Learning from Synchronization: Self-Supervised Uncalibrated Multi-View Person Association in Challenging Scenes

Supplementary Material

α	AP ↑	FPR-95 ↓	P ↑	R ↑	**ACC ↑**	**IPAA-100 ↑**	**IPAA-90 ↑**	**IPAA-80 ↑**
0	56.36	11.14	90.50	93.13	90.50	46.79	62.38	83.81
0.1	56.68	11.19	92.31	**94.34**	91.73	**54.64**	65.60	86.55
0.2	60.03	**9.39**	**93.26**	93.55	**93.14**	53.81	**72.97**	**91.07**
0.3	**63.06**	9.55	91.33	93.52	91.91	52.50	67.26	87.50
0.4	60.21	11.13	88.19	90.08	89.35	42.38	56.79	82.38

Table 7. Ablation study of the α in Eq. (6).

7. Additional experiments

7.1. Qualitative results

We show more qualitative multi-view association results of our approach on the WILDTRACK [6], MVOR [28] and SOLDIERS [11] datasets, as shown in Figs. 6 to 10.

7.2. Ablation study

Analysis of fusing Re-ID and geometric distances To investigate how the fusion of the Re-ID and geometric distances affects the performance, we conduct an ablation study of the α in Eq. (6). As shown in Tab. 7, when α increases to 0.3 and 0.4, the performance starts to decrease on the WILDTRACK dataset. Although different datasets may have different optimal α, the variance of the performance is not large, and thus it is easy to obtain decent performance by setting α hierarchically.

Analysis of the self-supervised tasks To further illustrate the roles of the self-supervised tasks, we display the ablation study on the MVOR and SOLDIERS datasets. As shown in Tab. 8, training with cross-view image synchronization task solely already achieves great performance, because these two datasets are relatively easier compared to the WILDTRACK dataset with a more crowded scene, leading to larger solution space for the association. Therefore, we need further linear constraints to reduce the solution space. Tab. 4 shows that adding multi-view re-projection constraint significantly improves the training.

Confidence threshold Treating the cross-view association as a maximum bipartite matching problem may introduce false positive matches, and thus we propose to set confidence threshold for scores computed in Eq. (9) to filter out the low-score matches. As shown in Tab. 9, setting the threshold to 0.4 effectively improves precision and maintains a similar recall rate.

Data	Syn.	Pro.	AP ↑	FPR-95 ↓	P ↑	R ↑	ACC ↑	IPAA-100 ↑
MVOR	✓	✗	80.08	92.29	91.90	**95.78**	88.46	**84.42**
	✗	✓	28.33	94.39	35.56	37.99	38.46	35.68
	✓	✓	**86.50**	**79.44**	**93.20**	93.93	**89.38**	83.92
SOLD.	✓	✗	69.66	**32.14**	92.17	92.17	92.17	77.82
	✗	✓	11.33	98.86	29.55	29.55	29.55	13.45
	✓	✓	**79.13**	32.26	**95.89**	**95.89**	**95.89**	**87.64**

Table 8. Ablation study of the self-supervised learning tasks, including cross-view image synchronization (Syn.) and multi-view re-projection (Pro.). On the MVOR and SOLDIERS (SOLD.) datasets, applying cross-view image synchronization task already achieves great performance.

Threshold	P ↑	R ↑	ACC ↑	IPAA-100 ↑	IPAA-90 ↑	IPAA-80 ↑
0	89.55	**94.54**	88.23	46.43	56.19	77.50
0.2	90.35	**94.54**	89.39	49.88	59.52	80.60
0.4	92.31	94.34	**91.73**	**54.64**	**65.60**	**86.55**
0.6	**96.00**	79.74	85.90	17.50	49.88	76.07

Table 9. Ablation study of the confidence threshold on the WILDTRACK dataset.

Figure 5. Examples of the failure cases: (a) the person at the edge of the image is incorrectly associated with the person standing next to him; (b) the two overlapping persons are associated with each other; (c) the two persons that only appear in one view are incorrectly associated.

7.3. Failure cases

Fig. 5 shows three types of failure cases from our approach on the WILDTRACK dataset. In Fig. 5a, the person at the edge of the image is severely occluded. Although Self-MVA manages to find its approximate position in the other view, it incorrectly associates him with the person standing next to him. In Fig. 5b, the two overlapping persons with obscure spatial relationships are incorrectly associated with each other. In Fig. 5c, the two persons that only appear in one view are incorrectly associated due to the logic of the Hungarian matching.

8. Limitations and future work

Although our work achieves great performance on different challenging benchmarks in a self-supervised way, there are some limitations.

First, we only solve the self-supervised multi-view person association with stationary cameras. For the other scenes where the cameras are continuously moving, our approach does not perform well, because the camera embeddings and the decoder for each view are fixed. To conduct self-supervised multi-view association with the moving cameras, there are two possible directions: (1) to model the continuous camera poses by predicting the relative camera poses in adjacent frames; (2) to solve the association in a one-shot setting by strong 3d geometric reasoning.

Second, we use a manually defined threshold to filter out the false matches for each dataset during the inference. Specifically, we select the threshold from $\{0.0, 0.1, 0.2, ..., 0.9\}$ based on the results on the validation set. For datasets without labels, we need to adjust the threshold value by manually observing the association results for better performance. Even if we have set a good threshold value, the false positive matches still exist as shown in Fig. 5c. Therefore, how to automatically and accurately remove the false positive matches during inference remains unexplored in our work. To effectively detect these false positive matches, explicitly calculating the epipolar constraint between the two views would be a possible solution.

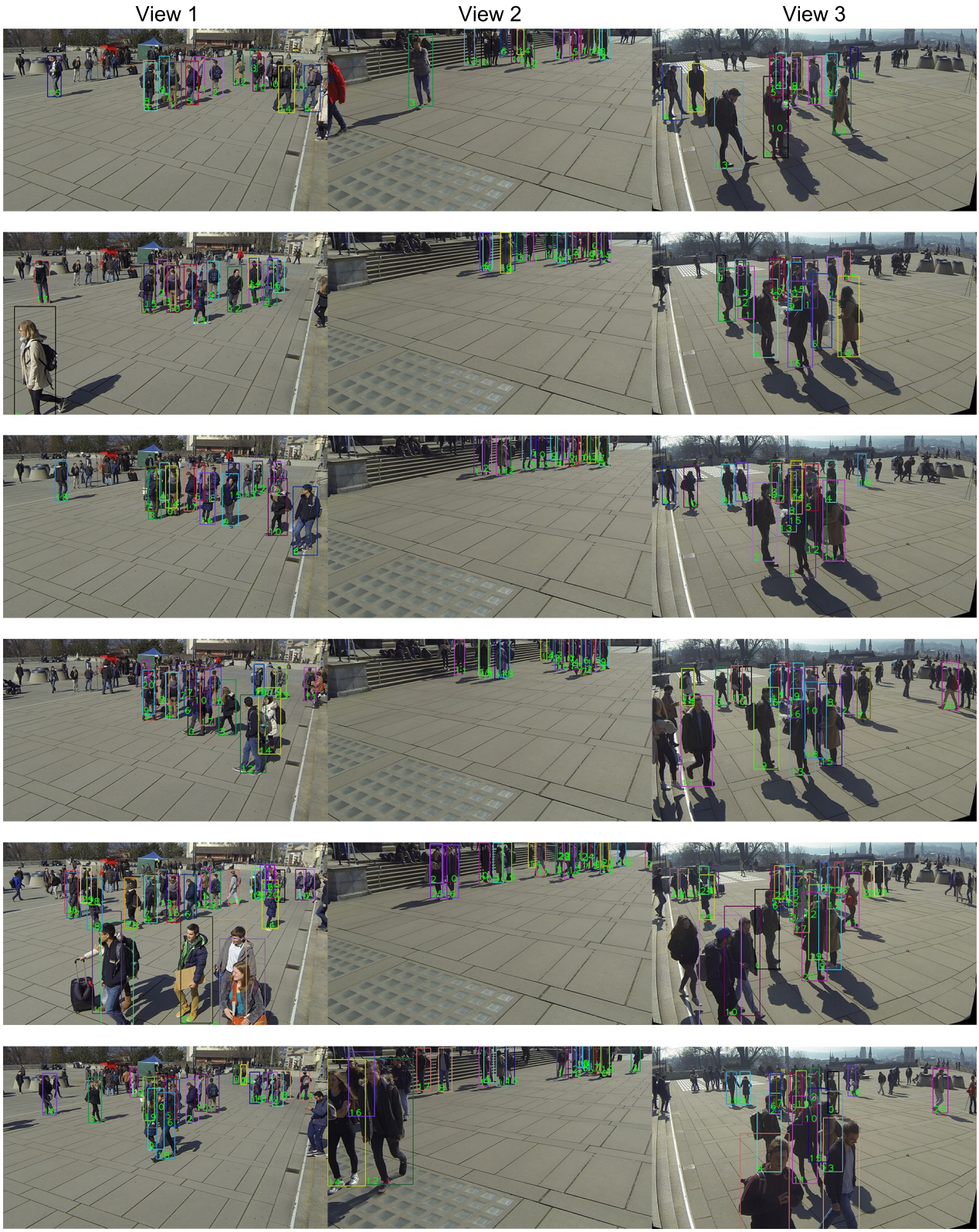

Figure 6. Qualitative results of our multi-view association approach on the WILDTRACK dataset (first 3 views).

Figure 7. Qualitative results of our multi-view association approach on the WILDTRACK dataset (last 3 views).

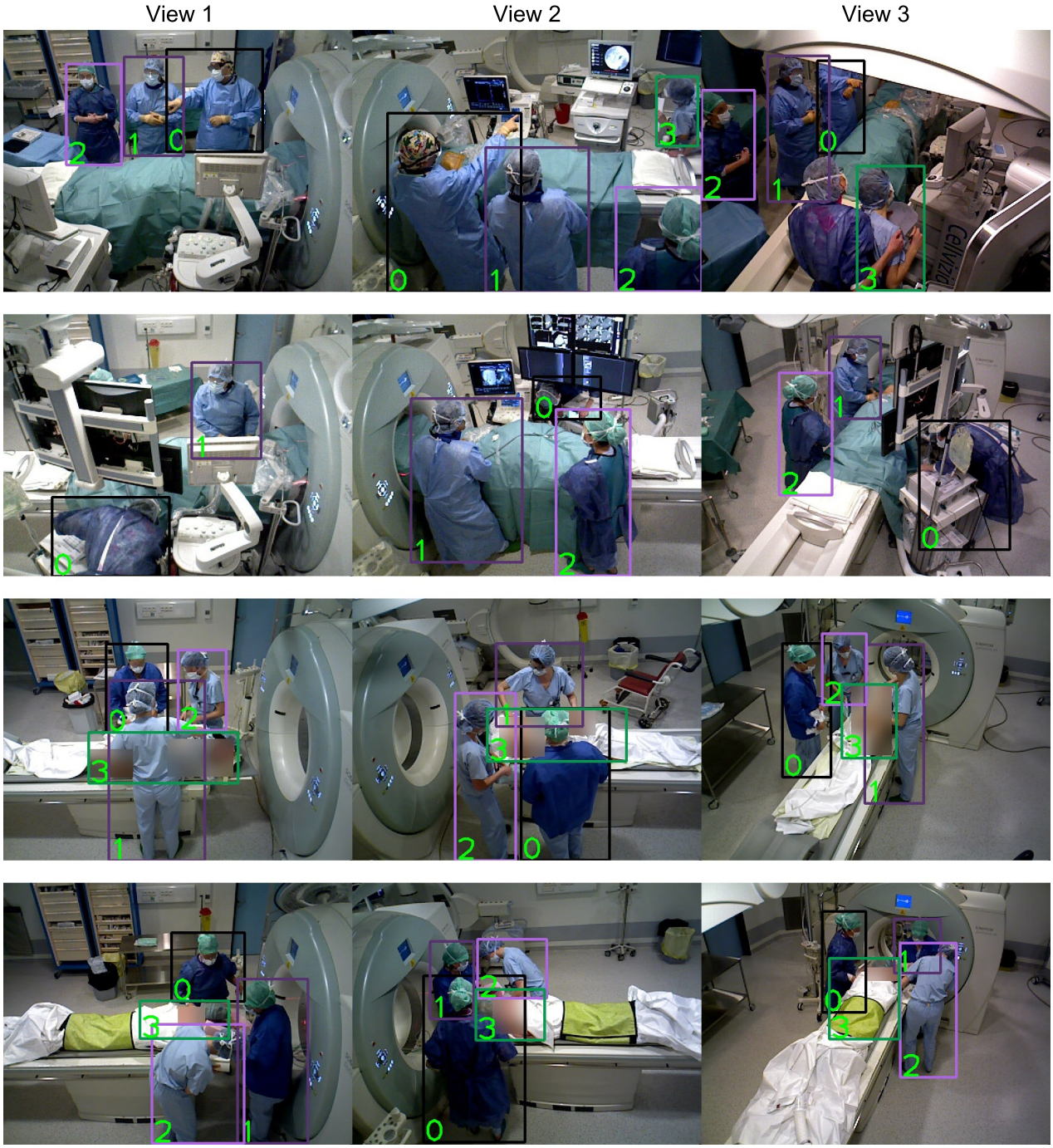

Figure 8. Qualitative results of our multi-view association approach on the MVOR dataset.

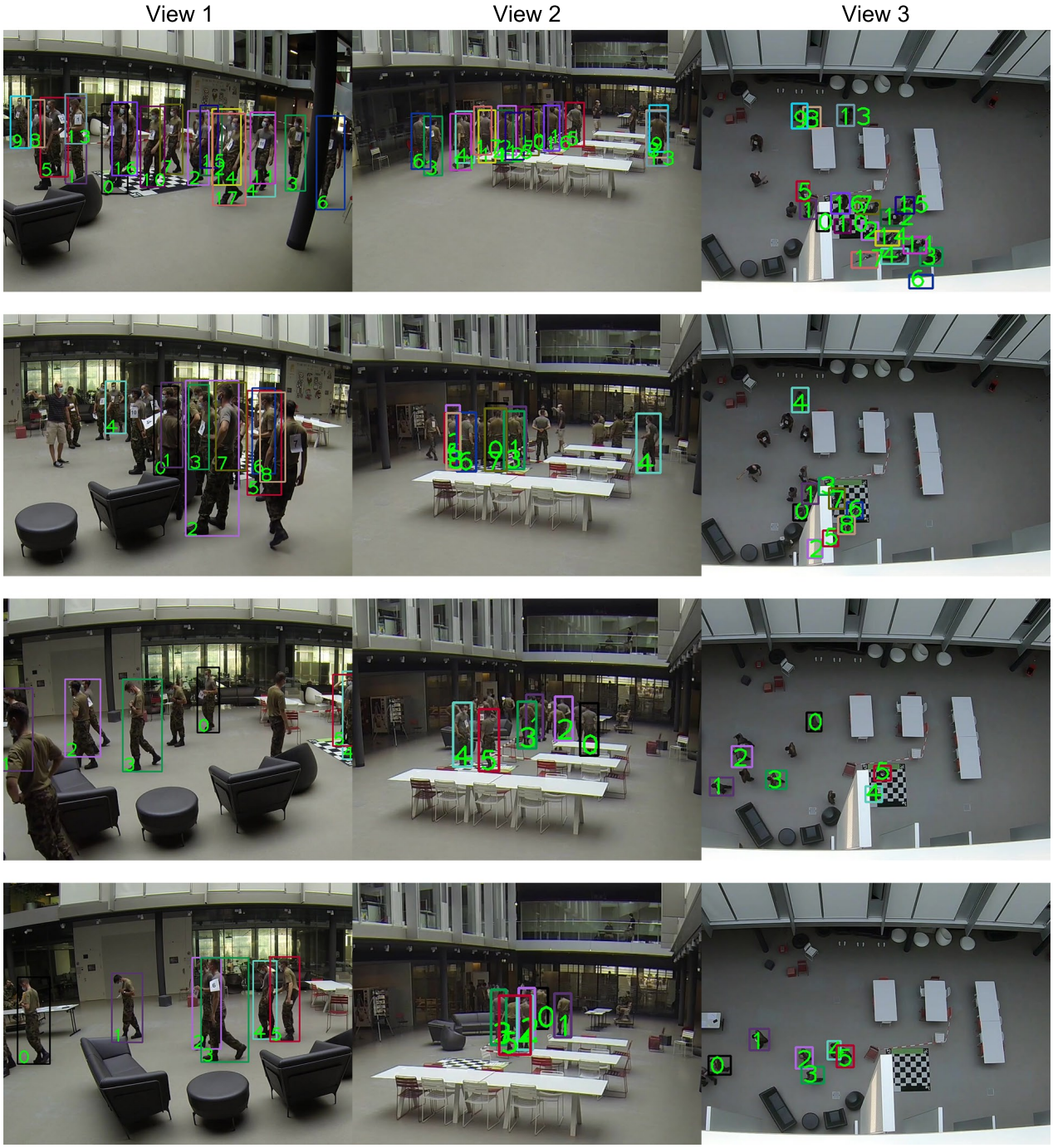

Figure 9. Qualitative results of our multi-view association approach on the SOLDIERS dataset (first 3 views).

Figure 10. Qualitative results of our multi-view association approach on the SOLDIERS dataset (last 3 views).