Deep Learning Based Occluded Person Re-Identification: A Survey

YUNJIE PENG, School of Computer Science and Technology, Beihang University, China
JINLIN WU and BOQIANG XU, Institute of Automation, Chinese Academy of Sciences, China
CHUNSHUI CAO and XU LIU, Watrix Technology Limited Co. Ltd., China
ZHENAN SUN, Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China
ZHIQIANG HE, School of Computer Science and Technology, Beihang University, China

Occluded person re-identification (Re-ID) focuses on addressing the occlusion problem when retrieving the person of interest across non-overlapping cameras. With the increasing demand for intelligent video surveillance and the application of person Re-ID technology, the real-world occlusion problem draws considerable interest from researchers. Although a large number of occluded person Re-ID methods have been proposed, there are few surveys that focus on occlusion. To fill this gap and help boost future research, this article provides a systematic survey of occluded person Re-ID. In this work, we review recent deep learning based occluded person Re-ID research. First, we summarize the main issues caused by occlusion as four groups: position misalignment, scale misalignment, noisy information, and missing information. Second, we categorize existing methods into six solution groups: matching, image transformation, multi-scale features, attention mechanism, auxiliary information, and contextual recovery. We also discuss the characteristics of each approach, as well as the issues they address. Furthermore, we present the performance comparison of recent occluded person Re-ID methods on four public datasets: Partial-ReID, Partial-iLIDS, Occluded-ReID, and Occluded-DukeMTMC. We conclude the study with thoughts on promising future research directions.

CCS Concepts: • Computing methodologies → Image representations; Object identification;

Additional Key Words and Phrases: Occluded person re-identification, partial person re-identification, literature survey, deep learning

ACM Reference format:
Yunjie Peng, Jinlin Wu, Boqiang Xu, Chunshui Cao, Xu Liu, Zhenan Sun, and Zhiqiang He. 2023. Deep Learning Based Occluded Person Re-Identification: A Survey. ACM Trans. Multimedia Comput. Commun. Appl. 20, 3, Article 73 (October 2023), 27 pages.
https://doi.org/10.1145/3610534

Authors’ addresses: Y. Peng and Z. He, School of Computer Science and Technology, Beihang University, 37 XueYuan Road, Haidian District, Beijing, China, 100191; emails: yunjiepeng@buaa.edu.cn, zqhe1963@gmail.com; J. Wu and B. Xu, Institute of Automation, Chinese Academy of Sciences, 95 ZhongGuanCun East Road, Haidian District, Beijing, China, 100190; emails: jinlin.wu@nlpr.ia.ac.cn, boqiang.xu@cripac.ia.ac.cn; C. Cao and X. Liu, Watrix Technology Limited Co. Ltd., 51 XueYuan Road, Haidian District, Beijing, China, 100191; emails: {chunshui.cao, xu.liu}@watrix.ai; Z. Sun, Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, 95 ZhongGuanCun East Road, Haidian District, Beijing, China, 100190; email: znsun@nlpr.ia.ac.cn.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
1551-6857/2023/10-ART73 $15.00
https://doi.org/10.1145/3610534

ACM Trans. Multimedia Comput. Commun. Appl., Vol. 20, No. 3, Article 73. Publication date: October 2023.
1 INTRODUCTION

Person Re-Identification (Re-ID) retrieves people of the same identity across non-overlapping cameras [2]. With the expanding deployment of surveillance cameras and the increasing demand for public safety, person Re-ID plays a crucial role in intelligent surveillance and has become a research hotspot in computer vision [81, 84, 86]. Academic datasets for person Re-ID usually provide full human detection boxes without any occlusions. However, in practical applications, occlusions are common and often lead to the dramatic accuracy drop of typical person Re-ID methods. To bridge the gap between academic study and real-world applications, occlusion has received considerable interest from researchers in recent years.

Occluded person Re-ID [103] and partial person Re-ID [95] aim to retrieve people of the same identity when given an occluded query. In this survey, following Xu et al. [75], we refer to these two research topics as occluded person Re-ID. With the advancement of deep learning, a large number of occluded person Re-ID methods have been proposed, whereas there are few surveys that focus on occlusion. To facilitate the understanding of the latest approaches and inspire new ideas in the field, this article summarizes the main challenges brought by occlusion and presents an in-depth analysis of occluded person Re-ID methods.

The challenges brought by occlusion are carefully summarized from the whole process of person Re-ID. Overall, there are four significant issues to be considered, namely position misalignment, scale misalignment, noisy information, and missing information. Each issue is illustrated in Figure 1(a), and details are introduced as follows.

Position Misalignment. Zheng et al. [95] point out the misalignment problem in occluded person Re-ID. Generally, the detected human boxes are resized by height to obtain the same size of input data for person retrieval. In the case of occlusion, the detected box contains only part of the human body while it undergoes the same data processing as that of a non-occluded person. Contents at the same position of the processed partial image and holistic image are likely to be mismatched, resulting in position misalignment.

Scale Misalignment. He et al. [16] raise the concern of the scale mismatching in occluded person Re-ID. Similar to position misalignment, scale misalignment also arises from the upstream data processing procedure. The occlusion may affect the height of the detected box and thus bring a bad effect on the ratio of resize operation in data processing, which leads to the scale misalignment between partial and holistic images.

Noisy Information. Wang et al. [71] point out the noise brought by occlusion. In the detected boxes of occluded pedestrians, obstacles are inevitably included in whole or in part and therefore bring noisy information for person Re-ID.

Missing Information. Hou et al. [23] raise the concern of information loss in occluded regions. In the detected boxes of occluded pedestrians, the identity information of occluded regions is lost, resulting in the missing information issue.

The preceding four issues are rarely considered in typical person Re-ID, which leads to the poor performance of methods under occlusion and promotes the specialized research on occluded person Re-ID. This article analyzes occlusion-related person Re-ID methods following six different types of solutions in Figure 1(b), namely matching, image transformation, multi-scale features, attention mechanism, auxiliary information, and contextual recovery. From an issue-oriented point of view, different types of solutions are briefly introduced as follows:

1. Matching-based solutions construct matching elements and designs matching strategies to obtain a valid comparison of components for Re-ID.
2. Image transformation-based solutions transform images from partial to holistic or from holistic to partial to reduce the discrepancy between partial and holistic ones for comparison.
Fig. 1. (a) The four issues caused by occlusion for person Re-ID. (b) The taxonomy of occluded person Re-ID methods from both the issue perspective and the solution perspective. Each issue is roughly connected with corresponding solutions by (dashed) lines.

(3) **Multi-scale features-based** solutions extract a feature at different scales to obtain robust representations to scale variation.

(4) **Attention mechanism-based** solutions learn to generate attention maps that provide position information for spatial alignment or noise suppression.

(5) **Auxiliary information-based** solutions mainly employ external information, such as pose locations and parsing masks, to guide the learning process for occluded person Re-ID.

(6) **Contextual recovery-based** solutions utilize contextual information to infer and recover occluded regions for Re-ID.

Specifically, we mainly review published publications of deep learning based occluded person Re-ID from top conferences and journals before July 2022. Meanwhile, we also introduce some methods from other conferences and journals as supplements. We discuss the issues caused by occlusion and provide an in-depth analysis of how these issues are addressed in recent works with evaluation results summarized accordingly. Particularly, we also summarize some typical person Re-ID methods that are closely related to occlusion for completing a more comprehensive survey.

The main contributions of this survey lie in three aspects:

- To help boost future research, we review occlusion-related person Re-ID methods and provide a systematic survey from the perspective of issues and solutions.
- We summarize and compare the performance of mainstream occluded person Re-ID approaches for researchers and industries to use based on their practical needs.
- We analyze and discuss the advantages and disadvantages of different types of solutions for occluded person Re-ID, and provide insights on promising directions for future research.

The rest of this article is organized as follows. Section 2 presents a summary of previous surveys and elaborates on efforts made by this work. Section 3 discusses the four significant issues caused by occlusion for person Re-ID. Section 4 reviews occlusion-related person Re-ID works following six different types of solutions. Section 5 summarizes the common datasets and evaluation metrics.
Table 1. Overview of Person Re-ID Surveys in Recent Years

| Survey   | Reference         | Occlusion Related          | Taxonomy                                                                 |
|----------|-------------------|----------------------------|--------------------------------------------------------------------------|
| Traditional |                  |                            |                                                                          |
| 2012 PRL [53] |                   | None                       | Feature Extraction/Cross-camera Calibration/Person Association          |
| 2014 IVC [3]  |                   | None                       | Contextual Methods (camera geometry info) Non-Contextual Methods (passive methods, active methods) |
| 2016 arXiv [39] |               | None                       | Image-Based Methods/Video-Based Methods                                |
| 2019 TCSVT [41] |               | None                       | Person Verification/Applications-Driven Methods (raw data, practical procedure, efficiency) |
| 2019 TPAMI [35] |               | None                       | Feature Extraction/Metric Learning/Multi-Shot Ranking                   |
| 2020 arXiv [40] |               | None                       | Identification Task/Verification Task                                  |
| 2021 arXiv [49] |               | None                       | Feature Learning/Model Architecture Design/Metric Learning/None function |
| 2021 IJCAI [70] |               | None                       | Low Resolution/Infrared/Sketch/Text                                   |
| Deep learning based |               |                            |                                                                          |
| 2022 TPAMI [82] | Several occluded Re-ID methods introduced in a small part of the Noise-Robust Re-ID | Closed-World Setting (deep feature representation learning, deep metric learning, ranking optimization) Open-World Setting (heterogeneous data, raw images or videos, unavailable or limited labels, open set, noisy annotation) |
| 2022 IVC [56]  | A few occluded Re-ID methods introduced in local feature or sequence-feature learning | Deep Metric Learning/Local Feature Learning/Generative Adversarial Networks/Sequence Feature Learning/Graph Convolutional Networks |

of occluded person Re-ID. Section 6 compares the performance of various methods. Section 7 provides insights on promising research directions. Section 8 presents our conclusion.

2 PREVIOUS SURVEYS

In previous literature, there are some surveys that have also reviewed the field of person Re-ID. To obtain a more comprehensive comparison, we summarize the surveys of person Re-ID since 2012. The occlusion-related contents and the taxonomies of these surveys are listed in Table 1. On the whole, previous surveys of person Re-ID can be roughly summarized into traditional surveys [3, 53] and deep learning based surveys [27, 35, 39–41, 48, 49, 56, 72, 82, 93].

Traditional surveys mainly review the person Re-ID methods that manually design feature extraction procedures and study better similarity measurements. Mazzon et al. [53] summarize four main phases for person Re-ID: multi-person detection, feature extraction, cross-camera calibration, and person association. Assuming that the first phase (i.e., multi-person detection) has been solved, methods of extracting color/texture/shape appearance features, grouping temporal information, conducting color/spatiotemporal cross-camera calibration, and distance/learning/optimization-based person association are reviewed accordingly. Bedagkar-Gala and Shah [3] regard the tracking across non-overlapping cameras and the identity retrieval as the open set matching problem and the close set matching problem, respectively. According to whether additional camera geometry/calibration information is available, methods are divided into contextual Re-ID for open set matching and non-contextual Re-ID for close set matching.

Deep learning based surveys mainly summarize the person Re-ID methods that manually design feature extraction techniques from different perspectives. A few of these surveys [35, 93] also review traditional methods for the sake of completeness. Leng et al. [41] regard person Re-ID as the task of person verification instead of person identification (i.e., the query person may not occur in the gallery set), and summarize relevant application-driven works from the aspects of raw data, practical procedure, and efficiency. Karanam et al. [35] incorporate the experimental protocol with the advances in both feature extraction and metric learning, and give an extensive review of single-shot and multi-shot Re-ID algorithms. Lavi et al. [40] summarize the deep learning based person Re-ID methods into single-stream feature learning and multi-stream feature learning, where the former considers the Re-ID as a standard classification problem and the latter learns more discriminative features in a pair or triplet units. Wang et al. [70] dedicate to heterogeneous person Re-ID
involving low-resolution, infrared, sketch, and text, and review relevant methods from the perspective of different application scenarios. Ye et al. [82] categorize person Re-ID into the closed-world settings with various research-oriented assumptions and the open-world settings with more challenging issues in real-world applications. They provide an in-depth analysis for closed-world person Re-ID from three different perspectives, including deep feature representation learning, deep metric learning, and ranking optimization, and discuss important open issues such as noisy annotation, unavailable labels, and heterogeneous data for open-world person Re-ID. In summary, previous deep learning based surveys have involved the loss design [27, 39], technical means [27, 56], data augmentation [39, 72], image and video [82, 93], classification and verification [39, 40, 72], open-world and closed-world [41, 82], multi-modality [41, 48], ranking optimization [35, 82], noisy annotation [82], unsupervised learning [49], and metric learning [27, 35, 56, 72, 82].

Despite such a number of surveys on person Re-ID, the occlusion that often happens in real-world scenarios and results in large performance drops has not drawn enough attention. Only a few surveys have a short glimpse of occluded person Re-ID. As far as we know, Ye et al. [82] have made a rough summary of occluded person Re-ID as a part of the Noise-Robust Re-ID. Ming et al. [56] have introduced several occluded person Re-ID methods in local feature learning or sequence feature learning. Considering the practical importance of occlusion for person Re-ID, a systematic investigation for occluded person Re-ID is essential. Therefore, we provide an in-depth survey of issues and solutions involved in occlusion-related person Re-ID works to help boost future research.

3 CHALLENGES BROUGHT BY OCCLUSION

Occluded person Re-ID aims at addressing the occlusion problem when retrieving the person of interest across non-overlapping cameras. In real-world applications, a person may be occluded by a variety of obstacles, such as cars, trees, and streetlights. The surveillance system often fails to capture the holistic person. The gap between holistic and occluded images brings great challenges to Re-ID.

In general, there are four issues to be considered when developing a solution for occluded person Re-ID: position misalignment, scale misalignment, noisy information, and missing information. The first two misalignment issues are mainly caused by upstream data processing: the detected box of a partial person undergoes the same alignment processing as that of a holistic person to obtain a consistent input size. The last two information issues directly raise from obstacles: in the detected boxes of occluded pedestrians, obstacles are inevitably included in whole or in part and the identity information of occluded regions is lost. Real-world examples are shown in Figure 1(a).

This section first summarizes the person Re-ID framework and gives the problem formulation based on metric learning. The four challenges brought by occlusion are then carefully discussed with formalized mathematical notations to help boost future research.

3.1 Person Re-ID Framework

Generally, person Re-ID adopts metric learning to construct a feature space where different samples of the same identity are closer than that of different identities. There are also classification-based [74] or verification-based [1] methods that regard person Re-ID as multi-class recognition or calculate pairwise similarities for Re-ID. However, the former generally requires that the people in the test set are trained before, whereas the latter performs pairwise comparisons in the test phase, which is time consuming. Therefore, we do not consider these two types of person Re-ID for illustration and use metric learning to formulate the person Re-ID instead.

Metric learning based person Re-ID aims at reducing the distance between samples with the same identity while enlarging the distance between samples with different identities in the feature
space. On the whole, the optimization goal of person Re-ID can be simplified as follows: for each sample in the feature space, the distance to a sample with the same identity is smaller than that to a sample with a different identity. Formally, the optimization objective of metric learning based person Re-ID can be formulated as

\[ \| f(x_i^a) - f(x_i^p) \|_2 < \| f(x_i^a) - f(x_i^n) \|_2, \]

(1)

where \( x_i \) denotes the \( i \)-th spatial part of sample \( x \). The \( a, p, \) and \( n \) denote anchor, positive, and negative, respectively—that is, \( x_i^a \) and \( x_i^p \) are with the same identity, whereas \( x_i^p \) and \( x_i^n \) are with different identities. The \( f(\cdot) \) indicates feature generation, and \( \| \cdot \|_2 \) represents L2 norm. In addition to L2 norm (i.e., Euclidean distance), other distance metrics can also be used (e.g., cosine distance and Manhattan distance). For clarity, we employ L2 norm for illustration.

### 3.2 Position Misalignment

In the case of occlusion, the detected box of a person contains only part of the human body while it undergoes the same alignment processing as that of a non-occluded person: the detected human boxes are resized by height to obtain the same size of input data for person retrieval. The contents at the same position of the processed partial image and holistic image are likely to be mismatched. Technically speaking, the position misalignment may lead to mismatches between samples in Equation (1), which can be formulated as

\[ P(x_i^a) \neq P(x_i^p) \mid P(x_i^a) \neq P(x_i^n), \]

(2)

where \( P \) determines the semantic meaning of the \( i \)-th spatial part for sample \( x \). The distance between mismatched semantic parts of different samples (i.e., \( \| f(x_i^a) - f(x_i^p) \|_2 \) or \( \| f(x_i^a) - f(x_i^n) \|_2 \) in Equation (1)) is meaningless for comparison.

### 3.3 Scale Misalignment

Similar to position misalignment, scale misalignment also arises from the upstream data processing procedure. The occlusion may affect the height of the detected box and thus influence the ratio of resize operation in data processing, resulting in scale mismatches between partial and holistic images. Formally, given \( S \) to determine the scale of the \( i \)-th spatial part for sample \( x \), the scale mismatches can then be formulated as

\[ S(x_i^a) \neq S(x_i^p) \mid S(x_i^a) \neq S(x_i^n). \]

(3)

In this case, scale misalignment requires extra efforts to associate features at different scales for comparison. It will be harder to learn a feature space that pulls the \( f(x_i^a) \) and \( f(x_i^p) \) together while pushing the \( f(x_i^a) \) and \( f(x_i^n) \) away.

### 3.4 Noisy Information

In detected boxes of occluded pedestrians, obstacles are inevitably included in whole or in part and bring noisy information for person Re-ID. Under the challenge of the noisy information issue, metric learning based person Re-ID can be reformulated as

\[ \| f(x_{i,o}^a, x_{i,r}^a) - f(x_{i,o}^p, x_{i,r}^p) \|_2 < \| f(x_{i,o}^a, x_{i,r}^a) - f(x_{i,o}^a, x_{i,r}^n) \|_2, \]

(4)

where \( x_{i,o} \) and \( x_{i,r} \) denote the occluded region and the clean region of the \( i \)-th spatial part for sample \( x \), respectively. Based on Equation (4), whether samples of the same identity are with different obstacles (i.e., the \( x_{i,o}^a \) is quite different from \( x_{i,o}^p \) or samples of different identities are with similar occlusions (i.e., the \( x_{i,o}^a \) is similar to \( x_{i,o}^n \)) can both have a negative impact on the optimization objective.
3.5 Missing Information

In detected boxes of occluded pedestrians, the identity information of occluded regions is missing. With the missing information issue caused by occlusion, metric learning based person Re-ID can be reformulated as

\[
\| f(x^a_i - x^a_{i,m}) - f(x^p_i - x^p_{i,m}) \|_2 < \| f(x^a_i - x^a_{i,m}) - f(x^n_i - x^n_{i,m}) \|_2,
\]

where \( x_i \) denotes the ideal clean \( i \)-th spatial part of sample \( x \) and \( x_{i,m} \) denotes the missing region under occlusion. From a semantic point of view, once the missing information of samples for comparison is not consistent (i.e., \( x^a_{i,m} \neq x^p_{i,m} \) or \( x^a_{i,m} \neq x^n_{i,m} \)), the difficulty of building a metric-based feature space for Re-ID will be increased.

4 METHODOLOGIES REVIEW

Deep learning based methods to address issues caused by occlusion in person Re-ID can be summarized following six different types of solutions: matching \([16, 17, 30, 33, 47, 68, 77, 79, 95]\), image transformation \([20, 21, 26, 47, 52, 97]\), multi-scale features \([16, 18, 79, 90]\), attention mechanism \([6, 29, 36, 37, 42, 45, 61, 63, 64, 67, 69, 71, 78, 88, 98, 102, 103]\), auxiliary information \([4, 10, 11, 14, 17, 21, 23, 24, 33, 34, 47, 50, 54, 55, 58, 65, 68, 75–77, 85, 87, 91, 99, 100]\), and contextual recovery \([22, 23, 51, 75]\). From an issue-oriented point of view, six types of solutions are briefly introduced as follows. First, the matching-based solution designs various matching elements, as well as matching strategies, to obtain a valid comparison of components for Re-ID. Second, the image transformation based solution transforms images from partial to holistic or from holistic to partial to reduce the discrepancy between partial and holistic images for comparison. Third, the multi-scale features based solution extracts a feature at different scales for the construction of robust representations to scale variation. Fourth, the attention mechanism based solution learns to generate attention maps that provide position information for spatial alignment or noise suppression. Fifth, the auxiliary information based solution mainly employs external information, such as pose locations and parsing masks, to guide the learning process for occluded person Re-ID. Sixth, the contextual recovery based solution utilizes contextual information to infer and recover occluded regions for Re-ID. Further details about the taxonomy are shown in Figure 2.

Following the preceding six different types of solutions, this section reviews occlusion-related methods and discusses the issues they addressed with theoretical formulations. It should be noted that some methods involving more than one technical route will be introduced multiple times from different perspectives accordingly.

4.1 Matching

The main points of a matching-based solution can probably be summarized into matching elements and matching strategies. Given \( G \) to denote the function that builds matching elements for sample
x, the optimization objective for a matching-based solution can be formulated as

$$M_i(G(x^a), G(x^p)) < M_i(G(x^a), G(x^n)),$$

where $M_i$ denotes the distance between the $i$-th semantic components of samples associated by matching strategies. Different from Equation (1), which assumes that samples are in good condition and directly compares features extracted from the same spatial position, the preceding formula matches components of the same semantic meaning across different samples for comparison.

Diverse definitions of the matching element, as well as the matching strategy, have been proposed to address issues caused by occlusion. On the whole, matching-based methods can be further grouped into sliding window matching [95], shortest path matching [77], reconstruction-based matching [16, 95], denoising matching [17, 33, 47], graph-based matching [68, 79], and set-based matching [30]. Examples are shown in Figure 3.

**Sliding window matching** [95] treats the partial probe image as a whole and slides it exhaustively over a full-body gallery image to match the most similar local region. The distance between a partial image and its most similar match on a full-body image is employed for comparison.

**Shortest path matching** [77] performs the matching by calculating the shortest path between two sets of local features and uses the matched local features to compute the similarity, explicitly accomplishing spatial alignment.

**Reconstruction-based matching** [16, 95] assumes that the identity information in an occluded image is a subset of that in a non-occluded image. The partial image can then be reconstructed from a holistic image with the same identity. AMC [95] decomposes images into regular grid patches and optimizes the selection of gallery patches for reconstructing each probe patch. The reconstruction difference of each probe patch based on gallery patches of different identities is employed to determine the identity. DSR [16] defines $c \times c$ pixels on a feature map as an independent block for matching. Each block of a partial probe image is assumed to be well reconstructed from the sparse linear combination of blocks of a full-body gallery image with the same identity.

**Denoising matching** [17, 33, 47] proposes to focus on foreground visible human parts while discarding noisy information during matching. GASM [17] designs a combination of pose-guided and mask-guided layers to generate saliency heatmaps, guiding spatial matching to assign the foreground human parts with larger weights adaptively. Co-attention [47] performs the matching between a partial and a holistic image under the guidance of body parsing masks. Particularly, the self-attention mechanism [66] is adopted in co-attention matching where the parsing mask of
the partial image is viewed as the query, and parsing masks and CNN features of the partial and the holistic images serve as the key and the value, respectively. ASAN [33] replaces segmentation masks of holistic gallery images with the mask of a partial probe image in each retrieval matching process to suppress the interference from useless parts.

Graph-based matching [68, 79] takes advantage of the graph structure and formulates the occluded person Re-ID as a graph matching problem. The graph structure is a powerful mathematical abstraction [32], which not only represents information about individuals but also captures the interactions between different parts of an individual for recognition. HOReID [68] regards the semantic local features extracted by key point heatmaps as nodes of a graph and designs a Graph Convolutional Network (GCN) with learnable adjacent matrices to capture high-order relation information. For measuring the similarity between two graphs, node features of the two graphs are further processed with the relevant information extracted from each other to learn topology information. Both the high-order relation information captured by the GCN and the topology information learned from one another are employed for final Re-ID. MGH [79] designs multiple hypergraphs with different spatial and temporal granularities to address the issues brought by occlusion for video-based person Re-ID. Different from conventional graphs, hypergraphs [31] can model high-order dependencies involving multiple nodes and are more suitable for extracting multi-granular correlations in a sequence.

Set-based matching [30] takes occluded person Re-ID as a set matching task. MoS [30] employs a CNN backbone to capture diverse visual patterns along the channel dimension as matching elements and adopts the Jaccard similarity coefficient to measure the distance between pattern sets of person images. Specifically, minimization and maximization are used to approximate the operations of intersection and union of sets, and the ratio of intersection over union is calculated to measure the similarity of two sets.

4.2 Image Transformation

Image transformation based methods aim to bridge the gap between partial and holistic images through partial transformation [20, 21, 47, 97] or holistic transformation [26, 52]. The problem formulation for an image transformation based solution can be formulated as

\[ \| f(T(x^a)_i) - f(T(x^n)_i) \|_2 < \| f(T(x^a)_i) - f(T(x^n)_i) \|_2, \]

where \( T \) denotes the image transformation module and \( i \) indicates the \( i \)-th spatial part of transformed sample \( T(x) \). The \( a, p, n, f(\cdot) \), and \( \| \cdot \|_2 \) are the same as in Equation (1). Through the image transformation, the \( i \)-th spatial parts across different samples are expected to be semantically aligned where issues caused by occlusion can be easily addressed (e.g., \( P(T(x)_a^p) = P(T(x)_a^n) \) and \( S(T(x)^a_p) = S(T(x)^a_n) \)) to meet the challenges in Equations (2) and (3), respectively.

Partial transformation [20, 21, 47, 97] aims to transform partial images to obtain the rectified results that are spatially aligned with holistic ones for Re-ID, as shown in Figure 4(a). APNet [97] constructs automatic data augmentation to train a bounding box aligner that predicts four offset values (i.e., top, bottom, left, and right) for shifting the detected bounding boxes to cover the holistic body. PPCL [20] randomly crops holistic images to generate partial images with known transformation coefficients and designs a gated transformation regression CNN module to infer the transformation parameters in a self-supervised manner. The image rescaler in the work of Lin and Wang [47] randomly crops the holistic image and masks the uncropped regions to supervise the prediction of 2D affine transformation parameters, transforming partial images into desirable distortion-free images for addressing the spatial misalignment issue. Differently, ACSAP [21] designs a pose-guided generator that utilizes extra pose information to guide the generation of aligned features for partial images.
Fig. 4. Image transformation based methods transform a holistic image to the target partial one conditioned on a given partial image (a) and transform a partial image to the target holistic one conditioned on a given holistic image (b). Attribute-based auxiliary associates the attribute annotations (e.g., gender, hat, and backpack) with Re-ID (c).

Holistic transformation [26, 52] aims to transform corresponding regions of holistic images to obtain the rectified results that are spatially aligned with partial images for Re-ID, as shown in Figure 4(b). STNReID [52] randomly crops holistic images to construct the self-supervised training of predicting 2D affine transformation parameters that are used to transform holistic images for matching with partial ones. APN [26] defines the cropping type of a partial image and randomly crops holistic images into different types for self-supervision. Based on the predicted cropping type of a partial probe image, APN crops the corresponding region of holistic gallery images for Re-ID.

4.3 Multi-Scale Features

Multi-scale features based methods [16, 18, 67, 79, 90] extract features at different scales to obtain robust representations for Re-ID (as shown in Figure 5), which can be reformulated as

\[
\left\| f_s(x^a_i) - f_s(x^p_i) \right\|_2 < \left\| f_s(x^a_i) - f_s(x^n_i) \right\|_2,
\]

where \( f_s(\cdot) \) generates features at scale \( s \). Compared with the single-scale features extracted by \( f(\cdot) \) in Equation (1), multi-scale features generated from \( \{ f_s(\cdot) | s = 1, 2, \ldots \} \) provide different scales of information, being more robust for person Re-ID.

In the literature, some multi-scale features based methods [79, 90] extract multi-scale pyramid features from global to local. The pyramidal model in the work of Zheng et al. [90] horizontally slices the feature map into \( n \) basic parts and builds new branches for every \( l \in \{1, 2, \ldots, n\} \) adjacent part to extract features. MGH [79] hierarchically divides the feature map into \( p \in \{1, 2, 4, 8\} \) horizontal strips and average pools each strip to obtain multi-granular spatial features. Other multi-scale features based methods [16, 18, 67] focus on restricted local regions and maintain features of different receptive fields at the same position. DSR [16] average pools the square area of \( s \times s \) pixels on a feature map, where \( s = \{1, 2, 3\} \), to obtain multi-scale block representations, alleviating the influence of scale mismatching. FPR [18] employs multiple max-pooling layers of different kernel sizes to capture diverse spatial features from small local regions to relatively large regions. SBPA [67] designs the scale-wise residual connection to maintain features at different scales on each pixel for comparison.

4.4 Attention Mechanism

The attention mechanism based solution learns to generate attention maps that provide position information for alignment or noise suppression. It should be noted that some methods have also
involved the attention mechanism but rely on external information provided by auxiliary models or use additional information for supervision. These methods are not included in this subsection and will be introduced in Section 4.5 on auxiliary information. The problem formulation in Equation (1) for an attention mechanism based solution can be reformulated as

$$\| f(x_i^a \cdot Att(x_i^a)) - f(x_i^p \cdot Att(x_i^p)) \|_2^2 < \| f(x_i^n \cdot Att(x_i^n)) - f(x_i^n \cdot Att(x_i^n)) \|_2^2,$$

(9)

where $Att$ generates the attention map of $x_i$ to figure out the position of semantic parts for alignment or to distinguish the identity information for noise suppression. The $a$, $p$, $n$, $f(\cdot)$, and $\| \cdot \|_2$ are the same as in Equation (1).

According to the main points of an attention learning process, the attention mechanism based solution can be further grouped into cropping-based [36], clustering-based [102], dropping-based [63], relation-based [37, 69], data augmentation [6, 29, 61, 71, 78, 88, 98, 103], and constraint-based [42, 45, 64, 67] methods. Examples are shown in Figure 6.

**Cropping-based attention** [36] crops images into different local regions and learns attention to find the same local regions across different images for similarity calculation. Specifically, DPPR [36] crops 13 pre-defined partial regions on holistic images and designs an attention module conditioned on the partial probe image to assign the partial regions with larger attention weights if the same body parts are included.

**Clustering-based attention** [102] generates pseudo-labels from clustering to supervise the attention learning. Based on the assumption that the foreground pixels have higher responses than the background ones, ISP [102] designs the cascaded clustering on CNN feature maps to gradually generate pixel-level pseudo-labels of human parts. The clustered pixel-level pseudo-labels of human parts are then employed to guide the partial attention learning.

**Dropping-based attention** [63] develops the training strategy based on dropping to guide the network to learn a more robust representation for Re-ID. Specifically, CBDB-Net [63] uniformly partitions the feature map into strips and continuously drops each strip from top to bottom. Trained with drop-based incomplete features, the model is forced to learn a more robust person descriptor.

**Relation-based attention** [19, 37, 69] mines the relation among different regions to refine features, alleviating the interference of occlusion. Specifically, OCNet [37] pre-defines the global region, top region (1/2 top horizontal strip), bottom region (1/2 bottom horizontal strip), and center region (1/3 center vertical strip) on an image, and extracts four region features through group convolution and a carefully designed attention mechanism. In OCNet, the relational adaptive module consisting of two fully connected shared layers is proposed to capture the relation between different region features. The relational weights are then used to refine region features for noise suppression.
suppression. Differently, SpaRs [69] argues that the global configuration of objects is more robust to problems, such as partial occlusions, and is more appropriate than previous attention-based approaches that attend to focus only on the most discriminative regions. In SpaRs, the proposed spatial rescaling layer first performs the summation over the channels for each spatial position, which represents the importance of that position. Then the summed activations are normalized and used to rescale the magnitude of the activations on the original feature map. In this way, the interdependencies between the spatial positions are explicitly captured to scatter the focus, being more robust to occlusions. Furthermore, TransReID [19] encodes an image as a sequence of patches and builds relations among different patches through a pure transformer-based ReID framework. To achieve robust feature learning in the context of transformers, patch embeddings are further rearranged via shift and patch shuffle operations, and learnable side information embeddings are introduced for mitigating feature bias toward camera/view variations.

Data augmentation based attention [6, 29, 61, 71, 78, 88, 98, 103] generates artificial occlusion to train the network in a self-supervised manner to focus on clean body parts and exclude noisy obstacles. Specifically, AFPB [103] constructs an occlusion simulator where a random patch of the background is used as an occlusion to cover a part of holistic persons. With the carefully designed multi-task loss, AFPB simultaneously learns to identify the person and classify whether the sample is from an occluded data distribution. Given a holistic image, VPM [61] defines $m \times n$ rectangle regions and randomly crops a patch in which every pixel is assigned with the region label accordingly. Based on the constructed self-supervision, VPM trains a region locator to generate a probability map for each region and sums them up to obtain the region visibility scores for noise suppression. RE [98] proposes the operation of random erasing, which randomly selects a rectangle region in an image and erases its pixels with random values, to generate images with various levels of occlusion for training the model to extract non-occluded discriminative identity information. IGOAS [88] designs an incremental generative occlusion block that randomly generates a uniform occlusion mask from small to large on images in a batch. With the synthesized occlusion data and their
corresponding occlusion masks, IGOAS is trained to focus more on foreground information by suppressing the response of generated occlusion regions to zero. SSGR [78] employs random erasing and batch-constant erasing, which equally divide images into horizontal strips and randomly erase the same strip in a sub-batch, to simulate occlusion for training the disentangled non-local (DNL [83]) attention network. OAMN [6] designs a novel occlusion augmentation scheme that crops a rectangular patch of a randomly chosen training image and scales the patch onto four pre-defined locations of the target image, producing diverse and precisely labeled occlusion. With the supervision of labeled occlusion data, OAMN learns to generate spatial attention maps that precisely capture body parts regardless of the occlusion. DRL-Net [29] exploits obstacles present in then training set to synthesize more realistic occluded samples for guiding the contrast feature learning to mitigate the interference of occlusion noise. Differently, FED [71] considers the occlusion from not only the non-pedestrian obstacles but also the non-target pedestrians for augmentation. To simulate non-pedestrian occlusions, FED manually crops the patches of backgrounds and occlusion objects from training images, and pastes them on pedestrian images. To synthesize non-target pedestrian occlusions, FED maintains a memory bank of the feature centers of different identities and diffuses the current pedestrian features with \( k \)-nearest features in the memory bank.

\textit{Constraint-based attention} [42, 45, 64, 67] builds constraints among different attention maps to help locate diverse body parts for Re-ID. The multiple spatial attentions in the work of Li et al. [42] employ a diversity regularization term on attention maps to ensure that each attention focuses on different spatial regions. SBPA [67] separates local attention maps by minimizing the L1 norm distance between the local attention and the masked global attention. PAT [45] maintains vectors of part prototypes to generate part-aware attention masks on contextual CNN features and designs the part diversity mechanism to help achieve diverse part discovery. MHSA-Net [64] proposes the feature diversity regularization term to encourage the diversity of local features captured by a multi-head self-attention mechanism.

### 4.5 Auxiliary Information

The auxiliary information based solution employs external information to guide the learning process for occluded person Re-ID. The problem formulation in Equation (1) for an auxiliary information based method can be reformulated as

\[
\left\| f(x^a_i, Aux^a_i) - f(x^p_i, Aux^p_i) \right\|_2 < \left\| f(x^a_i, Aux^a_i) - f(x^n_i, Aux^n_i) \right\|_2,
\]

where \( Aux_i \) denotes the auxiliary information of the \( i \)-th spatial part for sample \( x \). The \( a, p, n, f(\cdot) \), and \( \|\cdot\|_2 \) are the same as in Equation (1). With the help of auxiliary information \( Aux_i \) (e.g., pose locations or parsing masks), \( f(\cdot) \) is expected to employ position guidance for spatial alignment or prediction confidence for noise suppression as shown in Figure 7.

According to the type of employed auxiliary information, methods can be further summarized into pose-based [11, 21, 23, 50, 54, 55, 65, 68, 75–77, 85, 87, 91, 100], parsing-based [34, 47], segmentation-based [4, 24], attribute-based [33], and hybrid-based [10, 17, 58, 99] solutions. Examples are shown in Figure 4(c) and Figure 7.

**Pose-based auxiliary** [11, 21, 23, 50, 54, 55, 65, 68, 75–77, 85, 87, 91, 100] exploits additional pose information to assist in addressing challenges brought by occlusion. PGFA [54], PDVM [100], and PMFB [55] generate heatmaps consisting of a 2D Gaussian centered on key point locations to extract aligned pose features through a dot product with the CNN feature map. Based on the prediction confidence, these methods set a threshold to filter out occluded invisible pose landmarks, and only shared visible parts between probe and gallery images are used to compute the similarity, explicitly avoiding the disturbance from occlusion. Similarly, KBFM [14] utilizes shared visible key points between images to locate aligned rectangular regions for calculating the similarity.
Fig. 7. (a) Pose-based auxiliary uses key point coordinates to address the position misalignment issue and employs confidence scores to exclude the noisy information. (b) Parsing-based auxiliary employs semantic masks to locate positions and suppress noisy information.

DSAG [87] and PGFL-KD [91] use the features located by pose information to guide the learning process. DSAG constructs a set of densely semantically aligned part images with the external pose information provided by DensePose [13]. Taking the semantically aligned part images as the input, DSAG serves as a regulator to guide the feature learning on original images through the carefully designed fusion and loss. Similarly, PGFL-KD uses external key point heatmaps to extract semantically aligned features. The aligned features are then employed to regularize the global feature learning through knowledge distillation and interaction-based training. Differently, RFCNet [23] trains an adaptive partition unit supervised by external pose information to split the CNN feature map into different regions and extract region features for further processing. PVPM [11] uses key point heatmaps and part affinity fields estimated by OpenPose [5] to generate attention maps for extracting aligned part features. In PVPM, the characteristic of part correspondence between images of the same identity is employed to mine correspondence scores as pseudo-labels, training a visibility predictor that estimates whether a part suffers from occlusion. PGMANet [85] generates heatmaps of key point locations to calculate part attention masks on the CNN feature map. Based on the part features aggregated by part attention masks, PGMANet further computes the correlation among different part features to exploit the second-order information for more robust feature extraction. HOREID [68] employs the key point heatmaps to extract semantic local features on a target image. The local features of an image are taken as nodes of a graph, and HOREID designs a GCN with learnable adjacent matrices to exploit the more discriminative relation information among nodes for Re-ID. FRT [75] exploits the pose information to divide a person image into three parts: head part, torso part, and leg part, and employs a GCN to suppress the meaningless message of occluded parts while enhancing the meaningful features of shared regions. Specifically, the adjacent matrix of the GCN in FRT is specially designed in which the information of a node (i.e., a part) spreads less when its visibility score (i.e., averaged confidence scores for all key points in the corresponding part) is lower. CTL [50] proposes to divide the human body into three granularities and use key point heatmaps to extract multi-scale part features as graph nodes. Cross-scale graph convolution and 3D graph convolution are designed to capture structural information and hierarchical spatiotemporal dependencies for addressing spatial misalignment issues in video person Re-ID. ACSAP [21] employs external pose information to guide the adversarial generation of aligned feature maps. The confidence of pose estimation is used to decide the visibility of horizontally partitioned parts, and ACSAP assigns the shared visible parts with larger weights in similarity metrics for noise suppression. PFD [65] employs the Vision Transformer (ViT) [7] to generate local patch features and element-wisely multiplies them with the processed key point heatmaps to obtain pose-guided features. Through the measurement of set similarity, PFD performs the matching
between local features and pose-guided features to disentangle the pose information from patch features for position alignment. In addition, a set of learnable semantic views is introduced in the transformer decoder to implicitly enhance the disentangled body part features. Moreover, PFD designs a pose-guided push loss to encourage the difference between human body parts (i.e., the high-confidence features) and non-human body parts (i.e., the low-confidence features) to help focus on human body parts and alleviate the interference of occlusion. AACN [76] and DAReID [77] learn part attention maps to locate and extract part features with the ground truth built from external pose information. AACN [76] uses external pose information to supervise the part attention learning and computes part visibility scores based on the intensities of each part attention map to measure the occlusion extent of each body part. With the ground truth built from external pose information, DAReID [77] learns to predict upper and lower body masks to extract non-occluded part features. The heatmap of upper or lower features with large high-activation areas is regarded as reliable, and DAReID enhances these regions to suppress the occlusion noise.

Parsing-based auxiliary [34, 47, 59] takes advantage of human parsing masks and accomplishes spatial alignment or suppresses noisy obstacles for occluded person Re-ID. SPReID [34] trains an extra semantic parsing model on the human parsing dataset LIP [12] to predict probability maps that are associated with five pre-defined semantic regions of the human body. These probability maps are then used to extract different semantic features through the weighted sum operation on the feature map generated by a modified Inception-V3 [62]. Co-attention [47] employs parsing masks as the query in the self-attention mechanism to perform image matching on associated regions for alleviating the noisy information brought by occlusion. BPSReID [59] finds that local appearance is not necessarily discriminative and that standard ReID losses for global representation learning do not scale well to local representation learning. Based on the body part based features trained with labels generated by PifPaf [38], BPSReID designs a global-identity local-triplet loss to learn a set of local features that are each representative of their corresponding local parts while being discriminative when considered jointly.

Segmentation-based auxiliary [4, 24] utilizes extra segmentation masks provided by the segmentation model (e.g., the human parsing model or scene segmentation model) to guide the learning process for addressing challenges brought by occlusion. Specifically, MMGA [4] learns to generate the whole-body, upper-body, and lower-body attention maps with the parsing labels estimated by JPPNet [46]. The learned attention maps are then employed to extract global and local features for Re-ID. HPNet [24] introduces human parsing as an auxiliary task and employs the parsing masks to extract part-level features for addressing the position misalignment issue. The person Re-ID and the human parsing are learned in a multi-task manner where the pseudo parsing label is predicted by the scene segmentation model [9] trained on COCO DensePose [13] dataset.

Attribute-based auxiliary [33] leverages semantic-level attribute annotations of person Re-ID datasets to help suppress the noisy information brought by occlusion (shown in Figure 4(c)). ASAN [33] employs attribute annotations to guide the learning of occlusion-sensitive segmentation in a weakly supervised manner to extract non-occluded human body features for addressing the noisy information issue.

Hybrid-based auxiliary [10, 17, 58, 99] employs two or more types of auxiliary information for alleviating issues caused by occlusion. SSP-ReID [58] exploits the capabilities of both clues (i.e., the saliency and the semantic parsing) to guide the CNN backbone to learn complementary representations for Re-ID. Specifically, external off-the-shelf deep models [12, 44] are employed to generate the semantic parsing masks and the saliency masks. In addition, the element-wise product is applied between masks and a CNN feature map to obtain parsing features and saliency features for fusion. TSA [10] employs HRNet [60] and DensePose [13] to provide extra key point information and body part information. Based on key point locations, TSA divides the whole person into five
regions and obtains region features by soft region pooling. Based on the body part masks, TSA extracts corresponding region features on the texture image produced by a texture generator. The region features guided by the key points are then concatenated with the region features guided by the part masks accordingly to generate robust representations for Re-ID. GASM \cite{17} trains a mask layer and a pose layer with the ground truth predicted by the semantic segmentation model PSP-Net \cite{89} and the pose estimation model CenterNet \cite{101}. GASM then combines the mask heatmap predicted by the mask layer and the key point heatmaps estimated by the pose layer into a saliency map to extract salient features, explicitly excluding the noisy information brought by occlusion. FGSA \cite{99} mines fine-grained local features with the supervision of both the pose information and the attribute information to address the position misalignment issue. Specifically, FGSA designs a pose resolve net to provide part confidence maps and part affinity fields of the key parts. These part maps are then used to extract part features on a CNN feature map through compact bilinear pooling. Given the extracted part features, FGSA treats the attribute recognition as multiple classification tasks and trains an intermediate model for attribute classification along with the person Re-ID.

4.6 Contextual Recovery

The contextual recovery-based solution utilizes contextual information to infer and recover occluded regions for Re-ID, which can be formulated as

$$\| f(\text{Rec}(x_i^a, \text{Contex}_i^a)) - f(\text{Rec}(x_i^b, \text{Contex}_i^b)) \|_2 < \| f(\text{Rec}(x_i^a, \text{Contex}_i^a)) - f(\text{Rec}(x_i^n, \text{Contex}_i^n)) \|_2 ,$$

(11)

where \(\text{Rec}(\cdot)\) recovers occluded regions of \(x_i\) with its contextual information \(\text{Contex}_i\). The \(a, p, n\), \(f(\cdot)\), and \(\| \cdot \|_2\) are the same as in Equation (1). On the whole, contextual recovery based methods exploit contextual information to generate recovered results that are as identical to clean ones as possible—that is, minimizing the difference between holistic and occluded images for better comparison.

In the literature, most contextual recovery based methods \cite{22, 23, 51} employ spatial or temporal information to infer occluded regions for recovery (Figure 8). VRSTC \cite{22} designs an auto-encoder, which takes the image masked with white pixels as input, to generate the content of occluded regions. To improve the quality of generated content, VRSTC adopts both global and local discriminators that adversarially judge the reality and the contextual consistency of the synthesized content. In addition, VRSTC proposes a differentiable temporal attention layer that employs...
cosine similarity to determine where to attend from adjacent frames for recovering the content of the occluded parts. Assuming that the information of different frames at the same position in a continuous period of time can help recover the lost information at the current frame, the refining recurrent unit [51] implicitly refers to the appearance and motion information extracted from historical frames to remove noise and recover missing activation regions. RFCNet [23] exploits long-range spatial contexts from non-occluded regions to predict features of occluded regions, recovering the missing information at the feature level. Specifically, RFCNet estimates four key points to divide the feature map into six different regions. In RFCNet, the encoder-decoder architecture is adopted, in which the encoder models the correlation between regions through clustering, and the decoder utilizes the spatial correlation to recover occluded region features.

Differently, FRT [75] proposes to exploit the pedestrian information in feature sets of its $k$-nearest neighbors in the gallery to recover the occluded query feature. The concatenation of the query feature and its $k$-nearest neighbors features, which consists of position information, visibility score, and vision feature, is fed into the Transformer module for generating recovered features. RTGAT [25] leverages a GCN to fuzzily compensate the occluded image with its holistic counterpart for effective feature matching. Specifically, the semantic correlation between part features and the global feature is mined to reason the visibility scores of body parts. Then the visibility scores are regarded as the graph attention to guide the GCN to fuzzily suppress the noise of occluded part features and propagate the missing semantic information from the holistic image to the occluded image.

5 DATASETS AND EVALUATIONS

5.1 Datasets

We review eight widely used datasets for occluded person Re-ID, including three image-based partial Re-ID datasets, four image-based occluded Re-ID datasets, and one video-based occluded person Re-ID dataset. Examples of partial/occluded person Re-ID datasets are shown in Figure 9. Statistics of these datasets are summarized in Table 2, and details about each dataset are carefully reviewed as follows. It should be noted that Partial-REID, Partial-iLIDS, and Occluded-REID datasets do not provide the training set. Generally, methods are trained on the training set of Market-1501 [92] and tested on these three datasets for evaluation.

---

1 The partial person Re-ID assumes that the visible region of occluded person image is manually cropped for identification.

2 The occluded person Re-ID does not require the manually cropping process of occluded images. Unless otherwise specified, occluded person Re-ID in this survey includes both the partial person Re-ID and the occluded person Re-ID.
### Table 2. Occluded/Partial Person Re-ID Datasets

| Dataset                  | Training Set (ID/Image) | Test Set (ID/Image) | Gallery | Query |
|--------------------------|-------------------------|---------------------|---------|-------|
| Partial-REID             | –                       | 60/300              | 60/300  |       |
| Partial-iLIDS            | –                       | 119/119             | 119/119 |       |
| p-CUHK03                 | 1,160/15,080            | 100/300             | 100/1,000|      |
| P-ETHZ                   | 43/–                    | 42/–                | 42/–    |       |
| Occluded-REID            | –                       | 200/1,000           | 200/1,000|      |
| P-DukeMTMC-reID          | 650/–                   | 649/–               | 649/–   |       |
| Occ-DukeMTMC             | 702/15,618              | 1,110/17,661        | 519/2,210|      |
| Occ-DukeMTMC-Video       | 702/292,343             | 1,110/281,114       | 661/39,526|      |

For Partial-REID, Partial-iLIDS, and Occluded-REID, methods are generally trained on the training set of Market-1501 [92] and tested on these three datasets for evaluation.

**Partial-REID** [95] is an image-based partial Re-ID dataset with a variety of viewpoints, backgrounds, and occlusion types. It contains 600 images of 60 people, with five full-body images and five partial images per person. The partial observation is generated by manually cropping the occluded region in occluded images.

**Partial-iLIDS** [16] is an image-based simulated partial Re-ID dataset derived from iLIDS [94]. It is captured by multiple non-overlapping cameras in the airport and contains 238 images from 119 people, with one full-body image and one manually cropped non-occluded partial image per person.

**p-CUHK03** [36] is an image-based partial Re-ID dataset constructed from CUHK03 [43]. It contains 1,360 person identities captured in campus environment. In general, 1,160 person identities are used as training set, 100 person identities are used as validation set, and 100 person identities are used as test set. It selects five images with same view point from the raw dataset for each identity and generates 10 partial body probe images out of selected two images. The remaining three images of each identity are used as the full-body gallery image.

**P-ETHZ** [103] is an image-based occluded person Re-ID dataset modified from ETHZ [8]. It has 3,897 images of 85 person identities. Each identity has 1 to 30 full-body person images and occluded person images, respectively.

**Occluded-REID** [103] is an image-based occluded person Re-ID dataset captured by mobile cameras with different viewpoints and different types of severe occlusion. It consists of 2,000 images of 200 people, with five full-body images and five occluded images per person.

**P-DukeMTMC-reID** [103] is an image-based occluded person Re-ID dataset modified from DukeMTMC-reID [96]. It has 24,143 images of 1,299 person identities and contains images with target people occluded by different types of occlusion in public, such as people, cars, and guideboards. Each identity has both full-body images and occluded images.

**Occluded-DukeMTMC** [54] is an image-based occluded person Re-ID dataset built from DukeMTMC-reID [96]. It contains 15,618 images of 708 people for training while including 2,210 query images of 519 people and 17,661 gallery images of 1,110 people for testing. In addition, 9% of the training set, 100% of the query set, and 10% of the gallery set are occluded images.

**Occluded-DukeMTMC-VideoReID** [23] is a video-based occluded Re-ID dataset reorganized from DukeMTMC-VideoReID [73]. It includes a large variety of obstacles, such as cars, trees, bicycles, and other people. It contains 1,702 image sequences of 702 identities for training, and 661 query image sequences of 661 identities and 2,636 gallery image sequences of 1,110 identities for testing. More than 1/3 of the frames of each query sequence in the testing set contain occlusion.
5.2 Evaluation Metrics

The occluded person Re-ID evaluates the performance of an Re-ID system under the scenario of occlusion. Therefore, the settings of partial/occluded person Re-ID datasets are usually specially designed. In principle, the query images/videos for testing are all occluded samples, and the evaluation focuses on whether the correct identities can be retrieved when only occluded queries are provided. To evaluate an Re-ID system, **Cumulative Matching Characteristic (CMC)** curves and **Mean Average Precision (mAP)** are two widely used metrics.

CMC curves calculate the probability that a correct match appears in the top-\(k\) ranked retrieval results, \(k \in \{1, 2, 3, \ldots\}\). Specifically, the top-\(k\) accuracy of the query \(i\) is calculated as

\[
Acc^i_k = \begin{cases} 
1, & \text{if the top-}k\text{-ranked gallery samples contain the sample(s) of query } i; \\
0, & \text{otherwise.}
\end{cases}
\]

Supposing that there are \(N\) queries in the test set, the CMC-\(k\) (a.k.a. the rank-\(k\) accuracy) that calculates the probability of the top-\(k\) accuracy for all queries is computed as

\[
CMC-k = \frac{1}{N} \sum_{i=1}^{N} Acc^i_k.
\]

Since only the first match is concerned in the calculation, the CMC curves are acceptable when there are only one ground truth for each query or when we care more about the ground truth match in the top positions of the rank list.

The mAP measures the average retrieval performance that takes the order of all true matches in the ranked retrieval results into consideration. Specifically, the AP (average precision) of the query \(i\) is calculated as

\[
AP_i = \frac{1}{M_i} \sum_{j=1}^{M_i} \frac{j}{\text{Rank}_j},
\]

where \(M_i\) denotes the number of samples with identity \(i\) in the gallery set and \(\text{Rank}_j\) denotes the rank of the \(j\)-th ground truth in the retrieval gallery list for query \(i\). Supposing that there are \(N\) queries in the test set, the mAP is computed as

\[
mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i.
\]

Since the order of all true matches in the ranked retrieval results participates in the calculation of mAP, the mAP measures the average retrieval performance and is suitable for the gallery with multiple true matches.

On the whole, the mAP pays more attention to retrieval recall, whereas the CMC curves focus on the ability to retrieve a true match in candidate lists with different sizes. Consequently, the CMC curves and the mAP usually work together for the evaluation of an Re-ID system.

6 PERFORMANCE COMPARISON

We evaluate occluded person Re-ID methods on four widely used image-based datasets: Partial-REID [95], Partial-iLIDS [16], Occluded-DukeMTMC [54], and Occluded-REID [103]. Details about the four datasets are illustrated in Section 5.1. The performance comparisons on two partial person Re-ID datasets and two occluded person Re-ID datasets are summarized in Tables 3 and 4, respectively. In particular, methods are also marked with which of the four issues (i.e., position...
misalignment, scale misalignment, noisy information, and missing information) they have explicitly discussed. From the two tables, we can obtain the following observations.

First, effective solutions are rich and diverse, and there is not a dominant technical route for occluded person Re-ID. On Partial-ReID, the top rank-1 accuracy of the attention mechanism based method PAT [45] and the auxiliary information based (contextual recovery based) method FRT [75] are comparable, achieving superior results; on Partial-iLIDS, the matching-based method ASAN [33] has reached the best rank-1 and rank-3 accuracy; on Occluded-DukeMTMC and Occluded-ReID, auxiliary information based (contextual recovery based) methods FRT [75] and HPNet [24] have achieved the best rank-1 accuracy, respectively. Although none of the solution routes are dominant on all four datasets and some methods involve more than one technical route (e.g., FRT [75] and HORReID [68]), the advantages and disadvantages of different types of solutions can be summarized to boost future research (please refer to Section 7).

Second, the scale misalignment and missing information issues have drawn less attention in existing methods. Compared with methods that aim to address position misalignment or noisy information issues, the number of methods intended for alleviating scale misalignment or missing information issues is significantly smaller.

Third, a number of methods have considered more than one issue at the same time. For instance, PPCL [20] simultaneously addresses position and scale misalignment issues with carefully
Methods Publication Backbones Involved Issues Occluded-DukeMTMC Occluded-ReID

**Matching**

| Method          | Publication | Backbones     | P | S | N | M | Rank-1 | mAP | Rank-1 | mAP |
|-----------------|-------------|---------------|---|---|---|---|--------|-----|--------|-----|
| AMC+SWA [95]    | ICCV 2015   | ResNet-50     | ✓ | ✓ | - | - | 40.8   | 30.4 | 72.8   | 62.8 |
| DSR [16]        | CVPR 2018   | ResNet-50     | ✓ | ✓ | - | - | 77.4   | 72.8 | 62.8   | 62.8 |
| GASM [17]       | ECCV 2020   | ResNet-50     | ✓ | ✓ | - | - | 56.4   | 43.8 | 80.3   | 70.2 |
| HDRoID [68]     | CVPR 2020   | ResNet-50     | ✓ | ✓ | ✓ | ✓ | 78.3   | 72.8 | 62.8   | 62.8 |
| DaRoID [77]     | KBS 2021    | ResNet-50     | ✓ | ✓ | ✓ | ✓ | 55.1   | 43.8 | 80.3   | 70.2 |
| MoS [30]        | AAAI 2021   | ResNet-50+IBN | ✓ | ✓ | ✓ | ✓ | 66.6   | 55.1 | 72.8   | 62.8 |

**Multi-scale features**

| Method          | Publication | Backbones     | P | S | N | M | Rank-1 | mAP | Rank-1 | mAP |
|-----------------|-------------|---------------|---|---|---|---|--------|-----|--------|-----|
| DSR [16]        | CVPR 2018   | ResNet-50     | ✓ | ✓ | - | - | 40.8   | 30.4 | 72.8   | 62.8 |
| FPR [18]        | ICCV 2019   | ResNet-50     | ✓ | ✓ | - | - | 78.3   | 72.8 | 62.8   | 62.8 |

**Attention mechanism**

| Method          | Publication | Backbones     | P | S | N | M | Rank-1 | mAP | Rank-1 | mAP |
|-----------------|-------------|---------------|---|---|---|---|--------|-----|--------|-----|
| MiRSSD-Net [64] | TNNLS 2022  | ResNet-50+Transformer | ✓ | ✓ | - | - | 59.1   | 44.8 | 81.4   | 71.5 |
| IGOAS [67]      | TIP 2021    | ResNet-50     | ✓ | ✓ | - | - | 60.1   | 49.4 | 81.4   | 71.5 |
| OAIM [6]        | ICCV 2021   | ResNet-50     | ✓ | ✓ | - | - | 62.6   | 46.1 | 81.4   | 71.5 |
| ESP [102]       | ECCV 2020   | HRNet-W32     | ✓ | ✓ | - | - | 62.8   | 52.3 | 81.4   | 71.5 |
| OCNet [57]      | ICASSP 2022 | ResNet-50     | ✓ | ✓ | - | - | 64.3   | 54.4 | 81.4   | 71.5 |
| PAT [45]        | CVPR 2021   | ResNet-50+Transformer | ✓ | ✓ | - | - | 64.5   | 53.6 | 81.6   | 72.1 |
| SBPA [67]       | SPL 2021    | ResNet-50     | ✓ | ✓ | - | - | 64.5   | 54.0 | 81.6   | 72.1 |
| DRDL-Net [29]   | TMM 2021    | ResNet-50     | ✓ | ✓ | - | - | 65.8   | 53.9 | 81.6   | 72.1 |
| SSGR [79]       | ICCV 2021   | ResNet-50+DNL | ✓ | ✓ | - | - | 65.8   | 57.2 | 81.6   | 72.1 |
| TransRoID [19]  | ICCV 2022   | ViT           | ✓ | ✓ | - | - | 66.4   | 59.2 | 81.6   | 72.1 |
| FED [71]        | CVPR 2022   | ViT           | ✓ | ✓ | - | - | 68.1   | 56.4 | 86.3   | 79.3 |

**Auxiliary information**

| Method          | Publication | Backbones     | P | S | N | M | Rank-1 | mAP | Rank-1 | mAP |
|-----------------|-------------|---------------|---|---|---|---|--------|-----|--------|-----|
| PVPM [11]       | CVPR 2020   | ResNet-50     | ✓ | ✓ | - | - | 70.4   | 61.2 | 70.4   | 61.2 |
| GASM [17]       | ECCV 2020   | ResNet-50     | ✓ | ✓ | - | - | 74.5   | 65.6 | 74.5   | 65.6 |
| FPR [18]        | ICCV 2019   | ResNet-50     | ✓ | ✓ | - | - | 78.3   | 68.0 | 78.3   | 68.0 |
| PGFA [54]       | ICCV 2019   | ResNet-50     | ✓ | ✓ | - | - | 51.4   | 37.3 | 81.0   | 71.0 |
| PDVM [100]      | PRL 2020    | ResNet-50     | ✓ | ✓ | - | - | 53.0   | 38.1 | 81.0   | 71.0 |
| HRID [65]       | CVPR 2020   | ResNet-50     | ✓ | ✓ | - | - | 55.1   | 43.8 | 80.3   | 70.2 |
| PMFB [55]       | TNNLS 2021  | ResNet-50     | ✓ | ✓ | - | - | 56.3   | 43.5 | 80.3   | 70.2 |
| LKWS [80]       | ICCV 2021   | ResNet-50     | ✓ | ✓ | - | - | 62.2   | 46.3 | 81.0   | 71.0 |
| PGFL-KD [91]    | MM 2021     | ResNet-50     | ✓ | ✓ | - | - | 63.0   | 54.1 | 80.7   | 70.3 |
| DaRoID [77]     | KBS 2021    | ResNet-50     | ✓ | ✓ | - | - | 63.9   | 54.5 | 80.7   | 70.3 |
| RFCNet [23]     | TPAMI 2022  | ResNet-50     | ✓ | ✓ | - | - | 63.9   | 54.5 | 80.7   | 70.3 |
| PFD [65]        | AAAI 2022   | ViT           | ✓ | ✓ | - | - | 69.5   | 61.8 | 81.5   | 83.0 |
| HPNet [24]      | ICME 2020   | ResNet-50     | ✓ | ✓ | - | - | 87.3   | 77.4 | 87.3   | 77.4 |
| FRT [75]        | TIP 2022    | ResNet-50+Transformer | ✓ | ✓ | ✓ | ✓ | 77.0   | 61.3 | 80.4   | 71.0 |

**Contextual recovery**

| Method          | Publication | Backbones     | P | S | N | M | Rank-1 | mAP | Rank-1 | mAP |
|-----------------|-------------|---------------|---|---|---|---|--------|-----|--------|-----|
| RTGAT [25]      | TIP 2023    | ResNet-50     | ✓ | ✓ | - | - | 61.0   | 50.1 | 71.8   | 51.0 |
| RFCNet [23]     | TPAMI 2021  | ResNet-50     | ✓ | ✓ | - | - | 63.9   | 54.5 | 80.4   | 71.0 |
| FRT [75]        | TIP 2022    | ResNet-50+Transformer | ✓ | ✓ | ✓ | ✓ | 70.7   | 61.3 | 80.4   | 71.0 |

**Backbones:** ResNet-50 [15], IBN (Instance Batch Normalization) [57], HRNet-W32 [60], DNL (Disentangled Non-Local) [83], and ViT (Vision Transformer) [7].

The P, S, N, and M denote position misalignment, scale misalignment, noisy information, and missing information issues, respectively.

designed image transformation; PFD [65] focuses on position alignment and noise suppression with the help of auxiliary information; co-attention [47] takes the position misalignment, scale misalignment, and noisy information issues into consideration; and FRT [75] alleviates the position misalignment, noisy information, and missing information issues. Although methods addressing more than one issue show superior results in most cases, the four issues have not been considered simultaneously.

### 7 FUTURE DIRECTIONS

As shown in Tables 3 and 4, there have been consistent improvements in different types of solutions for addressing different issues over the past few years. Based on the analysis of issues and solutions, the following insights can be drawn for future research of occluded person Re-ID.

From the perspective of promising technical routes, it remains an open question since the evaluation results of state-of-the-art methods in most types of solutions are comparable, and different solution types can be combined with each other to boost performance. Despite this, the advantages and disadvantages of different types of solutions can be summarized and analyzed as follows to help inspire new ideas.

**Matching.** Well-designed matching elements and matching strategies greatly improve the performance of occluded person Re-ID. Local and scalable matching elements with the corresponding matching strategy can help address position misalignment, scale misalignment, and
noisy information issues. The matching can be easily integrated with other types of solutions, such as the repeatedly reported co-attention [47] in Table 3.

**Image Transformation.** The partial (holistic) image is transformed to obtain the image of consistent contents with the holistic (partial) image, addressing the position and the scale misalignment issues simultaneously (see Figure 4(a) and (b)). This technical route is close to the ideal process while requiring more computation costs for the conditional image transformation at the inference stage. Furthermore, the image transformation has not achieved a satisfying result in the current stage, and the performance of this technical route is a little bit lower than others.

**Attention Mechanism.** The attention mechanism has been widely studied in existing methods for its huge potential and flexibility. In the past 3 years, the methods [7, 29, 47, 64, 71] introducing self-attention (Transformer) to occluded person Re-ID have made remarkable improvements on public datasets. Similar to matching, the attention mechanism can also be easily integrated with other types of solutions, such as the repeatedly reported APN [26] in Table 3.

**Auxiliary Information.** In general, there are three main types of auxiliary information employed for occluded person Re-ID: poses, segments, and attributes. The position information as well as estimation confidence provided by pose estimation or segmentation models are used to help address position misalignment and noisy information issues accordingly (see Figure 7). The attribute information offered by person Re-ID datasets is generally used to formulate an extra task to help alleviate the issues caused by occlusion. The auxiliary information brings a lot of benefits and convenience but is dependent on extra labels or external models and requires more computation costs.

From the perspective of four significant challenges brought by occlusion, position misalignment and noisy information issues have been widely studied, whereas scale misalignment and missing information issues have been less discussed in existing methods. At the current stage, none of the methods have simultaneously considered all four issues for occluded person Re-ID. To build a more comprehensive solution, future research is encouraged to take all four issues into consideration and place more effort into innovation around scale alignment and occlusion recovery.

8 CONCLUSION
The aim of this article was to provide a systematic survey of occluded person Re-ID to help promote future research. We first analyzed and summarized four issues brought by occlusion in person Re-ID: position misalignment, scale misalignment, noisy information, and missing information. The published publications of deep learning based occluded person Re-ID from top conferences and journals were categorized and introduced accordingly. We summarized the performance comparison of recent occluded person Re-ID methods on four popular datasets: Partial-ReID, Partial-iLIDS, Occluded-ReID, and Occluded-DukeMTMC. Based on the analysis of evaluation results, we finally discussed promising future research directions.

REFERENCES
[1] Ejaz Ahmed, Michael Jones, and Tim K. Marks. 2015. An improved deep learning architecture for person re-identification. In Proceedings of CVPR. 3908–3916. https://doi.org/10.1109/CVPR.2015.7299016
[2] Jean-Paul Aïnam, Ke Qin, Guisong Liu, Guangchun Luo, and Brighter Agyemang. 2020. Enforcing affinity feature learning through self-attention for person re-identification. ACM Trans. Multimedia Comput. Commun. Appl. 16, 1 (March 2020), Article 16, 22 pages. https://doi.org/10.1145/3377352
[3] Apurva Bedagkar-Gala and Shishir K. Shah. 2014. A survey of approaches and trends in person re-identification. Image Vis. Comput. 32, 4 (2014), 270–286. https://doi.org/10.1016/j.imavis.2014.02.001
[4] Honglong Cai, Zhiguan Wang, and Jinxing Cheng. 2019. Multi-scale body-part mask guided attention for person re-identification. In Proceedings of CVPR Workshops. https://doi.org/10.1109/CVPRW.2019.00197
Deep Learning Based Occluded Person Re-Identification: A Survey

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2D pose estimation using part affinity fields. In Proceedings of CVPR. 7291–7299. https://doi.org/10.1109/CVPR.2017.143

[6] Peixian Chen, Wenfeng Liu, Pingyang Dai, Jianzhuan Liu, Qixiang Ye, Mingliang Xu, Qi’an Chen, and Rongrong Ji. 2021. Occlude them all: Occlusion-aware attention network for occluded person Re-ID. In Proceedings of ICCV. 11833–11842. https://doi.org/10.1109/ICCV48922.2021.01162

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jacob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020).

[8] Andreas Ess, Bastian Leibe, Konrad Schindler, and Luc Van Gool. 2008. A mobile vision system for robust multi-person tracking. In Proceedings of CVPR. IEEE, Los Alamitos, CA, 1–8. https://doi.org/10.1109/CVPR.2008.4587581

[9] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. 2019. Dual attention network for occluded person re-identification. In Proceedings of MM. 3771–3779. https://doi.org/10.1145/3394171.3413833

[10] Lijuan Huo, Chunfeng Song, Zhengyi Liu, and Zhaoxiang Zhang. 2021. Attentive part-aware networks for partial person re-identification. In Proceedings of MM. 3771–3779. https://doi.org/10.1145/3394171.3413833

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of CVPR. 770–778. https://doi.org/10.1109/CVPR.2016.91

[12] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. 2017. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In Proceedings of CVPR. 932–940. https://doi.org/10.1109/ICCV.2019.00899

[13] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. DensePose: Dense human pose estimation in the wild. In Proceedings of CVPR. 7297–7306. https://doi.org/10.1109/CVPR.2018.00762

[14] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. 2019. VRSTC: Occlusion-free occluded person re-identification. In Proceedings of CVPR. 9105–9115. https://doi.org/10.1109/CVPR.2019.00885

[15] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2D pose estimation using part affinity fields. In Proceedings of CVPR. 7291–7299. https://doi.org/10.1109/CVPR.2017.143

[16] Lingxiao He, Jianliang, Haiqing Li, and Zhenan Sun. 2018. Deep spatial feature reconstruction for partial person re-identification. In Proceedings of CVPR. 7073–7082. https://doi.org/10.1109/CVPR.2018.00739

[17] Lingxiao He and Wu Liu. 2020. Guided saliency feature learning for person re-identification in crowded scenes. In Proceedings of ECCV. 357–373. https://doi.org/10.1007/10978-3-030-58604-1_22

[18] Shunting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. 2021. TransReID: Transformer-based object re-identification. In Proceedings of ICME. IEEE, Los Alamitos, CA, 15013–15022. https://doi.org/10.1109/ICME46284.2020.9102789

[19] Tianyu He, Xu Shen, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. 2021. Partial person re-identification with part-part correspondence learning. In Proceedings of CVPR. 9105–9115. https://doi.org/10.1109/CVPR46437.2021.00899

[20] Yuanhang He, Hua Yang, and Lin Chen. 2021. Adversarial cross-scale alignment pursuit for seriously misaligned person re-identification. In Proceedings of ICIP. IEEE, Los Alamitos, CA, 226–230. https://doi.org/10.1109/ICIP40778.2020.9191196

[21] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. 2020. Pose-guided visible part matching for occluded person ReID. In Proceedings of CVPR. 11744–11752. https://doi.org/10.1109/CVPR42600.2020.01176

[22] Yuanhang He, Hua Yang, and Lin Chen. 2021.9506293

[23] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. 2017. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In Proceedings of CVPR. 932–940. https://doi.org/10.1109/ICCV.2019.00899

[24] Houjing Huang, Xiaotang Chen, and Kaiqi Huang. 2020. Human parsing based alignment with multi-task learning for occluded person re-identification. In Proceedings of ICME. IEEE, Los Alamitos, CA, 1–6. https://doi.org/10.1109/ICME46284.2020.9102789

[25] Meiyan Huang, Chunping Hou, Qingyuan Yang, and Zhipeng Wang. 2023. Reasoning and tuning: Graph attention network for occluded person re-identification. In Proceedings of MM. 3771–3779. https://doi.org/10.1145/3394171.3413833

[26] Lishuai Gao, Hua Zhang, Zan Gao, Weili Guan, Zhiyong Cheng, and Meng Wang. 2020. Texture semantically aligned with visibility-aware for partial person re-identification. In Proceedings of MM. 3771–3779. https://doi.org/10.1145/3394171.3413833

[27] Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jacob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020).

[28] Peixian Chen, Wenfeng Liu, Pingyang Dai, Jianzhuan Liu, Qixiang Ye, Mingliang Xu, Qi’an Chen, and Rongrong Ji. 2021. Occlude them all: Occlusion-aware attention network for occluded person Re-ID. In Proceedings of ICIP. IEEE, Los Alamitos, CA, 1–6. https://doi.org/10.1109/ICIP48806.2021.9412527

[29] Andreas Ess, Bastian Leibe, Konrad Schindler, and Luc Van Gool. 2008. A mobile vision system for robust multi-person tracking. In Proceedings of CVPR. IEEE, Los Alamitos, CA, 1–8. https://doi.org/10.1109/CVPR.2008.4587581

[30] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. 2019. Dual attention network for occluded person re-identification. In Proceedings of MM. 3771–3779. https://doi.org/10.1145/3394171.3413833

[31] Lingxiao He, Jianliang, Haiqing Li, and Zhenan Sun. 2018. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In Proceedings of CVPR. 7073–7082. https://doi.org/10.1109/CVPR.2018.00739

[32] Zhiyong Cheng, and Meng Wang. 2020. Texture semantically aligned with visibility-aware for partial person re-identification. In Proceedings of MM. 3771–3779. https://doi.org/10.1145/3394171.3413833

[33] Lingxiao He and Wu Liu. 2020. Guided saliency feature learning for person re-identification in crowded scenes. In Proceedings of CVPR. 932–940. https://doi.org/10.1109/ICCV.2019.00899
[27] Khawar Islam. 2020. Person search: New paradigm of person re-identification: A survey and outlook of recent works. *Image Vis. Comput.* 101 (2020), 103970. https://doi.org/10.1016/j.imavis.2020.103970

[28] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. 2015. Spatial transformer networks. In *Proceedings of NIPS*.

[29] Mengxi Jia, Xinhua Cheng, Shijian Lu, and Jian Zhang. 2022. Learning disentangled representation implicitly via transformer for occluded person re-identification. *IEEE Trans. Multimedia* 25 (2022), 1294–1305. https://doi.org/10.1109/TMM.2022.3141267

[30] Mengxi Jia, Xinhua Cheng, Y unpeng Zhai, Shijian Lu, Siwei Ma, Yonghong Tian, and Jian Zhang. 2021. Matching on sets: Conquer occluded person re-identification without alignment. In *Proceedings of AAAI*. 1673–1681. https://doi.org/10.1609/aaai.v35i12.16260

[31] Jianwen Jiang, Yuxuan Wei, Yifan Feng, Jingxuan Cao, and Yue Gao. 2019. Dynamic hypergraph neural networks. In *Proceedings of IJCAI*. 2635–2641. https://doi.org/10.24963/ijcai.2019/366

[32] Li cheng Jiao, Jie Chen, Fang Liu, Shuyuan Yang, Chao You, Xu Liu, Lingling Li, and Biao Hou. 2023. Graph representation learning meets computer vision: A survey. *IEEE Trans. Artif. Intell.* 4, 1 (2023), 2–22. https://doi.org/10.1109/TAI.2022.3194869

[33] Hanyang Jin, Shengqi Lai, and Xueming Qian. 2022. Occlusion-sensitive person re-identification via attribute-based shift attention. *IEEE Trans. Circuits Syst. Video Technol.* 32, 4 (2022), 2170–2185. https://doi.org/10.1109/TCSVT.2021.3088446

[34] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E. Kamasak, and Mubarak Shah. 2018. Human semantic parsing for person re-identification. In *Proceedings of CVPR*. 1062–1071. https://doi.org/10.1109/CVPR.2018.00117

[35] Srikrishna Karanam, Mengran Gou, Ziyan Wu, Angels Rates-Borras, Octavio Camps, and Richard J. Radke. 2019. A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 3 (2019), 523–536. https://doi.org/10.1109/TPAMI.2018.2807450

[36] Junyeong Kim and Chang D. Yoo. 2017. Deep partial person re-identification via attention model. In *Proceedings of ICIP*. IEEE, Los Alamitos, CA, 3425–3429. https://doi.org/10.1109/ICIP.2017.8296918

[37] Minjung Kim, MyeongAh Cho, Heansung Lee, Suhwan Cho, and Sangyoun Lee. 2022. Occluded person re-identification via relational adaptive feature correction learning. In *Proceedings of ICASSP*. IEEE, Los Alamitos, CA, 2719–2723. https://doi.org/10.1109/ICASSP43922.2022.9746734

[38] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. 2019. PifPaf: Composite fields for human pose estimation. In *Proceedings of CVPR*.

[39] Bahram Lavi, Mehdi Fat an Serj, and Ihsan Ullah. 2018. Survey on deep learning techniques for person re-identification task. *arXiv preprint arXiv:1807.05284* (2018).

[40] Bahram Lavi, Ihsan Ullah, Mehdi Fat an, and Anderson Rocha. 2020. Survey on reliable deep learning-based person re-identification models: Are we there yet? *arXiv preprint arXiv:2005.00355* (2020).

[41] Qingming Leng, Meng Ye, and Qi Tian. 2019. A survey of open-world person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* 30, 4 (2019), 1092–1108. https://doi.org/10.1109/TCSVT.2019.2898940

[42] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. 2018. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of CVPR*. 369–378. https://doi.org/10.1109/CVPR.2018.00046

[43] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. DeepReID: Deep filter pairing neural network for person re-identification. In *Proceedings of CVPR*. 152–159. https://doi.org/10.1109/CVPR.2014.27

[44] Xi Li, Liming Zhao, Ming-Hsuan Yang, Fei Wu, Yueting Zhuang, Haibin Ling, and Jingdong Wang. 2016. DeepSaliency: Multi-task deep neural network for salient object detection. *IEEE Trans. Image Process.* 25, 8 (2016), 3919–3930. https://doi.org/10.1109/TIP.2016.2579306

[45] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. 2021. Diverse part discovery: Occluded person re-identification via relational adaptive feature correction learning. In *Proceedings of ICIP*. IEEE, Los Alamitos, CA, 2299–2303. https://doi.org/10.1109/ICIP42928.2021.9506470

[46] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. 2018. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 4 (2018), 871–885. https://doi.org/10.1109/TPAMI.2018.2820063

[47] Ci-Siang Lin and Yu-Chiang Frank Wang. 2021. Self-supervised bodymap-to-appearance co-attention for partial person re-identification. In *Proceedings of ICIP*. IEEE, Los Alamitos, CA, 3425–3429. https://doi.org/10.1109/ICIP42928.2021.9506470

[48] Xiangtan Lin, Pengzhen Ren, Yun Xiao, Xiaojun Chang, and Alex Hauptmann. 2021. Person search challenges and solutions: A survey. In *Proceedings of IJCAI*. 1–10. https://doi.org/10.24963/ijcai.2021/613

[49] Xiangtan Lin, Pengzhen Ren, Chung-Hsing Yeh, Lina Yao, Andy Song, and Xiaojun Chang. 2021. Unsupervised person re-identification: A systematic survey of challenges and solutions. *arXiv preprint arXiv:2109.06057* (2021).
Deep Learning Based Occluded Person Re-Identification: A Survey

[50] Jiawei Liu, Zheng-Jun Zha, Wei Wu, Kecheng Zheng, and Qibin Sun. 2021. Spatial-temporal correlation and topology learning for person re-identification in videos. In Proceedings of CVPR. 4370–4379. https://doi.org/10.1109/CVPR46437.2021.00435

[51] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. 2019. Spatial and temporal mutual promotion for video-based person re-identification. In Proceedings of AAAI, Vol. 33. 8786–8793. https://doi.org/10.1609/aaai.v33i01.33018786

[52] Hao Luo, Wei Jiang, Xing Fan, and Chi Zhang. 2020. STNReID: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification. IEEE Trans. Multimedia 22, 11 (2020), 2905–2913. https://doi.org/10.1109/TMM.2020.2965491

[53] Riccardo Mazzon, Syed Fahad Tahir, and Andrea Cavallaro. 2012. Person re-identification in crowd. Pattern Recognit. Lett. 33, 14 (2012), 1828–1837. https://doi.org/10.1016/j.patrec.2012.02.014

[54] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. 2019. Pose-guided feature alignment for occluded person re-identification. In Proceedings of ICCV. 542–551. https://doi.org/10.1109/ICCV.2019.00063

[55] Jiaxu Miao, Yu Wu, and Yi Yang. 2022. Identifying visible parts via pose estimation for occluded person re-identification. IEEE Trans. Neural Netw. Learn. Syst. 33, 9 (2022), 4624–4634. https://doi.org/10.1109/TNNLS.2021.3059515

[56] Zhangqiang Ming, Min Zhu, Xiangkun Wang, Jianmin Zhu, Junlong Cheng, Chengrui Gao, Yong Yang, and Xiaoyong Wei. 2022. Deep learning-based person re-identification methods: A survey and outlook of recent works. Image Vis. Comput. 119 (2022), 104394. https://doi.org/10.1016/j.imavis.2022.104394

[57] Xingang Pan, Ping Luo, Jiamping Shi, and Xiaouo Tang. 2018. Two at once: Enhancing learning and generalization capacities via IBN-Net. In Proceedings of ECCV. 464–479.

[58] Rodolfo Quipe and Helio Pedrini. 2019. Improved person re-identification based on saliency and semantic parsing with deep neural networks. Image Vis. Comput. 92 (2019), 103809. https://doi.org/10.1016/j.imavis.2019.07.009

[59] Vladimir Somers, Christophe De Vleeschouwer, and Alexandre Alahi. 2023. Body part-based representation learning for occluded person re-identification. In Proceedings of WACV. 1613–1623.

[60] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In Proceedings of CVPR. 5693–5703. https://doi.org/10.1109/CVPR.2019.00584

[61] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. 2019. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In Proceedings of CVPR. 393–402. https://doi.org/10.1109/CVPR.2019.00048

[62] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In Proceedings of CVPR. 2818–2826. https://doi.org/10.1109/CVPR.2016.308

[63] Hongchen Tan, Xiuping Liu, Yuhao Bian, Huasheng Wang, and Baocai Yin. 2022. Incomplete descriptor mining with elastic loss for person re-identification. IEEE Trans. Circuits Syst. Video Technol. 32, 1 (2022), 160–171. https://doi.org/10.1109/TCSVT.2021.3061412

[64] Hongchen Tan, Xiuping Liu, Baocai Yin, and Xin Li. 2022. MHSA-Net: Multichannel self-attention network for occluded person re-identification. IEEE Trans. Neural Netw. Learn. Syst. Early access, March 21, 2022. https://doi.org/10.1109/TNNLS.2022.3144163

[65] Wang Tao, Liu Hong, Song Pinhao, Guo Tianyu, and Shi Wei. 2022. Pose-guided feature disentangling for occluded person re-identification based on transformer. In Proceedings of AAAI. https://doi.org/10.1609/aaai.v36i3.20155

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of NIPS. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dece919fd053c1c4a845a-Paper.pdf

[67] Guanshuo Wang, Xiong Chen, Jialin Gao, Xi Zhou, and Shiming Ge. 2021. Self-guided body part alignment with deep neural network models. IEEE Signal Process. Lett. 28 (2021), 1155–1159. https://doi.org/10.1109/LSP.2021.3087079

[68] Guan’an Wang, Shuo Yang, Hanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. 2020. High-order information matters: Learning relation and topology for occluded person re-identification. In Proceedings of CVPR. 6449–6458. https://doi.org/10.1109/CVPR42600.2020.00648

[69] Haoran Wang, Licheng Jiao, Shuyuan Yang, Lingling Li, and Zexin Wang. 2020. Simple and effective: Spatial rescaling for person re-identification. IEEE Trans. Neural Netw. Learn. Syst. 33, 1 (2020), 145–156.

[70] Zheng Wang, Zhixiang Wang, Yinqiang Zheng, Yang Wu, Wenjun Zeng, and Shin’ichi Satoh. 2021. Beyond intramodality: A survey of heterogeneous person re-identification. In Proceedings of IJCAI. 4973–4980. https://doi.org/10.24963/ijcai.2020/692

[71] Zhikang Wang, Feng Zhu, Shixiang Tang, Rui Zhao, Lihue He, and Jiangning Song. 2022. Feature erasing and diffusion network for occluded person re-identification. In Proceedings of CVPR. 4754–4763. https://doi.org/10.1109/cvpr52688.2022.00471
[72] Di Wu, Si-Jia Zheng, Xiao-Ping Zhang, Chang-An Yuan, Fei Cheng, Yang Zhao, Yong-Jun Lin, Zhong-Qiu Zhao, Yong-Li Jiang, and De-Shuang Huang. 2019. Deep learning-based methods for person re-identification: A comprehensive review. Neurocomputing 337 (2019), 354–371. https://doi.org/10.1016/j.neucom.2019.01.079

[73] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. 2018. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In Proceedings of CVPR. 5177–5186. https://doi.org/10.1109/CVPR.2018.00543

[74] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. 2017. Joint detection and identification feature learning for person search. In Proceedings of CVPR. https://doi.org/10.1109/CVPR.2017.360

[75] Boqiang Xu, Lingxiao He, Jian Liang, and Zhenan Sun. 2022. Learning feature recovery transformer for occluded person re-identification. IEEE Transactions on Image Processing 31 (2022), 4651–4662. https://doi.org/10.1109/TIP.2022.3186759

[76] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. 2018. Attention-aware compositional network for person re-identification. In Proceedings of CVPR. 2119–2128. https://doi.org/10.1109/CVPR.2018.00226

[77] Yunjie Xu, Liaoqing Zhao, and Feiwei Qin. 2021. Dual attention-based method for occluded person re-identification. Knowl. Based Syst. 212 (2021), 106554. https://doi.org/10.1016/j.knosys.2020.106554

[78] Cheng Yan, Guansong Pang, Jile Jiao, Xiao Bai, Xuetao Feng, and Chunhua Shen. 2021. Occluded person re-identification with single-scale global representations. In Proceedings of ICCV. 11875–11884. https://doi.org/10.1109/ICCV48922.2021.01166

[79] Yichao Yan, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, Ying Tai, and Ling Shao. 2020. Learning multi-granular hypergraphs for video-based person re-identification. In Proceedings of CVPR. 2899–2908. https://doi.org/10.1109/CVPR42600.2020.00297

[80] Jinrui Yang, Jiawei Zhang, Fufu Yu, Xinyang Jiang, Mengdan Zhang, Xing Sun, Ying-Cong Chen, and Wei-Shi Zheng. 2021. Learning to know where to see: A visibility-aware approach for occluded person re-identification. In Proceedings of ICCV. 11885–11894. https://doi.org/10.1109/ICCV48922.2021.01167

[81] Xun Yang, Meng Wang, Richang Hong, Qi Tian, and Yong Rui. 2017. Enhancing person re-identification in a self-trained subspace. ACM Trans. Multimedia Comput. Commun. Appl. 13, 3 (June 2017), Article 27, 23 pages. https://doi.org/10.1145/3089249

[82] Xun Yang, Meng Wang, Richang Hong, Qi Tian, and Yong Rui. 2017. Enhancing person re-identification in a self-trained subspace. ACM Trans. Multimedia Comput. Commun. Appl. 13, 3 (June 2017), Article 27, 23 pages. https://doi.org/10.1145/3089249

[83] Xun Yang, Meng Wang, Richang Hong, Qi Tian, and Yong Rui. 2017. Enhancing person re-identification in a self-trained subspace. ACM Trans. Multimedia Comput. Commun. Appl. 13, 3 (June 2017), Article 27, 23 pages. https://doi.org/10.1145/3089249

[84] You Zhai, Xianfeng Han, Wenzhuo Ma, Xinye Gou, and Guoqiang Xiao. 2021. PGMANet: Pose-guided mixed neural networks. In Proceedings of CVPR. 667–676. https://doi.org/10.1109/CVPR.2019.00076

[85] Caifong Zhao, Xinbi Lv, Shuguang Dou, Shanshan Zhang, Jun Wu, and Liang Wang. 2021. Incremental generative occlusion adversarial suppression network for person ReID. IEEE Trans. Image Process. 30 (2021), 4212–4224. https://doi.org/10.1109/TIP.2021.3070182

[86] Liang Zheng, Liuye Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In Proceedings of ICCV. 1116–1124. https://doi.org/10.1109/ICCV.2015.133
[93] Liang Zheng, Yi Yang, and Alexander G. Hauptmann. 2016. Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 (2016).

[94] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. 2011. Person re-identification by probabilistic relative distance comparison. In Proceedings of CVPR. IEEE, Los Alamitos, CA, 649–656. https://doi.org/10.1109/CVPR.2011.5995598

[95] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. 2015. Partial person re-identification. In Proceedings of ICCV. 4678–4686. https://doi.org/10.1109/iccv.2015.531

[96] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In Proceedings of ICCV. 3754–3762. https://doi.org/10.1109/ICCV42600.2020.00686

[97] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random erasing data augmentation. In Proceedings of AAAI, Vol. 34. 13001–13008. https://doi.org/10.1609/aaai.v34i07.7000

[98] Qinqin Zhou, Bineng Zhong, Xiangyuan Lan, Gan Sun, Yulun Zhang, Baochang Zhang, and Rongrong Ji. 2020. Fine-grained spatial alignment model for person re-identification with focal triplet loss. IEEE Trans. Image Process. 29 (2020), 7578–7589. https://doi.org/10.1109/TIP.2020.3004267

[100] Shuren Zhou, Jie Wu, Fan Zhang, and Paramjit Sehdev. 2020. Depth occlusion perception feature analysis for person re-identification. Pattern Recognit. Lett. 138 (2020), 617–623. https://doi.org/10.1016/j.patrec.2020.09.009

[101] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. 2019. Objects as points. arXiv preprint arXiv:1904.07850 (2019).

[102] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. 2020. Identity-guided human semantic parsing for person re-identification. In Proceedings of ECCV. 346–363. https://doi.org/10.1007/978-3-030-58580-8_21

[103] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. 2018. Occluded person re-identification. In Proceedings of ICME. IEEE, Los Alamitos, CA, 1–6. https://doi.org/10.1109/ICME.2018.8486568

Received 8 January 2023; revised 14 May 2023; accepted 15 June 2023