MULTISTAGE MULTISCALE INFERENCE NETWORK WITH VISIBILITY ATTENTION FOR OCCLUDED PERSON RE-IDENTIFICATION

YOON TAE KIM

A THESIS SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTERS OF SCIENCE

GRADUATE PROGRAM IN COMPUTER SCIENCE
YORK UNIVERSITY
TORONTO, ONTARIO

FEBRUARY 2021

Abstract

For occluded person re-identification this thesis presents the Multistage Multiscale Inference Network (MMI-Net) that leverages an inference framework based on multiscale representations with visibility guidance. MMI-Net consists of three sub-networks, i) global, ii) part-based and iii) integrated, to infer person re-identification. The global inference sub-network provides an overall holistic analysis of input images. The part-based sub-network captures more localized information. Both the global and part-based models make use of multiscale representation across multiple processing stages to capture a variety of complementary discriminative image structure. The integrated sub-network aggregates the global and part-based representations to obtain the final fusion of all extracted information. Pose guided attentional processing is used to provide robustness to occlusion. MMI-Net is unique in its integrated multistage inference architecture that accounts for local and global appearance with attentional processing. In empirical evaluation, MMI-Net outperforms current existing methods on multiple occluded person re-identification datasets. Notably, the proposed model shows superior performance by +9.3% Rank-1 and +5.8% mAP in terms of average score on Occluded-DukeMTMC, +6.2% Rank-1 and +0.7% Rank-3 in mean on Partial-iLIDS, and +5.3% Rank-1 and +7.3% mAP in max on P-DukeMTMC.

Acknowledgements

From the bottom of my heart, I would first like to express my deepest gratitude and appreciation to my supervisor, Professor Richard Wildes, for his patient support and guidance. Throughout my master's studies, he provided extensive knowledge, constructive advice, and enthusiastic encouragement. Without his unwavering support, this thesis would have never been accomplished.

I would like to thank my supervisory committee members, Professor Michael Jenkin and Professor Costas Armenakis, for their thoughtful comments and suggestions on this thesis.

I would also like to thank members in the Vision Lab for offering their support and inspiration.

Finally, I'm deeply indebted to my parents, especially to my mother for providing me with unconditional love and continuous encouragement throughout my life. Thank you mom.

Table of Contents

List of Abbreviations

ReID	Re-Identification
ConvNet	Convolutional Network
MMI-Net	Multistage Multiscale Inference Network
GVA	Gaussian Visibility Attention
VAF	Visibility Aware Fusion
GAP	Global Average Pooling
GPM	Global Pyramid Module
FC	Fully Connected Feature Fusion
VMP	Visibility Max Pooling
VAP	Visibility Average Pooling
PPM	Part-based Pyramid Module
PAP	Part-based Average Pooling
mAP	Mean Average Precision

List of Tables

List of Figures

1 Introduction

1.1 Motivation

The fast increasing threat of terrorism has become a global problem. Especially, public places are targets for terrorists. In this regard, multiple surveillance camera systems for public spaces have been necessary to prevent a terrorist attack. Moreover, in today's pandemic situation, surveillance systems plays an important role in contact tracing for public health. As more wide area, distributed camera systems are deployed for purposes of security and monitoring, autonomous target tracking systems are in increasing demand. Especially, crowded and complex environments of public spaces requires a tracker that is robust to occlusion. The person re-identification (ReID) task is to identify target pedestrians across multiple cameras with non-overlapping fields-of-view. Person ReID is a key component of such tracking systems in that it especially supports inter-camera target correspondences. Major challenges in person ReID center around appearance change due to variations in camera view, illumination, target pose and occlusion. In response, a wide variety of approaches have been developed, including those that rely on global ([1, 2, 3, 4]), local ([5, 6, 7]) as well combined local-global ([8, 9, 10]) feature representations. Recently, the particular challenge of occlusion, which can become dominant in realistic scenarios involving crowded and geometrically complicated environments, has especially caught the attention of researchers and led to a focus on occluded

person ReID ([11, 12, 13, 14, 15, 16]).

This thesis focuses on the challenge of occluded person ReID, which is a major requirement of desirable tracking systems in public places. In doing so, we make use of a two branch local-global approach, but offer novelty in three major ways. First, to improve the inference process, we introduce a Multistage Multiscale Inference Network (MMI-Net) that consists of three deep learning models that are comprised as i) global, ii) part-based and iii) integrated sub-networks. These models are complementary to each other: The global branch captures holistic representations of the images under analysis; the part-based branch captures more localized representations; the integrated model combines the global and local information. We hierarchically train each sub-part of the model to maintain the discriminative power of each individual feature set (global and local/part-based), and then have MMI-Net integrate visual information and capitalize on each stage's representation by doing separate inferences for final combination. Previous two branch person ReID architectures only make use of the final fusion of local and global features in their inference and thereby may miss intermediate level information. Second, we use multiresolution analysis in both branches of our system to avail it to a wide range of spatial scales. Alternative approaches to person ReID have restricted multiresolution analysis to the global branch and thereby less fully exploit the available information. Third, we make use of attentional mechanisms in both branches that integrate confidence scores generated by body pose analysis to help filter occluded vs. visible regions. While attentional processing has been used previously in occluded person ReID, it has not exploited pose estimation confidence scores in two branch visibility analysis. An overview of our approach is shown in Fig. 1.1.

1.2 Related work

Person ReID has been researched extensively over the past several years [17, 18]. Early works focus on two research directions to tackle the person ReID task. First, some approaches have been introduced to build a robust representation extracted from color and texture information. For example, Gray et al. [19] proposes the ensemble of localized features by merging eight color channel information (RGB, HS, and YCbCr) to identify pedestrian images captured in various viewpoint angles. Ma et al. [20] introduces a covariance feature based on biologically inspired information to overcome the challenge of background and illumination variation. Zhao et al. [21] captures appearance features from pairwise salient regions between images to alleviate body misalignment in matching pedestrian images. Bazzani et al. [22] designs a collection of local representations determined by symmetrical axes of body parts. Zhao et al. [23] proposes mid-level filters trained from image patch clusters to extract not only common but also discriminative visual attributes of pedestrian images. Ma et al. [24] uses Fisher vector framework to aggregate local descriptors. Matsukawa et al. [1] introduces a hierarchical covariance descriptor, which formulates each local patch in an image as a combination with different Gaussian distributions. Second, other early works delve into metric learning to overcome huge intra-class variance in high dimensional features. For instance, Dikmen et al. [25] introduce a cost function to reject nearest neighbor identification with large margin. Zheng et al. [26] proposes a different metric learning framework that finds optimal distance metric by maximizing inter-class variance in appearance features, instead of minimizing intra-class variation. Koestinger et al. [27] designs a simple metric learning building on equivalence constraints for computational efficiency. Hirzer et al. [28] proposes a pairwise metric learning framework to decrease computational complexity by relaxing the hard constraints in optimization. Roth et al. [29] uses Mahalanobis distance to learn a linear transformation matrix of the feature space to match the same pedestrian images.

Recently, with the rise of deep learning, most work has investigated deep neural network models to learn effective descriptors. One class of approaches focus on extraction of discriminative representations from full body appearance of pedestrian images. Extant approaches variously employ deep metric learning [2, 3, 4] and use attention modules [30, 31, 32]. For example. several methods [2, 3, 4] exploit the concept of distance metric learning to improve the embedding space over input data. Hermans et al. [2] demonstrates that a backbone CNN model with a variant of a triplet loss outperforms most other existing models. Chen et al. [3] designs a quadruplet loss, which is an extended version of the triplet loss and allows a deep neural network to learn better representation space in terms of inter-class variation and intra-class variation. Yu et al. [4] develops the point-to-set triplet loss to alleviate the sampling problem in the training process. Other approaches [30, 31, 32] use an attentive cue to learn more robust features. Lan et al. [30] designs a deep reinforcement learning model to effectively select an attention region in a bounding box image. Si et al. [31] introduces a dual attention framework to conduct both representation refinement and alignment in parallel. Wang et al. [32] designs an attention module that leverages both channel and spatial level of attention mechanisms to overcome visual ambiguity. While effective in capitalizing on global appearance information, these approaches easily miss fine scale visual cues. Furthermore, such approaches may pay equal attention to occluded and non-occluded regions, which limits their ability to perform correct identification as occlusion becomes prominent. Part-based approaches concentrate on fine level detailed information from local image regions to enhance the discriminative power of their representations. These approaches employ various strategies including horizontal partitioning [5, 6], segmentation [7], and two branch representations, one for global and another for local representation [8, 9, 10].

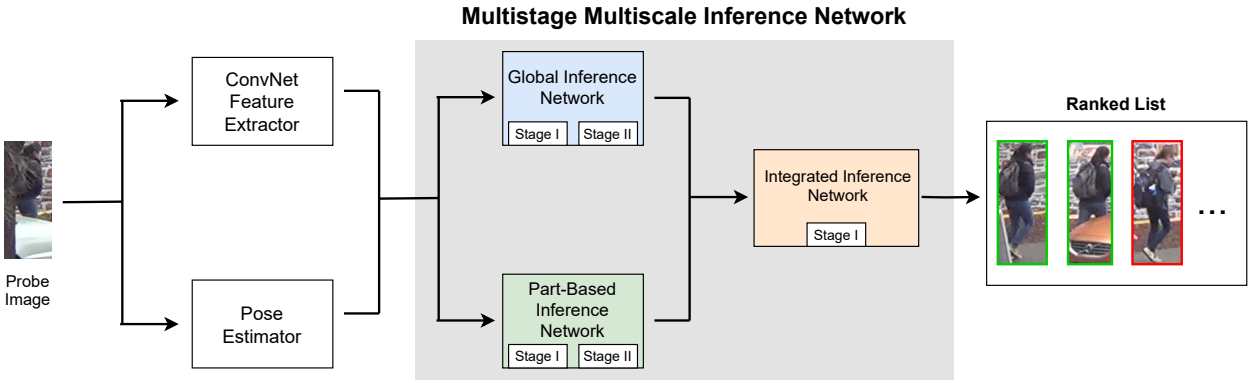


Figure 1.1: Overview of Multistage Multiscale Inference Network (MMI-Net) for occluded person ReID. An input image is preprocessed to extract two kinds of complementary features, ConvNet appearance features and body pose features. These features are passed to MMI-Net that is comprised of two branches, one to capture global information, the second to capture part-based (i.e., local information). Each of these branches provide ID inferences in two stages. Following the individual inferences, an integrated inference is performed that leads to a ranked list of gallery identifications. MMI-Net is unique in its integrated multistage inference architecture that accounts for local and global appearance with attentional processing.

Sun et al. [5] introduce a part-based convolutional model to capture fine scale information using a uniform partition strategy. Zheng et al. [6] develops a deep neural network to encode partial features of a target at multiple pyramid scales. Kalayeh et al. [7] incorporates a human semantic parsing strategy to effectively encode local visual information of different body parts.

Two branch approaches [8, 9, 10] exploit both global and local representations. For instance, Su et al. [8] proposes a deep learning architecture that conducts an affine transformation for each body part image to mitigate the pose variation problem. Li et al. [9] aggregates multi-scale representations using a stack of dilated convolutions with different ratios. Zhu et al. [10] present a viewpoint based loss with angular regularization to disentangle both identity and viewpoint sub-feature spaces. While purely local approaches can sacrifice global information, the two branch approaches have the merit of combining both local and global. However, to date these approaches only take advantage of the final fused feature stage and thereby miss potentially useful intermediate information. In response, our MMI model, while also having a two branch architecture, conducts multiple inferences across various processing stages. Moreover, in alternative two branch approaches multiresolution analysis is used only in the global branch; however, we make use of multiresolution in both branches to capture a wider range information across scales.

Recently, some research has especially concentrated on ReID in the context of occluded pedestrian images in crowded and complicated scenes. Extant approaches employ saliency maps [14, 33] and pose information [11, 12, 13] to pay more attention to visible parts. One approach formulates a pyramid reconstruction error between a probe and gallery image to compute foreground attention scores for occlusion and misalignment [14]. Another introduced a model that learns visible parts of pedestrian images from a self supervision signal [33]. A deep learning framework has been used to exploit topology of body keypoint nodes [11]. Pseudo visibility labels have been used to refine pose based attention maps [12]. Gaussian

attention maps for each body keypoint have been used to suppress occluded regions [13]. To various extents, these approaches ameliorate difficulties in dealing with heavily occluded scenarios, but fail to exploit confidence scores associated with pose estimation, which could further robustness to occlusions. Our MMI-Net also makes use of body pose and attentional processing, but is unique in proposing a Gaussian Visibility Attentive (GVA) mechanism to give more weight to local features with higher confidence scores. Furthermore, we employ a Visibility Aware Fusion (VAF) to aggregate partial features on shared visible parts between two images.

1.3 Contributions

We design a ReID system for operation in standard environments: a) all pedestrians have a normal walking pose; b) all cameras are upright; c) all cameras use a standard lens which yields an image that is approximately equal to what the human eye sees; d) Illumination variation over multiple cameras is small. Therefore, performance of the proposed system will degrade in non-standard environments that violate these assumptions. In this thesis, our main contributions are as follows:

- We present a Multistage Multiscale Inference Network (MMI-Net) that consists of three sub models for a bottom-up inference process with visibility guidance. No previous occluded person ReID approach has made use of such multistage inference; instead, they make a single inference that combines all intermediate processing results with the risk of missing some discriminative information.

- We introduce multiresolution analysis in both global and local sub-networks to benefit from coarse and fine level of visual information. Previous work has restricted multiresolution analysis to a global branch and thereby less thoroughly exploits visual information.

- We propose two attentional processing modules including a Gaussian Visibility Attentive (GVA) and a Visibility Aware Fusion (VAF). Previous work has not made use of such two stage visibility analysis to combat occlusion.

- Our model shows better performance than contemporary alternatives on three occluded person re-identification datasets: Occluded-DukeMTMC [13], Partial-iLIDS [34, 35] and P-DukeMTMC-reID [36]. Remarkably, our approach enhances performance by +9.3% Rank-1 and +5.8% mAP in terms of average score on Occluded-DukeMTMC, +6.2% Rank-1 and +0.7% Rank-3 in mean on Partial-iLIDS, and +5.3% Rank-1 and +7.3% mAP in max on P-DukeMTMC.

1.4 Outline of Thesis

This dissertation consists of four chapters. The first chapter has motivated the problem of occluded person re-identification (ReID), reviewed related research and highlighted major contributions of this thesis. Chapter 2 details the developed technical approach to occluded person ReID. Chapter 3 describes empirical evaluation of the developed approach. Finally, chapter 4 provides conclusions.

2 Technical approach

2.1 Overview

The core of our approach is the Multistage Multiscale Inference Network (MMI-Net), as depicted in Fig 1.1. MMI-Net is a two branch model that processes its input to extract separate global and part-based representations in two sub-networks, followed by an integrated sub-network that combines the representations from the global and part-based to yield final inference of person ReID. The global branch serves to capture a holistic representation, while the part-based branch serves to capture more localized representations. Both branches operate in two stages. The first stage processes the input feature maps in multiple ways to capture different aspects of the data; the second fuses these processed feature maps for more compact representation. Finally, the integrated sub-network combines the part-based and global representations. Body pose guided attentional processing is used in both branches to highlight visibility for robustness to occlusions. To best exploit the discriminative capability of each component of our model and maximize ReID performance in the end, we perform training in multiple stages, each focused on particular components.

Input to MMI-Net comes from preprocessing an image in two complementary ways. i) Appearance features are extracted to serve as the backbone representation of an image for re-identification. We consider two ConvNets for backbone feature extraction, and these

ConvNets provide primitive appearance features over which MMI-Net opertates. Primarily, we consider ResNet50 because of its generally strong performance in recognition tasks [37]; moreover, ResNet50 is a popular backbone for recent alternative person ReID approaches [13, 11, 12] and the choice thereby helps in empirical comparisons. We also considered HRNet-W32 [38] for the sake of comparison to one alternative ReID approach that used it as a backbone [39]. Notably, however, our approach could accommodate most any approach that provides dense feature maps. In our application, ResNet50 produces a dense map of dimensions $24 \times 8 \times 2048$, while HRNet-W32 produces a dense feature map of dimensions $96 \times 32 \times 256$. ii) Body pose features are extracted to guide attentional processing to particularly salient portions of the appearance feature maps. We make use of Alphapose because of its generally strong performance [40, 41]; although, once again, our approach could make use of other pose extractors that provide keypoint coordinates with confidence scores. Alphapose provides coordinates of 18 detected body keypoints as well as associated confidence scores. In the remainder of this section we detail all processing embodied by MMI-Net. Some details on ResNet50, HRNet-W32, and Alphapose are provided in the appendices A, B, and C, respectively.

2.2 Global inference sub-network

Figure 2.1 provides a detailed depiction of the global inference sub-network. It makes use of three different intermediate processing strategies to capture a holistic view: Global Average Pooling (GAP), Gaussian Visibility Attention (GVA), and Global Pyramid Module (GPM). GAP captures general visual information as a baseline representation. The GVA module constructs a visibility weighted representation to suppress occluded region features. The GPM exploits multiscale visual information. After these three different representations are obtained, the features are fused by fully connected layers. To exploit more information during

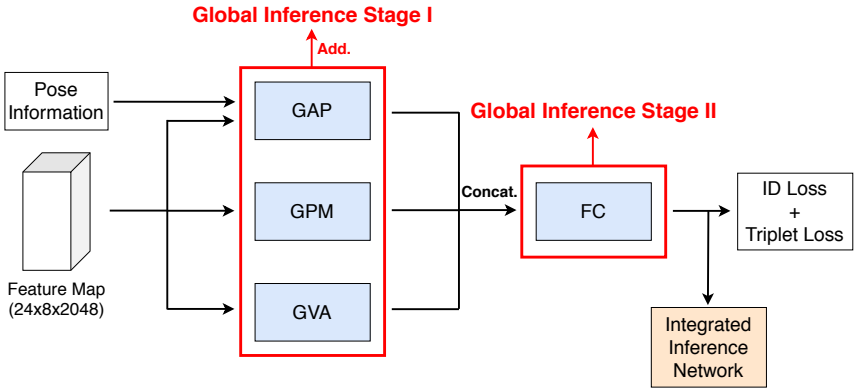


Figure 2.1: Global Inference Sub-Network. The global model makes use of three modules to capture visual information in a holistic view, but under three different representations: i) Global Average Pooling (GAP), ii) Global Pyramid Module (GPM), and iii) Gaussian Visibility Attention (GVA). Following fully-connected feature fusion (FC) the output features are fed to the Integrated Inference Sub-Network. In addition to sending output to the Integrated Sub-Network, there are two inference stages in the global branch indicated by red arrows: One operates on the simple sum of the representations; the other operates on representations that are more deeply fused using fully-connected processing. Training of the sub-network is performed under ID and triplet losses.

inference, we conduct two stages of inference in the global branch. As shown in Fig. 2.1, the first inference stage uses the sum of the three representations from the GAP, GVA, and GPM modules. The second stage leverages a deeper level of fusion via appeal to fully connected layers.

2.2.1 Global Average Pooling.

GAP operates by simple averaging over the input appearance feature maps, M, to yield a 1×2048 representation. For an input feature map, M, we symbolize the operation as $\mathcal{GAP}(M)$.

2.2.2 Gaussian Visibility Attention.

In the GVA module, we construct a visibility aware representation to concentrate attention on regions in the appearance feature map, M, that are most likely visible. Figure 2.2 provides a depiction. The approach is motivated to provide a degree of robustness to occluded body regions. We judge visibility in terms of the keypoints and confidences provided as input from the body pose preprocessing, [11, 12, 13]. In particular, inspired by [13], we employ 2D Gaussian maps centered on each keypoint as attentional weights. Unlike [13], we incorporate keypoint confidences in our processing for more selective attention. Let G_j be the resulting attentional map for keypoint j, c_j be its confidence score, and M be the backbone appearance feature map. Our attentional pooling layer captures a set of pose-based local representations by taking the Hadamard product of G_j weighted by c_j, with M and then applying average pooling, GAP, to the result to yield

$$\tilde{F}(j) = \mathcal{GAP}(c_j G_j \circ M) \quad (j = 1, ..., n_{pose}), \tag{2.1}$$

where n_{pose} is the number of body keypoints in pose estimation. We weight the attentional maps with confidence values so that the most reliable keypoints are most heavily considered

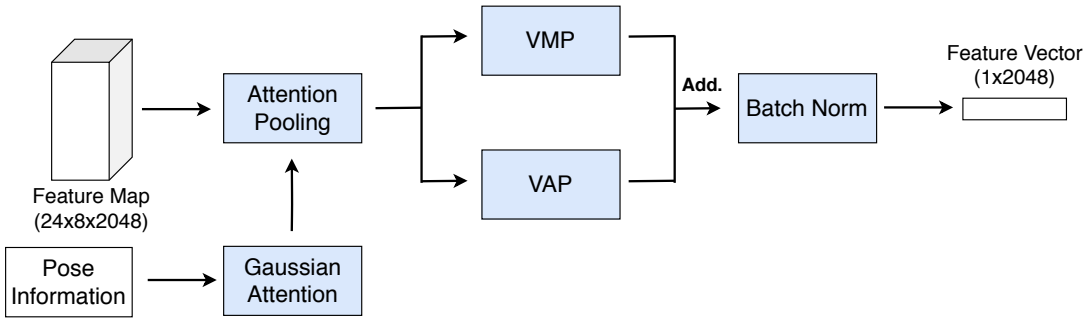


Figure 2.2: Gaussian Visibility Attention. An expanded view of the GVA module.

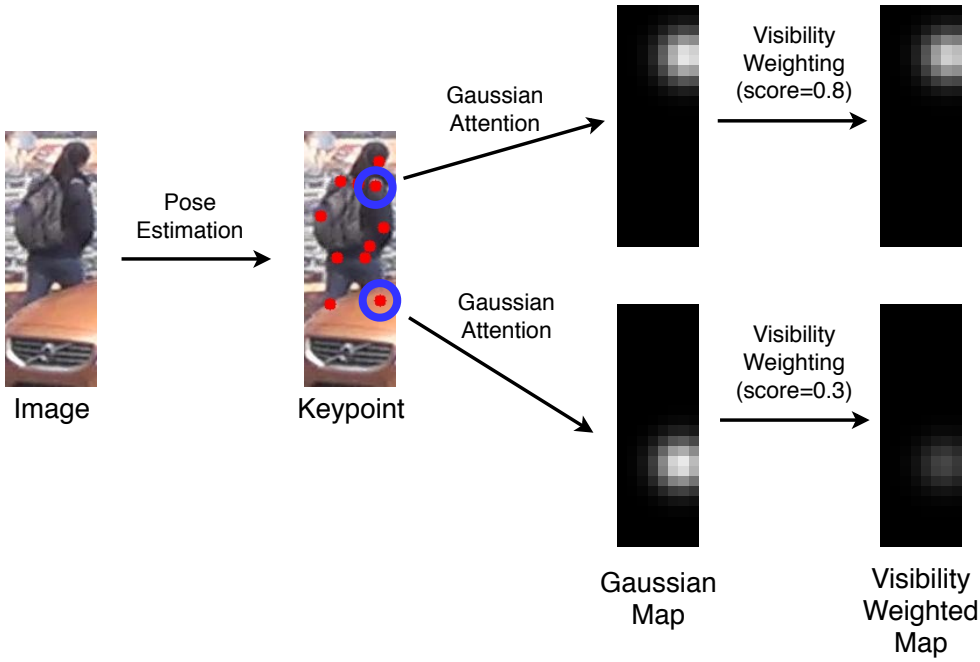


Figure 2.3: Visualization of Gaussian Visibility Attention. Given an input image (left) Visibility Weighted Maps are generated in a three step process. i) Body keypoints are extracted (shown as red dots overlaid on the keypoint image). ii) 2D Gaussian Maps are generated centered on each keypoint. In the example, maps for only two keypoints are shown (indicated with blue circles on the keypoint image) to avoid clutter; in actual operation a separate map would be generated for all keypoints. iii) The Gaussian Maps are weighted by the confidence scores of their corresponding keypoints to yield Visibility Weighted Maps. Subsequently, the Visibility Weighted Maps are applied to backbone features to guide attention to salient body points extracted with most confidence (not shown). Brighter image values indicate higher intensities in the maps shown.

to further strengthen robustness to occlusion. It appears that no previous person ReID approach has made use of pose confidence scores in visibility analysis. Figure 2.3 provides an illustration. In the example provided, notice the larger intensities around the visibility weighted map for fully visible right shoulder compared to the occluded right knee; thus, attention will be guided to the shoulder, as desired. Without use of confidence weights, equal attention would be paid to the right shoulder and knee in the example.

To assist further processing, we stack each $\tilde{F}(j)$ to yield a 2D matrix, $\tilde{F}$, where each row is a pose-based feature vector. Given the matrix $\tilde{F}$, we introduce Visibility Max Pooling (VMP) and Visibility Average Pooling (VAP) to discount unreliable keypoint features with low visibility scores and to enhance the robustness of the local representations to occlusion. In preliminary experimentation, we found that making use of both max and average pooling yielded superior results compared to using either alone. The VMP layer performs max pooling along the columns of $\tilde{F}$ to yield a 1×2048 representation,

$$F_{vmp} = \mathcal{MAX}(\tilde{F}). \tag{2.2}$$

In addition, the VAP layer performs average pooling along the columns of $\tilde{F}$ to yield another 1×2048 representation,

$$F_{vap} = \mathcal{AVG}(\tilde{F}). \tag{2.3}$$

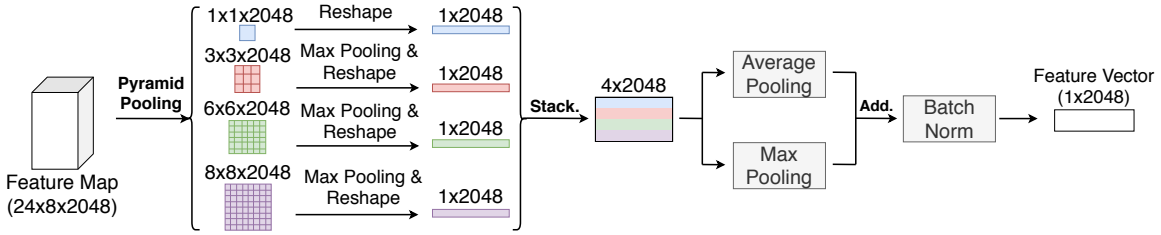
Finally, we sum the two representations from VMP and VAP, and apply normalization to the resulting vector to construct a visibility attentive feature representation, F_{gva}, according to

$$F_{gva} = \mathcal{BN}(F_{vmp} + F_{vap}), \tag{2.4}$$

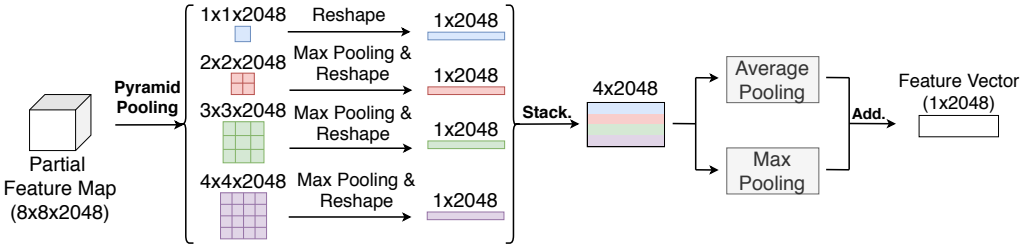
where $\mathcal{BN}$ performs batch normalization.

2.2.3 Global Pyramid Module.

The GPM captures multiscale information from a global perspective using pyramid pooling [42, 43]. As shown in Fig. 2.4-(a), backbone appearance features are average pooled across a range of spatial grids (1×1, 3×3, 6×6 and 8×8). Following pyramid pooling, max pooling



(a) Global Pyramid Module (GPM)



(b) Part-based Pyramid Module (PPM)

Figure 2.4: Two Pyramid Modules. Both global and part-based sub-networks use spatial pyramid processing to extract multiscale representations. Fewer grid cells are used in the part-based pooling because the input parts are smaller than the global input.

is applied to each pooled feature map, followed by reshaping and then stacking to yield a 4×2048 representation. Finally, max and average pooling are applied across the 4 resolutions and summed to yield a 1×2048 feature vector that provides a compact multiscale representation of the appearance features. As in the GVA processing, preliminary experimentation indicated that making use of both max and average pooling yielded superior results compared to using either alone.

2.3 Part-based inference sub-network

Figure 2.5 depicts the part-based inference sub-network. It operates in a patchwise fashion to capture localized detail. In particular, the input appearance feature map is equally subdivided into three nonoverlapping horizontal patches (each $8 \times 8 \times 2048$) to roughly capture the upper, middle and lower portions of the target. On occluded person ReID datasets [13, 34, 36], we observe that horizontal occlusion more frequently happens than vertical. Therefore, our part-based model mainly handles horizontal occlusion and divides the feature map into three horizontal patches. Notabely, however, the GVA module can handle both vertical and horizontal occlusion using local Gaussian attention. Each patch is processed separately with Part-based Average Pooling (PAP) to provide an overall baseline representation as well as with a Part-based Pyramid Module (PPM) to capture within patch multiscale information, analogous to how processing is done in the global branch. To ameliorate the challenge of part misalignment between query and gallery images during Re-ID matching, a feature smoothing module operates across adjacent patches. For better inference process in occluded pedestrian images, a Visibility Aware Fusion (VAF) layer is designed to recognize shared visible regions between query and gallery images and aggregate only features on related parts. This model operates with two inference stages using features before and after fusion.

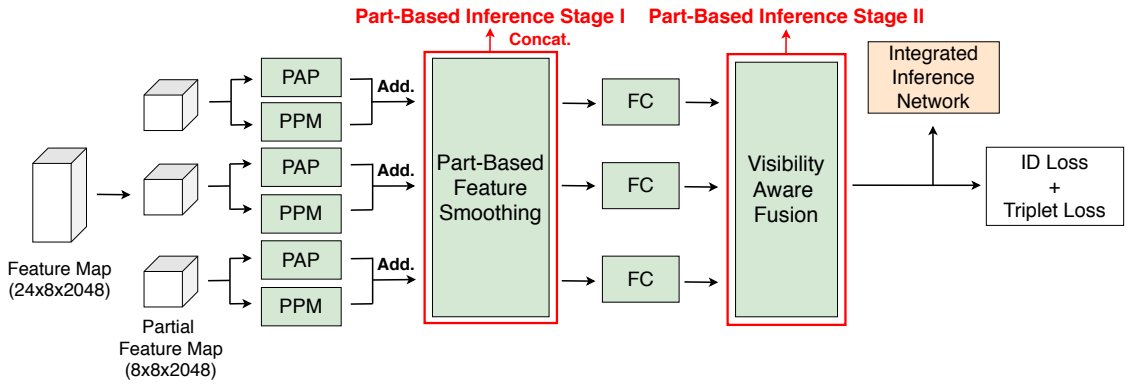


Figure 2.5: Part-Based Inference Sub-Network. The part-based model operates patchwise to extract localized information. The input feature map is decomposed into three horizontal patches that are each processed with a Part-based Average Pooling (PAP) and a Part-based Pyramid Module (PPM). Feature smoothing provides a degree of robustness to part misalignment between query and gallery images for improved matching. Visibility aware Fusion (VAF) aggregates related localized features. Features output by the Part-Based Sub-Network are fed to the Integrated Sub-Network. As in the global branch, there also are two inference stages indicated by red arrows in the part-based branch: One operates prior to fusion, the other operates after fusion. Training of the sub-network is performed under ID and triplet losses.

2.3.1 Part-based Average Pooling.

PAP operates by simple averaging over the input appearance feature maps of each of the three horizontal patches. This operation is exactly the same as GAP, but applied separately to each part to yield three 1×2048 representations.

2.3.2 Part-based Pyramid Module.

Analogous to the GPM, the PPM also captures multiscale representations using pyramid pooling [42, 43]; see Fig. 2.4-(b). Here, the pooling operates separately within each patch to capture multiscale structure on a more localized basis. Due to the individual patches being smaller than the global representation, a different set of pooling grids are used (1×1, 2×2, 3×3 and 4×4). As with the GPM, max and average pooling are applied across the 4 resolutions and summed to yield a 1×2048 feature vector that provides a compact multiscale representation of the appearance features, but now for each patch. Just as we sum the GAP and GPM representations for the global branch to create a unified representation, we sum the PAP and PPM representations, now doing so separately for each patch so that they each have their own unified representation of dimension 1×2048. It appears that previous person ReID approaches have not used pyramid pooling multiresolution analysis in part-based representation.

2.3.3 Part-Based Feature Smoothing.

Given differing capture conditions of a query image of a person to be re-identified and a set of gallery images with respect to which the re-identification is to be established, it is unlikely that the patches between the two are in precise spatial alignment. Correspondingly, the matches between patches across images will be compromised. In response, we apply a spatial

smoothing across the multiscale partwise feature representations. Since information becomes shared between neighboring patches, a degree of robustness to misalignment is achieved. In particular, let F_u, F_m and F_l be the multiscale feature representations extracted for the upper, middle and lower patches, respectively. The smoothed features are then calculated according to

$$\bar{F}_u = (F_u + 0.25F_m)/1.25, \tag{2.5}$$

$$\bar{F}_m = (0.25F_u + F_m + 0.25F_l)/1.5, \tag{2.6}$$

and

$$\bar{F}_l = (0.25F_m + F_l)/1.25 \tag{2.7}$$

where the weight value 0.25 is determined based on the preliminary experiment results. Finally, to extract more sophisticated features, each smoothed patch, $\bar{F}_j, j \in \{u, m, l\}$, is processed by its own fully connected layer to yield $\bar{F}'_j$, of dimension 1×2048, as its representation.

2.3.4 Visibility Aware Fusion.

We apply visibility scores to the extracted part representations to ignore features that likely arise from occluded body regions, which otherwise would negatively impact matching between query and gallery images. In particular, we make use of a body pose guided approach to visibility analysis that yields binary labels, $v_j \in \{1, 0\}$, to indicate visible and nonvisible, respectively [13]. Details of the binary label extraction are available in Appendix D. In the inference process, this fusion module takes each part visibility label of query and gallery, and selects shared visible parts by multiplying each pair of part labels. Given part visibility labels, v_j^q and v_j^g, for part j of query and gallery images, respectively, determined by pose information, this module uses the labels as masks and computes the weighted sum of partial features to fuse only common visible features between query and probes according to

$$\hat{F} = \sum_{j \in \{u,m,l\}} v_j^q v_j^g \bar{F}'_j. \tag{2.8}$$

Figure 2.6 provides a visualization. In the example shown, only the upper (u) and middle (m) parts of the query image (q) are visible, while in the gallery image (g) only the middle (m) and lower (l) parts are visible. Using the values for v_j^q and v_j^g of the example in the formula for feature fusion, (2.8), yields $\hat{F} = \hat{F}'_m$, i.e., only the feature derived from the common visible middle part results for use inference, as desired. Otherwise, the module fuses not only common visible part features but also occluded part features, and it results in huge performance decline. We validate its effectiveness in the ablation study.

2.4 Integrated inference sub-network

The integrated model combines the representations that result from both the global and part-based sub-networks to harness aggregated appearance information; see Fig. 2.7. The final fused features that are output from the individual branches are concatenated and processed by fully connected layers to yield a softmax identification, inference.

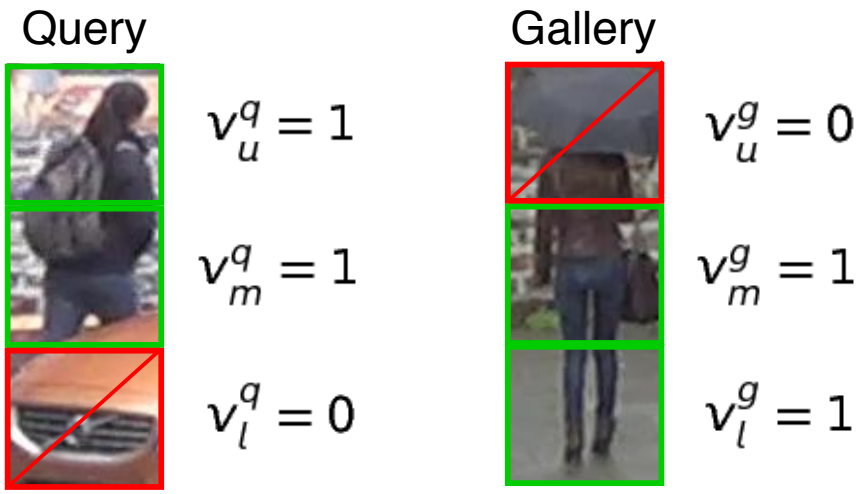


Figure 2.6: Visualization of Visibility Aware Fusion. Query and gallery images to be compared are shown. Extracted parts are overlaid as boxes. Green boxes indicate parts estimated as visible. Red boxes with diagonal lines indicate parts estimated as not visible. Corresponding binary visibility labels are indicated next to the boxes (1 for visible, 0 for not visible). The visibility labels are applied to the feature parts so that only parts that are visible in both the query and gallery are used in inference.

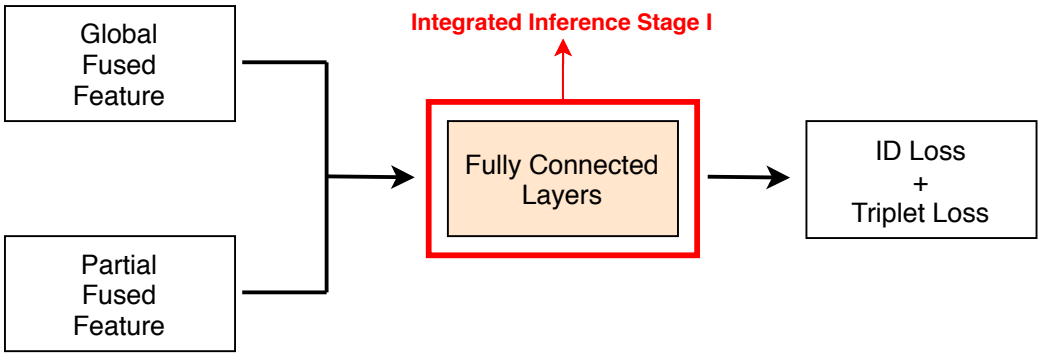


Figure 2.7: Integrated Inference Sub-Network. The integrated model combines the global and part-based visual representations and conducts an overall inference. The input global and part-based feature vectors are concatenated and then processed by fully connected layers. In contrast to the global and part-based sub-networks, the integrated network has only a single inference stage that combines all extracted information.

2.5 Multistage training

To preserve the discriminative capability of each stage's features and maximize ReID performance, we conduct three stage training: i) backbone (fine tuning of the preprocessing appearance ConvNet), ii) global and part-based and iii) integrated model. The training operates in a bottom up manner from backbone to integrated inference network. In each training stage, all parameters in the previously trained models are frozen. To train each model, the objective function is formulated as a sum of a triplet, $L_{triplet}$, and cross entropy loss, $L_{cross\_entropy}$,

$$Loss_{total} = L_{triplet} + L_{cross\_entropy}, \tag{2.9}$$

that is minimized with soft labels [44]. Additional details on multistage training are provided in Section 3.2.

2.6 Multistage inference

To harness different levels of information, we make use of a five stage inference process. Each of the global and part-based models has two inference stages to exploit visual representation before and after fusion of different within branch processing; see Figs 2.1 and 2.5. Additionally, the integrated model provides a combined inference; see Fig 2.7. For each inference stage, s, we measure the cosine distance, $\mathcal{C}$, between query, F_s^q, and gallery, F_s^g, features that result from that stage to yield

$$d_s = \mathcal{C}(F_s^q, F_s^g), \tag{2.10}$$

where $s \in \{\gamma_1, \gamma_2, \pi_{1u}, \pi_{1m}, \pi_{1l}, \pi_2, \iota\}$ indicates the features that result from global stage 1, global stage 2, part-based stage 1 upper patch, part-based stage 1 middle patch, part-based stage 1 lower patch, part-based stage 2 and integrated processing, respectively. To combine

all inference results, we take an average of distance scores, expressed as

$$d_{final} = \frac{\sum_{s\in\{\pi_{1u},\pi_{1m},\pi_{1l}\}} v_s^q v_s^g d_s + \sum_{s\in\{\pi_2,\gamma_1,\gamma_2,\iota\}} d_s}{\sum_{s\in\{\pi_{1u},\pi_{1m},\pi_{1l}\}} v_s^q v_s^g + 4} \tag{2.11}$$

where v_s^q and v_s^g are the part visibility labels of the query and gallery image parts, respectively, analogous to the visibility labels used in VAF. Distances $d_{\{\pi_{1u},\pi_{1m},\pi_{1l}\}}$ have the product of two labels, v_s^q and v_s^g, as weights to only count common visible part distances, whereas other distances $d_{\{\pi_2,\gamma_1,\gamma_2,\iota\}}$ have the same weight of 1. Note that use of visibility labels here and in VAF is not redundant: The former discounts likely occluded individual parts prior to fusion; the latter discounts likely occluded parts after fusion. Previous person ReID approaches have relied on inference based only on fused features, rather than multiple stage inference.

3 Empirical evaluation

3.1 Datasets and Evaluation Protocols

To evaluate MMI-Net on occluded person ReID, we test on three benchmark datasets specifically designed for that purpose, Occluded-DukeMTMC [13], Partial-iLIDS [34, 35] and P-DukeMTMC-reID [36], shown in Fig. 3.1. To verify performance additionally on the more general person ReID task, we use two more datasets, DukeMTMC-reID [45, 46] and Market-1501 [47], shown in Fig. 3.2. Both of the more general datasets also include occlusion challenges, but the portion of occluded query images in each of these datasets is much lower than the occluded person ReID datasets [13, 36].

3.1.1 Occluded-DukeMTMC.

This dataset is a modified version of the more general DukeMTMC-reID dataset [45, 46]. The modification is such that either or both of the query and gallery images have occlusion in computing pairwise feature distances and thereby yields a strong occlusion challenge. This dataset has 15,618 training images and 19,871 testing images, consisting of 17,661 gallery images and 2,210 query images.

3.1.2 Partial-iLIDS.

This dataset contains 238 images of 119 subjects from multiple cameras with non-overlapping fields-of-view. This dataset does not provide a standard train/test split. Training can be performed on a different dataset and testing performed on the entire Partial-iLIDS. We train on Market-1501, as do previous state-of-the-art approaches ([13, 11, 12]).

3.1.3 P-DukeMTMC.

This dataset is another occluded person ReID dataset that has been extracted from the more general DukeMTMC-reID dataset [45, 46]. This dataset contains 12,927 training images of 665 identities, 2,163 query images and 9,053 gallery images of 634 identities.

3.1.4 DukeMTMC-reID.

This dataset is one of the most widely used for general person ReID. It covers 1,404 identities with 16,522 training images, 2,228 query and 17,661 gallery images from eight cameras.

3.1.5 Market-1501.

This dataset also is widely used for general person ReID. The training set has 12,936 images of 751 identities and the testing set has 3,368 query images and 19,732 gallery images.

3.1.6 Evaluation Protocols.

We use the Cumulative Matching Characteristic (CMC) [48] and mean average precision (mAP) [49] to measure person ReID performance. The CMC tests how retrieval performance changes as more images are reacquired from ReID system [48]. For the CMC evaluation,

(a) Occluded-DukeMTMC

(b) Partial-iLIDS

(c) P-DukeMTMC

Figure 3.1: Example Images from Occluded Person ReID Datasets.

(a) DukeMTMC-reID

(b) Market-1501

Figure 3.2: Example Images from General Person ReID Datasets.

we report the same set of Rank-N scores as the second best performing approaches [11, 12]. Rank-N accuracy measures the portion of queries whose top N list from the gallery includes the same person image as the query, desribed as

$$A(q) = \begin{cases} 1 & if\ \exists \text{i}\ GT(q,i) == 1 \\ 0 & else \end{cases} \qquad (i = 1, ..., N),$$

so that

$$Rank\text{-}N = \frac{\sum_{q=1}^{Q} A(q)}{Q}, \tag{3.1}$$

where $A(q)$ specifies whether top N list contains one of ground truth images for the query q, and $GT(q,i)$ returns one if the ith image is the same person image as q, otherwise zero. The mAP measures an overall recall ability of a ReID system under multiple ground truth settings [49, 50], represented by the following equation

$$AP(q) = \frac{\sum_{i=1}^{G}(P(i) * GT(q,i))}{\sum_{j=1}^{G} GT(q,j)},$$

so that

$$mAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q}, \tag{3.2}$$

where G is the number of gallery images, $P(i)$ is the precision in the top i list, $GT(q,i)$ returns a binary value that indicates whether the ith image retrieved in response to the query q has the same identity as q, AP is average precision, and Q is the number of query images. We conduct all experiments under a single query setting. For thorough evaluation, we repeated the training and testing process 10 times and report the average, standard deviation and best scores. Not all alternative approaches provide all these statistics in their evaluations; in such cases we report what was provided in the original papers.

3.2 Implementation Details

We follow the training strategies for person ReID introduced elsewhere [51] to optimize ResNet50, except that we do not use a center loss in the object function. Specifically, we initialize ResNet50 [37] with parameters pre-trained on ImageNet [52] and then further train the entire model on the occluded person ReID datasets for 120 epochs which is what the paper suggests based on their experiment results. To build a batch of data, we randomly select 16 persons and 4 images per person. We use three data augmentation methods: 1) random horizontal flipping with 50% probability, 2) padding 10 pixels and random cropping into the original image size, and 3) random erasing [53]. We use only triplet and cross entropy loss as our cost function, because the contribution of center loss is minor and it significantly increases training time based on our preliminary experimentation. In training process, the learning rate is initialized to 0.00035 and decreased to its 10 % at 40 and 70 epochs, respectively.

In most cases, testing for all datasets made use of the training data provided for that dataset. However, for Partial-iLDS and a transfer learning experiment on P-DukeMTMC we train on Market-1501. Partial-iLDS does not specify its own training data and is overall too small to support training by itself. In a separate additional experiment with P-DukeMTMC we also tested using its own training data. To train our model, we use the same data augmentation methods used by an alternative current state-of-the-art performing model [11]. In detail, the data augmentation techniques are the same as first two methods used in training our backbone model. To evaluate on P-DukeMTMC [36] and Partial-iLIDS [34] with training on Market-1501 [49], random color jitter method is used to alleviate domain transferring challenges.

All fully connected modules consist of two layers. In the part-based and integrated models, the architecture of the fully connected module is the same: a fully connected layer with dropout followed by batch normalization [54], ReLU [54] and another fully connected layer

for classification output. In the global model, we use a simpler architecture that consists of a fully connected layer followed by batch normalization and another fully connected layer to provide a classification output. Input images are resized to 384x128, and the batch size is 64. In each training stage, sub-networks are trained for 60 epochs with a learning rate that starts at 0.1 and decrease to 0.01 after 40 epochs. We employ the Alphapose model [40, 41] pre-trained on the COCO dataset [55] to capture 18 body keypoint locations and confidence scores. In the GVA module, we use the same Gaussian variance as in related work [13], where the variance is the product of image height and image width divided by 1000.

3.3 Comparison with State-of-the-Art Approaches

To compare with extant state-of-the-art approaches, we report performance as given in the original papers.

3.3.1 Results on Occluded-DukeMTMC.

Table 3.1 shows results on the Occluded-DukeMTMC dataset. Our MMI-Net shows 66.4% mean Rank-1 and 66.1% maximum Rank-1 accuracy on this dataset. For mean average precision, our approach yields a 49.6 average score and a 52.5 maximum score. Moreover, across all Ranks as well as mAP, our approach outperforms all other compared approaches by considerable margins. Indeed, for cases where the original papers did not report whether they were providing mean or max scores, our mean always outperforms their reported number, which underlines our better performance as the mean cannot be larger than the max. Compared to the second best performing approach using ResNet50 as a backbone model, HOReID [11], our approach shows better performance by +9.3% Rank-1 and +5.8% mAP in terms of average scores. HOReID [11] also makes use of pose information and visibility, but our further operations (multiscale analysis) apparently captures richer discriminative information. In comparison with another second best performing model [39] using HRNet-

W32 as a backbone model, as described in Appendix B, our method still outperforms by +4.4% Rank-1 and 1.9% mAP using average. For fair comparison, we also use the same backbone model, HRNet-W32. Further, compared with the most recent state-of-the-art two branch approach, PGFA [13], we enhance performance by +14.7% max score on Rank-1 and +15.2% max score on mAP. We think this superior performance results from our multistage inference framework that makes better use of all intermediate visual information with visibility attentive processing. Finally, notice that in these and *all* experiments our standard deviations are small – an indication of the stability of MMI-Net.

Approach	Backbone	Rank-1	Rank-5	Rank-10	mAP
Part-Aligned* [56]	ResNet50	28.8	44.6	51.0	20.2
HACNN* [57]	ResNet50	34.4	51.9	59.4	26.0
Adver Occluded* [58]	ResNet50	44.5	-	-	32.2
PCB* [5]	ResNet50	42.6	57.1	62.9	33.7
Part Bilinear* [59]	ResNet50	36.9	-	-	-
FD-GAN* [60]	ResNet50	40.8	-	-	-
DSR* [61]	ResNet50	40.8	58.2	65.2	30.4
SFR* [62]	ResNet50	42.3	60.3	67.3	32.0
PGFA* [13]	ResNet50	51.4	68.6	74.9	37.3
HOReID [11]	ResNet50	55.1 (-) / -	-	-	43.8 (-) / -
ISP* [39]	HRNet-W32	62.8	78.1	82.9	52.3
MMI-Net (ours)	ResNet50	**64.4** (± 0.94) / **66.1**	**78.2** (± 0.96) / **80.3**	**83.3** (± 0.86) / **85.2**	**49.6** (± 1.09) / **52.5**
MMI-Net (ours)	HRNet-W32	**66.2** (± 0.58) / **67.2**	**79.4** (± 0.86) / **80.9**	**83.9** (± 0.74) / **85.2**	**53.0** (± 0.53) / **54.2**

Table 3.1: Performance comparison on Occluded-DukeMTMC. Under each performance column (Rank-1, Rank-5, Rank-10, mAP) numbers for each approach are reported as mean/max with ± indicating standard deviation across 10 runs. Statistic not reported is indicated with dash (-). * indicates it is not stated in original paper whether average or max scores were reported; in such cases we simply show the numbers given without mean/max formatting. Bold numbers in each column indicate best performance.

3.3.2 Results on Partial-iLIDS.

Table 3.2 shows results on the Partial-iLIDS dataset. It is seen that our approach achieves 78.8% Rank-1 and 87.1% Rank-3 according to mean scores, and 79.8% Rank-1 and 89.1% Rank-3 according to max scores. Again, we surpass all other comparison person ReID approaches. Compared with the second best performing approach, HOReID [11], our MMI-Net shows better performance by +6.2% Rank-1 and +0.7 % Rank-3 in mean. Also notable is our considerable improvement over the previous state-of-the-art two branch approach, PGFA [13], by +10.7% Rank-1 and +8.2% Rank-3 under max. Further, comparison with a previous multiscale approach, FPR [14], shows that our model improves Rank-1 performance by over 10%, likely due to its further enhancement of multistage inference. (Note that mAP typically is not given for Partial-iLIDS, as that statistic was not supplied when the dataset was originally released [34, 35].)

3.3.3 Results on P-DukeMTMC.

Tables 3.3 and 3.4 show the results of testing on P-DukeMTMC when training is on its own training data and on Market-1501, respectively. For training on its own training data, our model yields 90.1% Rank-1 and 76.7% mAP according to the mean and 90.4% Rank-1 and 77.2% mAP according to the max. Based on these results, our model shows better performance than the other approaches by at least 5.3% Rank-1 and 7.3% mAP. Additionally, our model outperforms all other comparison approaches under the transfer setting where training is performed on Market-1501. For example, our approach surpasses the second best performer, PVPM [12], by +1.1% Rank-1 and +2.3% mAP according to the max.

Approach	Rank-1	Rank-3
DSR* [61]	58.8	67.2
SFR* [62]	63.9	74.8
FPR* [14]	68.1	-
PGFA* [13]	69.1	80.9
HOReID [11]	72.6 (-) / -	86.4 (-) / -
MMI-Net (ours)	**78.8** (± 1.37) / **79.8**	**87.1** (± 1.21) / **89.1**

Table 3.2: Performance comparison on Partial-iLIDS. Table layout is analogous to that of Table 3.1.

Approach	Rank-1	Rank-5	Rank-10	mAP
PCB* [5]	79.4	87.1	90.0	63.9
IDE* [47]	82.9	89.4	91.5	65.9
PVPM* [12]	85.1	91.3	93.3	69.9
MMI-Net (ours)	**90.1** (± 0.34) / **90.4**	94.6 (± 0.27) / **94.9**	**95.8** (± 0.30) / **96.2**	**76.7** (± 0.50) / **77.2**

Table 3.3: Performance comparison on P-DukeMTMC with training on its own training data. Table layout analogous to that of Table 3.1.

Approach	Rank-1	Rank-5	Rank-10	mAP
IDE* [47]	36.0	49.3	55.2	19.7
HACNN* [57]	30.4	42.1	49.0	17.0
Part Bilinear* [59]	39.2	50.6	56.4	25.4
PCB* [5]	43.6	57.1	63.3	24.7
PCB+RPP* [5]	40.4	54.6	61.1	23.4
PGFA* [13]	44.2	56.7	63.0	23.1
PVPM* [12]	51.5	64.4	69.6	29.2
MMI-Net (ours)	**52.1** (± 0.81) / **52.6**	**66.2** (± 0.92) / **66.81**	**71.6** (± 0.70) / **71.8**	**31.0** (± 0.37) / **31.5**

Table 3.4: Performance comparison on P-DukeMTMC with training on Market-1501. Table layout analogous to that of Table 3.1.

3.3.4 Results on DukeMTMC-reID and Market-1501.

To evaluate on general person ReID, we also test our model on two general person ReID datasets, DukeMTMC-reID and Market-1501. Here, we conduct comparison not only with alternative occluded person ReID approaches, but also with more general person ReID approaches. Overall, Table 3.5 shows that our approach typically shows comparable or notably better performance compared to existing approaches. For example, MMI-Net and HOReID are within 1% of each other using the mean, but our method outperforms HOReID with big margin on Occluded-DukeMTMC (Table 3.1) and Partial-iLDS (Table 3.2). Especially, our approach shows much better performance than a previous two branch approach, PGFA [13], by +5.8% Rank-1 and 10.2% mAP on the DukeMTMC dataset. Also, in comparison to an alternative part-based approach, PCB and PCB+RPP [5], our approach once again outperforms by a considerable margin. In contrast, the holistic approach, ABD-Net [63], shows best overall performance, perhaps because of its use of both spatial and channel attention. Even compared to ABD-Net, however, MMI-Net is within ≈1% on Rank-1 on both datasets. Another interesting point of comparison is with occluded person ReID approach FPR [14]. On the general person ReID task it somewhat outperforms MMI-Net; however, in occluded person ReID per se it shows worse performance; on Partial-iLDS (Table 3.2) it trails MMI-Net by -11.7% Rank-1 according to the mean. This relative weakness may be due to its use of spatial foreground scores becoming unreliable in the presence of high occlusion. These combined results show that MMI-Net not only uniformly sets a new state-of-the-art on occluded person ReID, but also is a strong competitor on the general person ReID task.

Approach	DukeMTMC-reID		Market-1501	
	Rank-1	mAP	Rank-1	mAP
IDE* [47]	-	-	79.5	59.9
HACNN* [57]	80.5	63.8	91.2	75.7
Adver Occluded* [58]	79.1	62.1	86.5	78.3
PCB* [5]	81.8	66.1	92.3	77.4
PCB+RPP* [5]	83.3	69.2	93.8	81.6
OSNet* [64]	88.6	73.5	94.8	84.9
ABD-Net* [63]	**89.0**	**78.6**	**95.6**	**88.3**
FPR* [14]	88.6	78.4	95.4	86.6
PGFA* [13]	82.6	65.5	91.2	76.8
HOReID [11]	86.9 (-) / -	75.6 (-) / -	94.2 (-) / -	84.9 (-) / -
MMI-Net (ours)	87.9 (± 0.58) / 88.4	75.3 (± 0.76) / 75.7	94.0 (± 0.26) / 94.4	84.6 (± 0.24) / 85.0

Table 3.5: Performance comparison on DukeMTMC-reID and Market-1501. Layout for each dataset analogous to Table 3.1.

3.4 Ablation Studies

To confirm the effectness of our multistage inference framework, we conducted an ablation experiment on the Occluded-DukeMTMC dataset [13]. Results are shown in Table 3.6. The global sub-network shows 62.7% Rank-1 and 48.1% mAP in mean because it captures overall major visual information for strong discriminative power. The part-based sub-network yields 59.9% Rank-1 and 45.1% mAP in mean. Here, performance derives from successful extraction of locally discriminative information; however, lack of global context keeps it from equaling the global branch. Using both the global and part-based sub-networks together underlines their complementarity, as scores rise to 64.3% Rank-1 and 49.5% mAP in mean. Finally, including the integrated sub-network to bring us back to our full MMI-Net shows an additional, relatively small, yet consistent improvement to mean scores, which indicates that the integrated model helps bridge between the global and part-based models. Consideration of max scores shows the same patterns of relative performance. Taken together, these results support our motivation for including separate inferences from each of the sub-networks to complement each other and produce the overall best results.

We also examined the impacts of our multiresolution and visibility processing. Table 3.7 shows the results of ablating mulitresolution processing. The full MMI-Net makes separate use of multiresolution processing in each of its global and part-based branches. It is seen that removal of the processing from the global branch (- Global Pyramid Module) leads to a larger decrement in performance compared to removal of the processing from the part-based branch (- Part-Based Pyramid Module); however, they both show their separate effectiveness. Moreover, their complementarity is documented by the fact that performance is even more adversely impacted when both modules are ablated together (- GPM and PPM) than when either is ablated alone.

Table 3.8 shows the results of ablating visibility processing. The full MMI-Net makes

separate use of visibility processing in each of its global and part-based branches. It is seen that removal of the processing from the part-based branch (- Visibility Aware Fusion) has a larger impact compared to removal of the processing from the global branch (- Gaussian Visibility Attention); however, they both are effective separately. Also, when the visibility processing is removed from both branches (- VAF and GVA) performance is most severely decremented, which documents the complementarity of the individual modules. (Note that removal of VAF results in all three partial features being summed without multiplication by visibility labels; see Eq. (2.9).)

Finally, comparing across the results of Tables 3.7 and 3.8 it is seen that ablation of multiresolution processing generally has a bigger impact than ablation of visibility processing.

Approach	Rank-1	mAP
Baseline (ResNet50)	51.8 (± 1.42) / 53.2	40.8 (± 0.70) / 41.2
Global Inference Network (only)	62.7 (± 0.79) / 63.5	48.1 (± 0.80) / 48.7
Part-Based Inference Network (only)	59.9 (± 0.75) / 61.3	45.1 (± 0.98) /47.7
Global + Part-Based Inference (only)	64.3 (± 0.83) / 65.9	49.5% (± 1.05 / 52.3
MMI-Net (full)	**64.4** (± 0.94) / **66.1**	**49.6** (± 1.09) / **52.5**

Table 3.6: Ablation study of MMI-Net on Occluded-DukeMTMC dataset. Table layout analogous to that of Table 3.1.

Variation	Rank-1	mAP
MMI-Net (full)	64.4 (± 0.94) / 66.1	49.6 (± 1.09) / 52.5
- Global Pyramid Module (GPM)	60.6 (± 0.73) / 61.6	47.3 (± 1.04) / 49.8
- Part-Based Pyramid Module (PPM)	62.3 (± 0.81) / 63.9	47.7 (± 0.74) / 48.0
- GPM and PPM	55.5 (± 1.21) / 56.4	43.0 (± 0.99) / 44.0

Table 3.7: Multiresolution Processing Ablation Study. Under each performance column (Rank-1 and mAP) numbers for each variation are reported as mean/max with ± indicating standard deviation across 10 runs. - in front of a variation indicates that the corresponding module has been ablated from the full model.

Variation	Rank-1	mAP
MMI-Net (full)	64.4 (± 0.94) / 66.1	49.6 (± 1.09) / 52.5
- Visibility Aware Fusion (VAF)	62.8 (± 0.68) / 63.4	48.9 (± 0.96) / 51.3
- Gaussian Visibility Attention (GVA)	63.9 (± 0.92) / 65.3	49.3 (± 1.12) / 52.2
- VAF and GVA	62.3 (± 0.77) / 63.0	48.5 (± 1.00) / 51.0

Table 3.8: Visibility Processing Ablation Study. Under each performance column (Rank-1 and mAP) numbers for each variation are reported as mean/max with ± indicating standard deviation across 10 runs. - in front of a variation indicates that the corresponding module has been ablated from the full model.

4 Conclusion

4.1 Summary

We have introduced the Multistage Multiscale Inference Network (MMI-Net) for occluded person re-identification in the task of visual tracking. MMI-Net uses multiscale features with visibility analysis for multistage inference. Unique in comparison to alternative approaches, it makes use of multistage inference to avail itself to all incrementally extracted information, rather than relying only on a final fusion of extracted features. Empirical evaluation reveals that MMI-Net sets a new state-of-the-art on occluded person re-identification and also is competitive on the general person re-identification challenge. These results suggest that MMI-Net can serve as an especially valuable component to a visual tracking system that must cope with occlusion scenarios, even while also being useful in more general deployments.

4.2 Directions for future research

Several directions for future research could be considered. First, in this work, we concentrate on spatial attention with visibility guidance, but there is potential to employ channel attention to enhance its performance, c.f., [65]. Attention focus on feature channels that are most salient may provide further discriminative power to our approach. Second, early fusion

between global and local features in stage 1 inference could provide additional semantic insight into the target image. As in the VAF module, we can use the product of two confidence scores of query and gallery images to recognize shared visible keypoints, and this approach might enhance inference process and performance. Third, the developed algorithm could be incorporated into a complete multitarget, multicamera target tracking system. Fourth, both vertical and horizontal partitioning can be applied to feature map for handling not only horizontal occlusion but also vertical. Lastly, to validate further capability of our model, it could be evaluated on non-standard environments with more realistic scenarios.

A ResNet50

Previous works [66, 67] exploit exceptionally deep architecture to enhance discriminative power of appearance representation. However, the methods are still relatively shallow with maximum depth of thirty [67]. To build extremely deep models, one major problem that previous approaches couldn't solve is degradation where accuracy does not increase even with increased depth. To overcome the degradation problem, He et al. [37] introduce a residual learning approach that mitigates the training process of extremely deep neural networks. The residual learning framework is to stack nonlinear layers with shortcut connections. The basic block of a residual network is formulated as an element-wise addition of residual output $F(x)$ and input x,

$$y = F(x) + x \tag{A.1}$$

where F is a feed forward network and y is the final output of the residual block. In pratice, these connections ameliorate the problem of vashing gradient during training because the signal "skips over" intermediate layers. Based on empirical results, the residual networks can easily be trained and show better performance than simply stacked networks without any skip connection. ResNet50 is a relatively shallow version of residual networks that consists of 50 layers. The layers can be grouped into five stages that includes one simple convolution stage and four residual stages. The convolution stage consists of a 7x7 convolution and 3x3 max pooling layer. Afterward, each residual stage includes multiple residual blocks where 1×1, 3×3, 1×1 convolution layers are stacked with skip connection. Full details on ResNet50 can be found elsewhere [37].

B HRNet-W32

Many previous pose estimation approaches [68, 69] make use of an encoder-decoder structure to generate high resolution heatmap for body keypoint. However, to estimate more accurate body keypoints, Sun et al. [38] introduces the High-Resolution Network (HRNet) which consists of multiple different resolution branches in parallel and repeatedly adds multi-scale fusion connections to the branches. Each branch is a convolutional neural network connected to different branches, and it starts from one branch and adds one branch at each fusion stage up to four branches. The multi-scale fusion strategy is to exchange different resolution representations across branches so that each branch acquires the multi resolution information from other parallel branch. The fusion allows for capturing of rich appearance information through multiple aggregation processes of different resolution representations. In addition, preserving high resolution information through the whole architecture reduces localization error in human pose estimation. In the last fusion stage, all features from four branches are concatenated to obtain the final representation. Given the fused feature, a regression layer estimates heatmaps for body keypoints. HRNet-W32 is a small sized variant of HRNet that maintains a 32 dimensional representation in the last three fusion stages of the highest resolution branch. To use the HRNet-W32 as a backbone, we remove the regression module to obtain a feature map. Full details on HRNet can be found elsewhere [38].

C Alphapose

A two-step strategy in human pose estimation conducts top-down process; human detection model localizes all candidates of human beings and a single-person pose estimator (SPPE) estimates all body keypoints of a person from each detection [70, 71]. However, previous two-step pose estimation methods are vulnerable to human detection errors including the inaccurate localization and redundant detection. To mitigate the two challenges, Fang et al. [40] proposes Alphapose, a regional multi-person pose estimation system which consists of three modules: 1) Symmetric Spatial Transformer Network (SSTN), 2) Parallel Pose Non-Maximum-Suppression (NMS), and 3) Pose-Guided Proposals Generator (PGPG). First, the paper first makes use of Faster-RCNN [72] for human detection. The STTN module learns a 2D transformation to correct localization errors in human detection using spatial transformation networks [73], and then the SPPE model, Stacked Hourglass [68], performs more accurate single-person pose estimation through the image transformation. Given pose proposals from the STTN, the NMS filters out redundant keypoints based on pose similarity determined by its own pose distance metric. During the training process, the STTN module is optimized on the extant training data augmented by the PGPG module. The PGPG learns the distribution of bounding box proposals from a human detection model, and it generates human bounding boxes to augment training data. Full details on Alphapose are available elsewhere [40].

D Visibility label extraction

Miao et al. [13] propose a part binary label that estimates visibility of body parts to capture common visible regions between query and gallery images. To calculate the visibility labels, a pose estimation model first approximates body keypoint information of location and confidence. The confidence score is originally designed to describe reliability of keypoint estimation. However, in the work [13], the reliability information is used to measure which part region is occluded or not. For example, if one of body part is occluded, then confidence scores of landmarks on the occluded part region should be much lower than on visible parts. Given keypoint information, points with scores lower than threshold are removed. Eventually, the ith horizontal patch visibility label v_i (i=1, ..., P) is calculated by the following equation,

$$v_i = \begin{cases} 1 & if\ \exists \mathrm{cy}_j\ \in [\frac{i-1}{P}H, \frac{i}{P}H) \\ 0 & else \end{cases} (j = 1, ..., N), \tag{D.1}$$

where H is the image height, P is the number of horizontal patches, N is the number of points in the filtered set of body landmarks, and cy_j is the longitudinal coordinate of the jth point. In detail, we set the score threshold and the number of patches to 0.2 and 3, respectively. Full details on this approach for visibility detection are available elsewhere [13].

Bibliography

[1] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. "Hierarchical Gaussian descriptor for person re-identification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, Nevada). 2016.

[2] Alexander Hermans, Lucas Beyer, and Bastian Leibe. "In defense of the triplet loss for person re-identification". In: *arXiv preprint arXiv:1703.07737* (2017).

[3] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. "Beyond Triplet Loss: A Deep Quadruplet Network for Person Re-Identification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, Hawaii). 2017.

[4] Rui Yu, Zhiyong Dou, Song Bai, Zhaoxiang Zhang, Yongchao Xu, and Xiang Bai. "Hard-aware point-to-set deep metric for person re-identification". In: *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich, Germany). 2018.

[5] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)".

In: *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich, Germany). 2018.

[6] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. "Pyramidal person re-identification via multi-loss dynamic training". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Long Beach, California). 2019.

[7] Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E. Kamasak, and Mubarak Shah. "Human Semantic Parsing for Person Re-Identification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Salt Lake City, Utah). 2018.

[8] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. "Pose-driven deep convolutional model for person re-identification". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Venice, Italy). 2017.

[9] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. "Learning deep context-aware features over body and latent parts for person re-identification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, Hawaii). 2017.

[10] Zhihui Zhu, Xinyang Jiang, Feng Zheng, Xiaowei Guo, Feiyue Huang, Xing Sun, and Weishi Zheng. "Viewpoint-Aware Loss with Angular Regularization for Person Re-Identification". In: *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)* (New York City, New York). 2020.

[11] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. "High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-Identification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Online). 2020.

[12] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. "Pose-guided Visible Part Matching for Occluded Person ReID". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Online). 2020.

[13] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. "Pose-guided feature alignment for occluded person re-identification". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Seoul, South Korea). 2019.

[14] Lingxiao He, Yinggang Wang, Wu Liu, He Zhao, Zhenan Sun, and Jiashi Feng. "Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Seoul, South Korea). 2019.

[15] Lingxiao He and Wu Liu. "Guided Saliency Feature Learning for Person Re-identification in Crowded Scenes". In: *Proceedings of the European Conference on Computer Vision (ECCV)* (Online). 2020.

[16] Zhao Shizhen, Gao Changxin, Zhang Jun, Cheng Hao, Han Chuchu, Jiang Xinyang, Guo Xiaowei, Zheng Wei-Shi, Sang Nong, and Sun Xing. "Do Not Disturb Me: Person Re-identification Under the Interference of Other Pedestrians". In: *Proceedings of the European Conference on Computer Vision (ECCV)* (Online). 2020.

[17] Liang Zheng, Yi Yang, and Alexander Hauptmann. "Person re-identification: Past, present and future". In: *arXiv preprint arXiv: 1610.02984* (2016).

[18] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven Hoi. "Deep learning for person re-identification: A survey". In: *arXiv preprint arXiv: 2001.04193* (2020).

[19] Douglas Gray and Hai Tao. "Viewpoint invariant pedestrian recognition with an ensemble of localized features". In: *Proceedings of the European Conference on Computer Vision (ECCV)* (Marseille, France). 2008.

[20] Bingpeng Ma, Yu Su, and Frederic Jurie. "Covariance descriptor based on bio-inspired features for person re-identification and face verification". In: *Image and Vision Computing* 32.6-7 (2014), pp. 379–390.

[21] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. "Person re-identification by salience matching". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Sydney, Australia). 2013.

[22] Loris Bazzani, Marco Cristani, and Vittorio Murino. "Symmetry-driven accumulation of local features for human characterization and re-identification". In: *Computer Vision and Image Understanding* 117.2 (2013), pp. 130–144.

[23] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. "Learning mid-level filters for person re-identification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Columbus, Ohio). 2014.

[24] Bingpeng Ma, Yu Su, and Frédéric Jurie. "Local descriptors encoded by fisher vectors for person re-identification". In: *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)* (Florence, Italy). 2012.

[25] Mert Dikmen, Emre Akbas, Thomas S Huang, and Narendra Ahuja. "Pedestrian recognition with a learned metric". In: *Proceedings of the Asian Conference on Computer Vision (ACCV)* (Queenstwon, New Zealand). 2010.

[26] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. "Person re-identification by probabilistic relative distance comparison". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Colorado Springs, Colorado). 2011.

[27] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. "Large scale metric learning from equivalence constraints". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Providence, Rhode Island). 2012.

[28] Martin Hirzer, Peter M Roth, Martin Köstinger, and Horst Bischof. "Relaxed pairwise learned metric for person re-identification". In: *Proceedings of the European Conference on Computer Vision (ECCV)* (Florence, Italy). 2012.

[29] Peter M Roth, Martin Hirzer, Martin Köstinger, Csaba Beleznai, and Horst Bischof. "Mahalanobis distance learning for person re-identification". In: *Person re-identification*. Springer, 2014, pp. 247–267.

[30] Xu Lan, Hangxiao Wang, Shaogang Gong, and Xiatian Zhu. "Deep Reinforcement Learning Attention Selection For Person Re-Identification." In: *Proceedings of the British Machine Vision Conference (BMVC)* (London, United Kingdom). 2017.

[31] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C. Kot, and Gang Wang. "Dual Attention Matching Network for Context-Aware Feature Sequence Based Person Re-Identification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Salt Lake City, Utah). 2018.

[32] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. "MANCS: A multi-task attentional network with curriculum sampling for person re-identification". In: *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich, Germany). 2018.

[33] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. "Perceive Where to Focus: Learning Visibility-Aware Part-Level Features for Partial Person Re-Identification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Long Beach, California). 2019.

[34] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. "Person re-identification by probabilistic relative distance comparison". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Colorado Springs, Colorado). 2011.

[35] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. "Partial Person Re-Identification". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Santiago, Chile). 2015.

[36] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. "Occluded person re-identification". In: *IEEE International Conference on Multimedia and Expo (ICME)* (San Diego, California). 2018.

[37] Kaiming He, Zhang Xiangyu, Ren Shaoqing, and Sun Jian. "Deep residual learning for image recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, Nevada). 2016.

[38] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. "Deep High-Resolution Representation Learning for Human Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Long Beach, California). 2019.

[39] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. "Identity-Guided Human Semantic Parsing for Person Re-Identification". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2020.

[40] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. "RMPE: Regional Multi-Person Pose Estimation". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Venice, Italy). 2017.

[41] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. "Pose Flow: Efficient Online Pose Tracking". In: *Proceedings of the British Machine Vision Conference (BMVC)* (Newcastle, United Kingdom). 2018.

[42] Svetlana Lazebnik, Cordelia Schimd, and Jean Ponce. "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories". In: *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR) (New York City, New York). 2006.

[43] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. "Pyramid Scene Parsing Network". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, Hawaii). 2017.

[44] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. "Rethinking the Inception Architecture for Computer Vision". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, Nevada). 2016.

[45] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. "Performance measures and a data set for multi-target, multi-camera tracking". In: *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)* (Amsterdam, Netherlands). 2016.

[46] Zhedong Zheng, Liang Zheng, and Yi Yang. "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Venice, Italy). 2017.

[47] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. "Scalable person re-identification: A benchmark". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Santiago, Chile). 2015.

[48] Douglas Gray, Shane Brennan, and Hai Tao. "Evaluating appearance models for recognition, reacquisition, and tracking". In: *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)* (Rio de Janeiro, Brazil). 2007.

[49] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. "Scalable person re-identification: A benchmark". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Santiago, Chile). 2015.

[50] Wikipedia contributors. *Evaluation measures (information retrieval) — Wikipedia, The Free Encyclopedia*. [Online; accessed 18-February-2021]. 2021.

[51] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. "Bag of Tricks and a Strong Baseline for Deep Person Re-Identification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Long Beach, California). 2019.

[52] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252.

[53] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. "Random Erasing Data Augmentation". In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* (New York City, New York). 2020.

[54] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[55] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. "Microsoft COCO: Common objects in context". In: *European Conference on Computer Vision (ECCV)* (Zurich, Switzerland). 2014.

[56] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. "Deeply-Learned Part-Aligned Representations for Person Re-Identification". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Venice, Italy). 2017.

[57] Wei Li, Xiatian Zhu, and Shaogang Gong. "Harmonious Attention Network for Person Re-Identification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Salt Lake City, Utah). 2018.

[58] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. "Adversarially Occluded Samples for Person Re-Identification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Salt Lake City, Utah). 2018.

[59] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. "Part-Aligned Bilinear Representations for Person Re-Identification". In: *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich, Germany). 2018.

[60] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, and hongsheng Li. "FD-GAN: Pose-guided Feature Distilling GAN for Robust Person Re-

identification". In: *Conference on Neural Information Processing Systems (NeurIPS)* (Montreal, Canada). 2018.

[61] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. "Deep Spatial Feature Reconstruction for Partial Person Re-Identification: Alignment-Free Approach". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Salt Lake City, Utah). 2018.

[62] Lingxiao He, Zhenan Sun, Yuhao Zhu, and Yunbo Wang. "Recognizing partial biometric patterns". In: *arXiv preprint arXiv:1810.07399* (2018).

[63] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. "ABD-Net: Attentive but Diverse Person Re-Identification". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Seoul, South Korea). 2019.

[64] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. "Omni-Scale Feature Learning for Person Re-Identification". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Seoul, South Korea). 2019.

[65] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. "CBAM: Convolutional Block Attention Module". In: *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich, Germany). 2018.

[66] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *International Conference on Learning Representations (ICLR)*. 2015.

[67] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *International Conference on Machine Learning (ICML)*. 2015.

[68] Alejandro Newell, Kaiyu Yang, and Jia Deng. “Stacked hourglass networks for human pose estimation”. In: *Proceedings of the European Conference on Computer Vision (ECCV)* (Amsterdam, Netherlands). 2016.

[69] Bin Xiao, Haiping Wu, and Yichen Wei. “Simple baselines for human pose estimation and tracking”. In: *Proceedings of the European conference on computer vision (ECCV)* (Munich, Germany). 2018.

[70] Georgia Gkioxari, Bharath Hariharan, Ross Girshick, and Jitendra Malik. “Using k-Poselets for Detecting People and Localizing Their Keypoints”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Columbus, Ohio). 2014.

[71] Leonid Pishchulin, Arjun Jain, Mykhaylo Andriluka, Thorsten Thormählen, and Bernt Schiele. “Articulated people detection and pose estimation: Reshaping the future”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Providence, Rhode Island). 2012.

[72] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: *Conference on Neural Information Processing Systems (NeurIPS)* (Montreal, Canada). Vol. 28. 2015.

[73] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. “Spatial Transformer Networks”. In: *Conference on Neural Information Processing Systems (NeurIPS)* (Montreal, Canada). 2015.