Supplementary Material
A Gated Attention Transformer for Multi-Person Pose Tracking

Andreas Doering[1,2] Juergen Gall[1,2]
[1]University of Bonn [2]Lamarr Institute for Machine Learning and Artificial Intelligence

A. Implementation Details

Track Embedding and Pose Similarity Embedding Heads: Fig. 1 visualizes the structure of the proposed *Track Embedding Head* (Fig. 1a) and the *Temporal Pose Similarity Embedding Head* (Fig. 1b). The *New Track Embedding Head* (Fig. 2 in the main paper) has the same structure as the *Track Embedding Head*.

Re-Identification Model: We train the re-identification model for 244 epochs on the PoseTrack21 person search dataset with a batch size of 256 and a learning rate of 0.00035 that we decay to 0.000035 after 75 epochs. During the first 10 epochs, we apply a linear learning rate warm-up. Additionally, we apply data augmentation such as random scaling, random rotation and horizontal flipping.

Gated Attention Transformer: Our proposed gated attention transformer employs two encoder and two decoder stages. On the PoseTrack21 dataset, we train our transformer for 14 epochs with a learning rate of 0.0001, which is decayed by a factor of 10 after 13 epochs. We further incorporate a linear learning rate warm-up over the course of 16k iterations and optimize the network with the AdamW [7] optimizer. Each training sequence is split into subsequences of length three with an overlap of one frame. We follow the same settings on the PoseTrack 2018 dataset and train the gated attention transformer for 11 epochs.

Person Detector: For a fair comparison to related works on PoseTrack21, we utilize the same person detector and pose estimation model for all our experiments. In particular, we use the person detector from [3], which consists of a FasterRCNN [10] with a ResNet50-FPN [5] backbone. The detector was first pre-trained on MSCOCO [6] and further fine-tuned on PoseTrack21 for 30 more epochs. As pose estimator, we employ the released model [3], that was originally proposed in [9]. The pose estimation model was trained on MSCOCO and PoseTrack21 for 215 and 16 epochs, respectively. On PoseTrack 2018, we use Cascade R-CNN [1] as object detector and the pose estimator from CorrTrack [9].

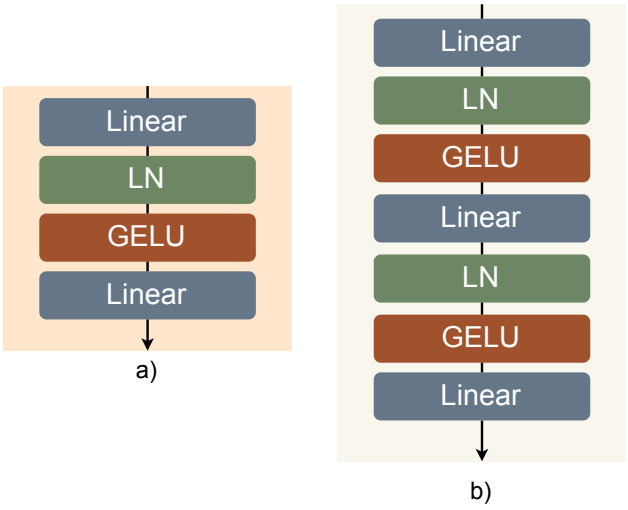


Figure 1. Illustration of the a) track embedding head and b) temporal pose similarity embedding head.

B. Ablation Studies

We perform additional ablation experiments to examine the influence of various building blocks of our proposed method. All experiments are conducted on the PoseTrack21 dataset.

B.1. Evaluation of the Network Architecture

New Track Embedding Head: Given detections in frame t that can not be matched to any existing track, we use the *New Track Embedding Head* (Fig. 2d) in the main paper) to generate new track embeddings from the unmatched detection embeddings. In Table 2, we evaluate the impact of the *New Track Embedding Head* compared to initializing new tracks directly from the detection embeddings. The *New Track Embedding Head* provides a better initialization of the track embeddings.

Model	GT Poses	Kernel Width	Sampling	pre-trained	mAP	loss
ResNet50				✓	64.24	triplet
ResNet50			✓	✓	68.31	triplet
ResNet50			✓		68.65	triplet
ResNet50 & back-of-tricks [8]			✓		73.38	triplet + ce + center
SPAPDE ResNet50	✓	2	✓		66.77	triplet
SPAPDE ResNet50	✓	5	✓		70.39	triplet
SPAPDE ResNet50	✓	10	✓		71.17	triplet
SPAPDE ResNet50	✓	15	✓		69.62	triplet
SPAPDE ResNet50 & bag-of-tricks [8]	✓	10	✓		**78.00**	triplet + ce + center
SPAPDE ResNet50 & bag-of-tricks [8]		10	✓		**74.42**	triplet + ce + center

Table 1. Person re-identification performance (mAP) on the PoseTrack21 dataset for various model settings.

Track Embedding Initialization	AssA	FragA	DetA	HOTA
Detection Embedding	61.87	60.73	47.15	53.76
New Track Embedding Head	62.20	60.93	47.20	53.94

Table 2. Impact of the *New Track Embedding Head* on the overall tracking performance on PoseTrack21.

Temporal Person Similarity	HOTA
IoU	53.59
OKS	53.81
IoU + OKS	53.94

Table 3. Impact of different temporal person similarities used for the calculation of the pose similarity embeddings.

Person Re-Identification We evaluate the person re-identification model on PoseTrack21 and measure the performance in terms of mean average precision (mAP), *i.e.*, we calculate the area under the Precision-Recall curve for each query and average over all queries. As we show in Table 1, training a ResNet50 [4] with proper data sampling results in a significant performance gain and results in a mAP score of 68.65. In more detail, we sample $K = 6$ different instances of the same person identity for every batch. Surprisingly, training the ResNet50 from scratch results in a better mAP performance (68.65) compared to training a ResNet50 pre-trained on ImageNet [2] (68.31). As discussed in Section 3.3 of the main paper, we follow [8], which further increases the performance to 73.78. To incorporate pose information, we replace each batch normalization layer by SPAPDE layers ((9) in the main paper), which incorporates pose information by the keypoint heatmaps. The best performance is achieved by using keypoint heatmaps with a kernel size of 10. Specifically, we denote the kernel size as the standard deviation of a Gaussian distribution. For each keypoint we calculate a Gaussian distribution with the mean set to the respective keypoint location and a standard deviation of 10. Further following [8], we achieve a total performance of 78.0 and 74.42 with ground truth and estimated poses, respectively. By adding SPAPDE, mAP thus increases from 73.78 to 74.42.

Impact of different pose similarity embeddings Table 3 shows the impact of different embedded pose similarities. IoU only provides coarse information and does not allow to distinguish between spatially closely located person instances. Even though a temporal similarity based on IoU achieves a HOTA score of 53.59 and outperforms all related works (Table 1 in the main paper), OKS-based temporal similarity performs better and the combination performs best.

B.2. Runtime Comparison

Finally, we measure the average runtime of our proposed Gated Attention Transformer in comparison to the baseline method CORRTRACK + REID. As both methods are not limited to a specific pose estimation framework, we estimate the tracking runtime independent of the person detector and pose estimation pipelines. While the pose estimation pipeline runs at 2.3 frames per second (fps), the tracking stage of CORRTRACK + REID achieves an average runtime of 6.54 fps. In contrast, our proposed tracker is **3.8** times faster and runs with an average runtime of 25.07 fps.

C. Qualitative Results

Fig. 2 shows some additional qualitative results. Videos are available at https://youtu.be/uFXD4mWPajo and https://youtu.be/lG6GkWSlNQU.

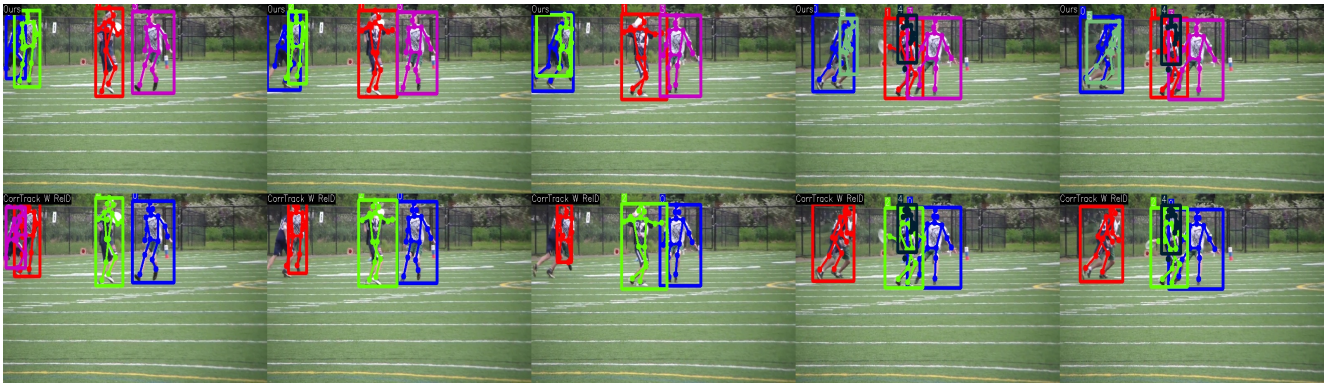
References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: High Quality Object Detection and Instance Segmentation. *arXiv preprint arXiv:1906.09756*, 2019.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[3] Andreas Doering, Di Chen, Shanshan Zhang, Bernt Schiele, and Juergen Gall. PoseTrack21: A Dataset for Person Search, Multi-Object Tracking and Multi-Person Pose Tracking. In *CVPR*, 2022.

(a)



(b)

Figure 2. Additional qualitative examples of our proposed method on the PoseTrack21 dataset. In both (a) and (b), the first row contains visual tracking results of our method and the second row shows visualizations of CorrTrack with ReID [3].

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

[5] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature Pyramid Networks for Object Detection. In *CVPR*, 2017.

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014.

[7] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICCV*, 2019.

[8] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of Tricks and a Strong Baseline for Deep Person Re-Identification. In *CVPRW*, 2019.

[9] Umer Rafi, Andreas Doering, Bastian Leibe, and Juergen Gall. Self-supervised Keypoint Correspondences for Multi-Person Pose Estimation and Tracking in Videos. In *ECCV*, 2020.

[10] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS*, 2015.