Keypoint Promptable Re-Identification: Supplementary Materials

In the upcoming sections, we present additional details regarding the novel Occluded PoseTrack-ReID dataset and our KPRTrack pose tracking method. Furthermore, we offer supplementary qualitative and quantitative results to provide a more comprehensive perspective on our research. Our code, annotations, and proposed dataset are available at https://github.com/VlSomers/keypoint_promptable_reidentification. Our codebase was forked from the **Torchreid**[10] framework and BPBreID[11] [34]. Pose-tracking results were obtained using the Tracklab framework for multi-object tracking [16].

Study on the Robustness to Multi-Person Occlusions

In Figure 7, we present a performance comparison between KPR and its non-promptable version under various occlusion levels on Occ-PTrack. To quantify the Multi-Person Occlusion Level (MPOL) of a query sample i in the query set Q, we compute the difference between the number of negative keypoints N_i and the number of positive keypoints P_i, and then normalize this value over the set Q to obtain a percentage as follows:

$$\mathrm{MPOL_i} = \frac{(\mathrm{N_i} - \mathrm{P_i}) - \min\limits_{\mathrm{j}\in\mathrm{Q}}(\mathrm{N_j} - \mathrm{P_j})}{\max\limits_{\mathrm{j}\in\mathrm{Q}}(\mathrm{N_j} - \mathrm{P_j}) - \min\limits_{\mathrm{j}\in\mathrm{Q}}(\mathrm{N_j} - \mathrm{P_j})}. \tag{3}$$

As shown in the plots, the utilization of prompts leads to at least 10% performance enhancement in scenarios with heavy multi-person occlusions. Consequently, these experiments demonstrate the efficacy of keypoints prompts to disambiguate the intended ReID target among multiple pedestrians. Furthermore, keypoint prompts contribute to better part-based feature extraction, which explains the performance boost in scenarios with low occlusion scores.

Occluded PoseTrack-ReID Dataset Details

As introduced in Section 5, Occluded PoseTrack-ReID (or simply Occ-PTrack), is a new ReID dataset we built out of the annotation available with PoseTrack21 [6]. PoseTrack21 is a popular video benchmark for multi-person pose tracking, that features keypoints and cross-video identity annotations. We provide a train/test split of Occluded PoseTrack-ReID that is based on the original train/validation split of PoseTrack21. Important numbers about each split are summarized in Table 4. Similar to the original dataset, the train and test splits do not have overlapping video scenes and therefore do not share identities. In order to build the Occ-PTrack dataset, we randomly sample 1000 identities from the PoseTrack21 train set and keep all 1411 identities from the test set, and uniformly

[10] https://github.com/KaiyangZhou/deep-person-reid
[11] https://github.com/VlSomers/bpbreid

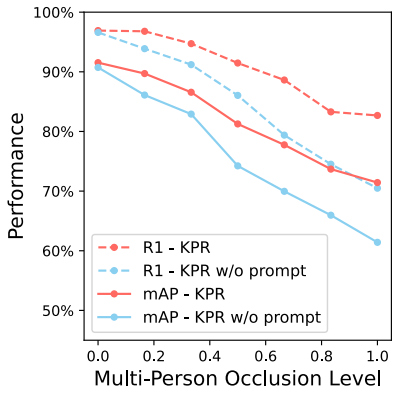

Fig. 7: Performance w.r.t. the multi-person occlusion level of the queries.

sample each tracklet, so that each identity has at least 4 image crops and maximum 20. Finally, we divide the test set into a gallery set and a query set, so that 20% of each identity images are added to the query set and the remaining 80% samples are added to the gallery set. Similar to existing dataset specialized on occlusions [29], we choose the 20% most occluded samples from each identity as query samples. The occlusion score of each sample is computed using Eq. (3). This sampling strategy renders the dataset more relevant to evaluate methods that are good at addressing occlusions and multi-person ambiguity. Compared to previous datasets that only compare query samples against gallery samples captured from a different camera viewpoint, we compare each query sample against all galleries, whether they are from a different video or not. Compared to previous sport re-identification datasets [11,38], PoseTrack-ReID contains identity labels that are valid across the entire dataset, since a given identity can be spotted in multiple videos. We provide numbers to compare the scale of our dataset with existing popular ReID benchmarks in Tab. 5. We encourage researchers to not use re-ranking [55] to evaluate their ReID method on this dataset.

Subset	# Ids	# Imgs	# Videos
train	1000	17898	474
query	1379	2581	163
gallery	1411	10831	170

Table 4: Characteristics of each PoseTrack-ReID subset.

Dataset	Type	# Ids	# Cams	# Imgs	Release	Train set	Occlusions
Market-1501 [53]	Street	1501	6	32,217	2015	✓	
MSMT17 [43]	Surveillance	4101	15	126,441	2018	✓	
Occluded-ReID [58]	Outdoor	200	-	2,000	2018		✓
SoccerNet-ReID [11]	Soccer	243,432	-	340,993	2022	✓	
DeepSportradar-ReID [38]	Basketball	486	-	9,529	2022	✓	
Occ-PTrack (ours)	Multi-sport	2411	-	31310	2024	✓	✓

Table 5: Comparing our proposed Occluded PoseTrack-ReID with other popular ReID datasets.

Keypoint and Human Parsing Annotations

Our method requires two types of annotations: the keypoints labels that are used for prompting at both train and test time, and the human parsing labels that are only used at training time by the part-prediction loss.

For Market-1501, Occluded-Reid and Partial-ReID, we generate keypoint labels with the PifPaf [19] pose estimation model as described in Sec. 4.2, and employ the human parsing labels provided by BPBreID [34].

For our proposed Occ-PTrack dataset, we simply employ the keypoint labels already available in the original PoseTrack21 dataset. To generate the human parsing labels, we employ the same methodology as the one described in [34], with a small difference regarding the generation of the segmentation mask of the ReID target. This segmentation mask is required to filter out the Pif and Paf confidence fields activations that belongs to other non-target persons. Different from BPBreID's authors who employed a MaskRCNN segmentation model, we employ Segment Anything [18] and prompt it with the target's person keypoints to generate a consistent segmentation mask. We also tried to employ SAM to directly generate the human parsing labels without PifPaf, but we found out SAM was bad at segmenting body parts, even when prompted with accurate keypoints. Combining SAM (to segment the target person) and PifPaf (to assign a body part to each pixel in the bounding box image) led to the most accurate human parsing labels. We refer readers to [34] for more details on the human parsing labels generation methodology used for existing ReID datasets.

KPRTrack Details

In this Section, we provide further details on our proposed pose-tracking method KPRTrack. KPRTrack employs two ByteTrack's components: 1) the tracklet life-cycle management mechanism and 2), the multi-stage association strategy based on the confidence level of the detections in the current frame. More details can be found in the original ByteTrack's paper [50].

To obtain strong person detection performance, we fine-tuned the YOLOX objector detector and the HrNet pose estimator on the PoseTrack21 train set. As stated in Section 4.5, our proposed tracker does not employ spatio-temporal information, as opposed to almost all existing state-of-the-art multi-object tracker.

However, our tracker can be easily extended with additional spatio-temporal cues (e.g. a Kalman Filter with bounding box IOU) to further improve performance.

As a re-identification method providing appearance cues to the tracking pipeline, we employ our proposed KPR method trained on the Occ-PTrack train set. Since KPR outputs part-based features, we made some adjustments to ByteTrack's tracklet management strategy, since it only supports global ReID methods outputting a single feature vector per detection. We therefore compute a tracklet's part-based features as a moving average of the part-based features of the underlying detections. If some body parts are not visible within a detection that has been matched to a given tracklet, the corresponding part-based features in the tracklet are not updated. Performance in Table 2 are reported on the PoseTrack21 [6] validation set.

Complete SOTA Table

We provide a more comprehensive state-of-the-art comparison in Tab. 6.

Qualitative Results

We provide additional qualitative results in Figure 8.

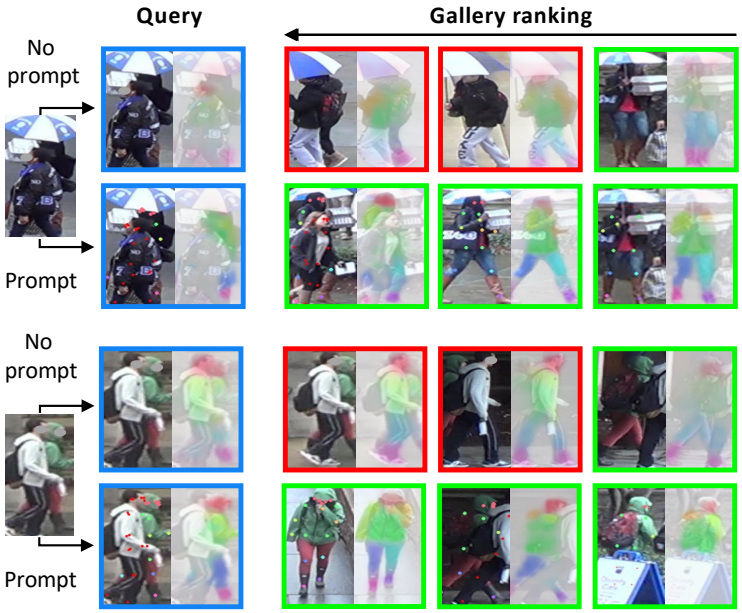

Fig. 8: Ranked gallery samples for a given query, when the input prompt is disabled/enabled. Green/red borders are correct/incorrect matches.

Table 6: Comparison of KPR with SOTA methods. Results in *Italic* are not provided in the original paper but reproduced by ourselves. The $1^{st}/2^{nd}/3^{rd}$ best scores are indicated with [1/2/3].

Datasets	Market-1501		Occluded-reID		Partial-reID		**Occluded-PoseTrack**	
Type	Holistic		Occluded					
Object Occlusions			✓		✓			
Person Occlusions							✓	
Methods	R-1	mAP	R-1	mAP	R-1	R-3	R-1	mAP
BoT [26]	94.5	85.9	58.4	52.3	-	-	*78.8*	*69.7*
SGAM [46]	91.4	67.3	-	-	74.3	82.3	-	-
PGFA [29]	91.2	76.8	-	-	68.0	80.0	-	-
PCB [36]	93.8	81.6	-	-	-	-	*81.7*	*71.2*
MHSA [37]	94.6	84.0	-	-	85.7	91.3	-	-
VGTri [45]	-	-	81.0	71.0	85.7	93.7[3]	-	-
OAMN [4]	93.2	79.8	-	-	86.0	-	-	-
HG [17]	95.6	86.1	-	-	74.8	87.3	-	-
PVPM [8]	-	-	66.8	59.5	78.3	-	-	-
HOReID [39]	94.2	84.9	80.3	70.2	85.3	91.0	-	-
ISP [57]	95.3	88.6	-	-	-	-	-	-
PAT [21]	95.4	88.0	81.6	72.1	88.0[3]	92.3	-	-
PGFL [52]	95.3	87.2	80.7	70.3	85.1	90.8	-	-
TRANS [12]	95.2	88.9	-	-	-	-	*83.5*	*73.4*
SOLIDER [5]	**96.9**[1]	**93.9**[1]	-	-	-	-	*84.4*	*61.9*
SSGR [44]	96.1	89.3	78.5	72.9	-	-	-	-
FED [41]	95.0	86.3	**86.3**[1]	79.3[3]	84.6	-	-	-
LDS [48]	95.8	90.3	-	-	-	-	-	-
BPBreid [34]	95.7	89.4	82.9	75.2	-	-	*84.9*	*75.5*
PFD [40]	95.5	89.7	83.0	81.5[2]	-	-	-	-
KPR_{IN} w/o prompt	95.6	88.7	83.3	78.2	81.7	86.0	85.3	75.4
KPR_{IN}	95.9	89.6	85.4[2]	79.1	86.0	90.0	**92.3**[1]	**82.3**[1]
KPR_{SOL} w/o prompt	96.6[3]	93.0[3]	80.0	78.5	90.3[2]	93.7[2]	86.1[3]	75.8
KPR_{SOL}	96.6[2]	93.2[2]	84.8[3]	**82.6**[1]	**90.7**[1]	**94.0**[1]	90.6[2]	81.2[2]

Additional Implementation Details

The SOLIDER [5] backbone was challenging to fine-tune and we found out 1) the "*semantic_weight*" parameter did not have much impact and was set to 0.2 in our experiments, and 2) it was necessary to freeze the keypoint tokenizer for the first 20 epochs. The label smoothing [1] regularization rate ε is set to 0.1, and the triplet loss margin [13] α is set to 0.3. The token input and output dimensions C_i and C_o are respectively set to 128 and 1024. All resulting part-based embeddings $\{f_1, ..., f_K\}$ are further downsampled to an output dimension of 256 with a simple linear layer followed by a batch normalization and ReLu operation, in order to reduce memory footprint. Images are first augmented with random cropping and 10 pixels padding, and then with random erasing [56] at 0.5 probability, before applying our propose BIPO data augmentation. A training batch consists of 64 samples from 16 identities with 4 images each. The model is trained in an end-to-end fashion for 120 epochs with the SGD optimizer on one

NVIDIA RTX8000 GPU. We employ a cosine annealing learning rate scheduler with 5 epochs warmup. For $\text{KPR}_{\text{IN}}/\text{KPR}_{\text{SOL}}$, images are resized to a width of 128 and a height of 256/384, and the learning rate is initialized to 0.008/0.0002.