Pose Guided Gated Fusion for Person Re-identification

Amran Bhuiyan[a,b], Yang Liu[b], Parthipan Siva[b], Mehrsan Javan[b], Ismail Ben Ayed[a], Eric Granger[a]
[a] LIVIA, Dept. of Systems Engineering, École de technologie supérieure, Montreal, Canada
[b]Sportlogiq Inc., Montreal, Canada
amran.apece@gmail.com, liuyoungwork@gmail.com, psiva7@gmail.com,
mehrsan@sportlogiq.com, ismail.benayed@etsmtl.ca, eric.granger@etsmtl.ca

Abstract

Person re-identification is an important yet challenging problem in visual recognition. Despite the recent advances with deep learning (DL) models for spatio-temporal and multi-modal fusion, re-identification approaches often fail to leverage the contextual information (e.g., pose and illumination) to dynamically select the most discriminant convolutional filters (i.e., appearance features) for feature representation and inference. State-of-the-art techniques for gated fusion employ complex dedicated part- or attention-based architectures for late fusion, and do not incorporate pose and appearance information to train the backbone network. In this paper, a new DL model is proposed for pose-guided re-identification, comprised of a deep backbone, pose estimation, and gated fusion network. Given a query image of an individual, the backbone convolutional NN produces a feature embedding required for pair-wise matching with embeddings for reference images, where feature maps from the pose network and from mid-level CNN layers are combined by the gated fusion network to generate pose-guided gating. The proposed framework allows to dynamically activate the most discriminant CNN filters based on pose information in order to perform a finer grained recognition. Extensive experiments on three challenging benchmark datasets indicate that integrating the pose-guided gated fusion into the state-of-the-art re-identification backbone architecture allows to improve their recognition accuracy. Experimental results also support our intuition on the advantages of gating backbone appearance information using the pose feature maps at mid-level CNN layers.

1. Introduction

Person re-identification is an important function required in many computer vision applications such as video surveillance, search and retrieval, pedestrian tracking for autonomous driving, and multi-camera multi-target tracking.

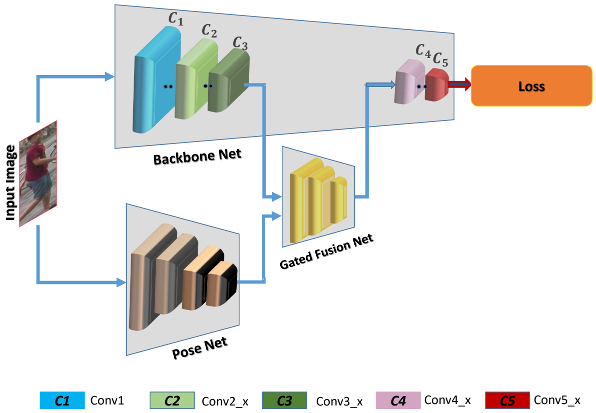


Figure 1. Illustration of the proposed DL architecture for pose-aligned re-identification. The pose networks learns the pose related features and the gated fusion network uses the pose features to dynamically select the relevant appearance features and assist the backbone network to learn the pose related appearance cues for the embedding space.

Given a query image of an individual, re-identification can be seen as a problem of ranking the similarity of all the previously observed images in the gallery. Generally, the aim of re-identification is to recognize individuals tracked over a set of distributed non-overlapping cameras, under the assumption that an individual's overall appearance preserved in all the viewpoints, i.e., no one changes the clothing. Despite of years of arduous efforts [17, 14, 11, 2, 3, 66, 39, 29, 60, 31, 32], person re-identification remains a challenging task due to the non-rigid structure of the human body, the different viewpoint/pose with which a pedestrian can be observed, and the variability of capture conditions (e.g., illumination, scale, motion blur).

State-of-the art approaches for person re-identification typically learning global appearance features in an end-to-end fashion through various metric learning losses [10, 33, 56]. Siamese or triplet-loss CNNs are often used to learn an embedding, where similar image pairs (with the same identity) are close to each other, and dissimilar image pairs (with different identities) are distant from each other [56]. More recent approaches try to incorporate spatial informa-

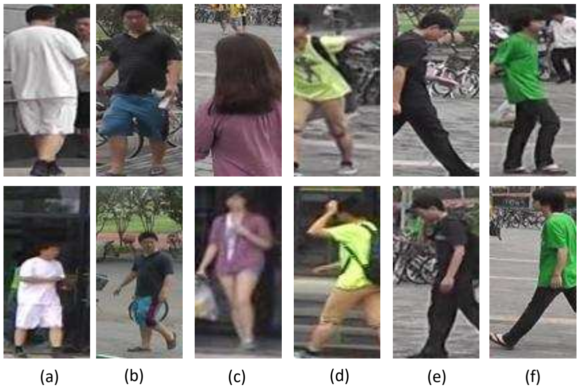


Figure 2. Examples of the misalignment challenge characterizing re-identification in two camera views: (a-c) due to inaccurate person detection, (d-f) due to pose/viewpoint variations.

tion about the human body into the re-identification process by first learning the local representations of predefined body parts, and then, aggregating the local and global representations to achieve robust appearance based features [50, 51, 67, 42, 7, 21, 55, 69].

This paper introduces a person re-identification framework able to co-jointly learn feature embedding that incorporate relevant spatial information from human body appearance with pose/viewpoint information. Figure 1 presents the overall architecture of the proposed pose-aligned re-identification system. It consists of two parallel streams, an appearance learning stream (backbone network) and a pose estimation stream (pose network) which serves as a context-based gating mechanism for re-identification. These streams are combined by the gating fusion network to integrate human body pose information into the metric learning process.

Most related to our proposed approach are the part-based models, which follow a late fusion approach, i.e., the local feature representations are fused together at the end of the network, which actually undermines the local representations throughout the networks [50, 51, 67, 64, 41, 65]. Although these approaches have achieved a high-level of accuracy, they suffer from misalignment which can be attributed to pose/viewpoint variation and person detection error, which are very common in the surveillance videos. A typical scenario of misalignment is depicted in Figure 2.

Feature aggregation approaches are commonly proposed to address the misalignment issue. These approaches include weighing [50, 67, 64] or maximum pooling as well as some advance techniques such as bilinear pooling [15] or gated recurrent unit (GRU) [12]. In particular, gating mechanisms allows for multiplicative interaction between input features and context gate features. During inference, this mechanism allows to dynamically increase the contextual information. Gating functions introduce an attention mechanism which can eventually help dynamically address misalignment issues by focusing on parts, rather than whole image. Given input image, most of the gating approaches [56, 57] dynamically select the most relevant units, layers, or other components in the main backbone (appearance-based) network. Gating features usually originate from a small sub-module that is integrated between different layers of the backbone network. These sub-modules are configured and trained within the network, and thereby allow for making decisions locally specific to the components being configured. Although these techniques are suitable for gated fusion, they employ complex dedicated part- or attention-based architectures that perform late fusion of the contextual information. Furthermore, they do not incorporate or propagate the contextual information through the backbone network during the metric learning process. In our proposed framework, the gated fusion network learns to co-jointly propagate pose and appearance information from mid-level CNN layers to the output, without additional mechanisms in the backbone network. During the training phase, the mechanisms inside the gating fusion network increases the contextual information that is back propagated gradients corresponding to the amplified local similarities, encouraging the lower and middle layers to learn filters to extract more locally similar patterns.

In this context, an important consideration is the best layer of a backbone network to integrate contextual pose-guided information. It is arguably better to fuse feature representations at the middle-layer. Indeed, lower layers of a deep neural network usually extract low-level features, whereas mid-level layers extract more abstract concepts (attributes) such as the parts or more complicated texture patterns. Features from these layers may be more compatible for integration with abstract structural pose information. Moreover, mid-level features are more informative than higher-level ones, and these finer details may be necessary for accurate pair-wide similarity matching.

Hence, in the proposed framework (Figure 1), pose-guided gating fusion is proposed to dynamically select the more informative portion of an individual's image from the mid-level layers of the backbone network. Given a query image, feature maps from the pose network and from mid-level CNN layers of the backbone network are combined by the gated fusion network to generate pose-based gates. This enhances local similarities along the higher layers, so that the backbone network propagates more relevant features to the higher-level layers.

Experimental results on three challenging benchmark datasets for re-identification show that the proposed gating fusion technique can effectively learn to gate informative pose-related features from network, and outperforms state-of-the-art methods upon which it is applied. In addition, our experiments indicate that the gated fusion at the middle layers of the networks are more effective than early or late fusion. This was expected, as the purpose of adding pose features in the re-identification process is to force the appearance embedding network to pick more relevant features and learn a more refined representation in the embedding space.

The main contributions of the paper are: (1) we propose a new pose-aligned re-identification framework with dynamic body pose-guided appearance feature learning and selection with a gating mechanism to address the misalignment issue; (2) we provide intuitive and experimental answers to a major research question about which layers are the proper ones to apply the gating mechanism; and (3) Experimental analysis on three benchmark datasets indicates that concerting pose feature into the state-of-the-art backbone networks assist to further increase their recognition accuracy.

2. Related Work

Classical approaches for re-identification can be regrouped into two major categories. One family of methods focused on searching of the most discriminative features and their combination so to design a powerful descriptor (or signature) for each individual regardless of the scene [14, 11, 2, 3, 66, 39, 29, 60, 31, 32] while the second family of the methods are trying to learn a discriminate distance metric learnt from the data in order to close the gap between a person's different feature [24, 40, 73, 38]. Like other computer vision systems that rely on the hand-crafted features, the former methods suffer from poor generalization, while the latter suffer from computational complexity when the number of cameras or individuals in the scene increases over time. Similar to other computer vision applications, deep learning methods for re-identification have been growing significantly and outperformed the classical methods [1, 19, 56, 8, 16, 10, 33]. This section provides an overview of the state-of-the-art re-identification including pose information and gated fusion methods.

Conventional re-identification using pose. Due to nonrigid structure of the human body, appearance models that consider individual parts or the pose of the body generate superior results compared to the holistic approaches. This observation was first exploited in symmetry-driven accumulation of local features by using two axes dependent on the body's pose to obtain pose invariant feature representation [14]. Similarly, [11, 2, 3] use a fine-grained pose representation to match features coming from a number of well-localized body parts and weighing them based on their salience. In [13], a pose-aware multi-shot matching techniques is proposed where efficient use of multi-shot matching is conducted based on the target pose information.

Deep re-identification. Deep re-identification methods originate from Siamese networks idea [4].The first attempt to use deep learning for re-identification was based on using three Siamese sub-networks for feature learning introduced in [62]. Following that direction various deep CNN based re-identification approaches has been introduced [1, 19, 56, 8, 16] to learn features in an end-to-end fashion through various metric learning losses such as contrastive loss [56], triplet loss [33], improved triplet loss [10], quadruplet loss [8], and triplet hard loss [19]. Most of these approaches are holistic, thereby learning a global feature representation without explicitly considering the spatial structure of the person. Unlike these methods, our proposed network is able to capture and propagate fine-grained body pose related details to generate a more robust feature embedding.

Deep re-identification using pose. More recently, incorporating contextual information into a CNN-based matching has proven to be successful to increase the re-identification accuracy. In this context, most of the re-identification approaches rely on pose-guided approaches [50, 51, 67, 64, 41, 65, 46] that detect body parts by using an off-the-shelf pose estimator. For instance, [50, 51, 64, 67] use pose-estimator to detect normalized part regions from an image, and then exploits different fusion techniques to fuse the features extracted from the original images and the part region images. In [46], confidence maps generated by the pose estimator is used as an additional channel to the input image. There are some attention maps based approaches [7, 21, 55, 69, 65, 36, 61] which are supposed to attend informative body parts, however they are estimated by the same backbone feature extractor network and hence, they often fail to produces reliable attention maps. Alternatively, we utilize a gated fusion techniques where pose feature map is used to gate the appearance feature map.

Gated fusion. Gated fusion is seen as an important component to regulate the flow of information through deep networks [20, 48]. In [20], the authors proposed the gating mechanisms on the input and output gates for information regulation. Similarly, highway networks idea [48] is proposed to overcome the difficulties associated with the increased depth in the networks. Other information regularization gating mechanisms are proposed to handle noise and occlusion [35, 43, 63]. More recently, gated fusion is used for dynamic selection of the feature by re-scaling or calibrating the different components in a model [22, 49]. In re-identification, gating mechanism is first introduced in [56], where a gating function is proposed as a similarity measure between the Siamese inputs. This method depends on the reliability of the appearance features from the lower layers for gating into upper layers. In contrast, our proposed network is comprised of a lightweight gating fusion module that enhances the representational power of the backbone feature embedding network. It learns to propagate pose and appearance information through the higher CNN layers with a minimal computational cost.

3. Proposed Approach

This DL model proposed in this paper contains three convolutional neural sub-networks, the backbone, the pose and the gated fusion networks. Ideally, all three networks would be trained in one step, however there is no re-identification dataset available with annotated pose data to be used in a joint training process. Therefore, the training process is performed in two steps. At first the pose network is pre-trained independently on a pose estimation dataset to estimate the pose related feature maps, and then the backbone and gated fusion networks are trained jointly for the re-identification task while freezing the weights of the trained pose network. Given an input, the gated fusion network relies on pose features as an attention mechanism to dynamically select the most discriminant convolutional filters from the backbone network for pairwise matching.

3.1. Backbone Network:

The backbone network is trained with a labelled dataset to extract appearance features for the given input image. Any CNN feature extractor networks such as ResNet [18], Inception [54] and DenseNet [23] can be used as the backbone network.

Considering the backbone network by itself, $\mathcal{A}$, it computes the appearance feature map, A^l, which is the output of the l-th layer of the backbone network for input image, $\mathbf{I}$:

$$\mathbf{A}^l = \mathcal{A}^l(\mathbf{I}) \tag{1}$$

where, $\mathcal{A}^l$ is the appearance feature extractor until l-th layer defined as $\mathcal{A}^l : \mathbf{I} \rightarrow \mathbf{f}, \mathbf{I} \in \mathbb{R}^{h \times w \times 3}, \mathbf{f} \in \mathbb{R}^{h' \times w' \times c_l}$, with h, and w are the height and width of an input image and h', w', c_l being the height, width and channel number of feature map size of the l-th layer.

3.2. Pose Network:

The pose network's role is to provide information about the human body parts and the overall pose to the gating network for regulating the learnt appearance feature in the backbone network. There are several networks that estimates the human joints effectively [5, 59]. However, it is difficult to precisely represent the individual portions by their joint coordinates. Here, we select the networks with the human pose confidence maps $\mathbf{S}$ and part affinity fields $\mathbf{L}$ outputs as pose network in our architecture. The confidence maps are the confidence distribution of the body joints, while the part affinity fields learn the vectorize association between the body part. Therefore, the pose network $\mathcal{P}$ generates the pose maps $\mathbf{P_{S,L}}$ to represents the informative portion of the individual body parts based on the given input images, which could be formulated as:

$$\mathbf{P_{S,L}} = \mathcal{P}(\mathbf{I}) \tag{2}$$

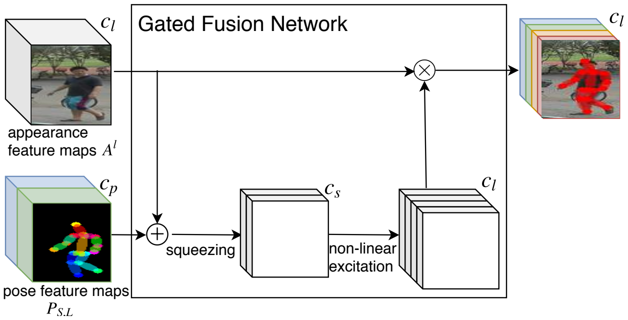


Figure 3. Schematic illustration of the function of Gated Fusion Network (Best viewed in colors).

where $\mathcal{P}$ is the pose map extractor defined as $\mathcal{P} : \mathbf{I} \rightarrow \mathbf{f}, \mathbf{I} \in \mathbb{R}^{h \times w \times 3}, \mathbf{f} \in \mathbb{R}^{h' \times w' \times c_p}$, with h, w, being the height and width of an input image and h', w' being the height and width of pose map. We use OpenPose [5] as the pose network and initialize the weight provided by [5]. Here, c_p represents the total number of confidence maps $\mathbf{S}$ of body part locations and vector fields $\mathbf{L}$ of part affinities, which encode the degree of affinities among the parts.

3.3. Gated Fusion Network:

The objective of the gated fusion network is to enable the backbone network to learn informative feature by fusing appearance and pose features within local receptive fields at fused layer. In this regards, we use a variant of the *squeeze and excitation* [22] as gated fusion which enables us to adaptively re-calibrate the channel-wise feature response to come up with informative feature. Note that, gated fusion network allows the backbone network to leverage pose information. It regulates the backbone network features to pay more attention to the pose-based informative portion. More specifically, the gating network learns a coefficient matrix:

$$\mathbf{G} = \mathcal{D}(E(\mathbf{I})) \tag{3}$$

where, $E \in \left\{\mathbf{A}^l, \mathbf{P_{S,L}}\right\}$ is the concatenated feature map of size $h' \times w'$, and each location is described by $c_g = (c_l + c_p)$. $\mathcal{D}$ is mapping function defined as $\mathcal{D} : \mathbf{f} \rightarrow \mathbf{g}, \mathbf{f} \in \mathbb{R}^{h' \times w' \times c_g}, \mathbf{g}_l \in \mathbb{R}^{h' \times w' \times c_l}$. This mapping function $\mathcal{D}$ is learned by proposed gate module, inspired by layer fusion methods [22] and multi-modal fusion techniques [43, 63].

The gate module in Figure 3 is designed to learn the spatial-wise and channel-wise attention guided by pose-based informative portion. It is composed by 2 operations: *squeeze* and *non-linear excitation*. The squeeze operation utilizes a 3×3 convolutional layer to aggregate the appearance features and pose features across their spatial domain, while non-linear excitation operation further captures the channel-wise dependencies between the appearance and pose feature internally and externally with a leaky rectified linear unit (LeakyReLU) and a 1 × 1 convolutional layer. All the feature maps in the gate module have the same h' and w'. The channel number of output attention keeps the same as the appearance features, which allows the output attention to scale and emphasis the appearance feature in pixel-wise level among all channels.

Once we have gated output from the gate module then we propose a simple and effective scheme to align the appearance feature. The resultant aligned appearance is propagating to the rest of the network. Specifically, we extract aligned feature map by applying Hadamard Product between the appearance feature, A^l and gated output, g_l, and the resulting features are then normalized to attain an aligned feature map for the rest of the layers on the backbone network. A schematic functionality of the gated fusion network is illustrated in Figure. 3. The gated aligned feature,f_g^l scheme is formulated as:

$$\mathbf{f}_g^l = \mathbf{A}^l \bigotimes \mathbf{g}_l, \quad \tilde{\mathbf{f}}_g^l = \frac{\mathbf{f}_g^l}{\|\mathbf{f}_g^l\|_2} \tag{4}$$

$\tilde{f}_g^l$ is the normalized aligned feature representation and $\bigotimes$ denote element-wise product (Hadamard product).

4. Experimental Results

This section presents the datasets, implementation details, and performance metrics used for validation. Then, qualitative and quantitative results with our method are shown, and compared to the state-of-the-art. Then experimental insights are also provided on the impact of applying gated fusion at different layers of the backbone network.

4.1. Datasets:

Our experiments are performed with 3 challenging video datasets for person re-identification – CUHK03-NP [26], Market-1501 [68] and DukeMTMC-reID [70].

Market-1501 [68] is one of the largest public benchmark datasets for person re-identification. It contains 1501 identities which are captured by six different cameras, and 32,668 pedestrian image bounding-boxes obtained using the Deformable Part Models (DPM) pedestrian detector. Each person has 3.6 images on average at each viewpoint. The dataset is split into two parts: 750 identities are utilized for training and the remaining 751 identities are used for testing. We follow the official testing protocol where 3,368 query images are selected as probe set to find the correct match across 19,732 reference gallery images.

CUHK03-NP [26] consists of 14,096 images of 1,467 identities. Each person is captured using two cameras on the CUHK campus, and has an average of 4.8 images in each camera. The dataset provides both manually labeled bounding boxes and DPM-detected bounding boxes. In this paper, both experimental results on labeled and detected data are presented. We follow the new training protocol proposed in [72], similar to partitions of Market1501 dataset. The new protocol splits the dataset into training and testing sets, which consist of 767 and 700 identities, respectively. In testing mode, one image is randomly selected from each camera as the query for each individual, and the remaining images are used to construct the gallery set.

DukeMTMC-reID [70] is constructed from the multi-camera tracking dataset DukeMTMC. It contains 1,812 identities. We follow the standard splitting protocol proposed in [70] where 702 identities are used as the training set and the remaining 1,110 identities as the testing set. During testing, one query image for each identity in each camera is used for query and the remaining as the reference gallery set.

4.2. Implementation details:

The proposed *Gated-fusion* is applied on *Trinet* [19] as a weak-baseline, PCB [53] and BOT [37] as strong baselines.

Network architecture. For *Trinet* [19] and BOT [37] architectures, images for all the baseline are resized to 256 × 128. For PCB [53] architectures, images are resized to 384 × 128, depending on the baseline. Although the proposed method can integrate a wide range of feature extractors, we chose all these state-of-the-art that uses ResNet50 [18] architecture as the backbone network due to its popularity in re-identification. Given an input image, the OpenPose network [5] is used to extract 50 pose-based feature maps, which are then used to gate the mid-layer features of the backbone architecture.

Training mode. The backbone and pose networks are initially pre-trained on ImageNet [45] and COCO [30] datasets. We adopt the OpenPose [5] network for low resolution image by augmenting the datasets with varying resolutions. The visualization of how good the pose estimation works for low resolution images are shown on Figure 4. The gated fusion network is initialized from a Gaussian distribution. During the fine-tuning with a re-identification dataset, the pose network is frozen since the task is completely different. For *Trinet* [19], the backbone and gating networks are trained using triplet loss, where the margin is empirically set to $m = 0.3$. The batch size is set to 128, with 4 randomly selected samples for each 32 identities. We train the network for 300 epochs using the Adam Optimizer. The initial learning rate is set to 2×10^{-4} and starts to decay from 151 epochs by a factor of 0.005. For PCB [53] and BOT [37] architectures, we follow the same training procedures as in [53] and [37] which include Cross-Entropy loss or ID-loss [53] and combination of ID-loss, triplet loss and center loss [37]. We use two NVIDIA GTX-1080Ti GPUs for all our experiments, and implement all code in PyTorch framework.

Testing mode. The proposed model is evaluated for its

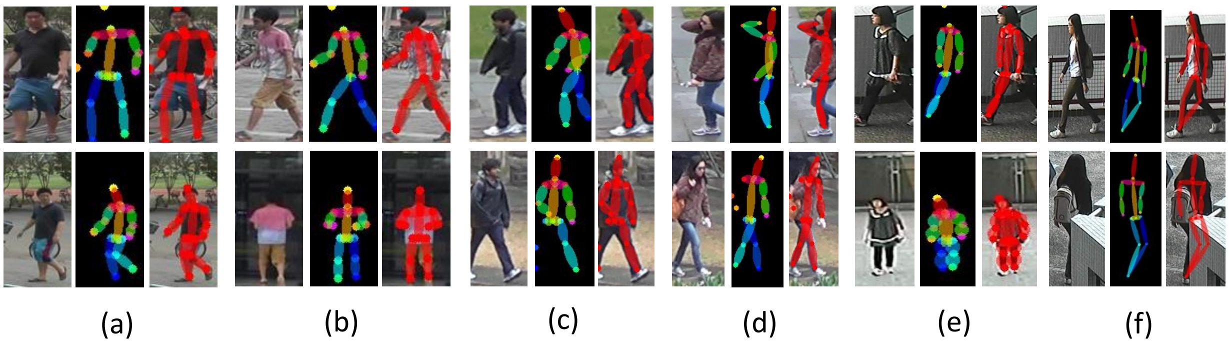


Figure 4. Pairs of samples from (a,b) Market-1501; (c,d) DukeMTMC-reID and (e,f) CUHK03-NP datasets, respectively. The first row shows the images of five persons (the input image, its pose map, and the corresponding schematic gated fusion output) captured in one view. The second row shows the images from the same five individuals, captured in another view.

ability to provide discriminant feature embeddings. OpenPose network extracts pose-based feature maps for each image to gate features of the backbone network. Feature extracted from query and gallery images are compared through pair-wise matching. Similarity between each pair of feature embeddings is measured using Euclidean distance. For each query image, all gallery images are thereby ranked according to the similarity between their embeddings in Euclidean space, and the label of the most similar gallery image is returned. *We did not use any re-ranking tricks to our experimental evaluation.*

4.3. Performance measures:

Following the common trend of evaluation [56, 68, 70], we measure the rank-1 accuracy of cumulative matching characteristics (CMC), and the mean average precision (mAP) to evaluate our proposed and baseline methods. The CMC represents the expectation of finding a correct match in the top n ranks. When multiple ground truth matches are available, then CMC cannot measure how well the gallery images are ranked. Thus, we also report the mAP scores. Higher values represent better performance.

4.4. Gated Fusion using Weak Re-identification Baselines:

Goal. The objective of this experiment is to analyze and compare our proposed approach with weak baseline such as *Trinet* [19]. In order to show the significance of our gated fusion, we compare it with some recent state-of-the-art methods those which considered some degree of contextual or spatial information on their architectures.

Table 1. Comparison of rank-1 accuracy and mAP of the proposed approach with weak baseline and state-of-the-art methods on **Market-1501** dataset [68]. The best and second best results are shown in red and blue, respectively.

Method	**Reference**	rank-1 (%)	mAP (%)
Gated Siamese [56]	ECCV, 2016	65.88	39.55
Spindle [64]	CVPR, 2017	76.90	-
PIE [67]	CVPR, 2017	78.65	53.87
MSCAN [25]	CVPR, 2017	80.31	57.53
HydraPlus [36]	ICCV, 2017	76.90	-
PAR [65]	ICCV, 2017	81.00	63.40
JLML [27]	IJCAI, 2017	85.10	65.50
PDC [50]	ICCV, 2017	84.14	63.41
SVDNet [52]	ICCV,2017	82.30	62.10
PAN [71]	TCSVT,2018	82.81	63.35
PSE [46]	CVPR, 2018	**87.70**	69.00
AACN [61]	CVPR, 2018	85.69	66.87
MGCAM [47]	CVPR, 2018	83.79	74.33
Pose Transfer [34]	CVPR, 2018	87.65	68.92
Pose Norm. [41]	ECCV, 2018	87.26	69.32
Part Aligned. [51]	ECCV, 2018	87.60	72.20
DaRe [58]	CVPR, 2018	86.40	69.30
AWTL [44]	CVPR, 2018	86.11	**71.76**
Trinet [19]	arxiv17	84.92	68.91
Gated Fusion (Trinet)	Proposed	**88.51**	**74.55**

Table 2. Comparison of the proposed method with weak baseline and state-of-the-art methods on **DukeMTMC-reID [70]** dataset. The best (second best) results are shown in red (blue).

Methods	**Reference**	**rank-1 (%)**	**mAP (%)**
PAN [71]	TCSVT, 2017	71.59	51.51
SVDNet [52]	ICCV, 2017	76.70	56.80
Pose Norm [41]	ECCV, 2018	72.80	52.48
AACN [61]	CVPR, 2018	76.84	**59.25**
Pose Transfer [34]	CVPR, 2018	**78.52**	56.91
DaRe [58]	CVPR, 2018	75.20	57.40
AWTL [44]	CVPR, 2018	75.31	57.28
Trinet [19]	arxiv,2017	74.91	56.65
Gated Fusion ((Trinet)	Proposed	**78.82**	**62.49**

Results. Tables 1 to 3 report the comparative performances of methods on Market-1501, DukeMTMC-reID, CUHK03-NP (detected) and CUHK03-NP (labeled) datasets, respectively. Integrating pose-guided gated fusion on weak baseline (*Trinet [19]*) shows a considerable improvements over it baseline performance on Market-1501 and DukeMTMC-reID dataset in both measures. The rank-01 and mAP performance improvements over baseline *Trinet* [19] are 3.59% and 5.5% on Market-1501 dataset, while 3.9% and 5.84% on DukeMTMC-reID datasets, respectively. Among the alternatives, *Gated Siamese [56]* performs worse while using their own network architecture. The performance of the state-of-the-

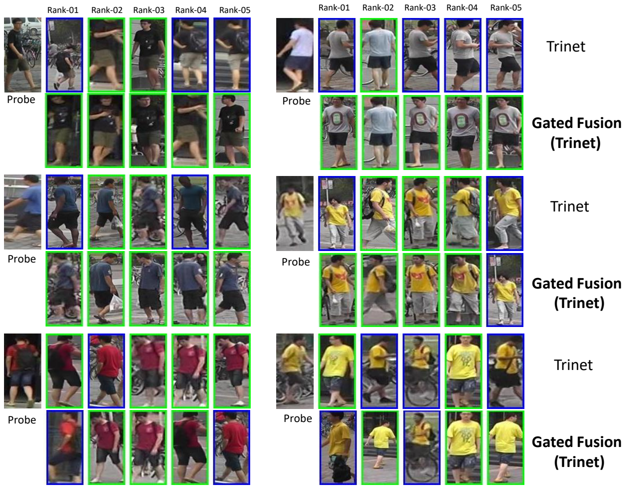


Figure 5. Visual comparison of probe-set images to top 5 matching images from the gallery-set for six random person in the Market-1501 dataset. For each probe, the first and second rows correspond to the ranking results produced by `Baseline (Trinet)` and `Gated Fusion (Trinet)` approaches, respectively. Images surrounded by a green box denotes a match between probe and gallery.

art methods varies significantly depending on their backbone networks and to make the results consistent and comparable, we demonstrate the state-of-the-art results in which the ResNet50 as well as some degrees of contextual information (i.e. pose, parts, attribute, segmentation mask) are uses as the backbone network. Nevertheless, our proposed approach consistently outperforms the considered state-of-the-art methods irrespective to their backbone architectures. We also present some qualitative examples from Market-1501 dataset which indicates that our proposed `Gated Fusion on Trinet [19]` approach effectively finds the true match in `rank-01` when there are cases of `misalignment, occlusions and body part missing`, while the `Baseline` approach finds it in later ranks. We also present some cases where our proposed `Gated Fusion` approach is not able to find the true match in `rank-01` when most of the images are well aligned, although they are eventually recognized within first few ranks (most cases in `rank-02`).

On the CUHK03-NP dataset, the margin of improvement of our approach is higher than the other datasets. The performance gap between baseline *Trinet* [19] and our gated fusion on baseline *Trinet* [19] are 7.42%/4.6% and 4.07%/3.35% of rank-01/mAP on CUHK03-NP(detected) and CUHK03-NP(labeled) dataset, respectively. We speculate that CUHK03-NP contains intensive alignment variations than other two datasets and thus, the effect of our proposed approach is more visible. In addition, the performance improvement of manually labeled data is comparatively higher than the detected ones which suggest that the manual annotations are good enough for the network to learn the feature embedding.

Table 3. Rank-1 accuracy and mAP of proposed approach compared to weak baseline and state-of-art methods on the CUHK03-NP (detected) and CUHK03-NP (labeled) datasets. The best and second best results are shown in red and blue, respectively.

	detected		**labeled**	
Methods	**rank-1 (%)**	**mAP (%)**	**rank-1 (%)**	**mAP (%)**
PAN [71]	36.30	34.00	36.90	35.00
DPFL [9]	40.70	37.00	43.00	40.50
SVDNet [52]	41.50	37.26	40.93	37.83
HA-CNN [28]	41.70	38.60	44.40	41.00
MGCAM [47]	46.71	46.87	50.14	50.12
MLFN [6]	52.80	47.80	54.70	49.20
Pose Transfer [34]	41.60	38.70	45.10	42.00
DaRe [58]	**55.10**	**51.30**	**58.10**	53.70
Trinet [19]	50.43	50.20	56.93	**55.64**
Gated Fusion (Trinet)	**57.85**	**54.80**	**61.00**	**58.99**

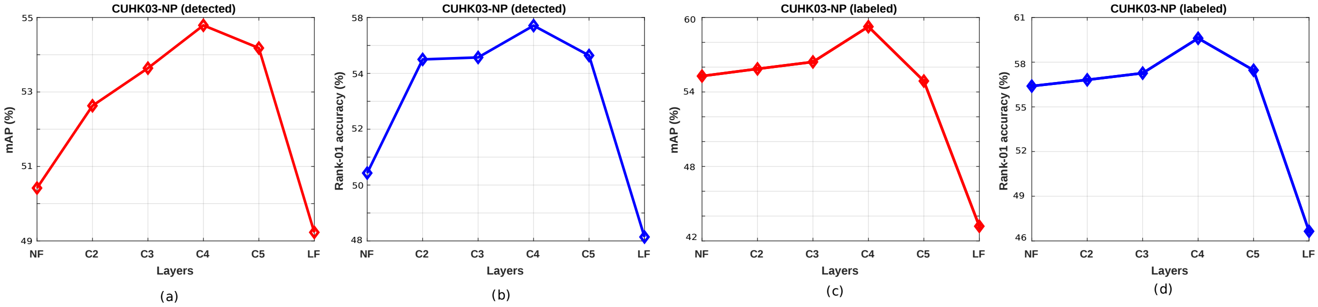
4.5. Gated Fusion using Strong Re-identification Baselines:

Goal. The aim of this experiment is to analyze the effectiveness of our proposed *Gated-Fusion* on strong baseline like PCB [53] and BOT [37]. We compared it with some recent remarkable works, including some alignment methods [46, 51, 69, 53], architecture [53], attention methods [61, 69, 7] and others on Market-1501 and DukeMTMC-reID datasets.

Table 4. Rank-1 accuracy and mAP of the proposed compared to strong baseline and state-of-the-art methods on **Market-1501** dataset [68]. The best/second results are shown in red/blue, resp.

Method	**Reference**	rank-1 (%)	mAP (%)
PSE [46]	CVPR, 2018	87.70	69.00
AACN [61]	CVPR, 2018	85.69	66.87
MGCAM [47]	CVPR, 2018	83.79	74.33
Pose Transfer [34]	CVPR, 2018	87.65	68.92
Pose Norm. [41]	ECCV, 2018	87.26	69.32
Part Aligned. [51]	ECCV, 2018	87.60	72.20
DaRe [58]	CVPR, 2018	86.40	69.30
AWTL [44]	CVPR, 2018	86.11	71.76
CASN+PCB [69]	CVPR, 2019	94.40	82.80
MHN-6 (PCB) [7]	ICCV. 2019	**95.10**	85.00
PCB [53]	ECCV, 2018	92.30	77.40
Gated Fusion (PCB)	Proposed	**92.90**	**78.50**
BOT [37]	CVPRWK, 2019	94.50	**85.90**
Gated Fusion (BOT)	Proposed	**94.60**	**87.10**

Results. Tables 4 and 5 report the comparative performances of methods on Market-1501 and DukeMTMC-reID datasets, respectively. The reported results on these tables suggest that margin of improvements over the strong baseline are relatively low compared to the improvement over the weak baseline as reported in the previous section (section 4.4). One noticeable thing is that gated fusion on BOT [37] outperforms all the state-of-the-art on mAP measurements which also suggests that where there is room for



(a) (b) (c) (d)

Figure 6. Impact on mAP accuracy of applying the proposed gated fusion at different layers of the backbone architectures: (a,b) evaluations for CUHK03-NP (detected) datasets, (c,d) evaluations for CUHK03-NP (labeled) datasets. NF: No Fusion and LF: Late Fusion.

improvements (as in mAP measurements), integrating gated fusion performs well over the considered baselines. Although our propose gated fusion does not outperform the state-of-the-art on rank-01, it serves our purpose to show how to improve the re-identification accuracy with contextual information (i.e pose) without considering complex architecture as in CASN+PCB [69] and MHN-6(PCB) [7]. Improvement over all the above frameworks suggest that proposed gated fusion is general and can be applied to multitude of different feature extractor and different loss functions.

Table 5. Performance of the proposed method using strong baseline and the state-of-the-art on **DukeMTMC-reID [70]** dataset. The best/second results are shown in red/blue, respectably.

Methods	Reference	rank-1 (%)	mAP (%)
PAN [71]	TCSVT, 2017	71.59	51.51
SVDNet [52]	ICCV, 2017	76.70	56.80
Pose Norm [41]	ECCV, 2018	72.80	52.48
AACN [61]	CVPR, 2018	76.84	59.25
Pose Transfer [34]	CVPR, 2018	78.52	56.91
DaRe [58]	CVPR, 2018	75.20	57.40
AWTL [44]	CVPR, 2018	75.31	57.28
CASN+PCB [69]	CVPR, 2019	87.70	73.70
MHN-6 (PCB) [7]	ICCV, 2019	**89.10**	**77.20**
PCB [53]	ECCV,2018	83.30	69.30
Gated Fusion (PCB)	Proposed	**84.50**	**71.10**
BOT [37]	CVPRWK,2019	86.40	76.40
Gated Fusion (BOT)	Proposed	**88.30**	**78.10**

4.6. Gated fusion at Different Backbone Layers:

Goal. The objective of this experiment is to verify the effectiveness of *Gated Fusion* approach by changing the location of the fusion in different layers. We conduct this experiment on CUHK03-NP (detected) and CUHK03-NP (labeled) datasets and apply the gated fusion on different layers of the backbone networks. To match the feature map dimensions a bi-linear interpolation is applied.

Results. Figure. 6 shows the results of our method by applying gated fusion on different layers of the backbone architecture (ResNet50) on CUHK03-NP (detected) and CUHK03-NP (labeled) datasets. Similar to the results in Section 4.5, the gated fusion method consistently works well when it fusion is done on the mid-level layers. This rank-01 performance improvements between when there is no fusion (NF) and the fusion at layer C4 are 7.42% and 4.07% on CUHK03-NP (detected) and CUHK03-NP (labeled) datasets, respectively.

Fusing the pose feature maps at the upper layers of the network, referred to as *late fusion (LF)*, degrades the performance and makes the results even worse than without using any fusion. We did not report the results for gated fusion on layer C1, as the performance is considerably worse compared to other fusion scenarios.

We also observed that there is a gap in performance for fusion at C3 and C4. For some datasets, fusion at C3 outperforms fusion at C4, however both C3 and C4 are considered as mid level CNN layers. Results support our claim that mid-level features (i.e. pose) can effectively be fused on mid-level layers on backbone architectures. Applying gated fusion simultaneously on multiple layers made the network unstable, possibly because the accumulation of the pose network output over multiple layers.

5. Conclusions

In this paper, a new framework is proposed for pose-aligned person re-identification, the aim of which is not to outperform the state-of-the-art but to assist the state-of-the-art CNN architectures. The key component of this framework is the gated fusion network that dynamically selects the more relevant convolutional filters of a state-of-the-art CNN architecture based on pose information, for enhanced feature representation and inference. This framework exploits the advantages of pose features to gate appearance information at mid-level CNN layers. Experimental results with three state-of-the-art methods on three benchmark datasets indicate that the proposed framework can outperform considered state-of-the-art methods. Moreover, our proposed architecture is general and can be applied with a multitude of different feature extractors and loss functions.

Acknowledgements: This work was supported by the Mitacs Elevate Postdoctoral Fellowship Program, and the Natural Sciences and Engineering Research Council of Canada.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015.

[2] A. Bhuiyan, A. Perina, and V. Murino. Person re-identification by discriminatively selecting parts and features. In *ECCV*, 2014.

[3] A. Bhuiyan, A. Perina, and V. Murino. Exploiting multiple detections for person re-identification. *Journal of Imaging*, 4(2):28, 2018.

[4] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a" siamese" time delay neural network. In *NIPS*, 1994.

[5] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[6] X. Chang, T. M. Hospedales, and T. Xiang. Multi-level factorisation net for person re-identification. In *CVPR*, 2018.

[7] B. Chen, W. Deng, and J. Hu. Mixed high-order attention network for person re-identification. *ICCV*, 2019.

[8] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, 2017.

[9] Y. Chen, X. Zhu, and S. Gong. Person re-identification by deep learning multi-scale representations. In *ICCVWK*, 2017.

[10] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, 2016.

[11] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, 2011.

[12] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[13] Y.-J. Cho and K.-J. Yoon. Improving person re-identification via pose-aware multi-shot matching. In *CVPR*, 2016.

[14] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[15] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell. Compact bilinear pooling. In *CVPR*, 2016.

[16] M. Geng, Y. Wang, T. Xiang, and Y. Tian. Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*, 2016.

[17] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, volume 3, pages 1–7. Citeseer, 2007.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[19] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[20] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[21] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen. Interaction-and-aggregation network for person re-identification. In *CVPR*, 2019.

[22] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[23] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[24] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[25] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017.

[26] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[27] W. Li, X. Zhu, and S. Gong. Person re-identification by deep joint learning of multi-loss classification. *IJCAI*, 2017.

[28] W. Li, X. Zhu, and S. Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018.

[29] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[31] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo. Person re-identification by iterative re-weighted sparse ranking. *CVPR*, 2015.

[32] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: What features are important? In *ECCV*, 2012.

[33] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 26(7):3492–3506, 2017.

[34] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu. Pose transferrable person re-identification. In *CVPR*, 2018.

[35] J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *ECCV*, 2016.

[36] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *ICCV*, 2017.

[37] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPRWK*, 2019.

[38] B. Mirmahboub, H. Kiani, A. Bhuiyan, A. Perina, B. Zhang, A. Del Bue, and V. Murino. Person re-identification using sparse representation with manifold constraints. In *ICIP*, 2016.

[39] R. Panda, A. Bhuiyan, V. Murino, and A. K. Roy-Chowdhury. Unsupervised adaptive re-identification in open world dynamic camera networks. In *CVPR*, 2017.

[40] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013.

[41] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue. Pose-normalized image generation for person re-identification. In *ECCV*, 2018.

[42] R. Quan, X. Dong, Y. Wu, L. Zhu, and Y. Yang. Auto-reid: Searching for a part-aware convnet for person re-identification. *ICCV*, 2019.

[43] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, W. Liu, and M.-H. Yang. Gated fusion network for single image dehazing. *CVPR*, 2018.

[44] E. Ristani and C. Tomasi. Features for multi-target multi-camera tracking and re-identification. In *CVPR*, 2018.

[45] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[46] M. Saquib Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *ICCV*, 2018.

[47] C. Song, Y. Huang, W. Ouyang, and L. Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, 2018.

[48] R. K. Srivastava, K. Greff, and J. Schmidhuber. Training very deep networks. In *NIPS*, pages 2377–2385, 2015.

[49] M. F. Stollenga, J. Masci, F. Gomez, and J. Schmidhuber. Deep networks with internal selective attention through feedback connections. In *NIPS*, 2014.

[50] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017.

[51] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee. Part-aligned bilinear representations for person re-identification. *ECCV*, 2018.

[52] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In *ICCV*, 2017.

[53] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018.

[54] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017.

[55] C.-P. Tay, S. Roy, and K.-H. Yap. Aanet: Attribute attention network for person re-identifications. In *CVPR*, 2019.

[56] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016.

[57] A. Veit and S. Belongie. Convolutional networks with adaptive inference graphs. In *ECCV*, 2018.

[58] Y. Wang, L. Wang, Y. You, X. Zou, V. Chen, S. Li, G. Huang, B. Hariharan, and K. Q. Weinberger. Resource aware person re-identification across multiple resolutions. In *CVPR*, 2018.

[59] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[60] Z. Wu, Y. Li, and R. J. Radke. Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features. *TPAMI*, 2015.

[61] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang. Attention-aware compositional network for person re-identification. In *CVPR*, 2018.

[62] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. In *ICPR*, 2014.

[63] X. Zhang, H. Dong, Z. Hu, W.-S. Lai, F. Wang, and M.-H. Yang. Gated fusion network for joint image deblurring and super-resolution. *BMVC*, 2018.

[64] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017.

[65] L. Zhao, X. Li, Y. Zhuang, and J. Wang. Deeply-learned part-aligned representations for person re-identification. In *CVPR*, 2017.

[66] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013.

[67] L. Zheng, Y. Huang, H. Lu, and Y. Yang. Pose invariant embedding for deep person re-identification. *arXiv preprint arXiv:1701.07732*, 2017.

[68] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[69] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke. Re-identification with consistent attentive siamese networks. In *CVPR*, 2019.

[70] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.

[71] Z. Zheng, L. Zheng, and Y. Yang. Pedestrian alignment network for large-scale person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[72] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017.

[73] X. Zhu, A. Bhuiyan, M. L. Mekhalfi, and V. Murino. Exploiting gaussian mixture importance for person re-identification. In *AVSS*, 2017.