Partial Person Re-identification with Part-Part Correspondence Learning

Tianyu He[1], Xu Shen[1], Jianqiang Huang[1], Zhibo Chen[2], and Xian-Sheng Hua[1]
[1]DAMO Academy, Alibaba Group
[2]University of Science and Technology of China
timhe.hty@alibaba-inc.com

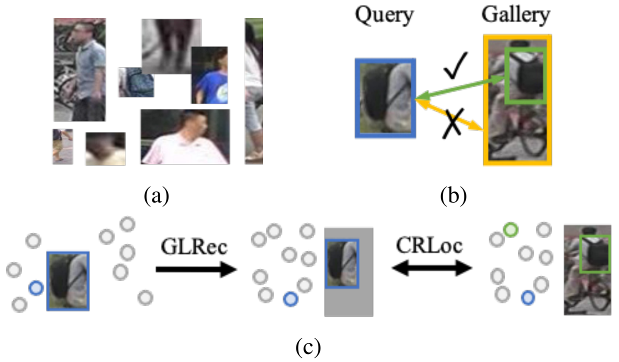
Abstract

Driven by the success of deep learning, the last decade has seen rapid advances in person re-identification (re-ID). Nonetheless, most of approaches assume that the input is given with the fulfillment of expectations, while imperfect input remains rarely explored to date, which is a non-trivial problem since directly apply existing methods without adjustment can cause significant performance degradation. In this paper, we focus on recognizing partial (flawed) input with the assistance of proposed Part-Part Correspondence Learning (PPCL), a self-supervised learning framework that learns correspondence between image patches without any additional part-level supervision. Accordingly, we propose Part-Part Cycle (PP-Cycle) constraint and Part-Part Triplet (PP-Triplet) constraint that exploit the duality and uniqueness between corresponding image patches respectively. We verify our proposed PPCL on several partial person re-ID benchmarks. Experimental results demonstrate that our approach can surpass previous methods in terms of the standard evaluation metric.

1. Introduction

For most of computer vision tasks, the input data for algorithms is commonly assumed to be complete or with adequate information that can be recognized [28, 48, 38, 44, 62]. But in the real-world scenario, especially for person re-identification (re-ID) in the surveillance video, this assumption can not always be satisfied due to flawed data process pipeline (*e.g.*, imperfect detectors [41], communication failures, pose variances, *etc.*). While as shown in Figure 5, trivially apply existing methods without adjustment causes significant performance degradation, since Convolutional Neural Networks (CNN) usually suffer from understanding part-whole relationships [27]. In this paper, we focus on this practical yet challenging problem, where the input image is represented by only a continuous part of the original one.

To alleviate the challenge of partial person re-ID, there are two main difficulties: 1) as shown in Figure 1a, the partial input images are typically provided with unknown size and scale; 2) as shown in Figure 1b, when computing the similarity with the reference image, the irrelevant partial regions may bring much noise for recognition. To our knowledge, the present studies for partial person re-ID have been developed from two perspectives: 1) learn a multi-scale feature to adapt to arbitrary input size [15, 16, 35]; 2) locate shared regions between partial image and the reference one [46, 36]. However, these methods typically address one of aforementioned difficulties and less effort has been made to solve both of them in an unified framework. This exactly motivates our work.



(a) (b) (c)

Figure 1: (a): Partial person re-ID is required to adapt to variances on input scales and regions. (b): The irrelevant region needs to be eliminated when measuring similarity between two images. (c): We employ GLRec to rectify arbitrary input and CRLoc to predict semantically corresponding partial region.

In this paper, we tackle partial person re-ID problem by proposing Part-Part Correspondence Learning (PPCL), as illustrated in Figure 1c, in which we propose a *gated layout rectifier* (GLRec) and a *corresponding region locator* (CRLoc) to cope with aforementioned difficulties respectively. Concretely, the GLRec module is a gated transformation regression CNN module that takes an arbitrary partial image x^p in, and outputs a rectified result x^r with predicted affine transformation coefficients. Through the GLRec module, we may reduce input space and obtain a rectified partial image with proper scale and layout. Then, to measure the similarity between the rectified partial image x^r and a reference image y, we extract spatial features of them through a backbone network, and employ the CRLoc module to produce a corresponding patch in y according to x^r. Finally, the similarity between two images are only calculated based on the shared semantically corresponding regions.

However, to train the CRLoc module, a key challenge here is that, for most of image recognition tasks, we only have image-level supervision while there is no part-level correspondence signal available during training. Therefore, we exploit the nature of part-part relationship, and accordingly propose two self-supervised training schemes that work coordinately to achieve PPCL:

- **Part-Part Cycle constraint (PP-Cycle)**: We leverage the duality of the functionality on finding corresponding regions according to the given input. We assume that, if the CRLoc module is able to predict a corresponding patch in y according to x^r, then it will also be able to translate back to achieve x^r. Therefore, we enable a self-supervised cycle consistency constraint in model optimization. The philosophy behind PP-Cycle is demonstrated in Figure 4a.
- **Part-Part Triplet constraint (PP-Triplet)**: We leverage the uniqueness of the optimal corresponding region between two image patches for a given partial input. We regard the given partial input as anchor, the output of the CRLoc module as positive exemplar. Along with the randomly sampled negative exemplar inside the reference image, we formulate a triplet constraint on those three patches. We show the principle of PP-Triplet in Figure 4b.

Basically, our methodology is expected to automatically find the semantically corresponding regions on reference images according to the given partial query image without any additional supervision. It also offers new insights on correspondence learning for part-part matching. In our experiments, we apply our PPCL to partial person re-ID problem, considering its impact in real-world applications. We boost the Rank-1 performance on Partial-REID benchmark from 67.7% to 83.7%, and obtain state-of-the-art performance on Partial-iLIDS. We also qualitatively demonstrate that PPCL indeed learns the semantic correspondence between image patches. In addition, we also provide extended experimental results on partial face recognition.

2. Related Work

2.1. Partial Person Re-identification

In 2015, Zheng *et al*. [64] formulated the partial person re-ID problem. To address this difficulty, the authors decomposed partial images into small patches and reconstructed each patch with the assistance of a gallery dictionary. Similarly, several works achieved corresponding regions alignment by solving a least square problem between the feature pairs [15, 16], or by densely predicting a probability map [46], or by regressing the full texture image [26]. Luo *et al*. [35] utilized the STN [25] which shares the same spirit of our GLRec module but did not learning correspondence between data pairs. To identify the missing regions, another route relates to equipping with an external well-trained pose estimator and focused on the shared partial regions when measuring the similarity [17, 36, 9, 50]. However, such approaches exploited additional supervision and thus have limited generalization capability for the real world application. Overall, the above methods either emphasize the exploration of shared regions or adapt model to arbitrary input, rather than taking care of both of them like ours.

2.2. Correspondence Learning

Learning correspondence between two pairs is a central topic in computer vision, and has been widely applied in image stitching [49], image restoration [5], object recognition [32], tracking [37, 7], medical imaging [1], *etc*. Basically, the correspondence can be learned from temporally adjacent frames (*i.e*., optical flow) [57, 24, 45], image pairs from the same instance [57] or category [43].

Accordingly, a natural learning paradigm that leverages the cycle consistency of training samples was proposed to regularize the structured data [12, 55, 19, 66, 18], including visual tracking [47, 58, 54], image alignment [65], 3D mapping [23, 29], depth estimation [10], *etc*. For example, recent studies leveraged inherent temporal consistency in video frames [51, 2, 54, 6, 4, 40, 30] since a pair of patches can be distributed into the same location after forward and backward tracking. Different from these works that leverage cycle consistency between language/image domains or video frames, we resort the cycle consistency to find semantic correspondence between image patches.

Early works proposed pairwise constraint [59, 11] and triplet constraint [56] that aimed at optimizing the similarity between samples and achieved great performance on various tasks [44, 53, 20]. Generally, all these works built the pairwise or triplet samples on image-level with the given supervised signal. In contrast, we construct triplet samples on the patches sampled from image pairs without the requirement of any part-level supervision.

3. Part-Part Correspondence Learning

Given a partial input person image $x^p \in \mathcal{X}^p$[1], our first step is to coarsely predict which region it belongs to in a holistic image. To achieve this, we employ the GLRec

[1] We use superscript p to represent the partial content in the rest of paper, in order to avoid confusion with the holistic one.

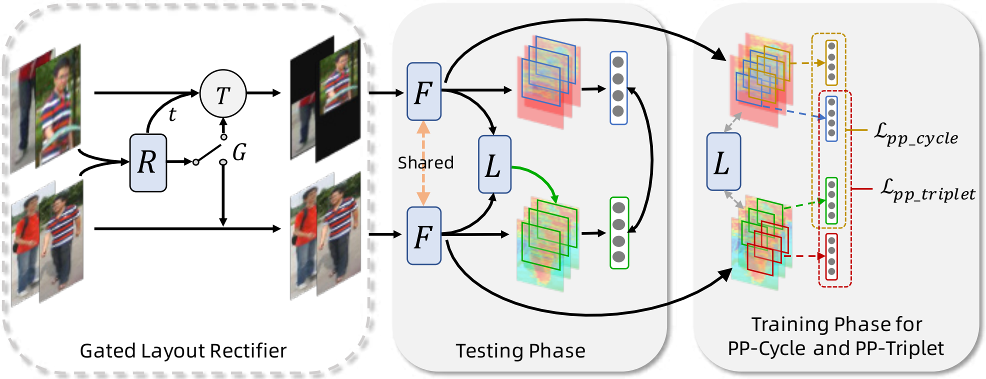


Figure 2: The proposed PPCL framework mainly comprises a *gated layout rectifier* including R, G, T (GLRec), a backbone network F and a *corresponding region locator* L (CRLoc). The GLRec module is a gated transformation regression CNN module that takes an arbitrary partial image in, and outputs a rectified result. Then after feature extraction by F, the CRLoc module is employed to learning correspondence for part-part matching. Both GLRec and CRLoc modules are trained in a self-supervised manner to obtain the corresponding patches between two images without any part-level supervision.

module to produce a set of affine transformation coefficients, indicating how the partial input should be transformed before feature extraction. After that, to locate the corresponding regions in reference images according to the given partial input x^p, we train a CRLoc module with the assistance of the proposed Part-Part Cycle (PP-Cycle) constraint and Part-Part Triplet (PP-Triplet) constraint collaboratively. In the end, the similarities between the partial input and the reference images are calculated among the corresponding partial regions. The overall pipeline is given in Figure 2 and Algorithm 1. Later in Section 4 and Section 5, we empirically show the advantage of our PPCL in partial re-ID.

3.1. Model Architecture

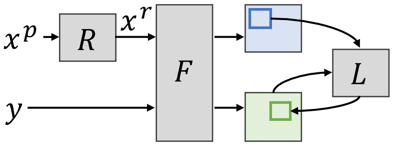


Figure 3: The overall pipeline for our PPCL framework.

Gated Layout Rectifier ($R + G + T$). For a partial input person image x^p, we first employ a *gated layout rectifier* (GLRec) to infer where it comes from then re-sampling the aligned output patch accordingly. It consists of a regression module R, a gate module G and a non-parametric geometric transformation module T. The regression module R consists of stacked convolutional layers and fully-connected layers, which is responsible for estimating affine transformation coefficients t and partial image confidence score η for the given x^p: $[t; \eta] = R(x^p)$, $R : \mathbb{R}^{h\times w\times c} \rightarrow \mathbb{R}^{n+1}$, where h, w, c are the height, width, the number of channels of x^p respectively, and n is the number of degrees of freedom for the geometric transformation, the additional output is a confidence score in $(0, 1)$, indicating whether the input is a partial image. Slightly different from [25], we use $n = 4$ in this paper, representing 4 degrees of freedom linear transformation capable of modeling translation and scaling. The gate module G acts as a switch. For inference, if confidence score $\eta < 0.5$, the input image is considered as a complete input and remains unchanged ($x^r = x^p$). Otherwise, the input image is considered as a partial input, the estimated affine transformation is then used to recover the partial input image x^p to a rectified counterpart x^r using the geometric transformation module T: $x^r = T(x^p, t)$.

A crucial advantage for the GLRec module is that, in the training stage, we are able to simulate partial inputs by randomly cropping training image samples, resulting in known transformation coefficients t. Let B as the batch size, a self-supervised loss function therefore can be defined as:

$$\mathcal{L}_R = \frac{1}{B}\sum_{i=1}^{B}\|R(x_i^p) - t\|_2^2, \qquad (1)$$

where η is omitted for simplicity, which is jointly optimized with a standard binary cross-entropy loss.

Feature Extraction (F). After obtaining the rectified input x^r, we compute spatial features with a classical backbone network[2] without fully-connected layers. The backbone network F outputs a down-sampled feature maps h_x^r for any input $x^r \in \mathcal{X}^r$, where $\mathcal{X}$ indicates a randomly sampled training batch and $\mathcal{X}^r$ is a rectified one. We train the backbone network with loss function $\mathcal{L}_F$, where $\mathcal{L}_F$ is determined by the specific task.

[2] Following the common practice in person re-ID, we use ResNet-50 as the backbone network in our experiments.

Algorithm 1 Part-Part Correspondence Learning Algorithm

```
Require: Batch size $B$; the pretrained gated layout rectifier $R(\cdot)$, the $G(\cdot)$, the $T(\cdot)$; the backbone network $F(\cdot)$; the corresponding region locator $L(\cdot)$.
repeat
    Sample one batch $\mathcal{X}$ with batch size $B$ from the holistic training image set.
    Shuffle batch $\mathcal{X}$ and set $\mathcal{Y} = \texttt{shuffle}(\mathcal{X})$.
    for each $i \in [1, B]$ do
        Randomly crop image patch $x_i^p$ from $x_i \in \mathcal{X}$ to simulate partial input.
        Generate the rectified input $x_i^r$ by:
            $x_i^r = T(G(R(x_i^p))).$    (2)
        Use $F(\cdot)$ to extract feature maps:
            $h_{i,x}^r = F(x_i^r), h_{i,x} = F(x_i), h_{i,y} = F(y_i),$    (3)
        where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$.
        Compute corresponding region $h_{i,x \to y}$ by:
            $h_{i,x \to y} = L(h_{i,x}^r, h_{i,y}).$    (4)
        Compute corresponding region $h_{i,y \to x}$ by:
            $h_{i,y \to x} = L(h_{i,x \to y}, h_{i,x}).$    (5)
        Sample negative partial region $h_{i,y*}$ from $h_{i,y}$.
        Compute $\mathcal{L}_R$ and $\mathcal{L}_F$.
        Compute $\mathcal{L}_{pp\_cycle}$ and $\mathcal{L}_{pp\_triplet}$ according to Equation 11 and 12 respectively.
        Update model according to the loss function:
            $\mathcal{L} = \mathcal{L}_R + \mathcal{L}_F + \lambda_{cyc}\mathcal{L}_{pp\_cycle} + \lambda_{tri}\mathcal{L}_{pp\_triplet}.$    (6)
    end for
until convergence
```

Corresponding Region Locator (L). To recognize a partial input image, a central challenge is to find the corresponding region, who may share the same semantics with the given partial image, within the reference images y, $y \in \mathcal{X}$. Along this line, our goal is to learn a CRLoc module L that takes h_x^r and h_y as input and outputs a corresponding region in h_y according to h_x^r:

$$h_{x \to y} = L(h_x^r, h_y), \tag{7}$$

where $h_{x \to y}$ is a partial region in h_y and shares the same semantics with h_x^r.

In order to achieve informative features, we first compute the correlation between h_x^r and h_y by a fusion layer $L_f(\cdot)$. Fortunately, there are lots of successful techniques to generate correlation maps, including parametric forms [52, 21] and non-parametric forms [43, 54]. Here we adopt the simplest one:

$$L_f(h_x^r, h_y) = \frac{\texttt{exp}(h_x^r(u) h_y(v)^{\mathsf{T}})}{\sum_v \texttt{exp}(h_x^r(u) h_y(v)^{\mathsf{T}})}, \tag{8}$$

where u and v are spatial positions in hidden representations. Note that the resulting correlation map is able to draw cross attention to the reference features, instead of being simply concatenated without any instruction.

Based on above analysis, we further design a region locator $L_l(\cdot)$ to accordingly yield corresponding region in h_y for the given h_x^r, with the correlation map calculated before. The region locator L_l outputs the exact coordinates that denotes corresponding region found in h_y. We thus obtain the semantically corresponding regions between the partial input image x^p and the reference image y.

In general, the CRLoc module can thus be formulated by:

$$L(h_x^r, h_y) = L_l(L_f(h_x^r, h_y)). \tag{9}$$

3.2. Part-Part Cycle Constraint

To train the CRLoc module, we here make an ingenious assumption that the optimal corresponding patch pairs can match themselves after forward-backward warping. Specifically, if CRLoc module is able to locate the corresponding region $h_{x \to y}$ in y according to x^r, then it will naturally can be translated back to obtain the original x^p. To achieve this, for each training batch $\mathcal{X}$, we randomly shuffle it to attain $\mathcal{Y}$, who has the same training samples with $\mathcal{X}$ but are arranged in a different order. For any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we also feed it into F to extract spatial features h_x and h_y respectively. Thus, we have:

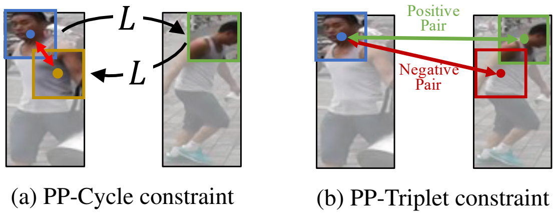
$$h_x^r = F(x^r);\ h_x = F(x);\ h_y = F(y). \tag{10}$$

It should be noticed that: 1) the samples in $\mathcal{X}$ and $\mathcal{Y}$ are raw image data in the training set, which are different from x^r that is randomly cropped and treated with GLRec module; 2) due to the shuffle operation, the sample y is probably different from x for a large batch size, which introduces diversity in PPCL.

Thus, we can make a cycle consistency by minimizing the reconstruction error:

$$\mathcal{L}_{pp\_cycle} = l(h_{x \to y \to x}, h_x), \tag{11}$$

where l is the Mean Squared Error (MSE) of the transformed grid points as used in [43]. The fundamental philosophy of PP-Cycle is demonstrated in Figure 4a.



(a) PP-Cycle constraint (b) PP-Triplet constraint

Figure 4: The illustration of the proposed constraints.

3.3. Part-Part Triplet Constraint

Triplet loss [44, 20] is a delicate technique, which aims to learn an representation of the data that keeps the distance between similar data points close and dissimilar data points far. Inspired by this, we propose Part-Part Triplet (PP-Triplet) loss that tends to make the similarity between corresponding region pairs larger than the irrelevant pairs. Acting in this way, as demonstrated in Figure 4b, we build our PP-Triplet loss on partial regions of spatial features between data pairs. Formally, let h_x^r be an anchor (blue box in Figure 4b), the corresponding region $h_{x\to y}$ found by L is regarded as a positive exemplar (green box), the target of PP-Triplet loss can be presented by:

$$\mathcal{L}_{pp\_triplet} = \frac{1}{B}\sum_{i=1}^{B}[\|h_x^r - h_{x\to y}\|_2^2 - \|h_x^r - h_{y*}\|_2^2 + \alpha], \tag{12}$$

where h_{y*} is a negative exemplar (red box in Figure 4b), and α represents the margin that is enforced between positive and negative pairs.

Note that, different from the original triplet loss that is capable of accessing image-level labels, our PP-Triplet is trained **without** any part-level ground truth. Therefore, the negative exemplar h_{y*} is randomly sampled from h_y with two principles: 1) the overlap between the positive exemplar and the negative one must lower than a pre-defined threshold β, which ensures that two exemplars come from different regions; 2) if one of them is involved in another, then the proportion of the overlapped area should also be lower than β, which reduces the noisy region that is irrelevant to the anchor. We empirically fix $\beta = 0.5$ in this paper. In general, the PP-Triplet loss allows the semantically corresponding regions to lie on the same manifold, while enlarging the distance to irrelevant regions.

3.4. Training and Inference

In general, the total training loss comprises $\mathcal{L}_R$, $\mathcal{L}_F$ and $\mathcal{L}_L$. We directly use the commonly adopted softmax cross-entropy loss and triplet loss [20] for $\mathcal{L}_F$. While for $\mathcal{L}_L$, we balance the two constraints with λ_{cyc} and λ_{tri}:

$$\mathcal{L}_L = \lambda_{cyc}\mathcal{L}_{pp\_cycle} + \lambda_{tri}\mathcal{L}_{pp\_triplet}. \tag{13}$$

During testing, we only measure the distance between semantically corresponding regions generated by GLRec and CRLoc modules as illustrated in Figure 2.

4. Evaluation on Person Re-Identification

4.1. Datasets and Evaluation Protocol

We conduct training on Market-1501 and test on commonly used Partial-REID and Partial-iLIDS dataset following [15, 46]. Market-1501 dataset [62] contains $32,668$ labeled images belonging to $1,501$ identities, each of them was captured by at most 6 cameras. The bounding-box of each person is detected by algorithm automatically. Partial-REID dataset [64] has 600 images of 60 identities, each identity consists of 5 full-body images and 5 partial images with different viewpoints and backgrounds. Especially, partial images in Partial-REID are cropped randomly with a small fraction (such as the left or the upper part of the body), yielding a challenge for partial re-ID algorithms. Partial-iLIDS dataset [64] is a simulated partial person dataset that is created from i-LIDS [63]. In the Partial-iLIDS dataset, there are 119 identities with a total of 238 images, each identity has 1 full-body image and 1 partial image. All partial images are generated by cropping the un-occluded part of the same person image.

During inference, we feed the model with a query list and search for the best matching reference (gallery) images for each query. Cosine distance is employed to measure the distance in the feature domain. We use the Cumulated Matching Characteristics (CMC) curve to evaluate the performance to align with the existing methods, which shows the probability that a query identity appears in different-sized candidate lists.

4.2. Model Configurations

For the GLRec module, we choose a lightweight ResNet-18 as backbone that is responsible for outputting 4 degrees of freedom geometric transformation, and 1 degree of gating instructor. Since the GLRec module is trained in a self-supervised manner, we pre-train it on Market-1501 with input images are randomly cropped to simulate partial input scenario. Note that, due to the self-supervised training style, our GLRec module does not depend on any specified crop method, and the sampling strategy can be determined according realistic condition as in Sun *et al*. [46].

For the feature extraction, we follow the common practice in person re-ID and adopt ResNet-50 [13] as the backbone network. To be consistent with previous works [15, 46, 16], we simultaneously adopt cross entropy loss and triplet loss to optimize the re-ID backbone. The triplet loss is equipped with the hard mining strategy [20]. As a result, the backbone network can achieve comparable results with previous works [15, 46] for fair comparison.

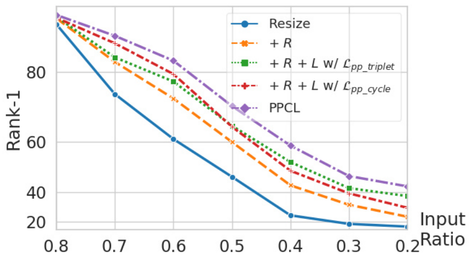


Figure 5: Our scheme consistently outperforms baselines for various input ratios.

4.3. Ablation Study

To evaluate each component of PPCL, we carry out detailed analysis in this section.

GLRec and CRLoc modules both contribute to the PPCL framework. To evaluate the effectiveness of GLRec and CRLoc modules, we implement four different settings: 1) The partial input is directly padding with constant and fed into the backbone network (Pad) 2) The partial input is directly resized and fed into the backbone network (Resize) 3) We only use the GLRec module R and remove CRLoc module L (+R). 4) We employ both GLRec and CRLoc modules (PPCL). The results are illustrated in Table 1a, from which we can make the following observations: 1) By adaptively recognizing and adjusting the scale and position of the partial input image, our GLRec module R significantly outperforms the baselines by a large margin, even though it is achieved in a self-supervised training manner.

PP-Cycle and PP-Triplet boost the re-ID performance individually. We use the GLRec and CRLoc modules trained only with $\mathcal{L}_{pp\_triplet}$ (+R+L w/$\mathcal{L}_{pp\_triplet}$) or $\mathcal{L}_{pp\_cycle}$ (+R+L w/$\mathcal{L}_{pp\_cycle}$), and demonstrate the results in Table 1a. It can be easily observed that individually apply one of the constraints both facilitate the learning of CRLoc module, resulting significant performance gain based on the GLRec module (+R).

PP-Cycle and PP-Triplet complements each other. From Table 1a we can also conclude that when combining the two constraints together (PPCL), they still boost each other with a significant margin, indicating the two items are complementary on part-part corresponding learning. We give a qualitatively analysis in the following.

Basically, for the PP-Cycle, we leverage the duality of the functionality on finding corresponding regions according to the given input. This makes sense since if one could find corresponding regions from one image to another, it must be able to perform locating reversely, and the corresponding patch pairs match themselves after forward-backward warping. While the PP-Cycle constraint is still not strong enough because we do not give clues of the optimal corresponding region in the reference image, which is typically unique for the given partial input. Therefore, inspired by the balanced training scheme in object detection [41], we assume the corresponding region generated by the CRLoc module is the optimal one, and thus the similarity between optimal corresponding pairs is always larger than the sub-optimal one. This intuitively motivates the creation of PP-Triplet, where the sub-optimal (negative) corresponding pairs are randomly sampled. Conversely, the PP-Cycle is more like a regularization term for PP-Triplet and make it stable in the training process. Therefore, the two constraints contribute together to PPCL.

PPCL can handle arbitrary inputs. Since the publicly available partial image recognition datasets contain fixed partial images of random input ratio (*e.g.*, $0.2 \sim 0.8$). Therefore, it is not clear on the characteristics of each component in PPCL when meets specific input ratio of partial image. Here we dive into the details and simulate the specific ratio of partial input by randomly cropping partial region in Market-1501 evaluation set.

We first illustrate the experimental results when varying the input ratio of partial (query) images in Figure 5, from which we can observe that our proposed scheme consistently improve the resize-based baseline. It is interesting to notice that the model trained with PP-Cycle constraint only (red line) slightly outperforms the model trained with PP-Triplet constraint only (green line) when the input ratio is large enough (*i.e.*, larger than 0.5). When the input ratio is small (*i.e.*, smaller than 0.5), the model trained with PP-Triplet constraint only (green line) shows great improvement compared with PP-Cycle (red line).

We also vary the number of partial images in query/gallery, as shown in Table 1b. We can achieve this functionality since our GLRec module is equipped with a gating scheme, that is able to automatically distinguish the partial and holistic image. More concretely, for the simulated partial scenario based on Market-1501, our gating scheme designed in the GLRec module is able to realize 94.3% accuracy on average.

Balance on two constraints. Table 1c presents performances of different weights for our PP-Cycle loss and PP-Triplet loss. These results show that $\lambda_{cyc} = 10$ and $\lambda_{tri} = 1$ achieves the optimal performance. Thus, we empirically follow this setting in all the experiments.

Case study. In order to investigate how GLRec and CRLoc modules work, we visualize the regions predicted by them in Figure 6. We can see that the GLRec module is capable of coarsely aligning the partial input to the holistic image, but the learned region is fixed for a given partial input and can not be customized to reference images, which bring much correspondence noise. In contrast, the CRLoc module

Methods	R-1	R-3	R-5
Pad	51.0	60.3	64.7
Resize	51.3	61.3	67.7
+ R	65.3	78.3	83.3
+ R + L w/ $\mathcal{L}_{pp\_triplet}$	71.3	84.3	88.0
+ R + L w/ $\mathcal{L}_{pp\_cycle}$	72.7	85.3	88.7
PPCL	79.0	87.3	90.7

(a) Model variations on Partial-REID.

	Gallery			
Query	20%	40%	60%	80%
20%	9.8	11.2	28.6	37.8
40%	10.5	18.7	39.1	54.8
60%	30.2	42.1	63.7	79.3
80%	37.3	57.7	83.2	92.1

(b) Adapt to arbitrary inputs.

	λ_{tri}			
λ_{cyc}	0.1	1	10	100
0.1	72.4	73.5	73.1	71.8
1	75.7	74.3	74.1	72.2
10	78.2	**79.0**	77.3	75.8
100	78.9	78.8	78.1	75.7

(c) Balance on two constraints.

Table 1: Ablation study of our PPCL framework. The details are provided in the text.

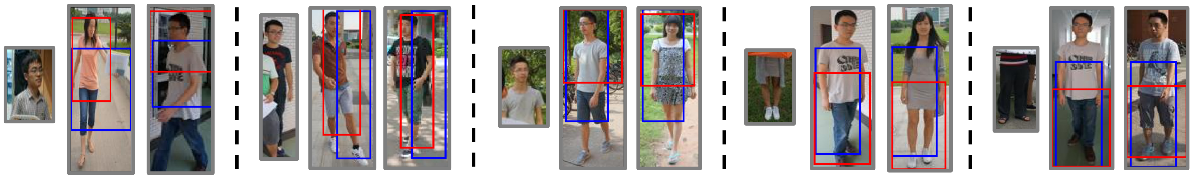


Figure 6: Visualization of our learned corresponding regions. The first column of each group is partial input, while the blue and red bounding-box are corresponding regions mapped from the outputs of GLRec and CRLoc module respectively. The results of GLRec in the first group are mismatched due to large variance of image resolution.

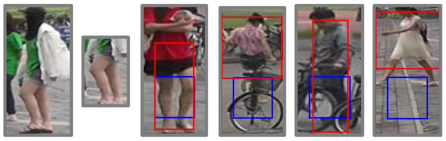


Figure 7: Failure cases.

eliminates the irrelevant partial regions accurately and adaptively. Figure 7 shows an example produced by our model. It can be observed that our GLRec module successfully predicts the layout (blue bounding-box) by the common sense (the legs are typically appeared in the bottom of the image), but fails on the cases of occlusion since it can not correct corresponding region according to reference image. While our PPCL (red bounding-box) is able to take the reference image into consideration and learn correspondence between image pairs adaptive to various scenarios.

Time analysis. Our GLRec is lightweight due to the ResNet-18 architecture, while the CRLoc only contains 3 convolutional layers and 1 fully-connected layer. When computing the similarity of each pair of images, we only execute CRLoc on the previously generated feature maps, instead of running the whole model. For a unified environment, where a resize-based baseline achieves 4.1s running time for Partial-REID. Our PPCL needs 4.8s running time with the GLRec module (+R), and 29.4s with both the GLRec and CRLoc module (+R+L) due to the pair-wise similarity calculation, which is also conducted in HOReID [50].

4.4. Compared with State-of-the-Art Schemes

We illustrate the experimental results in Table 2 and 3. On Partial-REID, we achieve 11.3% and 13.5% Rank-1 accuracy improvement over the state-of-the-art models. We also advance the Partial-iLIDS with a new record-breaking performance. It should be noted that, our PPCL is complementary to the advanced training techniques in person re-ID [34, 50] (PPCL+), therefore yields a higher performance (+4.7% and +1.7% Rank-1 accuracy on Partial-REID and Partial-iLIDS respectively). In general, the success of PPCL can be attributed to two factors: 1) the GLRec is able to adapt to arbitrary input and provide a rectified one; 2) by CRLoc, we associate corresponding regions between partial and holistic images, and eliminate the irrelevant regions when measuring similarity. For the holistic setting, since our GLRec contains a gate module that is designed for predicting partial/holistic input. Therefore, the PPCL will not harm performance on holistic datasets ($<$ 1% on both Market-1501 and DukeMTMC-reID [42]).

There is another line of works [36, 17, 8, 9, 50] that leveraged an external pose estimator or a well-trained pedestrian segmentation model to detect the fine-grained body part as **additional supervision** for partial re-ID. For example, Miao *et al.* [36] and Gao *et al.* [9] both encoded information from detected pose landmarks to align the corresponding region between partial and holistic images, achieving 68.0% and 75.3% Rank-1 accuracy on Partial-REID dataset respectively, which are comparable with ours. However, our method shows two crucial advantages: 1) we **do not** rely on any fine-grained labels like key point or segmentation map, or well-trained models which implicitly encodes additional supervision. 2) due to PPCL is pose-agnostic, we are able to generalize it to handle arbitrary inputs, even when the input ratio is lower than 30% (discussed in Section 4.3). In contrast, we empirically find that pose-based methods fail on extremely low input ratio. As a result, PPCL has the merit

Methods	Partial Images as Query Set			
	Partial-REID		Partial-iLIDS	
	R-1	R-3	R-1	R-3
MTRC [31]	23.7	27.3	17.7	26.1
AMC+SWM [64]	37.3	46.0	21.0	32.8
DSR [15]	50.7	70.0	58.8	67.2
SFR [16]	56.9	78.5	63.9	74.8
VPM [46]	67.7	81.9	67.2	76.5
STNReID [35]	66.7	80.3	54.6	71.3
PPCL	**79.0**	**87.3**	**69.7**	**84.0**
PPCL+	**83.7**	**88.7**	**71.4**	**85.7**

Table 2: Performance (%) comparison when partial images are regarded as query set.

Methods	Partial Images as Gallery Set			
	Partial-REID		Partial-iLIDS	
	R-1	R-3	R-1	R-3
MTRC [31]	26.0	37.0	28.6	43.7
AMC+SWM [64]	44.7	56.3	52.7	63.3
DSR [15]	58.3	82.0	59.7	79.0
SFR [16]	66.2	86.7	65.6	81.5
PPCL	**79.7**	**91.7**	**66.4**	**81.5**
PPCL+	**88.7**	**96.3**	**69.7**	**82.4**

Table 3: Performance (%) comparison when partial images are regarded as gallery set.

of being deployed in more complicated real-world scenarios, especially in cases of unsuccessful pose estimation.

5. Extension: Evaluation on Face Recognition

5.1. Datasets and Evaluation Protocol

In line with previous works [14, 16], the dataset we used for training is CASIA-Webface [60], which contains $494,414$ training images collected from $10,575$ identities. Following the common practice [33], we remove the overlapped images of identities appearing in testing sets and employ MTCNN [61] to perform similarity transformation. All images are resized to 224 × 224 [14].

For testing, since there is no publicly available face recognition datasets dedicated to partial image recognition, we imitate the partial scenario on LFW like [14], which is a common testing environment that contains $13,233$ images collected from $7,749$ individuals. Specifically, we select $1,000$ identities who have the largest number of images. Then we sample one image for each identity to form the holistic gallery set. The remaining images are randomly cropped and treated as query set. Therefore, the query set and gallery set share the same identities but with different images. Due to previous works do not release their testing set, we randomly generate the partial images with proximate area ratio which can achieve comparable performance with their methods.

Similar to person re-ID, we select Cumulative Match Characteristic (CMC) curves and Receiver Operating Characteristic (ROC) curves as evaluation metrics.

5.2. Model Configurations and Experiments

Methods	R-1	R-3	R-5	R-10
MKDSRC-GTP [31]	1.10	3.70	5.60	8.40
I2C [22]	6.8	8.3	11.20	14.60
VGGFace [39]	20.90	-	-	-
DFM [14]	27.30	34.40	39.20	47.58
SFR [16]	46.30	59.30	65.50	70.90
PPCL	**52.90**	**61.40**	**69.40**	**73.24**

Table 4: Performance (%) comparison on Partial-LFW.

We train the GLRec module in similar to person re-ID, with the same network architecture and training strategy described before. For easy comparison [14], we adopt commonly used VGGFace [39] as our backbone network. It should be noted that, in the common practice of face recognition, the spatial features outputted by network are directly re-shaped before being fed into FC layers, instead of downsampled to feature vectors by GAP. Therefore, we replace the first FC layers with Global Depthwise Convolution [3], which can achieve comparable performance but more compatible with our framework. The VGGFace model takes 224 × 224 images as input and outputs 4096-dimension feature descriptor. At test time, the feature descriptor are compared in Euclidean distance for the purpose of face verification. We carry out our method based on PyTorch. We demonstrate the results in Table 4, in which our PPCL consistently boosts the state-of-the-art performance.

6. Conclusions and Future Work

In this paper, we formulate Part-Part Correspondence Learning (PPCL). There are two main components involved in PPCL: a *gated layout rectifier* that is responsible for predicting an appropriate layout for the partial input image, and a *corresponding region locator* that learns to find the corresponding region in one data sample according to another. Especially, due to the absence of part-level labels, we accordingly present two constraints, which are Part-Part Cycle and Part-Part Triplet constraints, to train the *corresponding region locator* in a self-supervised manner.

In the future, we believe our proposed PP-Cycle and PP-Triplet can be widely applied to various tasks. Especially for the models that require a understanding of part-part relationship, or the tasks that are sensitive to negative partial features, such as object detection, tracking, etc.

Acknowledgements. This work was supported in part by NSFC under Grant U1908209 and the National Key Research and Development Program of China 2018AAA0101400. We thank all the anonymous reviewers for their valuable comments on our paper.

References

[1] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9252–9260, 2018. 2

[2] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. Recycle-gan: Unsupervised video retargeting. In *Proceedings of the European conference on computer vision (ECCV)*, pages 119–135, 2018. 2

[3] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Chinese Conference on Biometric Recognition*, pages 428–438. Springer, 2018. 8

[4] Zhibo Chen, Tianyu He, Xin Jin, and Feng Wu. Learning for video compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(2):566–576, 2019. 2

[5] Kevin Dale, Micah K Johnson, Kalyan Sunkavalli, Wojciech Matusik, and Hanspeter Pfister. Image restoration using online photo collections. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2217–2224. IEEE, 2009. 2

[6] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1801–1810, 2019. 2

[7] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *European conference on computer vision*, pages 834–849. Springer, 2014. 2

[8] Lishuai Gao, Hua Zhang, Zan Gao, Weili Guan, Zhiyong Cheng, and Meng Wang. Texture semantically aligned with visibility-aware for partial person re-identification. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3771–3779, 2020. 7

[9] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. Pose-guided visible part matching for occluded person reid. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11744–11752, 2020. 2, 7

[10] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017. 2

[11] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 2

[12] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *Advances in neural information processing systems*, pages 820–828, 2016. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[14] Lingxiao He, Haiqing Li, Qi Zhang, and Zhenan Sun. Dynamic feature learning for partial face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7054–7063, 2018. 8

[15] 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*, pages 7073–7082, 2018. 1, 2, 5, 8

[16] Lingxiao He, Zhenan Sun, Yuhao Zhu, and Yunbo Wang. Recognizing partial biometric patterns. *arXiv preprint arXiv:1810.07399*, 2018. 1, 2, 5, 8

[17] 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*, pages 8450–8459, 2019. 2, 7

[18] Tianyu He, Jiale Chen, Xu Tan, and Tao Qin. Language graph distillation for low-resource machine translation. *arXiv preprint arXiv:1908.06258*, 2019. 2

[19] Tianyu He, Yingce Xia, Jianxin Lin, Xu Tan, Di He, Tao Qin, and Zhibo Chen. Deliberation learning for image-to-image translation. In *IJCAI*, pages 2484–2490, 2019. 2

[20] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 2, 5

[21] Ruibing Hou, Hong Chang, MA Bingpeng, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. In *Advances in Neural Information Processing Systems*, pages 4005–4016, 2019. 4

[22] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Robust partial face recognition using instance-to-class distance. In *2013 Visual Communications and Image Processing (VCIP)*, pages 1–6. IEEE, 2013. 8

[23] Qi-Xing Huang and Leonidas Guibas. Consistent shape maps via semidefinite programming. In *Computer Graphics Forum*, volume 32, pages 177–186. Wiley Online Library, 2013. 2

[24] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 2

[25] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 2, 3

[26] Xin Jin, Cuiling Lan, Wenjun Zeng, Guoqiang Wei, and Zhibo Chen. Semantics-aligned representation learning for person re-identification. In *AAAI*, pages 11173–11180, 2020. 2

[27] Adam Kosiorek, Sara Sabour, Yee Whye Teh, and Geoffrey E Hinton. Stacked capsule autoencoders. In *Advances in Neural Information Processing Systems*, pages 15486–15496, 2019. 1

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[29] Nilesh Kulkarni, Abhinav Gupta, and Shubham Tulsiani. Canonical surface mapping via geometric cycle consistency. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2202–2211, 2019. 2

[30] Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. Joint-task self-supervised learning for temporal correspondence. In *Advances in Neural Information Processing Systems*, pages 317–327, 2019. 2

[31] Shengcai Liao, Anil K Jain, and Stan Z Li. Partial face recognition: Alignment-free approach. *IEEE Transactions on pattern analysis and machine intelligence*, 35(5):1193–1205, 2012. 8

[32] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):978–994, 2010. 2

[33] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017. 8

[34] 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*, pages 0–0, 2019. 7

[35] Hao Luo, Wei Jiang, Xing Fan, and Chi Zhang. Stnreid: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification. *IEEE Transactions on Multimedia*, 2020. 1, 2, 8

[36] 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*, pages 542–551, 2019. 1, 2, 7

[37] Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. Dtam: Dense tracking and mapping in real-time. In *2011 international conference on computer vision*, pages 2320–2327. IEEE, 2011. 2

[38] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 1

[39] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015. 8

[40] Fitsum A Reda, Deqing Sun, Aysegul Dundar, Mohammad Shoeybi, Guilin Liu, Kevin J Shih, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Unsupervised video interpolation using cycle consistency. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 892–900, 2019. 2

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1, 6

[42] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016. 7

[43] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. End-to-end weakly-supervised semantic alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6917–6925, 2018. 2, 4

[44] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 1, 2, 5

[45] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018. 2

[46] 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*, pages 393–402, 2019. 1, 2, 5, 8

[47] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *European conference on computer vision*, pages 438–451. Springer, 2010. 2

[48] 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*, pages 2818–2826, 2016. 1

[49] Richard Szeliski. Image alignment and stitching: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 2(1):1–104, 2006. 2

[50] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 6449–6458, 2020. 2, 7

[51] Naiyan Wang and Dit-Yan Yeung. Learning a deep compact image representation for visual tracking. In *Advances in neural information processing systems*, pages 809–817, 2013. 2

[52] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 4

[53] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2015. 2

[54] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2566–2576, 2019. 2, 4

[55] Yiren Wang, Yingce Xia, Tianyu He, Fei Tian, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. Multi-agent dual learning. In *Proceedings of the International Conference on Learning Representations (ICLR) 2019*, 2019. 2

[56] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244, 2009. 2

[57] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. Deepflow: Large displacement optical flow with deep matching. In *Proceedings of the IEEE international conference on computer vision*, pages 1385–1392, 2013. 2

[58] Hao Wu, Aswin C Sankaranarayanan, and Rama Chellappa. In situ evaluation of tracking algorithms using time reversed chains. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 2

[59] Eric P Xing, Michael I Jordan, Stuart J Russell, and Andrew Y Ng. Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, pages 521–528, 2003. 2

[60] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 8

[61] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 8

[62] 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*, pages 1116–1124, 2015. 1, 5

[63] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR 2011*, pages 649–656. IEEE, 2011. 5

[64] 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*, pages 4678–4686, 2015. 2, 5, 8

[65] Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A Efros. Learning dense correspondence via 3d-guided cycle consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 117–126, 2016. 2

[66] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2