2020

**Improving Semantic Part Features for Person Re-identification with Supervised Non-local Similarity**

Yifan Sun  
*the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.*

Zhaopeng Dou  
*the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.*

Yali Li  
*the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.*

Shengjin Wang  
*the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.*

Follow this and additional works at: [https://tsinghuauniversitypress.researchcommons.org/tsinghua-science-and-technology](https://tsinghuauniversitypress.researchcommons.org/tsinghua-science-and-technology)

Part of the Computer Sciences Commons, and the Electrical and Computer Engineering Commons

**Recommended Citation**

Yifan Sun, Zhaopeng Dou, Yali Li et al. Improving Semantic Part Features for Person Re-identification with Supervised Non-local Similarity. Tsinghua Science and Technology 2020, 25(05): 636-646.
Improving Semantic Part Features for Person Re-identification with Supervised Non-local Similarity

Yifan Sun, Zhaopeng Dou, Yali Li, and Shengjin Wang

Abstract: In person re-identification (re-ID) task, the learning of part-level features benefits from fine-grained information. To facilitate part alignment, which is a prerequisite for learning part-level features, a popular approach is to detect semantic parts with the use of human parsing or pose estimation. Such methods of semantic partition do offer cues to good part alignment but are prone to noisy part detection, especially when they are employed in an off-the-shelf manner. In response, this paper proposes a novel part feature learning method for re-ID, that suppresses the impact of noisy semantic part detection through Supervised Non-local Similarity (SNS) learning. Given several detected semantic parts, SNS first locates their center points on the convolutional feature maps for use as a set of anchors and then evaluates the similarity values between these anchors and each pixel on the feature maps. The non-local similarity learning is supervised such that: each anchor should be similar to itself and simultaneously dissimilar to any other anchors, thus yielding the SNS. Finally, each anchor absorbs features from all of the similar pixels on the convolutional feature maps to generate a corresponding part feature (SNS feature). We evaluate our method with extensive experiments conducted under both holistic and partial re-ID scenarios. Experimental results confirm that SNS consistently improves re-ID accuracy using human parsing or pose estimation, and that our results are on par with state-of-the-art methods.

Key words: person re-identification; non-local similarity; feature learning; semantic parts

1 Introduction

Given a query person-of-interest, person re-ID aims to retrieve all of the images containing that person from a gallery/database. Recently, many studies have explored deep learning of part features and in so doing, have improved re-ID performance compared with methods that learn global features. Part alignment is critical to learning discriminative part features and a popular solution for this is to leverage external cues from semantic partition with the assistance of pose estimation or human parsing.

In spite of significant progress in human parsing and pose estimation research, inaccurate part detection remains a problem when learning part features for re-ID purposes. There is a significant domain gap between re-ID images and the images used for training the human parser or the pose estimator such that applying the human parser to the re-ID images in an off-the-shelf manner inevitably deteriorates the performance. As an example, Fig. 1a shows a noisy human parsing example on Market-1501, which is a widely-adopted re-ID dataset. The head and feet are barely detected, and the lower and the upper bodies are not distinguished. Noisy human parsing leads to part extraction errors and thus compromises the discriminative ability of the learned part features.

To enhance robustness against noisy semantic parts, this paper proposes a part feature learning method for re-ID purposes based on SNS. We observe in Fig. 1a that,
Yifan Sun et al.: Improving Semantic Part Features for Person Re-identification with Supervised Non-local Similarity

Fig. 1 Motivation of SNS and the visualization of its effect.
(a) An example of noisy human parsing result. The head and the feet are barely detected, and the lower body is blurred with upper body. However, the center points, i.e., the red, yellow, green, and magenta dots lay within the ground truth regions. (b) In Supervised Non-local Similarity (SNS), each anchor (the center points) absorbs features from all the similar pixels (we only visualize the absorption from a single pixel for clarity) to generate a corresponding part feature. Higher brightness indicates larger similarity value as well as larger absorption probability. SNS is applied on the convolutional feature maps, and we visualize the SNS effect on the raw image for intuitive understanding.

despite the considerable number of part extraction errors, the geometric center of each detected part usually lays within the ground truth region. This motivates us to use the center points as cues to learn part features, instead of directly using the detected semantic parts. Specifically, we first extract several semantic parts using a human parser or a pose estimator, and then locate the center points of the semantic parts on convolutional feature maps (e.g., “conv5” in ResNet-50) for use as feature anchors. Each of these anchors learns which pixels are similar to itself and absorbs these similar pixels to generate a corresponding part feature, as shown in Fig. 1b. Assuming that a pixel on the feature maps is very similar to the “head” anchor, it is then absorbed into the “head” part with a relatively large probability. The similarity is learned using a supervised non-local mechanism, yielding the SNS. Finally, we concatenate all of the part features to form the final person descriptor.

There are two benefits of using SNS to learn part-level features for re-ID. First, SNS implicitly leverages human parsing or pose estimation for learning part features, and thus naturally benefits from semantic part information. Second, SNS employs a non-local mechanism to facilitate adaptive part learning, rather than directly using the semantic parts generated by human parsing or pose estimation, which makes it more robust against semantic part detection errors. For example, in Fig. 1b, SNS successfully recovers the head and the feet and distinguishes the lower body from the upper body.

We validate the effectiveness of SNS under both the holistic and partial re-ID scenarios. Under the holistic scenario, SNS surpasses prior works in learning semantic part features. Under the partial scenario, only partial observations of a person are available for retrieval. In this case, SNS further benefits from semantic part information, which permits comparing partial images by focusing on their shared parts (those parts that are visible in both images), and achieves performance on par with state-of-the-art methods. Under both scenarios, experimental results confirm that SNS enhances the discriminative ability of the learned part features, indicating higher robustness against human parsing (or pose estimation) errors.

The main contributions of this paper are summarized as follows:

- We propose a part feature learning method for re-ID tasks based on SNS. SNS achieves robustness against human parsing (or pose estimation) errors, and thus enhances the discriminative ability of the learned part features.
- We comprehensively evaluate SNS under the holistic re-ID scenario, by applying SNS on three types of human parser and on a pose estimator. The experiment results confirm that SNS consistently increases re-ID accuracy when compared with directly using the parsing or pose estimation results for feature extraction. The achieved performance is competitive with state-of-the-art methods.
- We also investigate SNS under the partial re-ID scenario. With semantic part information, SNS is able to compare two images with focus on their shared parts. We experimentally demonstrate that SNS also improves semantic part feature learning for partial re-ID tasks.

2 Related Work

2.1 Deeply-learned part features for re-ID

Over recent years, deep learning methods, especially those exploring part features, have been advancing the state of the art for re-ID tasks\cite{1, 6, 10}. A prerequisite for learning part features is part detection. strategies for which, can be roughly categorized into three approaches. One approach is to use semantic part
Non-local neural networks capture long-range dependencies by estimating the interactions between any two positions. Wang et al.\cite{wang2018non} defined a generic non-local operation and applied it to video classification, object detection, and pose estimation tasks. They also analyzed the connection between non-local operation and the self-attention mechanism, graphical models, etc.

SNS applies the non-local operation to the new task of feature learning for re-ID purposes, and makes several novel modifications. First, the non-local operation in SNS is learned under supervision, while the canonical non-local operation is generally unsupervised. Given a set of anchor points, we enforce a supervision ensuring that the similarity (interaction) between each anchor and itself should be strong, while the similarity between two different anchors should be weak. The ablation study presented in Section 4.4 demonstrates that this supervision is critical to the effectiveness of SNS. Second, SNS requires very little non-local computation when compared with the canonical non-local operation. In SNS, the non-local operation is conducted between each position and several anchors, rather than between every combination of two positions. This reduces the computational complexity from $O(N^2)$ to $O(N)$, where $N$ is the number of pixels in the feature maps. Moreover, SNS is applied only to a single layer (e.g., “conv5” in ResNet-50), rather than to all of the convolutional layers.

### 2.3 Partial re-ID

In real-world re-ID systems, cameras sometimes capture only partial observations of a person due to occlusion; it is then more challenging to make an accurate retrieval\cite{20,21}.

Currently, semantic part-based methods are generally inferior to methods based on heuristic partition, e.g., PCB\cite{1}. Although SNS is more robust against part extraction errors than other forms of semantic part extraction, its re-ID accuracy is slightly lower than PCB on holistic re-ID datasets. However, SNS surpasses PCB marginally under partial re-ID scenario, as addressed in Section 4.6 below, because SNS explores semantic part information and thus permits comparing partial person images with a focus on their common parts.

### 3 Proposed Method

#### 3.1 Overall structure

This work employs a human parser or pose estimator and a CNN combined with a non-local unit to learn part features for re-ID purposes, with the structure illustrated in Fig. 2. We crop a stack of convolutional layers from a canonical CNN (e.g., all of the convolutional layers prior to “global average pooling” in ResNet-50) as the backboned model. The backboned model transfers the input image of a pedestrian into a 3D tensor such that $T \in \mathbb{R}^{H \times W \times C}$ (where $H$, $W$, and $C$ are the height, width, and the channel number of $T$, respectively), which can be viewed as the ensemble of $W \times H$ column vectors $x$, where $x \in \mathbb{R}^C$.

Simultaneously, we detect $K$ semantic body parts with a human parser or pose estimator and then locate their geometric centers as the anchors on tensor $T$. When using the human parser, we directly use the predicted probability maps as the semantic parts and extract the “head”, “upper body”, “lower body”, and “feet” parts, which is a similar approach to that taken in Ref. [7]. When using the pose estimator, we follow GLAD\cite{2} in detecting the “head”, “upper body”, and “lower body” with bounding boxes inferred from the detected joints.
Fig. 2 Overall structure of SNS feature learning method. We input the pedestrian image into a human parser (or a pose estimator) and a stack of convolutional layers (“conv”) for feature learning in parallel. The human parser detects several semantic parts on the input image. The stack of convolutional layers (“conv”) transfers the input image into a 3D tensor $T$. We locate the center points of the detected parts as anchors on $T$ and append a non-local unit upon $T$. Through the proposed SNS loss, the non-local unit learns the non-local similarity between each anchor and all the column vectors $x$ on $T$. Each anchor then absorbs its similar column vectors with the similarity values as the absorption probabilities to generate an SNS feature.

and landmarks. The human parser and the pose estimator are used in an off-the-shelf manner, without being fine-tuned on a re-ID dataset.

Given the coordinates of the anchors and tensor $T$, a non-local unit first learns the similarity values between every column vector $x$ and each anchor. Based on these similarity values, each anchor absorbs the column vectors most similarity to it to generate an SNS feature. During training, the SNS features are fed into $K$ branches of ID classifiers, which are implemented with $K$ respective fully-connected layers. Each classifier learns to discriminate the identity of the training images through cross-entropy loss, which is a common practice in part feature learning methods for re-ID purposes\cite{1,2,5,7}. The non-local unit is trained with an SNS loss, which drives each anchor to be similar to itself and dissimilar to other anchors. Moreover, while Ref. \cite{9} learns the parameters of $W_\theta$ and $W_\phi$ using the unsupervised attention mechanism, SNS conducts supervised learning as follows.

**Supervision for learning the non-local unit.** Each anchor is calculated from a respective semantic part and is intended to be similar to itself, but dissimilar to other anchors. We define the SNS loss to supervise the learning of the non-local unit as

$$L_{\text{SNS}} = \sum_{j \neq i} [-f(a_i, a_j) + f(a_i, a_j) + \alpha]^+] \tag{2}$$

Equation (2) is similar to a triplet loss in form, and $\alpha$ is a margin. In Section 4.3 below, the impact of $\alpha$ on re-ID accuracy is analyzed experimentally. As assessed in Section 4.4 below, SNS loss is vital to the discriminative ability of the learned features.

**Non-local absorption.** Given the learned non-local similarity, each anchor absorbs features from its similar pixels on $T$ to generate part features by

$$g_i = \sum_{x \in T} f(x, a_i) x \tag{3}$$

where $g_i$ is the $i$-th part SNS feature. Equation (3) can
also be conceived of as assigning each pixel on $T$ to several SNS parts, as described in Section 4.5 below.

### 3.3 Multi-anchor SNS

For simplicity, the previous section describes SNS with a single anchor for each semantic part. We now consider multi-anchor SNS, which adopts $M$ (where $M > 1$) anchors around the center point for each part. For multi-anchor SNS, the SNS loss and the non-local absorption require corresponding modifications.

The multi-anchor SNS loss becomes

$$L_{\text{SNS}}^{\text{multi}} = \sum_{j \neq i} [F(A_i, A_j) - F(A_i, A_j) - \alpha]_+,$$

where $A_i$ is the set of anchors for the $i$-th semantic part, and $F$ is the function to accumulate the non-local similarities between anchor sets.

The non-local absorption for multi-anchor SNS becomes

$$g_i^j = \frac{1}{M} \sum_{u \in A_i} \sum_{x \in T} f(x, u)x$$

where $M$ is the number of anchors for each part. When a single anchor is used for each part (where $M = 1$), Eqs. (4) and (5) downgrade to Eqs. (2) and (3), respectively. We therefore consider multi-anchor SNS to be a generalized version that includes single-anchor SNS.

In multi-anchor SNS, multiple anchors “vote” and thereby jointly determine each SNS feature. Section 4.3 below analyzes the impact of $M$ on re-ID accuracy and recommends using four anchors for each part.

### 3.4 Employing SNS features

Given two images to be compared ($f^k$ and $f^l$), we extract their part features as $g_i^k$ and $g_i^l$, respectively.

**Holistic re-ID.** Under the holistic re-ID scenario, we concatenate all of the part features $g_i^k$ ($g_i^l$) to obtain the final descriptor $G^k$ ($G^l$). We then calculate the Euclidean distance between $G^k$ and $G^l$. This is the same procedure used in several prior part-based re-ID works\cite{1,3,7}.

**Partial re-ID.** Under the partial re-ID scenario, we first calculate the part-to-part distances by $d_{i,j}^{k,l} = \|g_i^k - g_j^l\|$. We then calculate the overall distance ($D_{i,j}^{k,l}$) between $G^k$ and $G^l$ by further analysis of the human parsing results. Human parsing returns $K$ probability maps $\mathcal{P}_i$ ($i = 1, 2, \ldots, K$), corresponding to $K$ semantic parts. We define the visibility score $s$ of a semantic part by

$$s_i = \sum_{p \in \mathcal{P}_i} p$$

where $p \in \mathcal{P}_i$ is the probability value of a pixel belonging to the $i$-th semantic part.

The overall distance $D_{i,j}^{k,l}$ is then calculated by

$$D_{i,j}^{k,l} = \frac{\sum_{i=1}^{K} s_i^k s_j^l d_{i,j}^{k,l}}{\sum_{i=1}^{K} s_i^k s_j^l}$$

Equations (6) and (7) allow for a comparison between partial images by focusing on their shared parts. If a semantic part is invisible in either of the compared images, its visibility score is small (approximating 0) and thus contributes little to the overall distance. In contrast, if a semantic part is visible in both images, the part-to-part distance is rated credited and dominates the calculation of overall distance.

### 4 Experiment

#### 4.1 Settings

**Datasets.** To evaluate our method, we use two large-scale (holistic) re-ID datasets, i.e., Market-1501\cite{8}, DukeMTMC-reID\cite{22,23} and two partial re-ID datasets, Partial-reID\cite{21} and Partial-iLIDS\cite{24}.

Most of our experiments, including those on parameter optimization, the ablation study and visualization are conducted on the holistic re-ID datasets. The Market-1501 dataset contains 1501 identities observed from six camera viewpoints with 19732 gallery images and 12936 training images detected by DPM\cite{25}. The DukeMTMC-reID dataset contains 1404 identities with 17661 gallery images, 16522 training images, and 2228 queries. On these two datasets, we adopt Cumulative Match Characteristic (CMC)\cite{26} and mean Average Precision (mAP)\cite{8} as the evaluation protocol and use the evaluation packages provided by Refs. [8, 23], respectively.

To further investigate our method under the partial-reID scenario, we also conduct experiments on the Partial-REID and Partial-iLIDS datasets. Partial-REID contains 600 images and 60 identities, each of which has five holistic images and five partial images. Partial-iLIDS contains 238 images and 119 identities, each of which has one holistic image and one partial image. Both Partial-REID and Partial-iLIDS offer only testing.
We therefore learn SNS features on the training set of Market-1501 for fair comparison with competing methods. On these two datasets, we use the evaluation packages provided by Ref. [20] and evaluate the CMC performance.

**Human parsers and pose estimator.** We implement SNS using three human parsers: EDANet\[^{27}\], DeepLab V2\[^{11}\], and DeepLab V3+\[^{28}\]. They achieve mean IoU of 38.2\%, 41.6\%, and 45.4\%, respectively, on the LIP human parsing dataset\[^{12}\]. We also build an SNS on Open Pose\[^{16}\], which is a popular pose estimator.

**Baselines.** We implement three baselines for comparison with our method:

- **IDE baseline** learns a global feature for a pedestrian image and is widely adopted by a number of works\[^{10,23,29–32}\]. A comparison between IDE and our method is used to demonstrate the capacity of SNS in learning part features.

- **H-baseline** directly uses the probability maps generated by a human parser to extract several semantic parts on tensor $T$. Specifically, the probability maps are down-sampled to the same size as $T$, and then multiplied with $T$ to generate the corresponding parts; this is similar to Ref. [7]. We extract four semantic body parts through human parsing: head, upper body, lower body, and feet.

- **P-baseline** infers several bounding boxes from the pose estimation results and learns a semantic part feature within each bounding box; this is similar to Ref. [2]. We extract three semantic body parts using pose estimation: head, upper body, and lower body.

Our implementation of the IDE baseline achieves an 87.3\% rank-1 accuracy and a 70.1\% mAP on Market-1501, which is consistent with Refs. [1, 32, 33]. H-baselines for the three human parsers all surpass the IDE baseline, indicating the effectiveness of extracting fine-grained information from the semantic parts, which is consistent with Ref. [7]. Our implementation of P-baseline does not improve over on the IDE baseline (and is even slightly inferior on Market-1501), mainly due to significant detection errors. The detailed performance is shown in Table 1.

### 4.2 Evaluation on holistic re-ID datasets

**The effectiveness of SNS.** We demonstrate the effectiveness of SNS by comparing it with all three baselines (the IDE baseline, P-baseline, and H-baseline) using three human parsers (EDANet, DeepLab-V2, and DeepLab-V3+) and one pose estimator (Open Pose). The results are shown in Table 1, from which we draw three observations.

**All versions of SNS surpass the IDE baseline.** On Market-1501, the four versions of SNS improve the rank-1 accuracy over IDE baseline by +2.3\%, +3.4\%, +4.0\%, and +4.2\%, respectively, implying that SNS benefits from fine-grained information by learning part features.

**SNS consistently improves semantic part feature recognition.** Comparing the four version of SNS with the corresponding H-baselines and P-baselines, we observe a consistent improvement in re-ID accuracy and mAP. For example, on Market-1501, the four versions of SNS surpass their corresponding baselines by +2.7\%, +2.5\%, +1.8\%, and +1.0\% in rank-1 accuracy and +7.1\%, +4.7\%, +3.1\%, and +1.7\% in mAP, respectively.

**SNS is more robust against semantic part detection errors.** Comparing the three human parsing version of SNS, we observe that they achieve very similar re-ID accuracy. For example, on Market-1501, SNS with EDANet has a rank-1 accuracy only 0.8\% lower than

### Table 1 Comparison between SNS and baselines. We employ a pose estimator (Open Pose) and three human parsers to build the part feature learning baselines and four editions of SNS. All editions of SNS use a single anchor per part. Setting appropriately more anchors slightly improves SNS, as to be accessed in Section 4.3.

| Model                  | Market-1501 | DukeMTMC-reID |
|------------------------|-------------|---------------|
|                         | rank-1     | rank-5       | rank-10  | mAP  | rank-1     | rank-5       | rank-10  | mAP  |
| IDE                    | 87.3        | 94.6         | 96.9     | 70.1  | 74.4        | 85.1         | 89.3     | 55.4  |
| P-baseline (Open Pose) | 86.9        | 94.4         | 96.4     | 66.7  | 74.6        | 85.3         | 89.4     | 55.8  |
| H-baseline (EDANet)    | 88.2        | 95.1         | 96.9     | 73.1  | 76.6        | 87.3         | 90.5     | 61.4  |
| H-baseline (DeepLab V2)| 89.5        | 95.7         | 97.3     | 75.1  | 79.0        | 87.9         | 91.1     | 64.9  |
| H-baseline (DeepLab V3+) | 90.5       | 96.3         | 97.5     | 76.9  | 80.2        | 88.5         | 91.7     | 66.2  |
| SNS (Open Pose)        | 89.6        | 96.3         | 97.5     | 73.8  | 78.1        | 87.3         | 90.2     | 60.5  |
| SNS (EDANet)           | 90.7        | 96.5         | 97.7     | 77.8  | 81.1        | 88.4         | 91.2     | 66.5  |
| SNS (DeepLab V2)       | 91.3        | 96.8         | 97.8     | 78.2  | 81.8        | 88.8         | 91.7     | 67.9  |
| SNS (DeepLab V3+)      | **91.5**    | **97.1**     | **98.0** | **78.6** | **82.4**    | **89.3**     | **92.1** | **68.7** |
SNS with DeepLab-V3+, SNS with EDANet is 2.3% lower in rank-1 accuracy on H-baseline than SNS with DeepLab-V3+. This characteristic allows SNS to work with a relatively fast and inaccurate human parser to achieve higher efficiency in a practical application.

Comparison with the state of the art. We compare SNS with state-of-the-art methods on Market-1501 and DukeMTMC-reID, with the results shown in Table 2. For these comparisons, we use DeepLab-V3+ for human parsing. Our method benefits from deeply-learned part features in surpassing all of the methods that learn global features. Our method is highly competitive with the state-of-the-art part feature learning methods. SNS leverages both semantic cues and the attention mechanism (marked with non-local operation), and surpasses most attention-based (marked with superscript “a” in Table 2) and semantic part based deep learning methods (with superscript “s”). PABR also uses semantic cues and achieves slightly higher performance than our method on DukeMTMC-reID. PABR trains a two-stream network, one of which learns to extract body part and achieves accurate part alignment. For holistic re-ID scenarios, part-based methods relying on heuristic partition, such as PCB, generally achieve a higher performance than semantic part based methods, including ours. Offsetting this are the advantages of SNS under partial re-ID scenarios, outlined in Section 4.6 below.

4.3 Parameter analysis

Using Market-1501, we analyze two important parameters: the margin $\alpha$ in SNS loss (Eq. (2)) and the number of anchors $M$ for each part. The results of this analysis are shown in Fig. 3.

The margin ($\alpha$) determines the force pushing the anchors for different semantic parts further away from each other, thus making them less similar. Adopting a large $\alpha$ forces anchors to mainly absorb from near-by points, whereas a small $\alpha$ enables anchors to absorb relatively more global information. Figure 3 shows that a setting of $\alpha = 0.6$ achieves the highest re-ID accuracy, indicating an optimized balance between local and global absorption.

We also compare three strategies for generating anchors from semantic parts: using the center point itself ($M = 1$), using its four adjoining points ($M = 4$), and using its eight adjoining points ($M = 8$). Figure 3 shows that a setting of $M = 4$ achieves the highest re-ID accuracy for most $\alpha$. This can be explained by a larger $M$ allowing more anchors to “vote” for determining an SNS part, thus increasing the robustness against semantic partition errors. However, if $M$ is too large, some points that are relatively far away from the center point may become noisy anchors (e.g., when $M = 8$, an anchor in “head” may be actually in the region of “upper body”), thus compromising the discriminative ability of

| Method           | Publication | Market-1501 rank-1 mAP | DukeMTMC-reID rank-1 mAP | mAP |
|------------------|-------------|-------------------------|--------------------------|-----|
| SVDNet[30]       | ICCV17      | 82.3                    | 76.7                     | 56.8|
| AOS[31]          | CVPR18      | 86.5                    | 79.2                     | 62.1|
| DML[34]          | CVPR18      | 87.7                    | -                        | -   |
| Cam-GAN[35]      | CVPR18      | 88.1                    | 75.3                     | 53.5|
| MLFN[36]         | CVPR18      | 90.0                    | 81.0                     | 62.8|
| HydraPlus[18]    | ICCV17      | 76.9                    | -                        | -   |
| PAR[5]           | ICCV17      | 81.0                    | 79.8                     | 62.0|
| PDC[3]           | ICCV17      | 84.4                    | -                        | -   |
| PSE[17]          | CVPR18      | 87.7                    | 73.9                     | -   |
| GLAD[2]          | MM17        | 89.9                    | 73.9                     | -   |
| HA-CNN[17]       | CVPR17      | 91.2                    | 80.5                     | 63.8|
| PABR[4]          | ECCV18      | 91.7                    | 84.4                     | 71.0|
| PCB[31]          | ECCV18      | 92.3                    | 81.8                     | 66.1|
| Manes[6]         | ECCV18      | 93.1                    | 84.9                     | 71.8|
| PCB + RPP[1]     | ECCV18      | 93.8                    | 83.3                     | 69.2|
| SNS (1 anchor)   |             | 91.5                    | 82.4                     | 68.7|
| SNS (4 anchors)  |             | 91.9                    | 82.7                     | 70.3|
the learned SNS feature.

### 4.4 Ablation study

We conduct an ablation study on Market-1501 using three exclusions as follows:

- The “No SNS” setting removes the SNS module and directly uses the anchors for feature representation;
- The “No supervision” setting cancels the supervision for SNS learning; and
- The “Abnormal anchors” setting uses several equidistant points aligned from top to bottom on tensor $T$ as the anchors, thus abandoning semantic cues for SNS learning.

The results are shown in Table 3, from which we draw three observations.

**Direct use of the anchors dramatically decreases re-ID accuracy.** The “No SNS” setting, which directly uses the anchors for feature representation, decreases the re-ID accuracy (by $-2.1\%$ rank-1 accuracy and $-5.2\%$ rank-1 mAP) below the H-baseline. This implies that directly employing the semantic parts from human parsing, although it incurs a considerable number of part errors, is still superior to using only the anchors (center points) for learning part features. Adding SNS modules results in a significant improvement ($+2.1\%$ rank-1 accuracy and $+6.9\%$ mAP) compared with “No SNS”.

**Supervision is critical to the method.** Comparing “No supervision” with “SNS”, we find that the supervision defined in Eq. (2) is critical for our method. Without supervision, the network learns each part’s features with self-attention; this improves on “No SNS” by $+3.4\%$ mAP, but SNS further increases the performance by $+1.7\%$ rank-1 accuracy and $+3.3\%$ mAP.

**Semantic cues are of benefit to SNS.** Comparing

| Setting             | rank-1 | rank-5 | rank-10 | mAP    |
|---------------------|--------|--------|---------|--------|
| IDE                 | 87.3   | 94.6   | 96.9    | 70.1   |
| H-baseline          | 90.5   | 96.3   | 97.5    | 76.9   |
| No SNS              | 88.4   | 95.1   | 97.3    | 71.7   |
| No supervision      | 89.8   | 95.7   | 97.8    | 75.3   |
| Abnormal anchors    | 90.3   | 96.0   | 97.6    | 75.1   |
| SNS                 | 91.5   | 97.1   | 98.0    | 78.6   |

“Abnormal anchors” with “SNS”, we find that semantic cues benefit SNS. The “Abnormal anchors” setting, which replaces semantic cues with heuristic ones (equidistant points), compromises SNS by $-3.5\%$ mAP.

### 4.5 Visualization of SNS parts

Each anchor absorbs similar pixels according to corresponding probabilities, which can alternatively be conceived of as assigning all of the pixels into several SNS parts. In Fig. 4, we visualize the SNS parts compared with the original semantic parts extracted by DeepLab-V3+, from which we draw two observations.

**SNS suppresses semantic part detection errors.** In Figs. 4a–4c, it can be seen that there are several errors in the detection of semantic parts by DeepLab-V3++; in particular the head and feet parts are barely detected. However, the center points of the semantic parts are correctly located in the corresponding part. With these center points as the anchors, SNS successfully recovers the head and feet parts. In Fig. 4b, the human parser blurs the distinction between the upper body and lower body, while SNS successfully separates them. **Correctly locating the anchors is vital for SNS.** In Fig. 4d, the human parsing results are too noisy, and the anchor for the “head” is wrongly located due to the over-large detection errors. SNS thus fails to recover the head region.
“head” is wrongly located. Consequentially, SNS fails to recover the head region. This observation supports that made in Section 4.3, which claims that noisy anchors compromise the discriminative ability of the learned SNS features.

4.6 Evaluation on partial re-ID datasets

We further evaluate our method on two partial re-ID datasets, Partial-REID and Partial-iLIDS. Relatively few results have been reported on these two datasets. We compare our method against several state-of-the-art partial re-ID methods, the IDE baseline, and PCB\cite{1}, which is a state-of-the-art method in holistic re-ID. The results are summarized in Table 4, from which we draw two observations.

SNS significantly surpasses IDE and PCB under the partial re-ID scenario. The superiority of SNS over PCB for partial re-ID originates from semantic part information. With semantic part information indicating which parts are visible, SNS is able to compare two partial images by focusing on their shared parts (Eq. (7)). In contrast, PCB heuristically divides a pedestrian image into a fixed number of parts and thus lacks the flexibility required for partial re-ID.

SNS is competitive with state-of-the-art partial re-ID methods. On Partial-REID and Partial-iLIDS, SNS surpasses DSR by +1.7% and +1.3% rank-1 accuracy, respectively. SNS is not specially designed for partial re-ID problems, and learning SNS features does not leverage the prior knowledge that some images are partial. SNS gains its partial re-ID capacity mainly through adaptive matching during testing stage, as detailed above in Section 3.4.

5 Conclusion

This paper proposes a part feature learning method for re-ID purposes that combines semantic part cues with SNS learning. Instead of using the detected semantic parts directly, SNS locates their geometric centers on the convolutional feature maps and uses these as anchors. Each anchor learns the non-local similarity between itself and other pixels on the convolutional feature maps and then absorbs similar pixels to generate a corresponding part feature. Experimental results confirm that SNS achieves high robustness against semantic part detection errors and thus enhances the discriminative ability of the learned part features. The achieved performance under both holistic and partial re-ID scenarios are competitive with state-of-the-art methods.

Acknowledgment

This work was supported by the National Key Research and Development Program of China (No. 2016YFB0801301) and the National Natural Science Foundation of China (No. 61771288).

References

[1] Y. F. Sun, L. Zheng, Y. Yang, Q. Tian, and S. J. Wang, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), presented at European Conference on Computer Vision (ECCV), Munich, Germany, 2018.
[2] L. H. Wei, S. L. Zhang, H. T. Yao, W. Gao, and Q. Tian, GLAD: Global-local-alignment descriptor for pedestrian retrieval, in Proceedings of the 25th ACM International Conference on Multimedia, 2017.
[3] C. Su, J. J. Li, S. L. Zhang, J. L. Xing, W. Gao, and Q. Tian, Pose-driven deep convolutional model for person re-identification, presented at IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017.
[4] Y. M. Suh, J. D. Wang, S. Y. Tang, T. Mei, and K. M. Lee, Part-aligned bilinear representations for person re-identification, presented at European Conference on Computer Vision (ECCV), Munich, Germany, 2018.
[5] L. M. Zhao, X. Li, Y. T. Zhuang, and J. D. Wang, Deeply-learned part-aligned representations for person re-identification, presented at IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017.
[6] C. Wang, Q. Zhang, C. Huang, W. Y. Liu, and X. G. Wang, Mancs: A multi-task attentional network with curriculum sampling for person re-identification, presented at European Conference on Computer Vision, Munich, Germany, 2018.
[7] M. M. Kalayeh, E. Basaran, M. Gokmen, M.E. Kamasak, and M. Shah, Human semantic parsing for person reidentification, presented at European Conference on Computer Vision (ECCV), Munich, Germany, 2018.
[8] L. Zheng, L. Y. Shen, L. Tian, S. J. Wang, J. D. Wang, and Q. Tian, Scalable person re-identification: A benchmark, presented at IEEE International Conference on Computer Vision, Santiago, Chile, 2015.
[9] X. L. Wang, R. Girshick, A. Gupta, and K. M. He, Non-local neural networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[10] L. Zheng, Y. Yang, and A. G. Hauptmann, Person reidentification: Past, present and future, arXiv preprint arXiv: 1610.02984, 2016.

[11] L. C. Chen, G. Papandreou, and I. Kokkinos, Semantic image segmentation with deep convolutional nets and fully connected CRFs, *Computer Science*, vol. 4, pp. 357–361, 2014.

[12] K. Gong, X. D. Liang, D. Y. Zhang, X. H. Shen, and L. Lin, Look into Person: Self-supervised structure-sensitive learning and a new benchmark for human parsing, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017.

[13] J. Long, E. Shelhamer, and T. Darrell, Fully convolutional networks for semantic segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2016.

[14] S. E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, Convolutional pose machines, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[15] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, DeeperCut: A deeper, stronger, and faster multi-person pose estimation model, presented at European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 2016.

[16] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, Realtime multiperson 2d pose estimation using part affinity fields, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[17] W. Li, X. T. Zhu, and S. G. Gong, Harmonious attention network for person re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[18] X. H. Liu, H. Y. Zhao, M. Q. Tian, L. Sheng, J. Shao, S. Yi, J. J. Yan, and X. G. Wang, HydraPlus-Net: Attentive deep features for pedestrian analysis, in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[19] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. S. Bengio, Show, attend and tell: Neural image caption generation with visual attention, *Computer Science*, in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

[20] X. L. He, J. Liang, H. Q. Li, and Z. N. 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*, 2018.

[21] W. S. Zheng, X. Li, T. Xiang, S. C. Liao, J. H. Lai, and S. G. Gong, Partial person re-identification, presented at IEEE International Conference on Computer Vision, Santiago, Chile, 2015.

[22] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, presented at European Conference on Computer Vision, Amsterdam, The Netherlands, 2016.

[23] Z. D. Zheng, L. Zheng, and Y. Yang, Unlabeled samples generated by GAN improve the person re-identification baseline in vitro, presented at IEEE International Conference on Computer Vision, Venice, Italy, 2017.

[24] W. S. Zheng and S. G. Gong, and T. Xiang, Person re-identification by probabilistic relative distance comparison, in *Computer Vision and Pattern Recognition*. Springer, 2011.

[25] P. Felzenszwalb, D. McAllester, and D. Ramanan, A discriminatively trained, multiscale, deformable part model, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[26] D. L. Gray, S. Brennan, and H. Tao, Evaluating appearance models for recognition, reacquisition, and tracking, in *Proceedings of the 10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, 2007.

[27] S. Y. Lo, H.M. Hang, S. W. Chan, and J. J. Lin, Efficient dense modules of asymmetric convolution for real-time semantic segmentation, arXiv preprint arXiv: 1809.06323, 2018.

[28] L. C. Chen, Y. K. Zhu, G. Papandreou, F. Schroff, and H. Adam, Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[29] T. Xiao, H. S. Li, W. Ouyang, and X. G. Wang, Learning deep feature representations with domain guided dropout for person re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[30] Y. F. Sun, L. Zheng, W. J. Deng, and S. J. Wang, SVDNet for pedestrian retrieval, presented at IEEE International Conference on Computer Vision, Venice, Italy, 2017.

[31] H. J. Huang, D. W. Li, Z. Zhang, X. T. Chen, and K. Q. Huang, Adversarially occluded samples for person re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[32] Z. Zhong, L. Zheng, and G. Kang, Random erasing data augmentation, arXiv preprint arXiv: 1708.04896, 2017.

[33] W. J. Deng, L. Zheng, Q. X. Ye, G. L. Kang, Y. Yang, and J. B. Jiao, Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[34] Y. Zhang, T. Xiang, T. M. Hospedales, and H. H. Lu, Deep mutual learning, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[35] Z. Zhong, L. Zheng, Z. Z. Zheng, S. Z. Li, and Y. Yang, Camera style adaptation for person re-identification, arXiv preprint arXiv: 1711.10295, 2017.

[36] X. B. Chang, T. M. Hospedales, and T. Xiang, Multilevel factorization net for person re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[37] S. M. Saquib, A. Schumann, A. Eberle, and R. Stiefelhagen, A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
Yifan Sun received the BEng degree in mechanical engineering from Tsinghua University, China, in 2005, and the MS degree in optical engineering from Tsinghua University, China, in 2008. He is pursuing PhD degree with the Department of Electronic Engineering in Tsinghua University, since 2015. His research interests are computer vision, person re-identification, and deep metric learning.

Shengjin Wang received the BE degree from Tsinghua University, China, and the PhD degree from the Tokyo Institute of Technology, Tokyo, Japan, in 1985 and 1997, respectively. From 1997 to 2003, he was a member of the research staff with the Internet System Research Laboratories, NEC Corporation, Japan. Since 2003, he has been a professor with the Department of Electronic Engineering, Tsinghua University, where he is currently the Director of the Research Institute of Image and Graphics. His current research interests include image processing, computer vision, video surveillance, and pattern recognition.

Zhaopeng Dou is an undergraduate senior in the Department of Electronic Engineering at Tsinghua University. His research interests include deep learning, object detection, and person re-identification.

Yali Li received the BE degree with Excellent Graduates Award from Nanjing University, China, in 2007 and the PhD degree from Tsinghua University, Beijing, China, 2013. Currently she is a research assistant in Department of Electronic Engineering, Tsinghua University. Her research interests include image processing, pattern recognition, computer vision, and video analysis etc.