

FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking

Yifu Zhang* , Chunyu Wang* , Xinggang Wang† , Wenjun Zeng, Wenyu Liu
Huazhong University of Science and Technology
Microsoft Research Asia

arxiv Wed, 9 Sep 2020

Presented by CHEN-NI CHEN
2020/09/21

A Simple Baseline for Multi-Object Tracking.

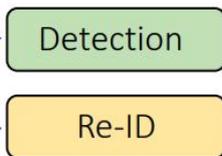
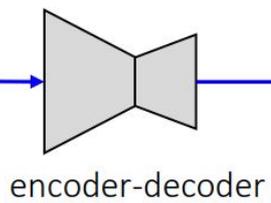
Yifu Zhang* , Chunyu Wang* , Xinggang Wang† , Wenjun Zeng, Wenyu Liu
Huazhong University of Science and Technology
Microsoft Research Asia

arxiv Wed, 9 Sep 2020

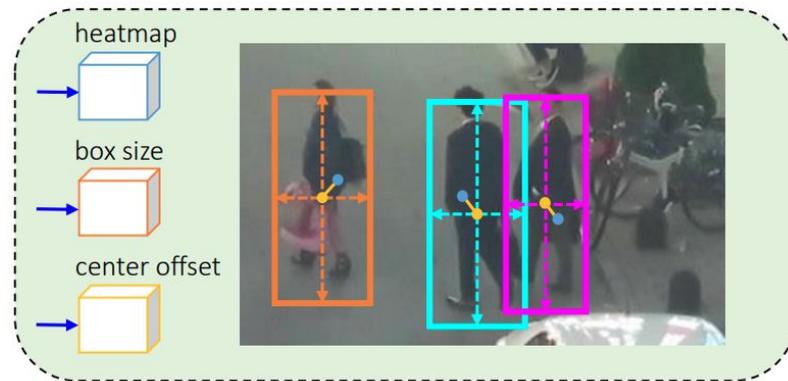
Presented by CHEN-NI CHEN
2020/09/21



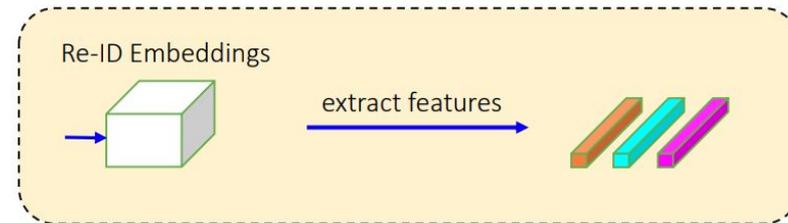
Image



Detection



Re-ID



INTRODUCTION

Multi-Object Tracking (MOT)

The existing methods such as [1], [2], [3], [4], [5], [6], [7] often address the problem by **two separate models**:

1. the detection model firstly localizes the objects of interest by bounding boxes in each frame
2. then the association model extracts re-identification (re-ID) features for each bounding box

[1] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in ICIP. IEEE, 2016, pp. 3464–3468.

[2] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in 2017 IEEE international conference on image processing (ICIP). IEEE, 2017, pp. 3645–3649.

[3] L. Chen, H. Ai, Z. Zhuang, and C. Shang, "Real-time multiple people tracking with deeply learned candidate selection and person re-identification," in 2018 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2018, pp. 1–6.

[4] F. Yu, W. Li, Q. Li, Y. Liu, X. Shi, and J. Yan, "Poi: Multiple object tracking with high performance detection and appearance feature," in ECCV. Springer, 2016, pp. 36–42.

[5] N. Mahmoudi, S. M. Ahadi, and M. Rahmati, "Multi-target tracking using cnn-based features: Cnnmtt," Multimedia Tools and Applications, vol. 78, no. 6, pp. 7077–7096, 2019.

[6] Z. Zhou, J. Xing, M. Zhang, and W. Hu, "Online multi-target tracking with tensor-based high-order graph matching," in 2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018, pp. 1809–1814.

[7] K. Fang, Y. Xiang, X. Li, and S. Savarese, "Recurrent autoregressive networks for online multi-object tracking," in WACV. IEEE, 2018, pp. 466–475.

INTRODUCTION

- With maturity(成熟) of multi-task learning, one-shot trackers which **estimate objects and learn re-ID features using a single network** have attracted more attention [14], [15].
- For example, Voigtlaender et al. [15] propose to **add a re-ID branch on top of Mask R-CNN to obtain proposals' re-ID features using ROI-Align**. It reduces inference time by re-using the backbone features for the re-ID network.
-> Unfortunately, the tracking accuracy drops remarkably compared to the two-step ones.

[14] Z. Wang, L. Zheng, Y. Liu, and S. Wang, "Towards real-time multi- object tracking," arXiv preprint arXiv:1909.12605, 2019.

[15] P.Voigtlaender,M.Krause,A.Osep,J.Luiten,B.B.G.Sekar,A.Geiger, and B. Leibe, "Mots: Multi-object tracking and segmentation," in CVPR, 2019, pp. 7942–7951.

INTRODUCTION

- The result suggests that **combining the two tasks is a non-trivial problem** and should be treated carefully.
- In this paper, we aim to deeply understand the reasons behind the failure, and present a simple yet effective approach. In particular, three factors are identified.

Three fairness issues

- Unfairness Caused by Anchors
- Unfairness Caused by Features
- Unfairness Caused by Feature Dimension

Unfairness Caused by Anchors

- The existing one-shot trackers such as Track R-CNN [15] and JDE [14] are mostly anchor-based since they are **directly modified from anchor-based object detectors** such as YOLO and Mask R-CNN.
- However, we find in this study that the **anchor-based** framework is not suitable for learning re-ID features which **result in a large number of ID switches in spite of the good detection results.**

Unfairness Caused by Anchors

Overlooked(忽視) re-ID task:

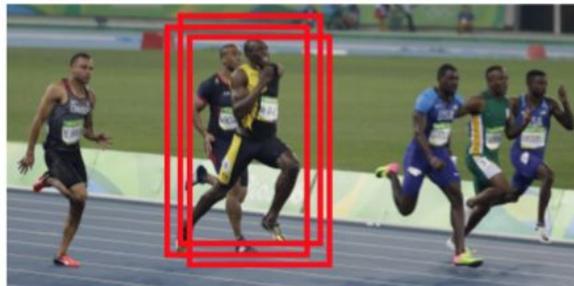
1. estimates object proposals (boxes)
2. pools re-ID features from the proposals to estimate the corresponding re-ID features.
 - -> the quality of re-ID features heavily depends on the quality of proposals.
 - As a result, in the training stage, the model is seriously biased to estimate accurate object proposals rather than high quality re-ID features.
 - **To summarize, this de facto standard “detection first, re-ID secondary” framework makes the re-ID network not fairly learned.**

Unfairness Caused by Anchors

FairMOT extracts re-ID features only at the object center and can mitigate the problems in (b) and (c).



(b) One anchor contains multiple identities



(c) Multiple anchors response for one identity



(d) One point for one identity

Unfairness Caused by Features

- For one-shot trackers, most features are shared between the object detection and re-ID tasks.
- But it is well known that **they actually require features from different layers** to achieve the best results.
- **object detection requires deep and abstract features** to estimate object classes and positions
but **re-ID focuses more on low-level** appearance features to distinguish different instances of the same class.

Unfairness Caused by Features

- We empirically find that **multi-layer feature** aggregation is effective to address the contradiction by **allowing the two tasks (network branches) to extract whatever features they need** from the multi-layer aggregated features.

Unfairness Caused by Feature Dimension

- The previous re-ID works usually learn very high dimensional features and have achieved promising results on the benchmarks of their field.
- However, we find that learning lower-dimensional features is actually better for one-shot MOT for three reasons:

Unfairness Caused by Feature Dimension

However, we find that learning lower-dimensional features is actually better for one-shot MOT for three reasons:

(1) high dimensional \rightarrow re-ID \uparrow , object detection accuracy \downarrow
object detection accuracy \downarrow \rightarrow tracking accuracy \downarrow

- \rightarrow So considering that the feature dimension in object detection is usually very low (class numbers + box locations), we propose to learn low-dimensional re-ID features to balance the two tasks;

Unfairness Caused by Feature Dimension

However, we find that learning lower-dimensional features is actually better for one-shot MOT for three reasons:

(2) when training data is small, learning low dimensional re-ID features reduces the risk(風險) of over-fitting.

(3) learning low dimensional re-ID features improves the inference speed as will be shown in our experiments.

Backbone Network

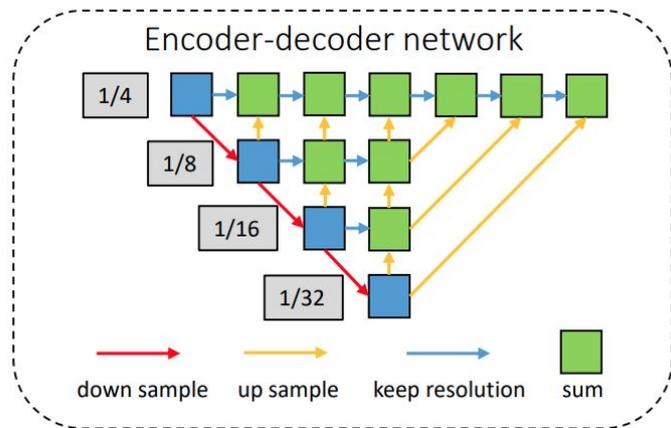
- We adopt ResNet-34 as backbone in order to strike a good balance between accuracy and speed. An enhanced version of Deep Layer Aggregation (DLA) [10] is applied to the backbone to fuse multi-layer features.

Backbone Network

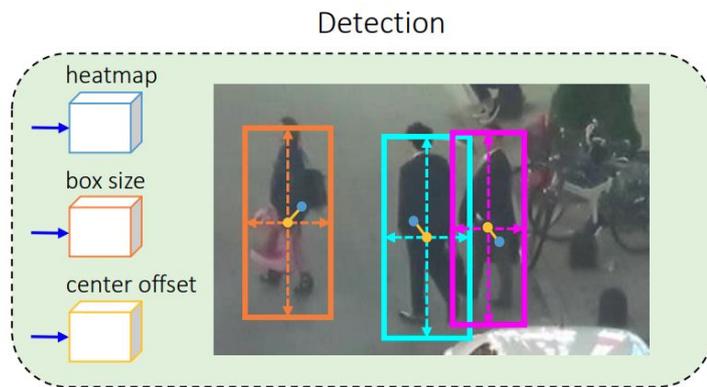
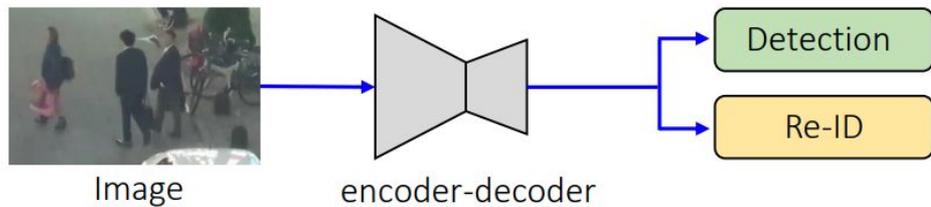
- Different from original DLA [20], it has more skip connections between low-level and high-level features which is similar to the Feature Pyramid Network (FPN) [45].
- In addition, convolution layers in all up-sampling modules are replaced by deformable convolution such that they can dynamically adjust the receptive field according to object scales and poses.

deformable convolution:

有點像attention在上一層conv layer 的 output, 只是在2d的區域上.



Detection Branch



Built on top of CenterNet [10]

- anchor-free
- 對物體的中心點位置進行預測，同時預測物體的大小



keypoint heatmap [C]



local offset [2]



object size [2]

Detection Branch: Heatmap Head

GT box $\mathbf{b}^i = (x_1^i, y_1^i, x_2^i, y_2^i)$

object center (c_x^i, c_y^i) as $c_x^i = \frac{x_1^i + x_2^i}{2}$

高斯: $M_{xy} = \sum_{i=1}^N \exp \frac{-(x - \tilde{c}_x^i)^2 + (y - \tilde{c}_y^i)^2}{2\sigma_c^2}$



$$CE(p, y) = \begin{cases} -\log(p), & \text{if } y = 1 \\ -\log(1 - p), & \text{O.W.} \end{cases}$$

參考Focal Loss:

$$L_{\text{heat}} = -\frac{1}{N} \sum_{xy} \begin{cases} (1 - \hat{M}_{xy})^\alpha \log(\hat{M}_{xy}), & M_{xy} = 1; \\ (1 - M_{xy})^\beta (\hat{M}_{xy})^\alpha \log(1 - \hat{M}_{xy}) & \text{otherwise,} \end{cases} \quad (1)$$

Detection Branch: Box Offset and Size Head

Ground Truth Box: $\mathbf{b}^i = (x_1^i, y_1^i, x_2^i, y_2^i)$

Size: $\mathbf{s}^i = (x_2^i - x_1^i, y_2^i - y_1^i)$

Offset: $\mathbf{o}^i = \left(\frac{c_x^i}{4}, \frac{c_y^i}{4}\right) - \left(\lfloor \frac{c_x^i}{4} \rfloor, \lfloor \frac{c_y^i}{4} \rfloor\right)$

L1 losses :

$$L_{\text{box}} = \sum_{i=1}^N \|\mathbf{o}^i - \hat{\mathbf{o}}^i\|_1 + \|\mathbf{s}^i - \hat{\mathbf{s}}^i\|_1. \quad (2)$$

Re-ID Branch

- Re-ID branch aims to generate features that can distinguish objects.
- Ideally, affinity among different objects should be smaller than that between same objects.
- To achieve this goal, we apply a convolution layer with 128 kernels on top of backbone features to extract re-ID features for each location.
- Denote the resulting feature map as $E \in \mathbb{R}^{128 \times W \times H}$.
The re-ID feature $E_{x,y} \in \mathbb{R}^{128}$ of an object centered at (x, y) can be extracted from the feature map.

Re-ID Loss

- We learn re-ID features through a classification task.
- GT box $\mathbf{b}^i = (x_1^i, y_1^i, x_2^i, y_2^i)$
- object center on the heatmap $(\tilde{c}_x^i, \tilde{c}_y^i)$
- learn to map it to a class distribution vector $\mathbf{P} = \{\mathbf{p}(k), k \in [1, K]\}$
K is the number of classes
- one-hot representation of the GT class label as $\mathbf{L}^i(k)$

$$L_{\text{identity}} = - \sum_{i=1}^N \sum_{k=1}^K \mathbf{L}^i(k) \log(\mathbf{p}(k)), \quad (3)$$

Training FairMOT

- we use the **uncertainty loss** proposed in [50] to automatically balance the detection and re-ID tasks:
- w_1 and w_2 are learnable parameters that balance the two tasks.

$$L_{\text{detection}} = L_{\text{heat}} + L_{\text{box}}, \quad (4)$$

$$L_{\text{total}} = \frac{1}{2} \left(\frac{1}{e^{w_1}} L_{\text{detection}} + \frac{1}{e^{w_2}} L_{\text{identity}} + w_1 + w_2 \right), \quad (5)$$

[51] N. Zhao, Z. Wu, R. W. Lau, and S. Lin, “Distilling localization for self-supervised representation learning,” arXiv preprint arXiv:2004.06638, 2020.

[52] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, “Crowdhuman: A benchmark for detecting human in a crowd,” arXiv preprint arXiv:1805.00123, 2018.

weakly supervised learning (self supervised learning ?)

- In addition to the standard training strategy presented above, we propose a weakly supervised learning method on image-level object detection datasets such as COCO.
- Inspired by [51], **we regard each object instance in the dataset as a separate class and different transformations of the same object as instances in the same class.** The adopted transformations include HSV augmentation, rotation, scaling, translation and shearing. We pre-train our model on the CrowdHuman dataset [52] and then finetune it on the MOT datasets. With this self supervised learning approach, we further improve the final performance.

EXPERIMENTS

Datasets

- [14] Z. Wang, L. Zheng, Y. Liu, and S. Wang, "Towards real-time multi- object tracking," arXiv preprint arXiv:1909.12605, 2019.
- [22] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "Mot16: A benchmark for multi-object tracking," arXiv preprint arXiv:1603.00831, 2016.
- [52] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, "Crowdhuman: A benchmark for detecting human in a crowd," arXiv preprint arXiv:1805.00123, 2018.
- [55] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, "A mobile vision system for robust multi-person tracking," in CVPR. IEEE, 2008, pp. 1–8.
- [56] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," in CVPR, 2017, pp. 3213–3221.
- [57] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in CVPR. IEEE, 2009, pp. 304–311.
- [58] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in CVPR, 2017, pp. 3415–3424.

There are six training datasets

- The ETH [55] and CityPerson [56] datasets **only provide box** annotations so we only train the detection branch on them.
- The CalTech [57], MOT17 [22], CUHK-SYSU [58] and PRW [12] datasets **provide both box and identity annotations** which allows us to train both branches.
- The overall training strategy is described in Section 3.4, which is the same as [14].
- For the **self-supervised training** of our method, we use the CrowdHuman dataset [52] which only contains object bounding box annotations.

Metrics

- Average Precision (AP) for evaluating detection performance
Precision = $TP / (TP + FP)$
- True Positive Rate (TPR) at a false accept rate of 0.1 for rigorously(嚴格) evaluating re-ID features with ground-truth detections.
- We use the CLEAR metric [59] and IDF1 [60] to evaluate overall tracking accuracy.

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IDSW_t)}{\sum_t GT_t}$$

[59] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," EURASIP Journal on Image and Video Processing, vol. 2008, pp. 1–10, 2008.

[60] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in

Ablative Studies

Fairness Issue in Anchors

Feature Extraction	Anchor	MOTA↑	IDF1↑	IDs↓	TPR↑
FairMOT (ROI-Align)	✓	68.7	71.0	331	93.1
FairMOT (POS-Anchor)	✓	69.0	70.3	434	93.9
FairMOT (Center)		69.1	72.8	299	94.4
FairMOT (Center-BI)		68.8	74.3	303	94.9
FairMOT (Two-Stage)	✓	69.0	68.2	388	90.5

- Comparison of different re-ID feature extraction (sampling) strategies on the validation set of MOT17.
- The only difference lies in how they sample re-ID features from detected boxes.

Fairness Issue in Anchors

Feature Extraction	Anchor	MOTA↑	IDF1↑	IDs↓	TPR↑
FairMOT (ROI-Align)	✓	68.7	71.0	331	93.1
FairMOT (POS-Anchor)	✓	69.0	70.3	434	93.9
FairMOT (Center)		69.1	72.8	299	94.4
FairMOT (Center-BI)		68.8	74.3	303	94.9
FairMOT (Two-Stage)	✓	69.0	68.2	388	90.5

- First, we can see that our approach (**Center**) obtains notably higher IDF1 score and True Positive Rate (TPR) than ROI-Align, POS-Anchor and the two-stage approach. This metric is **independent of object detection results and faithfully reflects the quality of re-ID features**.
- In addition, the number of ID switches (IDs) of our approach is also significantly smaller than the two baselines. The results validate that sampling features at object centers is more effective than the strategies used in the previous works.

Fairness Issue in Anchors

Feature Extraction	Anchor	MOTA↑	IDF1↑	IDs↓	TPR↑
FairMOT (ROI-Align)	✓	68.7	71.0	331	93.1
FairMOT (POS-Anchor)	✓	69.0	70.3	434	93.9
FairMOT (Center)		69.1	72.8	299	94.4
FairMOT (Center-BI)		68.8	74.3	303	94.9
FairMOT (Two-Stage)	✓	69.0	68.2	388	90.5

- Bi-linear Interpolation (Center-BI) achieves even higher TPR than Center because it samples features at more accurate locations.
- The two-stage approach harms(傷害) the quality of the re-ID features.

Fairness Issue in Feature

Backbone	MOTA↑	IDF1↑	IDs↓	AP↑	TPR↑
ResNet-34	63.6	67.2	435	75.1	90.9
ResNet-50	63.7	67.7	501	75.5	91.9
ResNet-34-FPN	64.4	69.6	369	77.7	94.2
HRNet-W18	67.4	74.3	315	80.5	94.6
DLA-34	69.1	72.8	299	81.2	94.4

Comparison of different backbones on the validation set of MOT17 dataset. The best results are shown in bold.

Fairness Issue in Feature

Backbone	MOTA \uparrow	IDF1 \uparrow	IDs \downarrow	AP \uparrow	TPR \uparrow
ResNet-34	63.6	67.2	435	75.1	90.9
ResNet-50	63.7	67.7	501	75.5	91.9
ResNet-34-FPN	64.4	69.6	369	77.7	94.2
HRNet-W18	67.4	74.3	315	80.5	94.6
DLA-34	69.1	72.8	299	81.2	94.4

- By comparing the results of ResNet-34 and ResNet-50, we surprisingly find that **using a larger network only slightly improves the overall tracking result measured by MOTA.**
- In particular, the quality of re-ID features barely benefits from the larger network. For example, IDF1 only improves from 67.2% to 67.7% and TPR improves from 90.9% to 91.9%, respectively. In addition, the number of ID switches even increases from 435 to 501. All these results suggest that using a larger network adds very limited values to the final tracking accuracy

Fairness Issue in Feature

Backbone	MOTA \uparrow	IDF1 \uparrow	IDs \downarrow	AP \uparrow	TPR \uparrow
ResNet-34	63.6	67.2	435	75.1	90.9
ResNet-50	63.7	67.7	501	75.5	91.9
ResNet-34-FPN	64.4	69.6	369	77.7	94.2
HRNet-W18	67.4	74.3	315	80.5	94.6
DLA-34	69.1	72.8	299	81.2	94.4

- By comparing the results of ResNet-34 and ResNet-50, we surprisingly find that **using a larger network only slightly improves the overall tracking result measured by MOTA.**
- In particular, the quality of re-ID features barely benefits from the larger network. For example, IDF1 only improves from 67.2% to 67.7% and TPR improves from 90.9% to 91.9%, respectively. In addition, the number of ID switches even increases from 435 to 501. All these results suggest that using a larger network adds very limited values to the final tracking accuracy

Fairness Issue in Feature

Backbone	MOTA↑	IDF1↑	IDs↓	AP↑	TPR↑
ResNet-34	63.6	67.2	435	75.1	90.9
ResNet-50	63.7	67.7	501	75.5	91.9
ResNet-34-FPN	64.4	69.6	369	77.7	94.2
HRNet-W18	67.4	74.3	315	80.5	94.6
DLA-34	69.1	72.8	299	81.2	94.4

- By comparing the results of ResNet-34 and ResNet-50, we surprisingly find that using a larger network only slightly improves the overall tracking result measured by MOTA.
- In particular, the quality of re-ID features barely benefits from the larger network. For example, IDF1 only improves from 67.2% to 67.7% and TPR improves from 90.9% to 91.9%, respectively.
- In addition, the number of ID switches even increases from 435 to 501. All these results **suggest that using a larger network adds very limited values to the final tracking accuracy.**

Fairness Issue in Feature

Backbone	MOTA↑	IDF1↑	IDs↓	AP↑	TPR↑
ResNet-34	63.6	67.2	435	75.1	90.9
ResNet-50	63.7	67.7	501	75.5	91.9
ResNet-34-FPN	64.4	69.6	369	77.7	94.2
HRNet-W18	67.4	74.3	315	80.5	94.6
DLA-34	69.1	72.8	299	81.2	94.4

- In contrast, ResNet-34-FPN, which actually has fewer parameters than ResNet-50, achieves a larger MOTA score than ResNet-50. More importantly, TPR improves significantly from 90.9% to 94.2% which **suggests that multi-layer feature fusion has clear advantages over simply using larger networks.**

Fairness Issue in Feature

Backbone	MOTA↑	IDF1↑	IDs↓	AP↑	TPR↑
ResNet-34	63.6	67.2	435	75.1	90.9
ResNet-50	63.7	67.7	501	75.5	91.9
ResNet-34-FPN	64.4	69.6	369	77.7	94.2
HRNet-W18	67.4	74.3	315	80.5	94.6
DLA-34	69.1	72.8	299	81.2	94.4

- In addition, **DLA-34**, which is also built on top of ResNet- 34 but has **more levels of feature fusion**, achieves an **even larger MOTA score**.
- In particular, **TPR increases significantly from 90.9% to 94.4%** which in turn **decreases the number of ID switches (IDs) from 435 to 299**.

Fairness Issue in Feature

Backbone	MOTA↑	IDF1↑	IDs↓	AP↑	TPR↑
ResNet-34	63.6	67.2	435	75.1	90.9
ResNet-50	63.7	67.7	501	75.5	91.9
ResNet-34-FPN	64.4	69.6	369	77.7	94.2
HRNet-W18	67.4	74.3	315	80.5	94.6
DLA-34	69.1	72.8	299	81.2	94.4

- The results validate that **feature fusion (both FPN and DLA) effectively improves the discriminative ability of re-ID features.**

Fairness Issue in Feature

AP -> detection performance

True Positive Rate (TPR) -> re-ID features with ground-truth detections.

CLEAR and IDF1-> overall tracking accuracy.

Backbone	MOTA↑	IDF1↑	IDs↓	AP↑	TPR↑
ResNet-34	63.6	67.2	435	75.1	90.9
ResNet-50	63.7	67.7	501	75.5	91.9
ResNet-34-FPN	64.4	69.6	369	77.7	94.2
HRNet-W18	67.4	74.3	315	80.5	94.6
DLA-34	69.1	72.8	299	81.2	94.4

- Although ResNet-34-FPN obtains equally good re-ID features (TPR) as DLA-34, its detection results (AP) are significantly worse than DLA-34.
- We think the use of **deformable convolution** in DLA-34 is the main reason because it enables more flexible receptive fields for objects of different sizes - it is very important for our method since FairMOT only extracts features from object centers without using any region features.
- We can **only** get 65.0 MOTA and 78.1 AP when **replacing all the deformable convolutions** with normal convolutions in DLA-34.

Fairness Issue in Feature

AP -> detection performance
True Positive Rate (TPR) -> re-ID features with ground-truth detections.
CLEAR and IDF1-> overall tracking accuracy.

Backbone	MOTA↑	IDF1↑	IDs↓	AP↑	TPR↑
ResNet-34	63.6	67.2	435	75.1	90.9
ResNet-50	63.7	67.7	501	75.5	91.9
ResNet-34-FPN	64.4	69.6	369	77.7	94.2
HRNet-W18	67.4	74.3	315	80.5	94.6
DLA-34	69.1	72.8	299	81.2	94.4

Backbone	AP ^S	AP ^M	AP ^L	TPR ^S	TPR ^M	TPR ^L	IDs ^S	IDs ^M	IDs ^L
ResNet-34	40.6	57.8	85.2	91.7	85.7	88.8	190	87	118
ResNet-50	39.7	59.4	86.0	91.3	85.3	89.0	248	91	124
ResNet-34-FPN	45.9	61.0	85.4	90.7	91.5	93.3	166	71	90
HRNet-W18	51.1	63.7	85.7	94.2	92.5	93.1	168	55	56
DLA-34	46.8	65.1	88.8	92.7	91.2	91.8	134	64	70

- DLA-34 mainly outperforms HRNet-W18 on middle and large size objects.

但 TRP: $(92.7 a + 91.2 b + 91.8 c)/d = 94.4$???
數字好像怪怪的

Fairness Issue in Feature Dimensionality

Fairness Issue in Feature Dimensionality

Backbone	dim	MOTA \uparrow	IDF1 \uparrow	IDs \downarrow	FPS \uparrow	AP \uparrow	TPR \uparrow
DLA-34	512	68.5	73.7	312	24.1	80.9	94.6
DLA-34	256	68.5	72.5	337	26.1	81.1	94.3
DLA-34	128	69.1	72.8	299	26.6	81.2	94.4
DLA-34	64	69.2	73.3	283	26.8	81.3	94.3

- 512 achieves the highest IDF1 and TPR scores which indicates that higher dimensional re-ID features lead to stronger discriminative ability.
- In our experiments, we set the feature dimension to be 64 which strikes a good balance between the two tasks.

Data Association Methods

Box IoU	Re-ID Features	Kalman Filter	MOTA ↑	IDF1 ↑	IDs ↓
✓			67.8	67.2	648
	✓		68.1	70.3	435
	✓	✓	68.9	71.8	342
✓	✓	✓	69.1	72.8	299

Visualization of Re-ID Similarity



Query Image



Target Image



ResNet-34-det + Center



ResNet-34 + Center



DLA-34 + Center



DLA-34 + ROI-Align



DLA-34 + POS-Anchor



DLA-34 + Center



DLA-34 + Center-BI



DLA-34 + Two-Stage

- By comparing the similarity maps of ResNet-34 and ResNet-34-det, we can see that training the re-ID branch is important.

Visualization of Re-ID Similarity



Query Image



Target Image



ResNet-34-det + Center



ResNet-34 + Center



DLA-34 + Center



DLA-34 + ROI-Align



DLA-34 + POS-Anchor



DLA-34 + Center



DLA-34 + Center-BI



DLA-34 + Two-Stage

- By comparing DLA-34 and ResNet-34, we can see that multi-layer feature aggregation can get more discriminative re-ID features.

Visualization of Re-ID Similarity



Query Image



Target Image



ResNet-34-det + Center



ResNet-34 + Center



DLA-34 + Center



DLA-34 + ROI-Align



DLA-34 + POS-Anchor



DLA-34 + Center



DLA-34 + Center-BI



DLA-34 + Two-Stage

- Among all the sampling strategies, the proposed Center and Center-BI can better discriminate the target object from surrounding objects in crowded scenes (擁擠的場景).



Query Image

Target Image

ResNet-34-det + Center

ResNet-34 + Center

DLA-34 + Center



DLA-34 + ROI-Align

DLA-34 + POS-Anchor

DLA-34 + Center

DLA-34 + Center-BI

DLA-34 + Two-Stage



Query Image

Target Image

ResNet-34-det + Center

ResNet-34 + Center

DLA-34 + Center



DLA-34 + ROI-Align

DLA-34 + POS-Anchor

DLA-34 + Center

DLA-34 + Center-BI

DLA-34 + Two-Stage

Self-supervised Learning

Training Data	MOTA \uparrow	IDF1 \uparrow	IDs \downarrow	AP \uparrow	TPR \uparrow
MOT17	67.5	69.9	408	79.6	93.4
CH*+MOT17	71.1	75.6	327	83.0	93.6
MIX+MOT17	69.1	72.8	299	81.2	94.4

- “CH” and “MIX” stand for CrowdHuman and the composed five datasets introduced in Section 4.1.
- * means no identity annotations are used.
- We first pre-train FairMOT on the CrowdHuman dataset. In particular, we assign a unique identity label for each bounding box and train FairMOT using the method described in section 3.4 (一般的supervised learning). Then we finetune the pre-trained model on the target dataset MOT17.

Self-supervised Learning

Training Data	MOTA \uparrow	IDF1 \uparrow	IDs \downarrow	AP \uparrow	TPR \uparrow
MOT17	67.5	69.9	408	79.6	93.4
CH*+MOT17	71.1	75.6	327	83.0	93.6
MIX+MOT17	69.1	72.8	299	81.2	94.4

- First, pre-training via self-supervised learning on CrowdHuman outperforms directly training on the MOT17 dataset by a large margin.
- Second, the self-supervised learning model even outperforms the fully-supervised model trained on the “MIX” and MOT17 datasets.
- **The results validate the effectiveness of the proposed self-supervised pre-training, which saves lots of annotation efforts and makes FairMOT more attractive in real applications.**

Comparing with One-Shot SOTA MOT Methods

Training Data	Method	MOTA \uparrow	IDF1 \uparrow	IDs \downarrow	FP \downarrow	FN \downarrow	FPS \uparrow
MIX	JDE [14]	67.5	66.7	218	1881	2083	26.0
	FairMOT(ours)	77.2	79.8	80	757	2094	30.9
MOT17 Seg	Track R-CNN [15]	69.2	49.4	294	1328	2349	2.0
	FairMOT(ours)	70.2	64.0	96	1209	2537	30.9

Comparing with Two-Step SOTA MOT Methods

Dataset	Tracker	MOTA \uparrow	IDF1 \uparrow	MT \uparrow	ML \downarrow	IDs \downarrow	FPS \uparrow
MOT15	MDP_SubCNN [25]	47.5	55.7	30.0%	18.6%	628	<1.7
	CDA_DDAL [64]	51.3	54.1	36.3%	22.2%	544	<1.2
	EAMTT [65]	53.0	54.0	35.9%	19.6%	7538	<4.0
	AP_HWDPL [66]	53.0	52.2	29.1%	20.2%	708	6.7
	RAR15 [7]	56.5	61.3	45.1%	14.6%	428	<3.4
	TubeTK* [44]	58.4	53.1	39.3%	18.0%	854	5.8
	FairMOT (Ours)*	60.6	64.7	47.6%	11.0%	591	30.5
MOT16	EAMTT [65]	52.5	53.3	19.9%	34.9%	910	<5.5
	SORTwHPD16 [1]	59.8	53.8	25.4%	22.7%	1423	<8.6
	DeepSORT_2 [2]	61.4	62.2	32.8%	18.2%	781	<6.4
	RAR16wVGG [7]	63.0	63.8	39.9%	22.1%	482	<1.4
	VMaxx [67]	62.6	49.2	32.7%	21.1%	1389	<3.9
	TubeTK* [44]	64.0	59.4	33.5%	19.4%	1117	1.0
	JDE* [14]	64.4	55.8	35.4%	20.0%	1544	18.5
	TAP [6]	64.8	73.5	38.5%	21.6%	571	<8.0
	CNNMTT [5]	65.2	62.2	32.4%	21.3%	946	<5.3
	POI [4]	66.1	65.1	34.0%	20.8%	805	<5.0
	CTrackerV1* [68]	67.6	57.2	32.9%	23.1%	1897	6.8
	FairMOT (Ours)*	74.9	72.8	44.7%	15.9%	1074	25.9
MOT17	SST [69]	52.4	49.5	21.4%	30.7%	8431	<3.9
	TubeTK* [44]	63.0	58.6	31.2%	19.9%	4137	3.0
	CTrackerV1* [68]	66.6	57.4	32.2%	24.2%	5529	6.8
	CenterTrack* [70]	67.3	59.9	34.9%	24.8%	2898	22.0
	FairMOT (Ours)*	73.7	72.3	43.2%	17.3%	3303	25.9
MOT20	FairMOT (Ours)*	61.8	67.3	68.8%	7.6%	5243	13.2

Training Data Ablation Study

Training Data	Images	Boxes	Identities	MOTA \uparrow	IDF1 \uparrow	IDs \downarrow
MOT17	5K	112K	0.5K	69.8	69.9	3996
MOT17+MIX	54K	270K	8.7K	72.9	73.2	3345
MOT17+MIX+CH	73K	740K	8.7K	73.7	72.3	3303

- We can achieve 69.8 MOTA when only using the MOT17 dataset for training, which already outperforms other methods using more training data.
- When we use the same training data as JDE [14], we can achieve 72.9 MOTA, which remarkably outperforms JDE.
- self supervised learning on the CrowdHuman dataset, the MOTA score improves to 73.7.
- The results suggest that our approach is not data hungry which is a big advantage in practical applications.

Qualitative Results

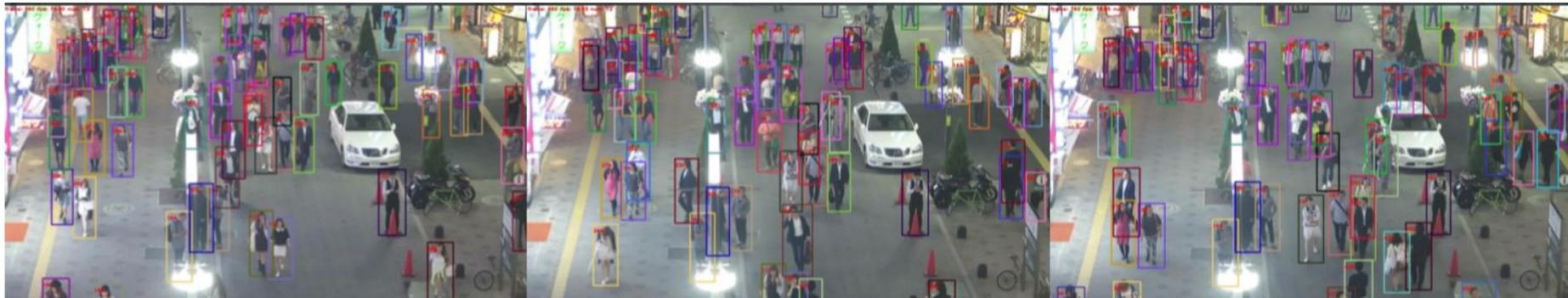
MOT17-01



MOT17-01, we can see that our method can assign correct identities with the help of high-quality re-ID features when two pedestrians cross over each other. Trackers using bounding box IOUs [1], [24] usually cause identity switches under these circumstances.

Qualitative Results

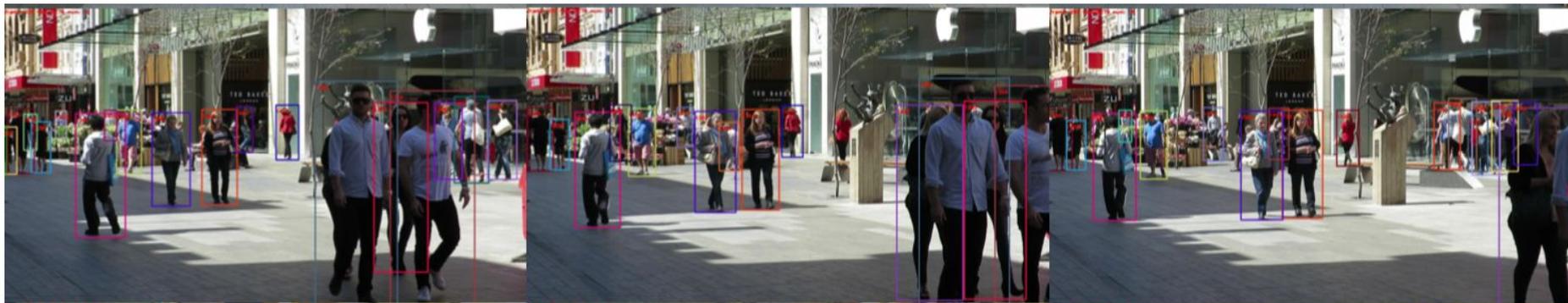
MOT17-03



perform well under crowded scenes

Qualitative Results

MOT17-08



keep both correct identities and correct bounding boxes when the pedestrians are heavily occluded.

Qualitative Results

MOT17-06



MOT17-12



The results of MOT17-06 and MOT17-12 show that our method can deal with large scale variations.

Qualitative Results

MOT17-07



MOT17-14



our method can detect small objects accurately.

CONCLUSION

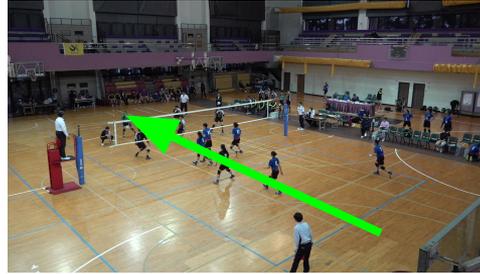
- we find that the **use of anchors** in object detection and identity embedding is the main reason for the degraded results.
- multiple nearby anchors, which correspond to different parts of an object, may be responsible for estimating the same identity which causes ambiguities for network training.
- Further, **we find the feature unfairness issue** and feature conflict issue between the detection and re- ID tasks in previous MOT frameworks.
- By addressing these problems in an anchor-free single-shot deep network, we propose FairMOT. It outperforms the previous **state-of-the-art** methods on several benchmark datasets by a large margin in terms of both tracking accuracy and inference speed.

Progress Report

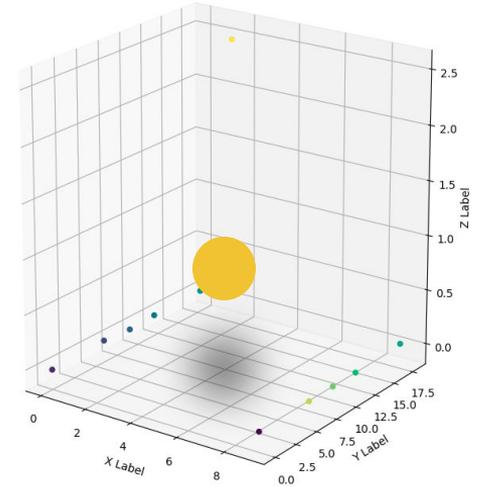
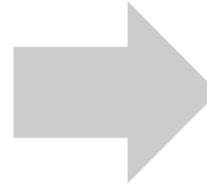
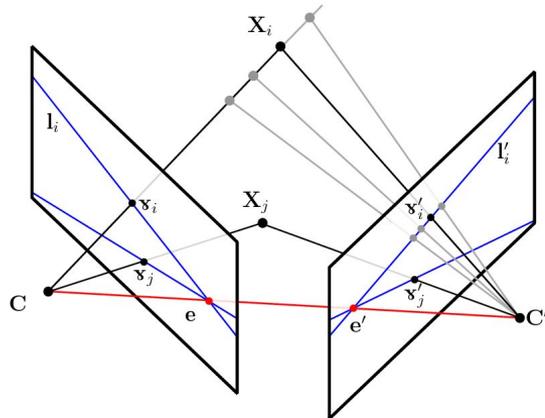
3D position measurements



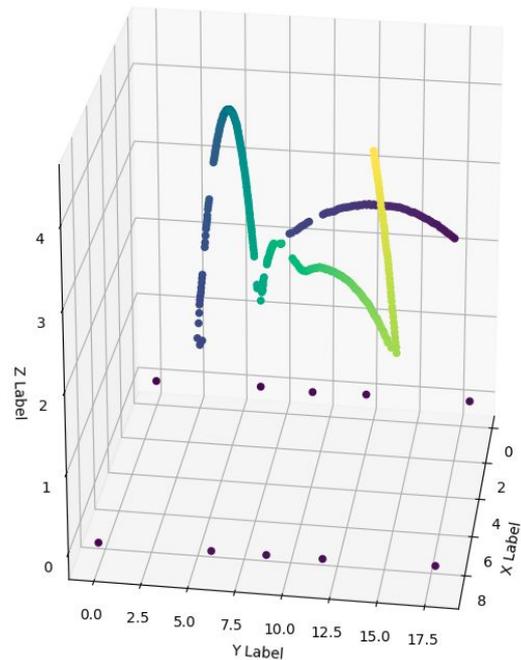
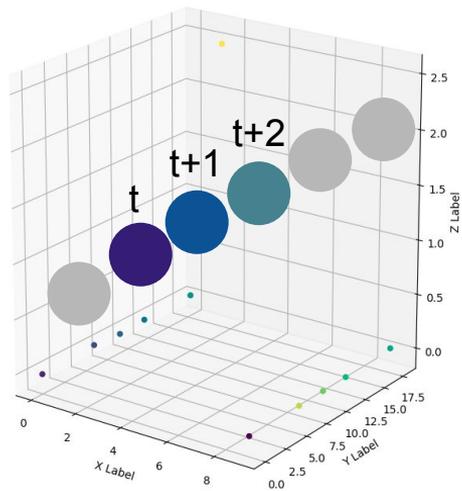
View 1

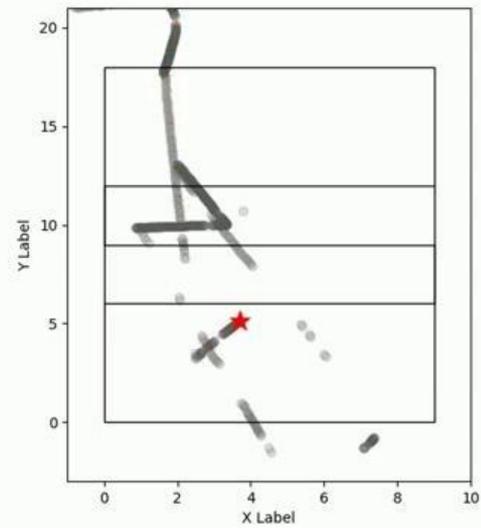
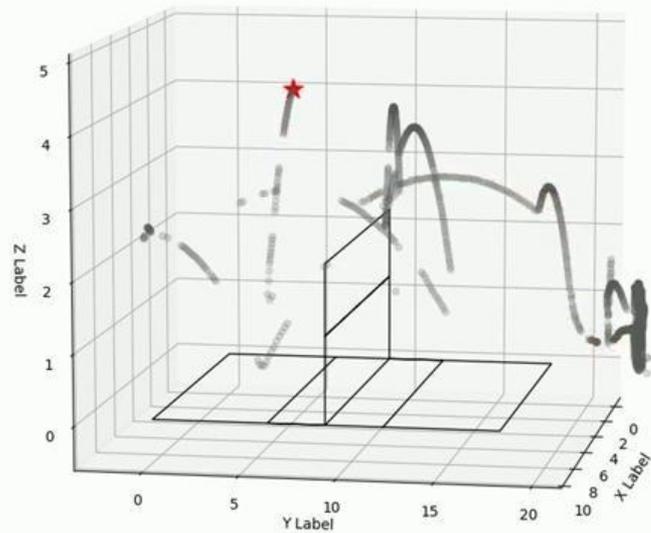


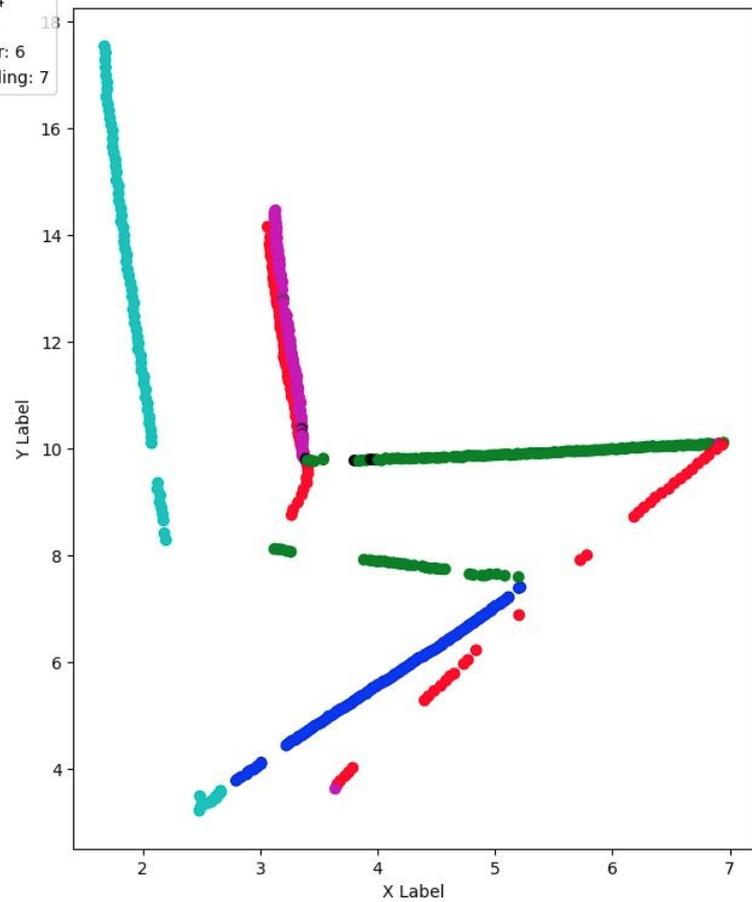
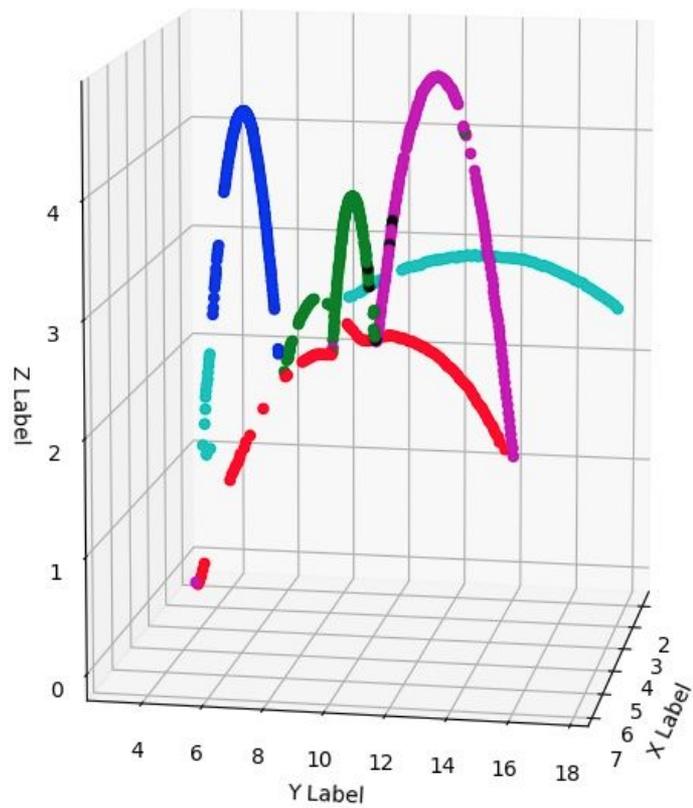
View 2

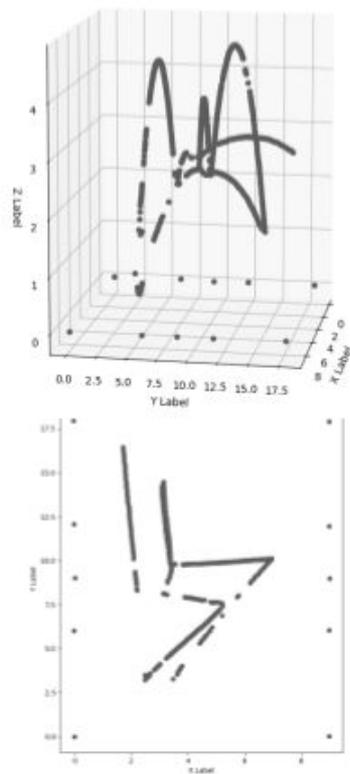


Frames to trajectories

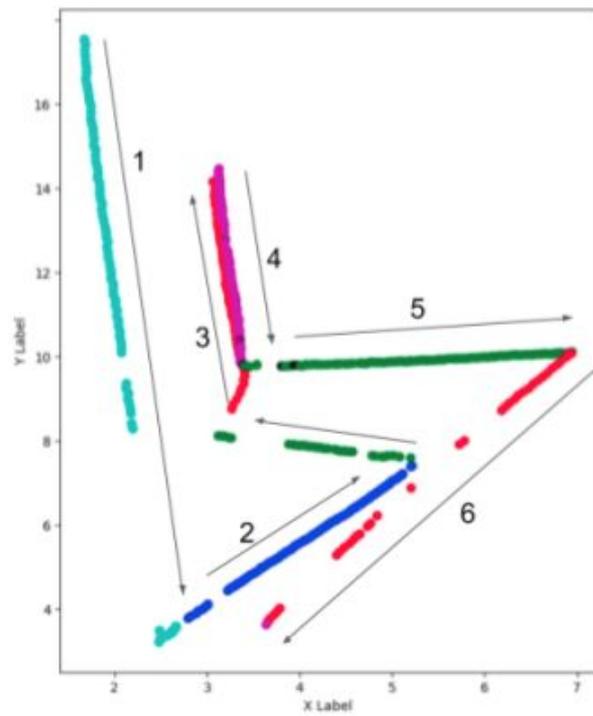




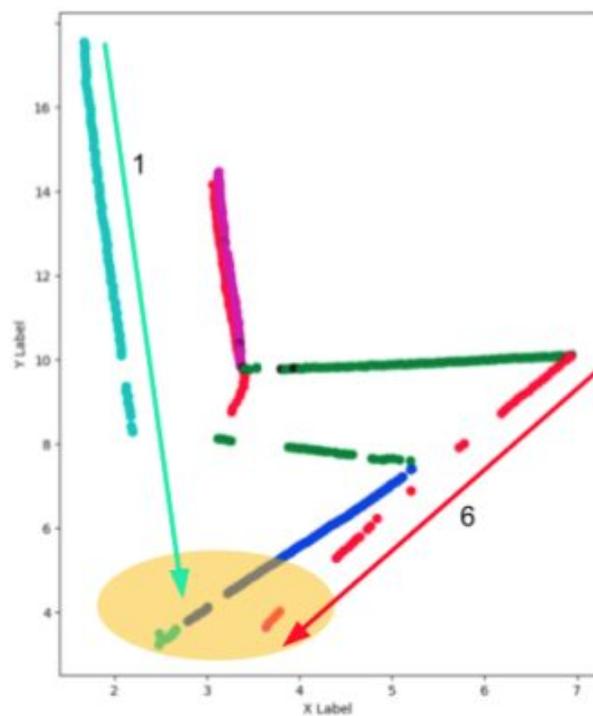




(a)



(b)



(c)

Encoder

- The encoder is composed of a stack of $N = 6$ identical layers.

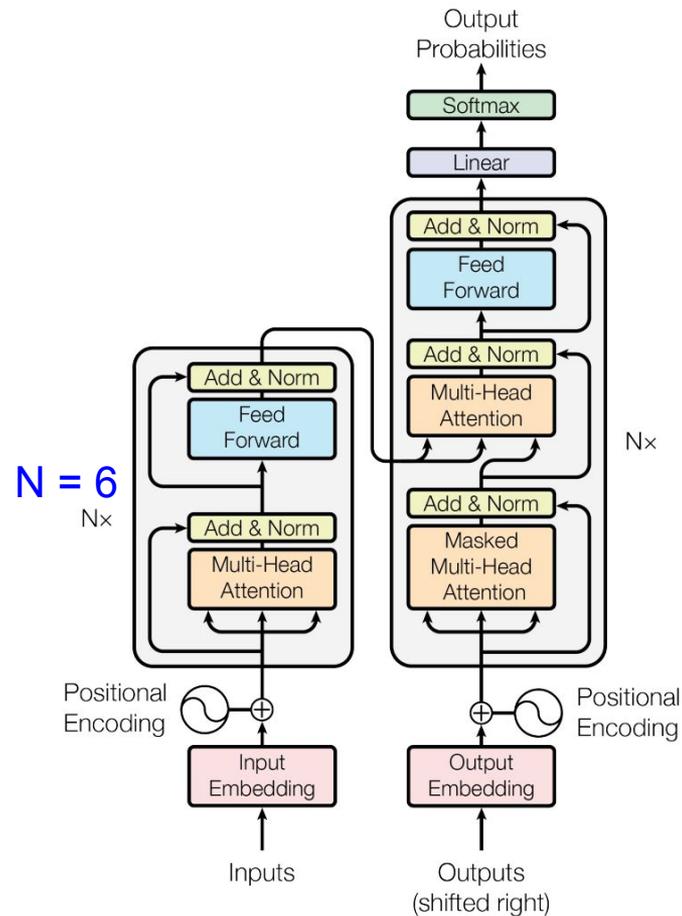


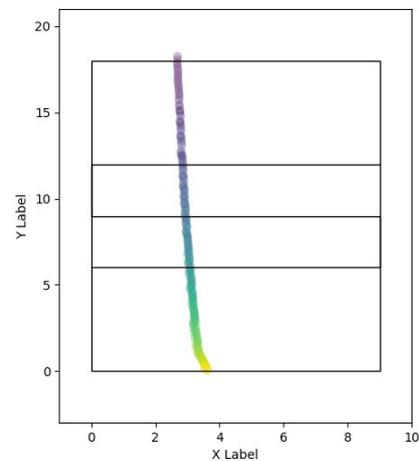
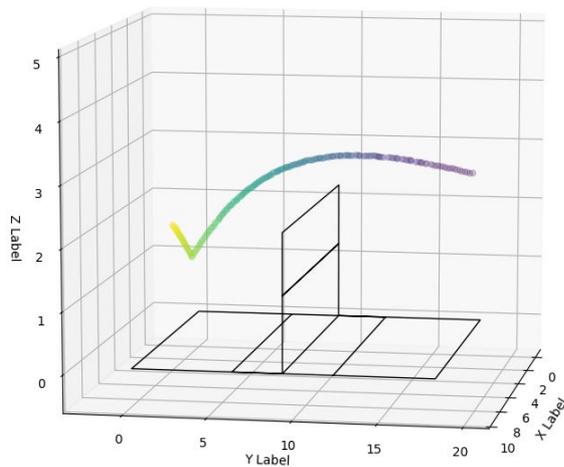
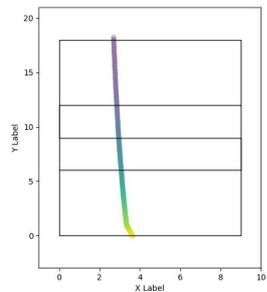
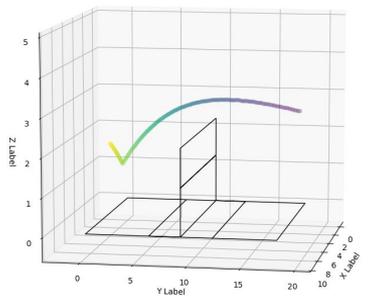
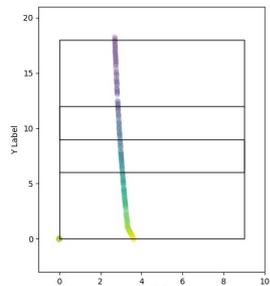
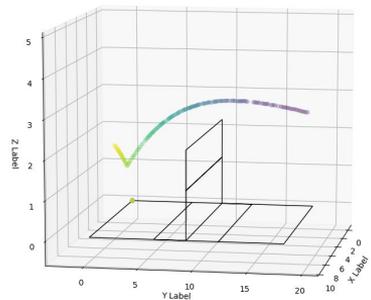
Figure 1: The Transformer - model architecture. 70

BERT

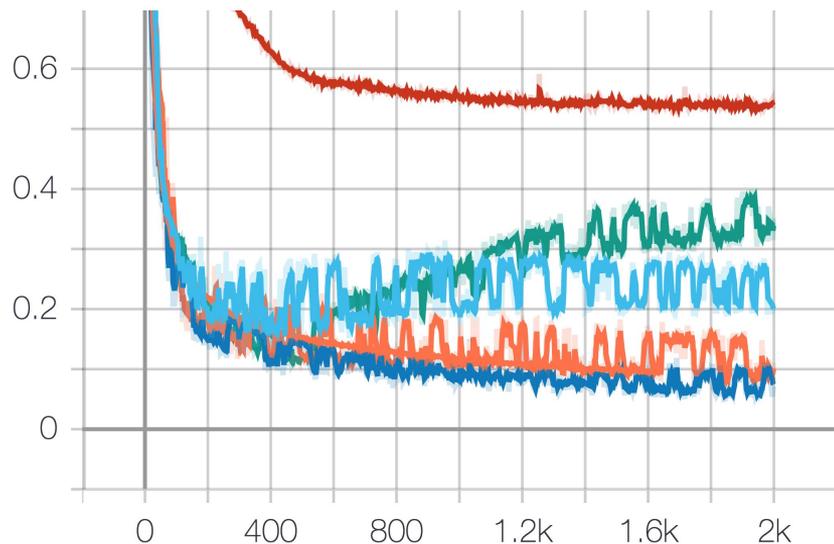
2 training approach:

1. Masked Language Modeling
2. Next Sentence Prediction

Result



Loss Value = RMSE + RMSE_onlyMasked



Name	Smoothed	Value	Step
exp_dim128_vocabSize200/tensorboard_logs/validation	0.2006	0.196	1.999k
exp_dim128_vocabSize20000/tensorboard_logs/validation	0.0744	0.07131	1.999k
exp_dim256_vocabSize20000/tensorboard_logs/validation	0.1007	0.1111	1.999k
exp_dim4_vocabSize4/tensorboard_logs/validation	0.5386	0.5375	1.999k
exp_dim64_vocabSize20000/tensorboard_logs/validation	0.3388	0.3511	1.999k