.75
7	Host Team	39.17	31.91	50.42	66.53	57.24	56.23	67.96

3 Conclusion

Based on Unicorn, we win the second place of MOTS challenge and the third place of MOT challenge. Furthermore, many thanks to Unicorn for providing excellent work.

References

1. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the clear mot metrics. EURASIP Journal on Image and Video Processing **2008**, 1–10 (2008)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
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. Advances in Neural Information Processing Systems **34**, 1192–1203 (2021)
4. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
5. Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., Leibe, B.: Hota: A higher order metric for evaluating multi-object tracking. International journal of computer vision **129**(2), 548–578 (2021)
6. 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)
7. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: European conference on computer vision. pp. 17–35. Springer (2016)
8. Yan, B., Jiang, Y., Sun, P., Wang, D., Yuan, Z., Luo, P., Lu, H.: Towards grand unification of object tracking. In: ECCV (2022)
9. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2636–2645 (2020)