Appendix

In this file, we provide the ablation study of diversity loss and the visualization of masks generated by the *Multi-head Part Mask Generator* (MPMG).

I. Analysis of Diversity Loss

Since the MOT Challenge does not provide validation sets, we split the MOT17 dataset into two parts. The first half of each video serves as the training set and the second half as the validation set.

Method	Metrics	
	Rank-1↑	mAP↑
1 FineTrack(w/o diversity loss)	91.1	58.9
2 FineTrack(w diversity loss)	**92.3**	**61.8**

Table 1. **Ablation study of diversity loss.** We verify the effectiveness of diversity loss by training the same model (FineTrack) with and without it.

To fairly measure the performance of identity embedding, we utilize the Ground Truth of bboxes as input, which can eliminate the interference caused by false detection. Based on this, we analyze the effectiveness of diversity loss, as shown in Tab. 1. Benefiting from diversity loss, FineTrack can achieve improvement (1.2% on Rank-1 and 2.9% on mAP). Diversity loss propels multiple parallel branches in the *Multi-head Part Mask Generator* (MPMG) to focus on different parts of the target, resulting in a more comprehensive and robust appearance embedding.

Method	MOTA↑	IDF1↑	IDs↓
1 FineTrack(w/o diversity loss)	76.0	78.8	209
2 FineTrack(w diversity loss)	**77.0**	**81.1**	**115**

Table 2. **Ablation study of diverse loss when tracking.** The same association strategy is adopted for FineTrack trained with and without diversity loss to obtain the tracking results.

To verify the effect of FineTrack trained with and without diversity loss in tracking, we utilize the association strategy introduced in Section 3.4 of this paper. As reported in Tab. 2, when FineTrack is trained with diversity loss, there is a significant advantage over row 1 (1.0% on MOTA, 2.3% on IDF1, and the IDs decreases from 209 to 115) , demonstrating the effectiveness of diversity loss.

II. Visualization

FineTrack does not employ the mask Ground Truth of the target for training but compares the appearance embedding of targets with their positive and negative samples to ensure the quality of the mask. Thus the training mode for masks generated by MPMG belongs to weakly supervised learning.

The effectiveness of MPMG was demonstrated by the ablation study in Section 4.3 of the paper. Further, to more intuitively verify the quality of masks generated by MPMG, representative target masks in MOT17 and part of MOT20 scenes were intercepted.

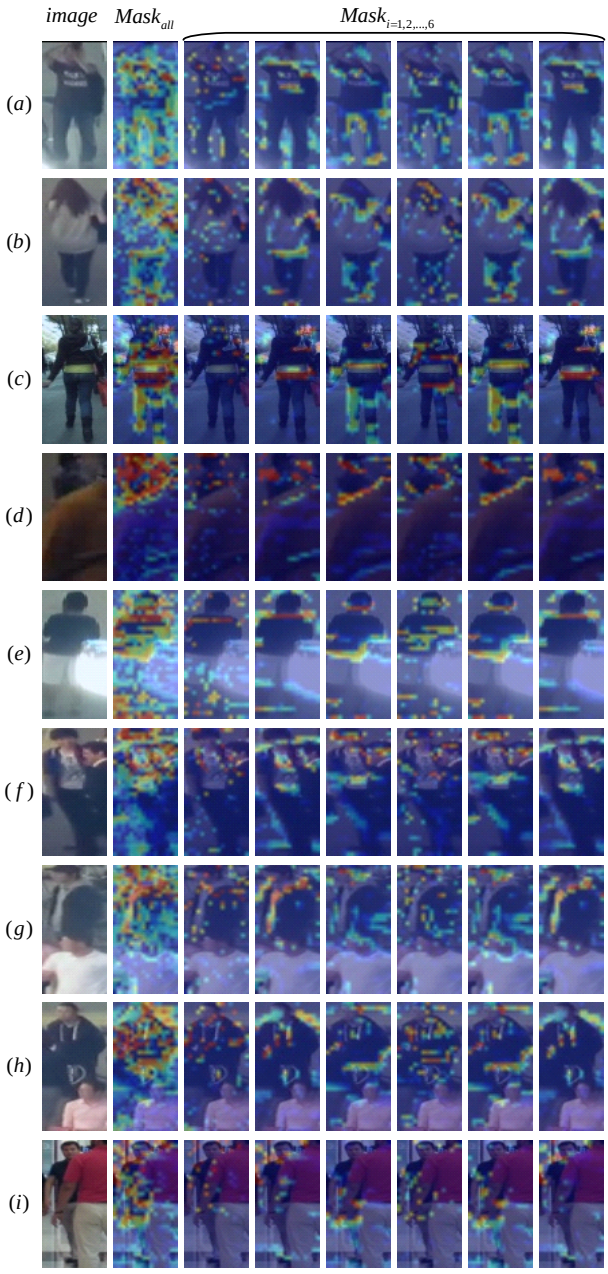

Figure 1. **Visualization of masks generated by MPMG on MOT17.**

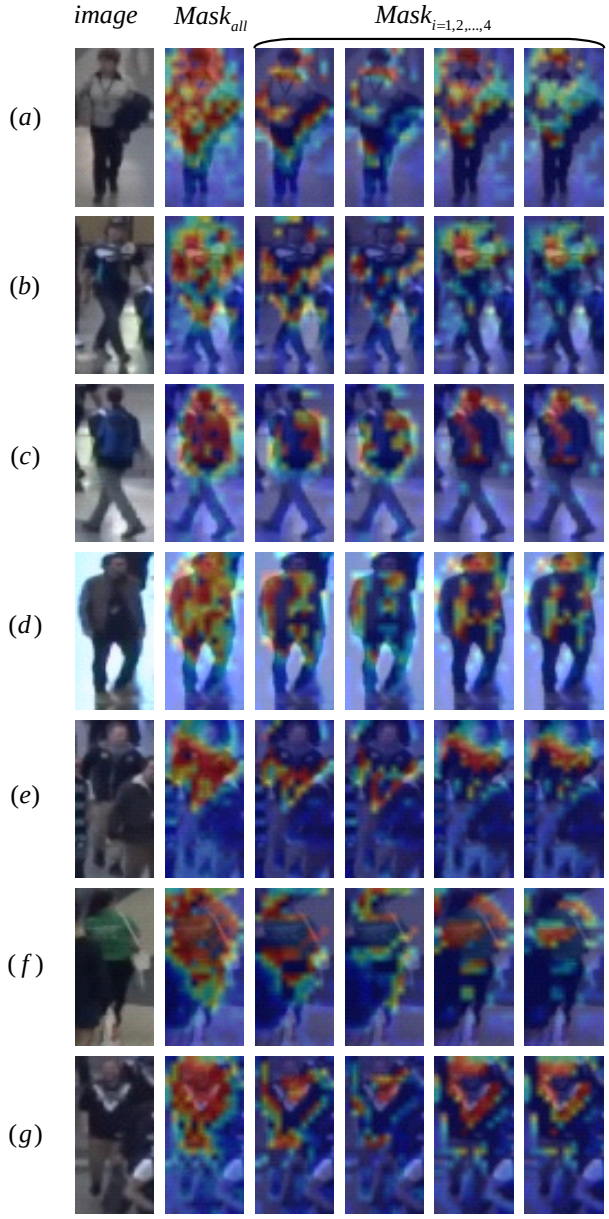

Figure 2. **Visualization of masks generated by MPMG on MOT20.**

As shown in Fig. 1 and Fig. 2, $image$ represents the target image clipped from the video frame and $Mask_{all}$ is obtained by concatenating $Mask_{i=1,2,\dots,K}$ along the channel dimension and maximizing it. It is worth noting that K is 6 in MOT17 and 4 in MOT20.

When the target is not occluded (such as (a)-(c) in Fig. 1 and (a)-(d) in Fig. 2) , the masks generated by MPMG focus on this target to suppress background noise. When the target is unfortunately occluded, whether it is occluded by background or other targets (such as (d)-(i) in Fig. 1 and (e)-(g) in Fig. 2) , the masks generated by MPMG can still focus on this target itself well. Thanks to the *Shuffle-Group Sampling* (SGS) training strategy of balanced positive and negative samples, the model can compare the target with positive and negative samples, so that the branches in MPMG can focus on the target while keeping the differences.