Keypoint Promptable Re-Identification

Vladimir Somers[1,2,3], Alexandre Alahi[1], and Christophe De Vleeschouwer[2]

[1] Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[2] Université Catholique de Louvain (UCLouvain), Belgium
[3] Sportradar, Switzerland

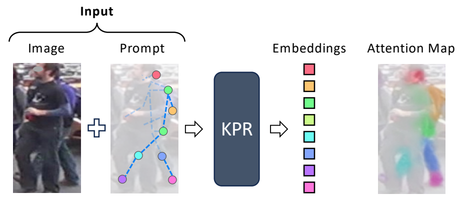
Abstract. Occluded Person Re-Identification (ReID) is a metric learning task that involves matching occluded individuals based on their appearance. While many studies have tackled occlusions caused by objects, multi-person occlusions remain less explored. In this work, we identify and address a critical challenge overlooked by previous occluded ReID methods: the Multi-Person Ambiguity (MPA) arising when multiple individuals are visible in the same bounding box, making it impossible to determine the intended ReID target among the candidates. Inspired by recent work on prompting in vision, we introduce Keypoint Promptable ReID (KPR), a novel formulation of the ReID problem that explicitly complements the input bounding box with a set of semantic keypoints indicating the intended target. Since promptable re-identification is an unexplored paradigm, existing ReID datasets lack the pixel-level annotations necessary for prompting. To bridge this gap and foster further research on this topic, we introduce Occluded PoseTrack-ReID, a novel ReID dataset with keypoints labels, that features strong inter-person occlusions. Furthermore, we release custom keypoint labels for four popular ReID benchmarks. Experiments on person retrieval, but also on pose tracking, demonstrate that our method systematically surpasses previous state-of-the-art approaches on various occluded scenarios. Our code, dataset and annotations are available at `https://github.com/VlSomers/keypoint_promptable_reidentification`.

Keywords: Person Re-Identification · Vision Prompting · Pose Tracking

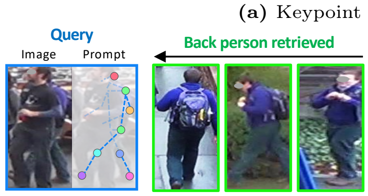
1 Introduction

Person re-identification (ReID) [47] is a challenging retrieval task that involves matching a query image of a person of interest with other person images. ReID finds wide-ranging applications in multi-object tracking [32], pedestrian flow analysis [2], sport understanding [11, 14, 28, 33, 38] and video-surveillance [53]. However, ReID is a difficult task due to various factors including inaccurate bounding boxes, luminosity changes, poor image quality, and occlusions.

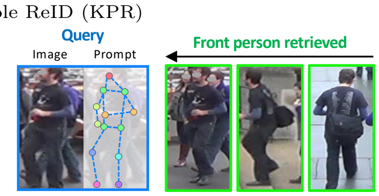
In recent years, there has been a growing interest in addressing the occluded re-identification task [30], where the ReID target might be occluded by various objects or other people. In this work, we identify and explicitly address a critical challenge that has been overlooked by previous occluded ReID methods [34, 41]:



(a) Keypoint Promptable ReID (KPR)



(b) Back (occluded) person prompted



(c) Front person prompted

Fig. 1: Overview of our proposed Keypoint Promptable ReID (KPR) method. KPR takes an image with keypoints prompts as input and produces part-based features of the prompted target. The prompt instructs the model to focus on a specific individual, i.e. the back blue jacket man (b) or the front black t-shirt man (c) in this example. Colored dots illustrate the positive keypoints prompt, with one color per body part.

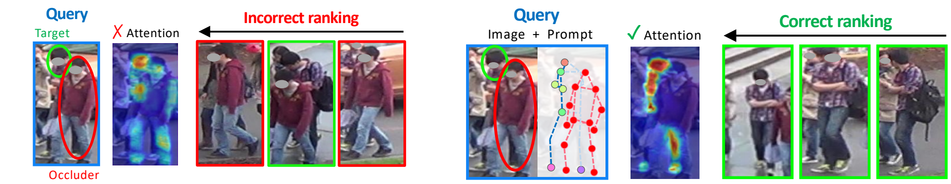
the presence of **Multi-Person Ambiguity** (MPA) inherent to bounding box annotations. The multi-person ambiguity arises when a bounding box image contains multiple individuals, making it challenging even for humans to accurately identify the intended ReID target among all the candidates. Figure 2a shows how MPA can negatively impact the re-identification model, causing feature mixing among different individuals or even a focus on an unintended person.

To tackle this multi-person ambiguity, it is necessary to incorporate additional pixel-level information, such as keypoints or segmentation masks, provided by an upstream human operator or vision model. This additional data plays a vital role in assisting the downstream ReID method to distinguish the intended target from other candidates.

Inspired by recent advances in promptable[4] methods for vision transformers (e.g., SAM [18]), we propose to explicitly use additional semantic[5] keypoints as inputs to disambiguate images with multiple individuals. Semantic keypoints are suitable for two crucial use cases: 1) manual prompting by a human operator with a few clicks on an image, and 2) automated prompting by an upstream pose estimation model. Nevertheless, our architecture could be easily extended to support segmentation masks as prompts.

[4] A "prompt" typically refers to an input instruction that guides the model's response. In vision tasks, e.g. segmentation, prompts can be any kind of data that specifies what the model should focus on [18].

[5] Semantic keypoints are distinct points within a person image that carry semantic information about specific body parts, typically obtained with pose estimation [22].



(a) Existing ReID models (e.g. BPBreID [34]) are confused by **MPA**

(b) KPR explicitly addresses MPA with positive (pastel) and negative (red) keypoints prompt

Fig. 2: Person retrieval with Multi-Person Ambiguity (**MPA**). Green/red borders are correct/incorrect matches. Red/pastel dots indicate negative/positive prompts.

Prior research [8,39,40] on *pose-guided methods* has explored the use of additional information from external pose estimation models. However, none of these works explicitly tackled the multi-pedestrian ambiguity issue.

Following above insights, we propose **KPR**, a novel ***K**eypoint-**P**romptable transformer for part-based person **R**e-identification*, illustrated in Fig. 1a. KPR is a transformer-based model that takes an image along with *semantic keypoints* as input. It then produces *part-based features*, each representing a distinct body part of the ReID target, along with their respective visibility scores. Part-based methods have demonstrated superior performance [34,57] in re-identifying occluded individuals because they utilize only the visible parts for comparison. Our method can process both *positive* and *negative* keypoints, which respectively represent the target and non-target pedestrians. As illustrated in Fig. 1b, prompts can be used to re-identify the desired (front or back) person in case of occlusion. Furthermore, KPR is designed to be *prompt-optional* to offer more practical flexibility. This means the same model can be used without prompt on non-ambiguous images, or with prompt when dealing with occlusions, while consistently achieving state-of-the-art performance in both cases. To demonstrate the advantages of keypoint prompts for re-identification, we evaluate KPR on two tasks: *person retrieval* and *multi-person pose tracking*.

Moreover, since no current ReID dataset is provided with explicit target identification in case of multi-person ambiguity, we introduce a new dataset called ***Occluded PoseTrack-ReID***, nicknamed "*Occ-PTrack*", and derived from the PoseTrack21 dataset [6]. This dataset offers high-quality keypoints annotations and features a substantial number of images with multi-person occlusions.

Overall, we summarize the main contributions of our work as follows:

1. We propose *KPR*, a novel prompt-optional model for part-based ReID. To our knowledge, we are the first to introduce a visual prompting mechanism for person re-identification, and the first to explicitly address the ambiguity caused by multi-person occlusions.
2. We introduce *Occ-PTrack*: the first multi-person occluded ReID dataset with explicit target identification through manual keypoints annotations. Furthermore, we propose new keypoint annotations for four popular re-identification datasets.

3. Our method outperforms all previous state-of-the-art methods on the occluded ReID task, demonstrating the effectiveness of the keypoint promptable approach.

Our ambition is to advocate for a shift from the ambiguous bounding-box approach to ReID and promote a keypoint-centric paradigm. To this end, we publicly release our codebase, proposed dataset, and annotations, to provide a common setup for evaluating promptable ReID methods.

2 Related Work

Multi-Person Ambiguity: A previous study [41] addressed a related issue called "Non-Target Pedestrians (NTP)," similar to our concept of "Multi-Person Ambiguity" (MPA). However, their focus was on mitigating the negative impact of NTP during training. In contrast, we provide a method to explicitly disambiguate the intended ReID target during both training and testing phases.

Pose-guided ReID: We detail here the fundamental difference between our novel keypoint promptable approach and the traditional pose-guided methods. Although none of the previous works explicitly address the multi-person ambiguity (MPA) issue, we identify two categories within these pose-guided approaches: 1) those inherently unable to address MPA [4,8,15,27,29,34,44,45], and 2) those with the potential to overcome MPA [39,40]. In the first category, methods like BPBreID [34] or Pirt [27] utilize pose labels only during training, and can not leverage additional information at test time to disambiguate multiple persons. Moreover, methods relying on horizontal stripes (VGTri [45], PGFA [29]) or the affinity fields of a pose model (PVPM [8]), are meant to tackle object-occlusions, but struggle to disambiguate multiple persons. In the second category, methods like HOReID [39] and PFD [40] directly employ keypoints from a single person, enabling focus on a specified target. However, HOReID [39] is restricted to utilizing spatial features at the exact keypoints locations, thus overlooking critical appearance cues on entire body regions. Finally, PFD [40], and all previous pose-guided methods including HOReID, incorporate pose information *after appearance encoding has occurred*, e.g., for local pooling of a spatial feature map, but do not leverage pose data to guide the appearance encoding process as our method do. Indeed, our approach implements a *prompt-aware appearance encoding*, since it conditions feature extraction on the input keypoints. This enables *better feature disentanglement between the target and the occluders* from the early feature extraction stages, resulting in more accurate representations. Furthermore, our method is *prompt-optional*, setting it apart for its flexibility from these previous pose-guided methods. Finally, our method can ingest additional *negative keypoints*, to further disambiguate a multi-person occlusion scenario.

Prompting in Vision: Recent works have introduced promptable transformers to tackle object segmentation. For instance, SAM [18] employs two networks: a transformer encoder to generate generic spatial features and a lightweight promptable decoder to integrate user inputs for interactive segmentation. In contrast, our approach directly prompts a feature encoder, so that *feature extrac-*

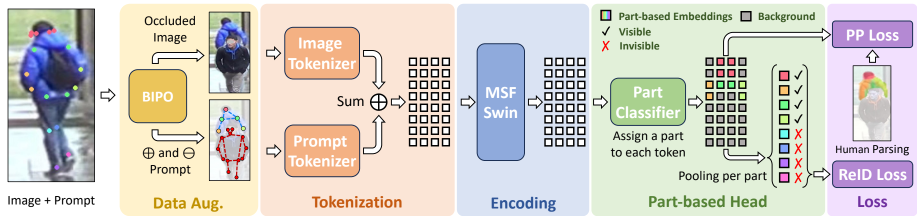


Fig. 3: Architecture overview of our proposed Keypoint Promptable ReID (KPR) model. The Batch-wise Inter-Person Occlusion (BIPO) data augmentation is first applied to generate artificial inter-person occlusions (Sec. 3.5). The image and the optional positive/negative ($\oplus/\ominus$) prompts are then tokenized and summed (Sec. 3.1). Tokens are then fed to our proposed Multi-Stage feature Fusion (MSF) Swin transformer to generate high-resolution feature maps (3.2). The feature map is then fed to a Part-based Head (PBH), which assigns a part (or the background) to each token with a part classifier, and then averages all tokens of the same part to produce the final K part-based embeddings $\{f_1, ..., f_K\}$ and their binary visibility scores $\{v_1, ..., v_K\}$, with $v_i \in \{0, 1\}$, and visually denoted by {✗, ✓} (Sec. 3.3). A part i with no token assigned is considered invisible (i.e., $v_i = 0$), and ignored when computing two samples' similarity. KPR is illustrated here for K=8, with a unique color for each body part: {*head*, *torso*, *right/left arm*, *right/left leg*, and *right/left feet*}. Finally, the entire pipeline is trained with two losses: a Part-Prediction (PP) Loss and a ReID Loss (Sec. 3.4).

tion is explicitly driven by the prompts. Unlike SimpleClick [23], which prompts a transformer with a single 2D point and a previous mask for iterative segmentation, our model ingests multiple keypoints concurrently, enhancing body part localization and representation by *leveraging their semantic information.*

3 Methodology

The overall architecture of our model KPR is illustrated in Fig. 3 and is based on a Swin transformer backbone [24]. The next subsections first describe the various components of the architecture. We then present the training process, including our proposed data augmentation. Finally, we describe the inference procedure and conclude with an intuitive discussion of our design choices.

3.1 Tokenization

The tokenization procedure coverts raw image and keypoints information into C_i-dimensional tokens, to be subsequently processed by the Swin encoder.

Image tokenization: The input image, with dimensions $H \times W$, undergoes tokenization using the conventional Swin patch embed module. This process generates $H_t \times W_t = \frac{H}{4} \times \frac{W}{4}$ spatial tokens of dimensions C_i by embedding each 4×4 image patch through a linear layer.

Keypoints tokenization: As previously mentioned, our model uses two types of semantic keypoints as prompts: *positive keypoints* indicating the ReID target and *negative keypoints* indicating other individuals. We organize these prompt keypoints into K+1 groups: the first group contains all negative keypoints, and the remaining K groups contain subsets of the positive keypoints. Each subset of positive keypoints actually corresponds to a semantic body region. For instance, with $K = 8$, we define groups like: {*head*, *torso*, *right/left arm*, *right/left leg*, and *right/left feet*}. For each group $i \in \{1, ..., K + 1\}$, a unique heatmap M_i of size $H \times W$ is created by drawing a Gaussian 2D kernel at each of its keypoint's center location (x, y). The kernel's standard deviation is set to a fraction α_G of the image width. Invisible keypoints are ignored. These $K + 1$ heatmaps are then concatenated into a global spatial tensor M of size $H \times W \times (K + 1)$. Finally, each 4×4 patch in M is linearly projected to produce prompt tokens of dimension C_i. Like the image tokenizer, the keypoints tokenizer outputs therefore $H_t \times W_t$ spatial tokens of dimension C_i.

Token fusion: Images and keypoints tokens are fused by summing tokens at the same spatial location, producing $H_t \times W_t$ spatial tokens with dimension C_i. These tokens are then forwarded to the Swin-based encoder.

Design motivations: Prompted keypoints are organized into $K+1$ channels to preserve semantic meaning before tokenization. It allows the model to differentiate positive and negative prompts, and to distinguish prompt information of different body parts, for better part localization. *Prompt-aware appearance encoding*: Moreover, feeding the network with keypoint information before the encoding stage is critical in enabling the network to disentangle the ReID target features from other persons' features. Finally, this design offers the flexibility to support segmentation mask prompts by replacing M with a double-channel map of positive and negative segmentation masks. Regarding token fusion, the sum operation maintains the input dimension for the Swin backbone, permitting the use of pre-trained weights from the image-only model. *Prompt optionality*: This design also offers the flexibility of running the model without prompts, by processing the image tokens alone, with no architectural adjustments required.

3.2 MSF Swin Encoder

A *Swin-transformer* [24] is employed as the backbone feature extractor for its great performance in vision tasks and its ability to capture high-frequency details, which is beneficial to ReID [49]. Moreover, we propose to enhance Swin with a Multi-Stage feature Fusion strategy, denoted as "MSF", to output feature maps with a high spatial resolution. Indeed, similar to CNN, each Swin stage reduces the number of tokens by a factor of 4 while doubling the channel dimension. To obtain a high-resolution token map, the outputs of all stages are bilinearly upsampled to the same resolution $H_t \times W_t$ as the output of the tokenization module, then concatenated into a single feature map along the channel dimension, and finally passed through a linear layer to obtain the final spatial token map of size $H_t \times W_t \times C_o$. Different from the standard Swin, our MSF Swin therefore preserves the input tokens resolution. As pointed out in prior works [26,57], a high-resolution feature map enriches the granularity of the features and brings significant performance improvements (see Tab. 3, Exp. 3).

3.3 Part-based Head

The encoder yields $H_t \times W_t$ tokens, which are sent to the Part-based Head (PBH) module, as illustrated in Fig. 3. Each token is first labeled either as one of the K body parts or as background, by a *token-wise part classifier*. The part classifier is implemented as a linear layer with $K+1$ output logits followed by the softmax operation, and is applied individually on each token. The assignment of each token to one of the K body parts (or the background) is therefore formulated as a classification task with $K+1$ classes. We term the probability map of size $H_t \times W_t \times (K+1)$ produced by this part classifier over all tokens as "*part-attention maps*". These maps are showcased as colored heatmaps in Fig. 1a and Fig. 6b, with one color per channel (the background channel is ignored). Finally, tokens having the same assigned body part are averaged together, resulting in K part-based embeddings $\{f_1, ..., f_K\}$ of size C_o. Background tokens are ignored. Parts with no tokens assigned are considered invisible and given a visibility score $v_i = 0$. The corresponding invisible embedding f_i, with $v_i = 0$, is therefore ignored for downstream computation. The PBH module thus produces K part-based embeddings $\{f_1, ..., f_K\}$ and their binary visibility scores $\{v_1, ..., v_K\}$.

Human parsing labels: To train the token-wise part classifier, each spatial token is assigned a ground-truth class, i.e., part label, derived from the coarse human parsing labels provided by [34]. These human parsing labels are illustrated on the right of Fig. 3 and in Fig. 4a.

3.4 Training Process

The overall objective function used to train the model is:

$$L = L_{\mathrm{ReID}} + \lambda_{pp} L_{pp} \ , \tag{1}$$

where L_{ReID} is the ReID loss supervised with identity labels, L_{pp} is the token-wise part-prediction loss, and λ_{pp} is a weight parameter empirically set to 0.3.

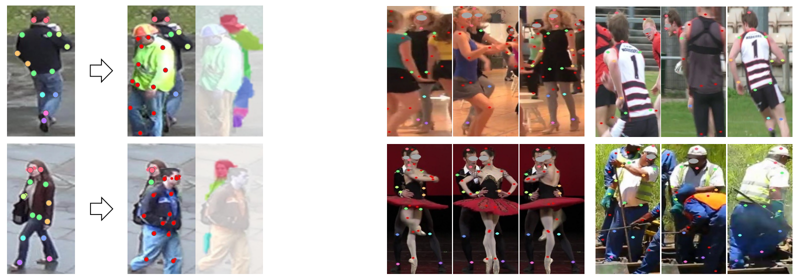
Part Prediction Loss: The token-wise part classifier introduced in Sec. 3.3 is supervised with a Part Prediction (PP) Loss that is actually a cross-entropy loss applied on each spatial token as in [34].

Re-identification loss: As a ReID objective, we employ the GiLt Loss [34], that is specifically designed to train part-based ReID models. It has the benefit of being robust to occlusions and non-discriminative body parts. GiLt combines an identity loss [26] and a batch-hard triplet loss [13]. The distance function employed for the triplet loss is the average of all local part-based distances.

3.5 BIPO Data Augmentation

Since most ReID datasets feature limited occlusion in their training set, a model often struggles to handle the occlusions it encounters during testing. To address this, we propose a novel Batch-wise Inter-Person Occlusion data augmentation (BIPO)[6], to generate artificial inter-person occlusions on training images,

[6] Inspired by the copy-paste technique for segmentation [10].



(a) BIPO (b) Occluded PoseTrack-ReID

Fig. 4: (a) Our proposed Batch-wise Inter-Person Occlusion (BIPO) data augmentation creates artificial person occlusions that are consistent across image, prompt, and human parsing labels. BIPO is crucial to enforce the model to rely on the input prompts. (b) Four identities with their keypoints labels from our Occ-PTrack dataset.

prompts, and human parsing labels. When provided with a training sample, BIPO randomly selects a different person's image (the occluder) from the same training batch, contours it with a segmentation mask derived from the human parsing labels, and overlays it on the main image. The human parsing label and keypoints prompt of the training image are then updated accordingly. Finally, all positive keypoints from the occluder are added to the training prompt to serve as additional negative points.

Design motivations: Given that a triplet loss with batch-hard mining [13] is employed during training, using images from the same training batch to generate occluders yields hard positive and negative samples for mining triplets, thereby creating challenging scenarios for contrastive learning. BIPO is illustrated in Fig. 4a and its significance for KPR is further discussed in Sec. 3.7.

3.6 Inference

At inference, KPR processes each input image together with their respective keypoints prompts, and produces K part-based representations $\{f_1, ..., f_K\}$ of the ReID target and their binary visibility scores $\{v_1, ..., v_K\}$. Similar to previous part-based works [45,57], the global distance $dist^{qg}_{\text{global}}$ between a query image q and a gallery image g is computed as the average cosine distance $dist_{\cos}$ of the part-based embeddings f_i that are visible in both samples:

$$dist^{qg}_{\text{global}} = \frac{\sum\limits_{i \in \{1,\ldots,K\}} \left(v^q_i \cdot v^g_i \cdot dist_{\cos}(f^q_i, f^g_i) \right)}{\sum\limits_{i \in \{1,\ldots,K\}} \left(v^q_i \cdot v^g_i \right)} \, . \tag{2}$$

3.7 KPR Pipeline Motivation and Intuition

Our pipeline comprises two closely related architectural components. First, the prompt tokenizer, which encodes semantic keypoints information to disambiguate the ReID target, thereby providing prior information about body part location to drive feature extraction. Second, the Part-based Head (PBH), responsible for localizing the body parts of the target and constructing its part-based features.

While the PBH is explicitly supervised with human parsing labels to learn human topology, there is no dedicated loss incentivizing the model to leverage the input prompt. Additionally, multi-person occlusions are rare in most ReID training sets, giving the model little incentive to rely on these prompts to localize the target person. For this reason, just adding a prompt tokenizer and training the pipeline as is produces a KPR model with weak reliance on the input prompts, and therefore low performance in multi-person occlusion scenarios.

To address these issues, our proposed BIPO data augmentation plays a critical role. By generating artificial multi-person occlusions during training, the model is compelled to rely more on the prompts, as they represent the sole reliable source of information to distinguish the true ReID target from the occluder.

Consequently, incorporating BIPO during training is crucial to produce a model that is proficient at leveraging prompt information and effectively handling multi-person occlusions. Finally, designing the Part-based Head without explicit reliance on the prompt offers an additional advantage: it renders our model *prompt-optional*, enhancing robustness against noisy or missing prompts while still enabling effective localization of the target's body parts.

4 Experiments

4.1 Evaluation Benchmarks and Metrics

We assess our model's performance on three types of *person retrieval* setups: non-occluded (i.e., holistic) on **Market-1501** [53], object occlusions on **Occluded-ReID** [58] and **Partial-ReID**[7] [54], and finally multi-person occlusions on our below introduced **Occluded PoseTrack-ReID** dataset. **Occluded-Duke** [29] was taken down and removed from our study; check out our GitHub for more information. Two standard person retrieval metrics are reported: the CMC at *Rank-1* and the *mAP*. Performances are evaluated in a single query setting and without re-ranking [55]. Furthermore, we evaluate our model on *multi-person pose tracking* with the **PoseTrack21** [6] dataset. We report the *HOTA* [25] performance metric, and relevant sub-metrics such as association accuracy (*AssA*), detection accuracy (*DetA*), and the number of identity switches (*IDs*) [20].

4.2 Proposed Occluded PoseTrack-ReID Dataset and Annotations

We introduce Occluded PoseTrack-ReID (or simply Occ-PTrack), a new ReID dataset we built out of the annotation available with PoseTrack21 [6], a popular video benchmark for multi-person pose tracking, that features keypoints and cross-video identity annotations. Unlike previous ReID datasets focused on street surveillance, Occ-PTrack consists of images from everyday life videos, primarily

[7] We use the occluded (non-cropped) images as queries.

Table 1: 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*
PCB [36]	93.8	81.6	-	-	-	-	*81.7*	*71.2*
VGTri [45]	-	-	81.0	71.0	85.7	93.7^{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	-	-
TRANS [12]	95.2	88.9	-	-	-	-	*83.5*	*73.4*
SOLIDER [5]	**96.9**^{1}	**93.9**^{1}	-	-	-	-	*84.4*	*76.1*^{3}
SSGR [44]	96.1	89.3	78.5	72.9	-	-	-	-
FED [41]	95.0	86.3	**86.3**^{1}	79.3^{3}	84.6	-	-	-
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}

from sports activities, as illustrated in Fig. 4b. Occ-PTrack is divided into a train/test that includes 1000/1411 identities with 17.898/13.412 images from 474/170 videos, which is roughly equivalent in terms of scale to other popular ReID datasets [29,53] (Tab. 5 of the supp. materials). To assess the ReID model's performance in multi-person occlusion scenarios, we select the most cluttered images of each identity in the test set as query samples, and the remaining test images as gallery samples. Cluttered images corresponds to multi-persons occlusions scenarios where either the front (occluding) or back (occluded) person is the ReID target, to evaluate the model ability to re-identify individuals in both scenarios. We provide further details about our proposed dataset in the supp. materials. Occ-PTrack is challenging as persons within the same video exhibit a high degree of visual resemblance since they often wear similar sports kits.

Keypoint Annotations for Standard ReID Datasets: Since conventional ReID datasets [29,53,58] do not contain keypoint annotations, we generate pseudo-labels for them with the PifPaf [19] pose estimator. These pseudo-labels include keypoint annotations in COCO [22] format for each image, as illustrated in Fig. 4a. When multiple skeletons are detected in an image, we make the assumption the one with its head closer to the top center part of the image is the intended target, and mark it with a "*is_ target*" attribute. Other skeletons are used as negatives. More details are provided in the supplementary materials.

4.3 Implementation Details

We evaluate our model with both the ImageNet [31] (**$\mathbf{KPR_{IN}}$**) and the SOLIDER [5] (**$\mathbf{KPR_{SOL}}$**) pre-trained Swin backbones. SOLIDER is a human-centric foundation model trained on the large-scale LUPerson [7] dataset. Following [34], the number of body parts K is set to 8 for Occ-PTrack, and 5 otherwise. The training procedure is mainly adopted from TransReID [12]. We apply our *BIPO* data augmentation (Sec. 3.5) with a 0.3 probability. The model is trained for 120 epochs with a batch size of 64 and a cosine annealing *lr* scheduler. For $\mathrm{KPR_{IN}}/\mathrm{KPR_{SOL}}$, images are resized to a width of 128 and a height of 256/384. The Gaussian kernel standard deviation α_G (Sec. 3.1) is set to 0.1.

4.4 Comparison with State-of-the-Art Methods

In Table 1, we present a comparative analysis of KPR, when **using prompts** (*KPR*) and **not using prompts**[8] (*KPR w/o prompt*), against all leading occluded ReID methods. Our method emerges as the top performer overall.

Occluded PoseTrack-ReID: Unlike the other three datasets, Occ-PTrack contain a substantial number of images with multi-person occlusions, making them suitable environments to assess the effectiveness of our promptable method. On this challenging occluded dataset, our proposed model KPR outperforms all previous methods. Our *keypoint promptable* approach offers several key advantages over the previous *pose-guided* methods, including: 1) the ability to handle multi-person occlusions; 2) the capability to process both positive and negative keypoints; 3) a *prompt-aware appearance encoding*, leading to better feature disentanglement between multiple persons; and 4) the flexibility of prompts being optional. We refer readers to Sec. 2 for a detailed discussion about these key differences and a comparison to previous methods. On our proposed Occ-PTrack, we can see that the best-performing method is the foundation model SOLIDER, but it still lies about 6% behind our KPR solution. Even BPBreID [34], which is specialized for occluded scenarios, demonstrates far lower performance, since it is not designed to address the Multi-Person Ambiguity. Our experiments show that integrating keypoint prompts boosts performance by at least 7.0% Rank-1 and 6.2% mAP, since it helps in mitigating the negative impact of MPA.

Market-1501, Occluded-ReID and Partial-reID: These three datasets primarily feature single individuals in their images, which means multi-person ambiguities are rare, and therefore prompts are not exploited to their full potential. On Market-1501, our method performs on par with the state-of-the-art. However, performance levels have plateaued in recent years on this dataset. In contrast, on Occluded-ReID and Partial-reID, using a prompt noticeably improves performance by helping distinguish the individual features from occluding objects. Additionally, using prompt is beneficial when facing this cross-domain[9] setting. These findings are consistent with earlier studies [8, 39–41] that also leveraged pose estimation to attain strong performance on this dataset.

[8] The prompt tokenizer is complitely removed at both training and inference.

[9] Occ-ReID and Partial-ReID have no train set, so Market-1501 is used for training.

Table 2: Multi-person pose tracking performance in videos on PoseTrack21 [6].

Method	HOTA↑	DetA↑	AssA↑	IDs↓
TRMOT [42]	46.8	40.9	55.0	-
FairMOT [51]	53.5	47.4	61.4	-
Tracktor++ [3]	58.3	52v.7	65.4	-
CorrTrack + ReID [6]	56.9	51.3	64.2	-
KPRTrack w/o prompt	61.5	58.4	65.8	3830
KPRTrack	**63.0** (+1.5)	**59.0** (+0.6)	**68.3** (+2.5)	**3344** (-486)

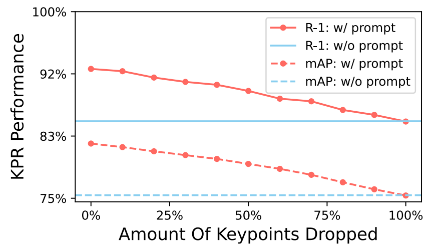


Fig. 5: Prompt optionality study

4.5 Multi-Person Pose Tracking in Videos

We introduce a simple yet robust pose tracker, referred to as *KPRTrack*, built upon our promptable ReID method. KPRTrack combines the YOLOX [9] object detector with the HRNet [35] pose estimator. For each detected person, it extracts part-based ReID features utilizing their keypoints as positive prompts and keypoints from other individuals *within the detected person bbox* as negative prompts. Detections are matched in an online fashion with the Hungarian algorithm based on the distance function introduced in Eq. (2). Only tracklet-detection pairs with a distance below a threshold of 0.2 are considered for matching. Therefore, our proposed tracker relies solely on appearance information, meaning spatio-temporal cues are omitted. Pose tracking results are reported in Tab. 2. Despite relying on ReID only, KPRTrack achieves state-of-the-art performance, demonstrating that an occlusion-robust ReID model has strong tracking capabilities. Finally, our two last experiments in Tab. 2 reveal that employing keypoint prompts significantly enhances tracking performance. This improvement can be attributed to the enhanced communication between the detector and the ReID via the keypoint prompts. This leads to improved association accuracy and reduced identity switches, especially when multiple targets intersect paths (i.e., multi-person occlusion scenarios). Further information regarding our baseline tracker is provided in the supplementary materials.

4.6 Ablation Studies

In this section, we analyze the performance impact of our architectural choices.

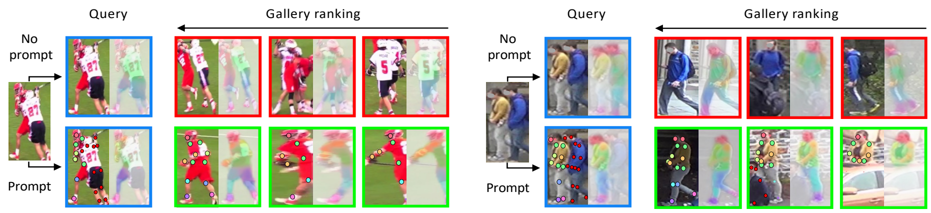
Components of KPR As demonstrated in Table 3, all our proposed modules introduced in Sec. 3 play a crucial role in improving the overall ReID performance. Exp. 2 and Exp. 8 demonstrate the importance of the Part-based Head to compare only mutually visible body parts when facing occluded persons, compared to using a single global feature (as in Exp. 1 and 7). PBH degrades performance when applied alone (Exp. 2 vs Exp. 1), but is beneficial when jointly

Table 3: Ablation study of our proposed KPR architecture. PBH stands for "Part-based Head", meaning we use multiple part-based embeddings with visibility scores instead of a single global embedding (Sec. 3.3). MSF stands for the "Multi-Scale Features" module added to Swin (Sec. 3.2). Prompt $\oplus/\ominus$ refers to positive/negative prompts (Sec. 3.1). BIPO stands for the "Batch-wise Inter-Person Occlusion" data augmentation (Sec. 3.5). SOL stands for the pre-trained weights from SOLIDER [5] (Sec. 4.3).

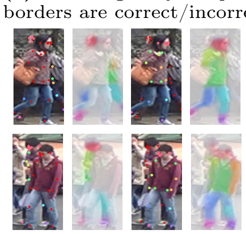
Id	Main components of KPR						Dataset	
	PBH	MSF	Prompt		BIPO	SOL	Occluded-PoseTrack	
			$\oplus$	$\ominus$			R-1	mAP
1							84.8	74.7
2	✓						83.3 (-1.5)	72.2 (-2.5)
3	✓	✓					84.8 (+0.0)	76.1 (+1.4)
4	✓	✓	✓				88.8 (+4.0)	80.4 (+5.7)
5	✓	✓	✓	✓			90.0 (+5.2)	80.7 (+6.0)
6	✓	✓			✓		86.1 (+1.3)	76.7 (+2.0)
7		✓	✓	✓	✓		80.4 (+5.1)	89.9 (+5.7)
8	✓	✓	✓	✓	✓		**92.3** (+6.5)	**82.3** (+7.6)
9	✓	✓	✓	✓	✓	✓	90.6 (+5.8)	81.2 (+6.5)

enabled with other modules (Exp. 8 vs Exp 7). Exp. 3 is consistent with claims from previous works [26,57] that increasing the feature map resolution is crucial to achieve fine-grained re-identification and better overall performance. Exp. 4 and 5 illustrate the benefits of positive and negative prompts to disambiguate the intended ReID target from other persons. Exp. 6 demonstrates the importance of our BIPO augmentation to generate artificial occlusions at training, since most occluded ReID benchmarks feature mainly occluded samples in their test set and not in their training set. Exp. 8, in comparison with Exp. 5 and 6, demonstrates how generating these artificial multi-person occlusions at training is crucial to teach the model to rely on the input prompt to identify the intended ReID target among all candidates. The usage of the SOLIDER foundation model as pre-trained weights is illustrated in Exp. 9. Unfortunately, SOLIDER does not enhance performance on the Occ-PTrack dataset, likely because of a domain gap: SOLIDER focuses on street surveillance, while Occ-PTrack predominantly features sports images from handheld cameras. However, as demonstrated in Tab. 1, SOLIDER does improve performance for all other datasets.

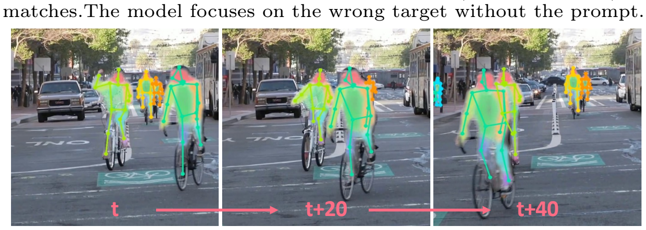
A Prompt-Optional Method In Figure 5, we analyze the test-time robustness of KPR against noisy or missing prompts on Occ-PTrack, by increasingly removing a higher percentage of keypoints from the input prompt. The red curve represents the performance of our proposed promptable KPR model, while the blue constant line illustrates the performance of the non-promptable version, which is unaffected by the proportion of keypoints removed from the input prompt. As illustrated, KPR significantly outperforms the non-promptable baseline with a full prompt and remains competitive with few or no keypoints, highlighting its prompt-optional capability. We provide an additional study of the performance w.r.t. the amount of multi-person occlusion in the supplementary materials.



(a) Ranked gallery samples for a given query, when the input prompt is disabled/enabled. Green/red borders are correct/incorrect matches.The model focuses on the wrong target without the prompt.



(b) Part-attention maps for various input prompts.



(c) Pose tracking results: all track IDs, depicted by unique skeleton colors, are maintained after a long occlusion by the front cyclist.

Fig. 6: Qualitative results on multi-person pose tracking and person retrieval. Red/pastel dots indicate negative/positive prompts. Colored heatmaps illustrates the part-attention maps of KPR described in Sec. 3.3, with one color per body part.

Qualitative Assessment Figure 6 demonstrates the impact of keypoint prompting on the part-attention maps, on person retrieval, and on pose tracking. As illustrated, KPR effectively leverages the input prompts to localize the intended target, and is therefore robust to occlusions and multi-person ambiguities.

5 Conclusion

In this work we propose KPR, the first promptable ReID model, designed to address occluded scenarios and multi-person ambiguities arising from bounding-box-based inputs. Our model offers practical flexibility by being prompt-optional, achieving state-of-the-art performance without prompts, and outperforming this baseline with prompts. Furthermore, our model outperforms the state-of-the-art on three popular ReID benchmarks and on our novel Occ-PTrack dataset. We finally demonstrate KPR's potential for pose tracking in videos. Our codebase, keypoint labels and proposed dataset will be released to encourage further research on this new promptable ReID paradigm.

Acknowledgment: This work has been funded by Sportradar, by the Walloon region project ReconnAIssance, and by the FNRS.

References

1. Adaimi, G., Kreiss, S., Alahi, A.: Deep visual re-identification with confidence. Transportation Research Part C: Emerging Technologies **126**, 103067 (May 2021). https://doi.org/10.1016/j.trc.2021.103067

2. Alahi, A., Ramanathan, V., Fei-Fei, L.: Tracking millions of humans in crowded spaces. Group and Crowd Behavior for Computer Vision pp. 115–135 (2017). https://doi.org/10.1016/b978-0-12-809276-7.00007-2
3. Bergmann, P., Meinhardt, T., Leal-Taixe, L.: Tracking without bells and whistles. In: IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 941–951. Inst. Electr. Electron. Eng. (IEEE) (Oct 2019). https://doi.org/10.1109/iccv.2019.00103
4. Chen, P., Liu, W., Dai, P., Liu, J., Ye, Q., Xu, M., Chen, Q., Ji, R.: Occlude them all: Occlusion-aware attention network for occluded person re-ID. In: IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 11813–11822. Inst. Electr. Electron. Eng. (IEEE), Montreal, QC, Canada (Oct 2021). https://doi.org/10.1109/iccv48922.2021.01162
5. Chen, W., Xu, X., Jia, J., Luo, H., Wang, Y., Wang, F., Jin, R., Sun, X.: Beyond appearance: A semantic controllable self-supervised learning framework for human-centric visual tasks. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 15050–15061. Inst. Electr. Electron. Eng. (IEEE), Vancouver, Can. (Jun 2023). https://doi.org/10.1109/cvpr52729.2023.01445
6. Doering, A., Chen, D., Zhang, S., Schiele, B., Gall, J.: PoseTrack21: A dataset for person search, multi-object tracking and multi-person pose tracking. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 20931–20940. Inst. Electr. Electron. Eng. (IEEE), New Orleans, LA, USA (Jun 2022). https://doi.org/10.1109/cvpr52688.2022.02029
7. Fu, D., Chen, D., Bao, J., Yang, H., Yuan, L., Zhang, L., Li, H., Chen, D.: Unsupervised pre-training for person re-identification. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 14745–14754. Inst. Electr. Electron. Eng. (IEEE), Nashville, TN, USA (Jun 2021). https://doi.org/10.1109/cvpr46437.2021.01451
8. Gao, S., Wang, J., Lu, H., Liu, Z.: Pose-guided visible part matching for occluded person ReID. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 11741–11749. Inst. Electr. Electron. Eng. (IEEE), Seattle, WA, USA (Jun 2020). https://doi.org/10.1109/cvpr42600.2020.01176
9. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: YOLOX: Exceeding YOLO series in 2021. arXiv **abs/2107.08430** (2021). https://doi.org/10.48550/arXiv.2107.08430
10. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 2917–2927. Inst. Electr. Electron. Eng. (IEEE), Nashville, TN, USA (Jun 2021). https://doi.org/10.1109/cvpr46437.2021.00294
11. Giancola, S., Cioppa, A., Deliège, A., Magera, F., Somers, V., Kang, L., Zhou, X., Barnich, O., De Vleeschouwer, C., Alahi, A., Ghanem, B., Van Droogenbroeck, M., Darwish, A., Maglo, A., Clapés, A., Luyts, A., Boiarov, A., Xarles, A., Orcesi, A., Shah, A., Fan, B., Comandur, B., Chen, C., Zhang, C., Zhao, C., Lin, C., Chan, C.Y., Hui, C.C., Li, D., Yang, F., Liang, F., Da, F., Yan, F., Yu, F., Wang, G., Chan, H.A., Zhu, H., Kan, H., Chu, J., Hu, J., Gu, J., Chen, J., Soares, J.V.B., Theiner, J., De Corte, J., Brito, J.H., Zhang, J., Li, J., Liang, J., Shen, L., Ma, L., Chen, L., Santos Marques, M., Azatov, M., Kasatkin, N., Wang, N., Jia, Q., Pham, Q.C., Ewerth, R., Song, R., Li, R., Gade, R., Debien, R., Zhang, R., Lee, S., Escalera, S., Jiang, S., Odashima, S., Chen, S., Masui, S., Ding, S., Chan, S.w., Chen, S., El-Shabrawy, T., He, T., Moeslund, T.B., Siu, W.C., Zhang, W., Li, W., Wang, X., Tan, X., Li, X., Wei, X., Ye, X., Liu, X., Wang, X., Guo, Y., Zhao, Y., Yu, Y., Li, Y., He, Y., Zhong, Y., Guo, Z., Li, Z.: SoccerNet 2022

challenges results. In: Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports. ACM (Oct 2022). https://doi.org/10.1145/3552437.3558545
12. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: TransReID: Transformer-based object re-identification. In: IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 14993–15002. Inst. Electr. Electron. Eng. (IEEE), Montreal, QC, Canada (Oct 2021). https://doi.org/10.1109/iccv48922.2021.01474
13. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv **abs/1703.07737** (2017). https://doi.org/10.48550/arXiv.1703.07737
14. Istasse, M., Somers, V., Elancheliyan, P., De, J., Zambrano, D.: DeepSportradar-v2: A multi-sport computer vision dataset for sport understandings. In: Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports. ACM (Oct 2023). https://doi.org/10.1145/3606038.3616160
15. Jia, M., Cheng, X., Zhai, Y., Lu, S., Ma, S., Tian, Y., Zhang, J.: Matching on sets: Conquer occluded person re-identification without alignment. In: AAAI Conf. Artif. Intell. vol. 35, pp. 1673–1681. Association for the Advancement of Artificial Intelligence (AAAI) (May 2021). https://doi.org/10.1609/aaai.v35i2.16260
16. Joos, V., Somers, V., Standaert, B.: TrackLab. https://github.com/TrackingLaboratory/tracklab (2024)
17. Kiran, M., Praveen, R.G., Nguyen-Meidine, L.T., Belharbi, S., Blais-Morin, L.A., Granger, E.: Holistic guidance for occluded person re-identification. arXiv **abs/2104.06524** (2021). https://doi.org/10.48550/arXiv.2104.06524
18. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv **abs/2304.02643** (2023). https://doi.org/10.48550/arXiv.2304.02643
19. Kreiss, S., Bertoni, L., Alahi, A.: PifPaf: Composite fields for human pose estimation. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 11969–11978. Inst. Electr. Electron. Eng. (IEEE), Long Beach, CA, USA (Jun 2019). https://doi.org/10.1109/cvpr.2019.01225
20. Leal-Taixé, L., Milan, A., Reid, I., Roth, S., Schindler, K.: MOTChallenge 2015: Towards a benchmark for multi-target tracking. arXiv **abs/1504.01942** (2015). https://doi.org/10.48550/arXiv.1504.01942
21. Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y., Wu, F.: Diverse part discovery: Occluded person re-identification with part-aware transformer. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 2897–2906. Inst. Electr. Electron. Eng. (IEEE), Nashville, TN, USA (Jun 2021). https://doi.org/10.1109/cvpr46437.2021.00292
22. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Eur. Conf. Comput. Vis. (ECCV). Lect. Notes Comput. Sci., vol. 8693, pp. 740–755. Springer Int. Publ. (2014). https://doi.org/10.1007/978-3-319-10602-1_48
23. Liu, Q., Xu, Z., Bertasius, G., Niethammer, M.: SimpleClick: Interactive image segmentation with simple vision transformers. In: IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 22233–22243. Inst. Electr. Electron. Eng. (IEEE), Paris, Fr. (Oct 2023). https://doi.org/10.1109/iccv51070.2023.02037
24. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 9992–10002. Inst. Electr. Electron. Eng. (IEEE), Montreal, QC, Canada (Oct 2021). https://doi.org/10.1109/iccv48922.2021.00986

25. Luiten, J., Osuep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., Leibe, B.: HOTA: A higher order metric for evaluating multi-object tracking. Int. J. Comput. Vis. **129**(2), 548–578 (Oct 2020). https://doi.org/10.1007/s11263-020-01375-2
26. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW). pp. 1487–1495. Inst. Electr. Electron. Eng. (IEEE), Long Beach, CA, USA (Jun 2019). https://doi.org/10.1109/cvprw.2019.00190
27. Ma, Z., Zhao, Y., Li, J.: Pose-guided inter- and intra-part relational transformer for occluded person re-identification. In: Proceedings of the 29th ACM International Conference on Multimedia. ACM (Oct 2021). https://doi.org/10.1145/3474085.3475283
28. Mansourian, A.M., Somers, V., De Vleeschouwer, C., Kasaei, S.: Multi-task learning for joint re-identification, team affiliation, and role classification for sports visual tracking. In: Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports. ACM (Oct 2023). https://doi.org/10.1145/3606038.3616172
29. Miao, J., Wu, Y., Liu, P., Ding, Y., Yang, Y.: Pose-guided feature alignment for occluded person re-identification. In: IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 542–551. Inst. Electr. Electron. Eng. (IEEE) (Oct 2019). https://doi.org/10.1109/iccv.2019.00063
30. Peng, Y., Hou, S., Cao, C., Liu, X., Huang, Y., He, Z.: Deep learning-based occluded person re-identification: A survey. arXiv **abs/2207.14452** (2022). https://doi.org/10.48550/arXiv.2207.14452
31. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (Apr 2015). https://doi.org/10.1007/s11263-015-0816-y
32. Seidenschwarz, J., Brasó, G., Serrano, V.C., Elezi, I., Leal-Taixé, L.: Simple cues lead to a strong multi-object tracker. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 13813–13823. Inst. Electr. Electron. Eng. (IEEE), Vancouver, Can. (Jun 2023). https://doi.org/10.1109/cvpr52729.2023.01327
33. Somers, V., Joos, V., Giancola, S., Cioppa, A., Ghasemzadeh, S.A., Magera, F., Standaert, B., Mansourian, A.M., Zhou, X., Kasaei, S., Ghanem, B., Alahi, A., Van Droogenbroeck, M., De Vleeschouwer, C.: SoccerNet game state reconstruction: End-to-end athlete tracking and identification on a minimap. In: IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports. Seattle, WA, USA (Jun 2024). https://doi.org/10.48550/ARXIV.2404.11335
34. Somers, V., Vleeschouwer, C.D., Alahi, A.: Body part-based representation learning for occluded person re-identification. In: IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV). pp. 1613–1623. Inst. Electr. Electron. Eng. (IEEE), Waikoloa, HI, USA (Jan 2023). https://doi.org/10.1109/wacv56688.2023.00166
35. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 5686–5696. Inst. Electr. Electron. Eng. (IEEE), Long Beach, CA, USA (Jun 2019). https://doi.org/10.1109/cvpr.2019.00584
36. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Eur. Conf. Comput. Vis. (ECCV). Lect. Notes Comput. Sci., vol. 11208, pp. 501–518. Springer Int. Publ. (2018). https://doi.org/10.1007/978-3-030-01225-0_30

37. Tan, H., Liu, X., Yin, B., Li, X.: MHSA-net: Multihead self-attention network for occluded person re-identification. IEEE Trans. Neural Networks Learn. Syst. **34**(11), 8210–8224 (Nov 2023). https://doi.org/10.1109/tnnls.2022.3144163
38. Van Zandycke, G., Somers, V., Istasse, M., Don, C.D., Zambrano, D.: DeepSportradar-v1: Computer vision dataset for sports understanding with high quality annotations. In: Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports. ACM (Oct 2022). https://doi.org/10.1145/3552437.3555699
39. Wang, G., Yang, S., Liu, H., Wang, Z., Yang, Y., Wang, S., Yu, G., Zhou, E., Sun, J.: High-order information matters: Learning relation and topology for occluded person re-identification. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 6448–6457. Inst. Electr. Electron. Eng. (IEEE), Seattle, WA, USA (Jun 2020). https://doi.org/10.1109/cvpr42600.2020.00648
40. Wang, T., Liu, H., Song, P., Guo, T., Shi, W.: Pose-guided feature disentangling for occluded person re-identification based on transformer. In: AAAI Conf. Artif. Intell. vol. 36, pp. 2540–2549. Association for the Advancement of Artificial Intelligence (AAAI) (Jun 2022). https://doi.org/10.1609/aaai.v36i3.20155
41. Wang, Z., Zhu, F., Tang, S., Zhao, R., He, L., Song, J.: Feature erasing and diffusion network for occluded person re-identification. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 4744–4753. Inst. Electr. Electron. Eng. (IEEE), New Orleans, LA, USA (Jun 2022). https://doi.org/10.1109/cvpr52688.2022.00471
42. Wang, Z., Zheng, L., Liu, Y., Li, Y., Wang, S.: Towards real-time multi-object tracking. In: Eur. Conf. Comput. Vis. (ECCV). Lect. Notes Comput. Sci., vol. 12356, pp. 107–122. Springer Int. Publ. (2020). https://doi.org/10.1007/978-3-030-58621-8_7
43. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer GAN to bridge domain gap for person re-identification. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 79–88. Inst. Electr. Electron. Eng. (IEEE), Salt Lake City, UT, USA (Jun 2018). https://doi.org/10.1109/cvpr.2018.00016
44. Yan, C., Pang, G., Jiao, J., Bai, X., Feng, X., Shen, C.: Occluded person re-identification with single-scale global representations. In: IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 11855–11864. Inst. Electr. Electron. Eng. (IEEE), Montreal, QC, Canada (Oct 2021). https://doi.org/10.1109/iccv48922.2021.01166
45. Yang, J., Zhang, J., Yu, F., Jiang, X., Zhang, M., Sun, X., Chen, Y., Zheng, W.S.: Learning to know where to see: A visibility-aware approach for occluded person re-identification. In: IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 11865–11874. Inst. Electr. Electron. Eng. (IEEE), Montreal, QC, Canada (Oct 2021). https://doi.org/10.1109/iccv48922.2021.01167
46. Yang, Q., Wang, P., Fang, Z., Lu, Q.: Focus on the visible regions: Semantic-guided alignment model for occluded person re-identification. Sensors **20**(16), 1–15 (Aug 2020). https://doi.org/10.3390/s20164431
47. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.H.: Deep learning for person re-identification: A survey and outlook. IEEE Trans. Pattern Anal. Mach. Intell. **44**(6), 2872–2893 (Jun 2022). https://doi.org/10.1109/tpami.2021.3054775
48. Zang, X., Li, G., Gao, W., Shu, X.: Learning to disentangle scenes for person re-identification. Image Vis. Comput. **116**, 104330 (Dec 2021). https://doi.org/10.1016/j.imavis.2021.104330
49. Zhang, G., Zhang, Y., Zhang, T., Li, B., Pu, S.: PHA: Patch-wise high-frequency augmentation for transformer-based person re-identification. In: IEEE/CVF Conf.

Comput. Vis. Pattern Recognit. (CVPR). pp. 14133–14142. Inst. Electr. Electron. Eng. (IEEE), Vancouver, Can. (Jun 2023). https://doi.org/10.1109/cvpr52729.2023.01358
50. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: ByteTrack: Multi-object tracking by associating every detection box. In: Eur. Conf. Comput. Vis. (ECCV). Lect. Notes Comput. Sci., vol. 13682, pp. 1–21. Springer Nat. Switz. (2022). https://doi.org/10.1007/978-3-031-20047-2_1, https://doi.org/10.1007/978-3-031-20047-2_1
51. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: FairMOT: On the fairness of detection and re-identification in multiple object tracking. Int. J. Comput. Vis. **129**(11), 3069–3087 (Sept 2021). https://doi.org/10.1007/s11263-021-01513-4
52. Zheng, K., Lan, C., Zeng, W., Liu, J., Zhang, Z., Zha, Z.J.: Pose-guided feature learning with knowledge distillation for occluded person re-identification. arXiv **abs/2108.00139** (2021). https://doi.org/10.48550/arXiv.2108.00139
53. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: IEEE Int. Conf. Comput. Vis. (ICCV). pp. 1116–1124. Inst. Electr. Electron. Eng. (IEEE), Santiago, Chile (Dec 2015). https://doi.org/10.1109/iccv.2015.133
54. Zheng, W.S., Li, X., Xiang, T., Liao, S., Lai, J., Gong, S.: Partial person re-identification. In: IEEE Int. Conf. Comput. Vis. (ICCV). pp. 4678–4686. Inst. Electr. Electron. Eng. (IEEE), Santiago, Chile (Dec 2015). https://doi.org/10.1109/iccv.2015.531
55. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 3652–3661. Inst. Electr. Electron. Eng. (IEEE), Honolulu, HI, USA (Jul 2017). https://doi.org/10.1109/cvpr.2017.389
56. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: AAAI Conf. Artif. Intell. vol. 34, pp. 13001–13008. Association for the Advancement of Artificial Intelligence (AAAI) (Apr 2020). https://doi.org/10.1609/aaai.v34i07.7000
57. Zhu, K., Guo, H., Liu, Z., Tang, M., Wang, J.: Identity-guided human semantic parsing for person re-identification. In: Eur. Conf. Comput. Vis. (ECCV). Lect. Notes Comput. Sci., vol. 12348, pp. 346–363. Springer Int. Publ. (2020). https://doi.org/10.1007/978-3-030-58580-8_21
58. Zhuo, J., Chen, Z., Lai, J., Wang, G.: Occluded person re-identification. arXiv **abs/1804.02792** (2018). https://doi.org/10.48550/arXiv.1804.02792