IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 28, 2018, accepted November 22, 2018, date of publication November 27, 2018,
date of current version December 27, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2883560

Person Re-Identification by Multi-Camera Networks for Internet of Things in Smart Cities

SHILIN ZHANG[1] AND HANGBIN YU[2]
[1]Beijing Key Laboratory of Urban Road Intelligent Traffic Control, North China University of Technology, Beijing 100144, China
[2]College of Electrical and Control Engineering, North China University of Technology, Beijing 100144, China

Corresponding author: Shilin Zhang (zhangshilin@126.com)

This work was supported by the National Natural Science Foundation of China (No. 61303004).

ABSTRACT As one of the most important areas of public safety and security, intelligent video surveillance is an indispensable part of the urban Internet of Things infrastructure. Person re-identification (person re-ID), which aims to track and recognize a person in a multi-camera scene, is mostly viewed as an image retrieval problem, and this task has been greatly boosted by deep convolutional neural networks (CNNs) in recent years. In practice, person re-ID usually adopts automatic detectors to obtain cropped pedestrian images, and CNNs are inherently limited to model geometric transformations due to the fixed geometric structures in their building modules. We incorporate the deformable convolution module to the traditional baseline to enhance the transformation modeling capability without additional supervision. The new module can readily replace their plain counterparts in the existing CNNs and can be easily trained end-to-end by standard backpropagation. Experiments on two large-scale re-ID datasets confirm the performance of our approach. The experiments also show that learning dense spatial transformation in deep CNNs is effective for person re-ID task and has a bright future in the intelligent video surveillance area.

INDEX TERMS Neural networks, video surveillance, Internet of Thing, identification of persons.

I. INTRODUCTION

The urban road intelligent visual surveillance systems are fully developed in recent years. The cameras deployed as sensing devices are connected with each other to form a multi-antenna sensing network. This requires all sensing devices not only to feel but also to identify and analyze, and to transmit the result to the command center for decision making. The communication and real-time performance are big challenges, so scholars put forward new network architectures such as Heterogeneous Internet of Things [1], Event-Aware Backpressure Scheduling [2] and Data-Emergency-Aware Scheduling Scheme [3] to address these issues. Deng *et al.* [4] focus on these problems by applying a service cache policy and by adopting mobility enabled approach [5] in mobile environments. Dependability is another important factor to influence the distributed intelligent systems, and in literature [6] and [7] the problem is thoroughly discussed. Besides, the author of literature [8] addresses the information sharing and visualization issues of IoT services.

The intelligent visual surveillance systems are a vital infrastructure for smart cities. Cameras deployed in urban areas can monitor not only cars but also persons, and the multi-camera networks can serve as an IoT service for city traffic and safety. Zhao *et al.* [9], [10] and Zhao [11] investigate the human tracking and recognition problems and build an intelligent monitoring system using Kinect. In view of the importance of intelligent monitoring, many other research scholars and industry giants have recently seen the prospects of the video-awareness application, and have turned their research focus to the intelligent recognition technology. Person re-identification (person re-ID) is a critical task in most surveillance and security applications, and has increasingly attracted attention from the computer vision community. Person re-ID aims at spotting the target person in different cameras and it is mostly viewed as an image retrieval problem, searching for the queried person in a large image pool (gallery). The application, as an illustration, is shown in Figure 1. The latest progress mainly consists in

2169-3536 © 2018 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

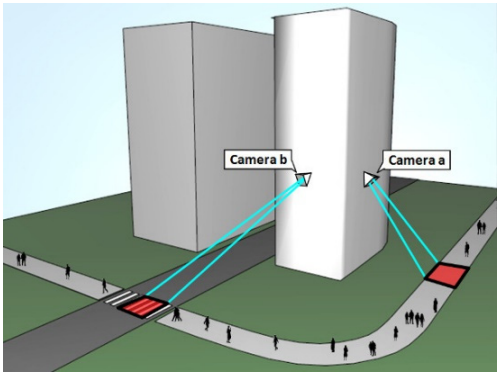


FIGURE 1. The application in a realistic application environment.

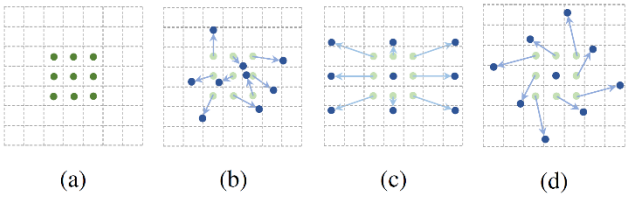


FIGURE 2. Illustration of the sampling locations in 3 × 3 standard and deformable convolutions. (a) regular sampling grid (green points) of standard convolution. (b) deformed sampling locations (dark blue points) with augmented offsets (light blue arrows) in deformable convolution. (c)(d) are special cases of (b), showing that the deformable convolution generalizes various transformations for scale, (anisotropic) aspect ratio and rotation.

the discriminatively learned embeddings using the convolutional neural network (CNN) on large-scale datasets. CNNs have achieved significant success for person re-ID task. The learned embeddings extracted from the fine-tuned CNNs are superior to the handcrafted features.

In real cases, the hand-drawn bounding boxes, existing in some previous datasets such as CUHK01 and CUHK02 [12], are infeasible to acquire when millions of bounding boxes are to be generated. So recent large-scale benchmarks such as CUHK03, Market1501 [13] and MARS [14] adopt the Deformable Part Model [15] to automatically detect pedestrians. However, when detectors are used, detection errors are inevitable, which may lead to two common noisy factors: excessive background and part missing. For the former, the background may take up a large proportion of a detected image. For the latter, a detected image may contain only part of a human body, with missing parts. Among the many influencing factors, misalignment is a critical one in person re-ID.

In order to tackle the above difficulties, we try to build a model with enough desired variations for the misalignment problem. CNNs are inherently limited to model geometric transformations due to the fixed geometric structures in their building modules. Therefore, we propose to incorporate the new module called deformable convolution [16] into the person re-ID architecture. It adds 2D offsets to the regular grid sampling locations in the standard convolution. It enables free-form deformation of the sampling grid [16]. It is illustrated in Figure 2. The offsets are learned from the preceding feature maps, via additional convolutional layers. Thus, the deformation is conditioned on the input features in a local, dense, and adaptive manner.

This module only adds a small number of parameters and can achieve higher accuracy and considerable efficiency. It can be easily trained. Our experiments show that Deformable CNN is superior to the traditional CNN and can make great contribution to the person re-ID task.

II. RELATED WORK

A. HAND-CRAFTED SYSTEMS FOR Re-ID

Person re-ID needs to find robust and discriminative features among different cameras. Before deep learning methods dominated the re-ID research community, hand-crafted algorithms have developed many approaches to learn part or local features.

Several pioneering approaches have explored person re-ID by extracting local hand-crafted features such as Local Binary Pattern [17], Gabor [18] and etc. In a series of works [19]–[21] the 32-dim LAB color histogram and the 128-dim SIFT descriptor are extracted from every 10×10 patches. In the paper of [13], the method uses color name descriptor for each local patch and aggregate them into a global vector through the Bag of Words model. Approximate nearest neighbor search [22] is employed for fast retrieval but accuracy compromise. Paper [23] also deploys several different hand-crafted features extracted from overlapped body patches. Differently, paper [24] localize the parts first and calculate color histograms for part-to-part correspondences. This line of works is beneficial from the local invariance in different viewpoints.

Besides finding robust feature, metric learning is nontrivial for person re-ID. Kostinger *et al.* [25] propose KISSME based on Mahalanobis distance and formulate the paired comparison as a log-likelihood ratio test. Further, paper [13] extend the Bayesian face and KISSME (keep it simple and straightforward metric) [25] to learn a discriminant subspace with a metric. Aside from the methods using Mahalanobis distance, Prosser *et al.* [18] apply a set of weak Rank SVMs to assemble a strong ranker. Gray and Tao [26] propose using the AdaBoost algorithm to fuse different features into a single similarity function. Paper [27] proposes a cross canonical correlation analysis for the video-based person re-ID.

B. DEEP LEARNING FOR PERSON re-ID

Deep learning models have been popular since Krizhevsky *et al.* [28] won ILSVRC12 by a large margin. Deeply learned person representations are producing state-of-the-art re-ID performance recently. It extracts features and learns a classifier in an end-to-end system. The first two works in re-ID to use deep learning were [29] and [30]. Generally speaking, from the perspective of the model, classification models as used in image classification [28] and Siamese models that use image pairs [31] or triplets [32] are two types of CNN models that are commonly employed in re-ID. At the beginning when the training datasets are

not big enough, such as VIPeR [33] that provides only two images for each identity, the Siamese model dominates the re-ID community. As the scale of re-ID dataset becomes large, such as Market-1501 [13], the classification model is widely employed.

On the other hand, comparing with deep similarity learning methods [34]–[36] the representation learning methods are more extendable for large galleries. For example, Hermans *et al.* [37] uses an efficient variant of the triplet loss based on the distance between samples. Another popular choice consists in training an identification network and extracting the intermediate output as a discriminative embedding, [38]–[40]. For example, Xiao *et al.* [41] propose an online instance matching loss to address the problem of having only a few training samples for each class. Lin *et al.* [42] combine attribute classification losses and the re-ID loss for embedding learning. In order to exploit more training data, Zheng *et al.* [43] employ the generative adversarial network to generate samples which are expected to generate uniform prediction probabilities in the softmax layer.

In this paper, we focus on a different goal of finding robust pedestrian embedding for person re-identification, and thus our method can be potentially combined with the previous methods to further improve the performance. Our approach shares the similar high-level spirit with spatial transform [44] networks and its application in the field of person re-ID. A key difference in deformable convolution is that it deals with dense spatial transformations in a simple, efficient and end-to-end manner. In the next part we will discuss in details the relation of our work to previous works and analyze the superiority of the deformable convolution.

Zheng's paper is the first work to learn spatial transformation from data in a deep learning framework use in the large-scale person re-ID. It warps the feature map via a global parametric transformation such as affine transformation. Such warping is expensive and learning the transformation parameters is known difficult. STN has shown successes in small-scale image classification problems. The inverse STN [45] method replace the expensive feature warping by efficient transformation parameter propagations.

The offset learning in deformable convolution can be considered as an extremely light-weight spatial transformer in STN [44]. However, deformable convolution does not adopt a global parametric transformation and feature warping. Instead, it samples the feature map in a local and dense manner. To generate new feature maps, it has a weighted summation step, which is absent in STN. Besides, deformable convolution is easy to be integrated into any CNN architectures. Its training process is easy. It shows effectiveness for complex vision tasks. But these tasks are difficult for STN.

III. DEFORMABLE CONVOLUTION

A. THE OVERALL ARCHITECTURE

The feature maps and convolution in CNNs are 3D. The deformable convolution module operates on the 2D spatial domain. The operation remains the same across the channel dimension. Without loss of generality, the module is described in 2D here for notation clarity. Extension to 3D is straightforward.

The 2D convolution consists of two steps:

1. sampling using a regular grid $\mathcal{R}$ over the input feature map X;

2. summations of sampled values weighted by w. The grid R defines the receptive field size and dilation. For example,

$$\mathcal{R} = \{(-1, -1), (-1, 0), \ldots\ldots, (0, 1)(1, 1)\} \tag{1}$$

defines a 3×3 kernel with dilation 1.

For each location p_0 on the output feature map y, we have

$$\mathbf{y}(p_0) = \Sigma_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n) \tag{2}$$

where p_n enumerates the locations in $\mathcal{R}$.

In deformable convolution, the regular grid $\mathcal{R}$ is augmented with offsets $\{\Delta p_n \mid n = 1, \ldots, N\}$ where $N = |\mathcal{R}|$. The (2) becomes,

$$\mathbf{y}(p_0) = \Sigma_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \tag{3}$$

Now, the sampling is on the irregular and offset locations $p_n + \Delta p_n$. As the offset Δp_n is typically fractional, (3) is implemented via bilinear interpolation as

$$\mathbf{x}(p) = \Sigma_q G(\mathbf{q}, \mathbf{p}) \cdot x(q) \tag{4}$$

Where p denotes an arbitrary location $\mathrm{p} = p_0 + p_n + \Delta p_n$ for (3), q enumerates all integral spatial locations in the feature map X, and $G(\cdot, \cdot)$ is the bilinear interpolation kernel.

Note that G is two dimensional. It is separated into two one dimensional kernels as

$$G(\mathbf{q}, \mathbf{p}) = g(q_x, p_x) \cdot g(q_y, p_y) \tag{5}$$

Where $g(a, b) = \max(0, 1 - |a - b|)$.

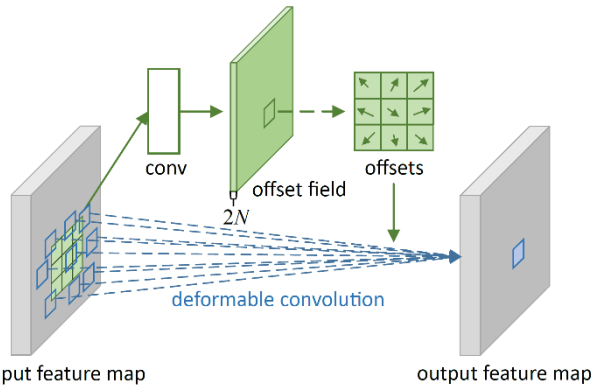


FIGURE 3. Illustration of 3 × 3 deformable convolution.

As illustrated in Figure 3, the offsets are obtained by applying a convolutional layer over the same input feature map. The output offset fields have the same spatial resolution with the input feature map. The channel dimension 2N corresponds to N 2D offsets. During training, both the convolutional kernels

for generating the output features and the offsets are learned simultaneously. To learn the offsets, the gradients are back-propagated through the bilinear operations in (4) and (5).

B. UNDERSTANDING DEFORMABLE CONVOLUTION

This work is built on the idea of augmenting the spatial sampling locations in convolution with additional offsets and learning the offsets from target tasks.

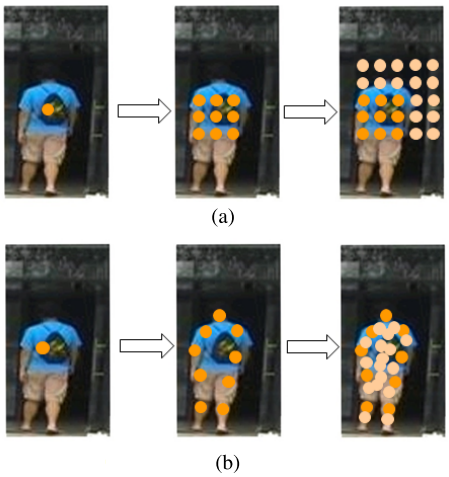


FIGURE 4. Illustration of the fixed receptive field in standard convolution (a) and the adaptive receptive field in deformable convolution (b).

When the deformable convolution module is used, the effect of the composited deformation is profound. This is illustrated in Figure 4. The receptive field and the sampling locations in the standard convolution are fixed all over the top feature map (up). The offsets are adaptively adjusted according to the objects' scale and shape in deformable convolution (down). The adaptive deformation is enhanced especially for nonrigid objects such as a pedestrian.

IV. EXPERIMENTS

To evaluate our system's effectiveness, the experiment is carried on two large-scale datasets Market 1501 and CUHK03 for evaluation. The experimental platform of this algorithm is Caffe and MATLAB R2016b. The OS system is Ubuntu 16.04, which is running on a desktop computer with an Intel Core i5-4590 3.3GHz CPU and TitanX GPU and 16GB of memory. All experiments were repeated 10 times with different test/train splits and the results averaged to ensure stable results. The feature extraction time is 46.38 frames per second, and the feature matching time is 1.03 million pairs of photos per second.

A. DATASETS

Market-1501 [13] is a large-scale image-based re-ID benchmark dataset collected in front of a supermarket in Tsinghua University from six cameras. All images are automatically detected by the DPM [15] detector. The dataset is split into two parts: 12,936 images for training and 19,732 images for testing. There are 751 identities in the training set and 750 identities in the testing set without overlapping. In testing, 3,368 hand-drawn images with 750 identities are used as a probe set to identify the correct identities on the testing set. A single query is to use one image of one person as a query, and multiple queries mean to use several images of one person under one camera as a query. In this paper, we report the single-query evaluation results for this dataset.

CUHK03 [46] contains 14,096 images of 1,467 identities. Each identity is captured from two cameras in the CUHK campus and has an average of 9.6 images. In addition to manually cropped pedestrian images, samples detected with DPM-detected bounding boxes is also provided. This is a more realistic setting considering misalignment, occlusions and body part missing. In this paper, both experimental results on 'labeled' and 'detected' data are presented.

B. EVALUATION METRICS

We use two evaluation metrics to evaluate the performance of re-ID methods on all datasets. The first one is the Cumulated Matching Characteristics (CMC). Considering re-id as a ranking problem, we report the cumulated matching accuracy at rank-1. The rank-1 accuracy denotes the probability of whether one or more correctly matched images appear in top-1. CMC represents the probability that a query identity appears in different sized candidate lists. No matter how many ground truth matches there are in the gallery, only the first match is counted in the CMC calculation. So basically, CMC is accurate as an evaluation method only when one ground truth for each query exists. This measurement is acceptable, in practice, when people care more about returning the ground truth match in the top positions of the rank list.

The other one is the mean average precision (mAP), which reflects the precision and recall rate of the performance. For research integrity, when multiple ground truths exist in the gallery, Zheng *et al.* propose using the mean average precision (mAP) for evaluation. The motivation is that a perfect re-ID system should be able to return all true matches to the user. The case might be that two systems are equally competent at spotting the first ground truth, but have different retrieval recall ability. In this scenario, CMC does not have enough discriminative ability but mAP does. Therefore, the mAP is used together with CMC for the Market-1501 dataset where multiple ground truths from multiple cameras exist for each query.

C. FEATURE REPRESENTATIONS

In this paper, we adopt the baseline identification model, named "ID-discriminative Embedding" (IDE). The IDE extractor is effectively trained on classification model ResNet-50 [47]. It generates a 2,048-dim vector for each image, which is effective in large-scale re-ID datasets. We use ResNet-50 pre-trained on ImageNet as our basic CNN model.

D. IMPLEMENTATION DETAILS

Following the practice, baselines using ResNet-50 are fine-tuned with the default parameter settings except that the

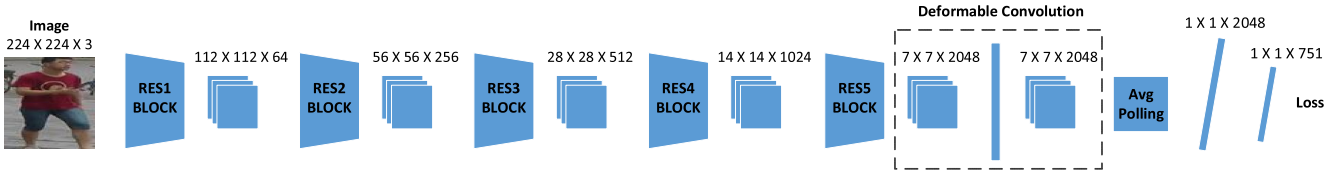


FIGURE 5. The architecture of our model. It adds the Deformable Convolution between the Res5 Block and the Avg Polling. In the training phase, the model minimizes the identification losses. In the test phase, we concatenate a 1 × 1 × 2048-dim pedestrian descriptor for retrieval.

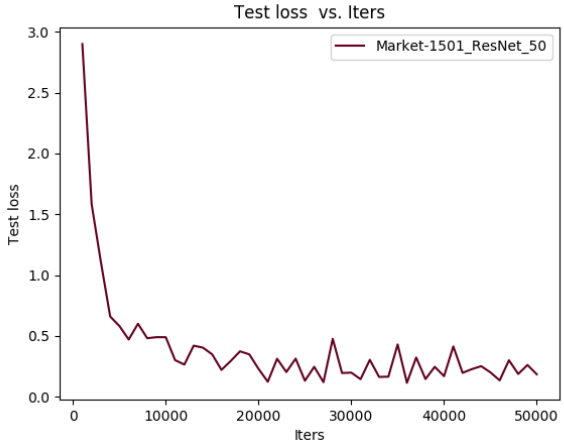


FIGURE 6. Training losses keep declining as iteration continues.

output dimension of the last FC layer is set to the number of training identities and add our Deformable Convolution module. We use the stochastic gradient descent algorithm to train the whole network based on Caffe [48].

The deformable convolution module has the same input and output as their plain versions. Hence, it can readily replace its plain counterparts in existing CNNs. In the training, these added convolution layers for offset learning are initialized with zero weights. It trained via backpropagation through the bilinear interpolation operations. The deformable convolution is applied to the last few convolutional layers (with kernel size > 1). As shown in the [16] we can find that this way can achieve the best performance. Figure 5 briefly illustrates the architecture of our model.

The image feature map extraction part is initialized using the ResNet-50 model, pre-trained over ImageNet. The Baseline is trained for 60 epochs with learning rate initialized at 0.001 and reduced by 10 on 25 and 50 epochs. During testing, the Pool5 or FC descriptor of ResNet-50 is used for feature representation to learns a global descriptor. Specifically, at the beginning of the last block, the stride is changed from 2 to 1. To compensate, the dilation of all the convolution filters in this block (with kernel size > 1) is changed from 1 to 2.

E. EXPERIMENTS ON Market-1501

We first evaluate our method on the Market-1501, in this dataset, we trained the IDE feature on ResNet-50. For each sequence, we first extract feature for each image and use max pooling to combine all features into a fixed length vector. From Figure 6 we can see that when we add the new module, the training loss also keeps declining. Moreover, experiments conducted with two metrics, KISSME [25]

TABLE 1. Comparison of various methods with our approach on the Marke1501 dataset.

Method	Rank1	mAP
IDE ResNet 50 + XQDA	77.58	56.06
IDE ResNet 50 + XQDA + Ours	78.59	57.52
IDE ResNet 50 + XQDA + Re-ranking	80.70	69.98
IDE ResNet 50 + XQDA + Re-ranking + Ours	81.95	72.35
IDE ResNet 50 + KISSME	73.60	49.05
IDE ResNet 50 + KISSME + Ours	80.52	57.78
IDE ResNet 50 + KISSME + Re-ranking	77.11	63.63
IDE ResNet 50 + KISSME + Re-ranking + Ours	**83.61**	**74.60**

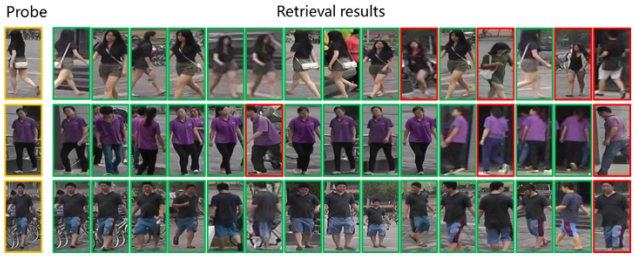


FIGURE 7. Example results of three probes on the Market-1501 dataset. The images in the first column are the query images. The retrieved images are sorted according to the similarity score from left to right. The correct and false matches are in the green and red rectangles, respectively.

and XQDA (Cross-view Quadratic Discriminate Analysis) metrics verify the effectiveness of our method on different distance metrics. The performance of our method on different metrics are reported in Table 1, as we can see, our method consistently improves the rank-1 accuracy and mAP of the two different metrics, even after using the Re-ranking method.

Our method's best metric learning partner is KISSME, further enhanced by the Re-ranking. As we can see from Table 1, when using the KISSME metrics, our method gains 6.92% improvement in rank-1 accuracy and gains 10.97% improvement in mAP. Our best method impressively outperforms the previous work and achieves large margin advances compared with the state-of-the-art results in rank-1 accuracy, particularly in mAP.

In Figure 7, we visualize three retrieval results on the Market-1501 dataset. We can see from the picture, the proposed method can effectively rank true persons in the top of the ranking list.

F. EXPERIMENTS ON CUHK03

For better comparison with the past work, in this paper, we adopt the new training/testing protocol which splits the CUHK03 dataset into a training set and testing set similar to that of Market-1501, proposed by Zheng *et al.*

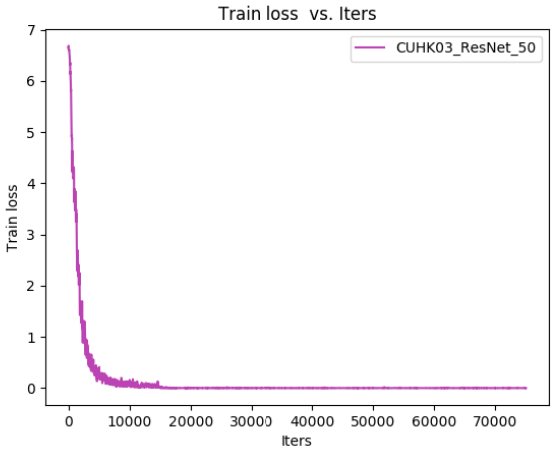


FIGURE 8. The loss of training on the CUHK03.

TABLE 2. Comparison of various methods with our approach on the CUHK03 dataset.

Method	Labeled		Detected	
	Rank1	mAP	Rank1	mAP
IDE + KISSME	28.4	26.1	27.8	25.4
IDE + KISSME + Ours	29.6	26.6	27.8	25.6
IDE + KISSME + Re-ranking	33.9	36.1	32.5	35.8
IDE + KISSME + Re-ranking + Ours	34.8	38.6	33.2	35.5
IDE + XQDA	32.0	29.6	31.1	28.2
IDE + XQDA + Ours	34.2	30.7	31.7	28.8
IDE + XQDA + Re-ranking	38.1	40.3	34.7	37.4
IDE + XQDA + Re-ranking + Ours	**39.1**	**41.7**	**36.6**	**39.1**

FIGURE 9. Sample retrieval results on the CUHK03 dataset. The images in the first column are queries. A person surrounded by green box denotes the same person as the probe.

The new protocol splits the dataset into a training set and testing set, which consist of 767 identities and 700 identities respectively. This method is more in line with the actual situation and avoids repeating training and testing multiple times. CUHK03 originally adopts 20 random train/test splits, which is time-consuming for deep learning. The same as above datasets the experiment evaluates the single-query setting.

From Figure 8, we can see that when we add the new module, there is no bad impact on the ability of the network to learn, the training loss is rapidly declining. As can be seen from Table 2, the results also prove that our feature is in the lead. The new method also gains an increase of 1% in rank-1 accuracy and 1.4% in mAP of the KISSME method, even though in the 'detected' setting enhancing is not obvious. Moreover, our method can improve the rank-1 accuracy and mAP in all cases while XQDA method is used. In particular, our method improves the rank-1 accuracy from 34.7% to 36.6% and the mAP from 37.4% to 39.1% for IDE (R) + XQDA. We believe that the results of this problem will be further improved by combining a more sophisticated feature model with our method.

As shown in Figure 9, we visualize some retrieval results on the CUHK03 dataset both in 'labeled' and 'detected' setting. It turns out that this method works equally well on different datasets.

V. CONCLUSION

In this paper, we present a deformable convolution based scheme for the person re-ID task, which is a simple, efficient, deep and end-to-end solution to model dense spatial transformations. Except for the identity labels, we do not need any extra annotation and supervision.

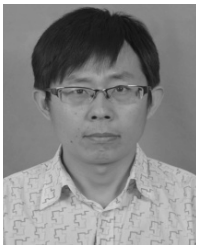
In the experiment, we show that it is feasible and effective to learn dense spatial transformation in CNNs for re-ID tasks. We extensively compared our methods with many state-of-the-art methods on the two most popular and challenging datasets. From the rank-1 results of our proposed system, we can see that the deformable convolutional networks based residual network perform better for every dataset. In addition to the automatically detected datasets, our network also improves the re-ID performance on the datasets with hand-drawn bounding boxes. Our system shows superior performance against other re-ID systems. The proposed method effectively ranks more true persons in the top of the ranking list.

At this stage, video surveillance, especially person re-ID, plays a major role in the intelligent transportation applications. With the popularity of the Internet of Vehicles and the Internet of Things, person re-ID will gradually occupy a dominant position in intelligent transportation and smart city project. In the future, person re-ID will have a wide application prospect and we will continue to investigate the deformable convolution model and apply our model to other related fields.

REFERENCES

[1] T. Qiu, N. Chen, K. Li, M. Atiquzzaman, and W. Zhao, "How can heterogeneous Internet of Things build our future: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2011–2027, 3rd Quart., 2018.
[2] T. Qiu, R. Qiao, and D. Wu, "EABS: An event-aware backpressure scheduling scheme for emergency Internet of Things," *IEEE Trans. Mobile Comput.*, vol. 17, no. 1, pp. 72–84, Jan. 2018.
[3] T. Qiu, K. Zheng, M. Han, C. L. P. Chen, and M. Xu, "A data-emergency-aware scheduling scheme for Internet of Things in smart cities," *IEEE Trans. Ind. Informat.*, vol. 14, no. 5, pp. 2042–2051, May 2018.
[4] S. Deng, Z. Xiang, J. Yin, J. Taheri, and A. Y. Zomaya, "Composition-driven IoT service provisioning in distributed edges," *IEEE Access*, vol. 6, pp. 54258–54269, 2018.
[5] S. Deng, L. Huang, D. Hu, J. L. Zhao, and Z. Wu, "Mobility-enabled service selection for composite services," *IEEE Trans. Services Comput.*, vol. 9, no. 3, pp. 394–407, May 2016.
[6] W. Zhao, *Building Dependable Distributed Systems*. Hoboken, NJ, USA: Wiley, 2014.

[7] W. Shen, "Distributed manufacturing scheduling using intelligent agents," *IEEE Intell. Syst.*, vol. 17, no. 1, pp. 88–94, Jan. 2002.
[8] Z. Shusheng, W. Shen, and H. Ghenniwa, "A review of Internet-based product information sharing and visualization," *Comput. Ind.*, vol. 54, no. 1, pp. 1–15, 2004.
[9] W. Zhao, H. Feng, R. Lun, D. D. Espy, and M. A. Reinthal, "A Kinect-based rehabilitation exercise monitoring and guidance system," in *Proc. IEEE 5th Int. Conf. Softw. Eng. Service Sci.*, Jun. 2014, pp. 762–765.
[10] W. Zhao, D. D. Espy, and A. Reinthal, "A validation study of rehabilitation exercise monitoring using Kinect," in *Encyclopedia of Information Science and Technology*. Hershey, PA, USA: IGI Global, 2018, pp. 5941–5954.
[11] W. Zhao, "A concise tutorial on human motion tracking and recognition with Microsoft Kinect," *Sci. China Inf. Sci.*, vol. 59, pp. 233–237, Sep. 2016.
[12] W. Li and X. Wang, "Locally aligned feature transforms across views," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3594–3601.
[13] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1116–1124.
[14] L. Zheng *et al.*, "MARS: A video benchmark for large-scale person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 868–884.
[15] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
[16] J. Dai *et al.*, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 764–773.
[17] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2666–2672.
[18] B. Prosser, W. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *Proc. Brit. Mach. Vis. Conf.*, 2010, pp. 21.1–21.11.
[19] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 144–151.
[20] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2528–2535.
[21] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3586–3593.
[22] J. Wang and S. Li, "Query-driven iterated neighborhood graph search for large-scale indexing," in *Proc. ACM Int. Conf. Multimedia*, 2012, pp. 179–188.
[23] D. Chen, Z. Yuan, J. Wang, B. Chen, G. Hua, and N. Zheng, "Exemplar-guided similarity learning on polynomial kernel feature map for person re-identification," *Int. J. Comput. Vis.*, vol. 123, no. 3, pp. 392–414, 2017.
[24] S. C. Dong, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *Proc. Brit. Mach. Vis. Conf.*, 2011, pp. 68.1–68.11.
[25] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2288–2295.
[26] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. Eur. Conf. Comput. Vis.*, vol. 5302, 2008, pp. 262–275.
[27] C. C. Loy, T. Xiang, and S. Gong, "Time-delayed correlation analysis for multi-camera activity understanding," *Int. J. Comput. Vis.*, vol. 90, no. 1, pp. 106–129, 2010.
[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, Jun. 2012.
[29] D. Yi, Z. Lei, and S. Z. Li, "Deep metric learning for practical person re-identification," in *Proc. Int. Conf. Pattern Recognit.*, 2014, pp. 34–39.
[30] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.
[31] F. Radenovi, G. Tolias, and O. E. Chum, "CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 3–20.
[32] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.
[33] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany, 2008, pp. 262–275.
[34] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 791–808.
[35] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017.
[36] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1320–1329.
[37] A. Hermans, L. Beyer, and B. Leibe. (2017). "In defense of the triplet loss for person re-identification." [Online]. Available: https://arxiv.org/abs/1703.07737
[38] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1249–1258.
[39] L. Zheng *et al.*, "MARS: A video benchmark for large-scale person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 868–884.
[40] L. Zheng, Y. Huang, H. Lu, and Y. Yang. (2017). "Pose invariant embedding for deep person re-identification." [Online]. Available: https://arxiv.org/abs/1701.07732
[41] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3376–3385.
[42] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang. (2017). "Improving person re-identification by attribute and identity learning." [Online]. Available: https://arxiv.org/abs/1703.07220
[43] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline *in vitro*," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3774–3782.
[44] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.
[45] C. Lin and S. Lucey, "Inverse compositional spatial transformer networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 2252–2260.
[46] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.
[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[48] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.



SHILIN ZHANG was born in Dezhou, Shandong, China, in 1980. He received the Ph.D. degree in computer science from the Chinese Academy of Sciences, China, in 2012. He is currently with the Beijing Key Laboratory of Urban Road Intelligent Traffic Control, North China University of Technology, Beijing. His current research interests include pedestrian detection and recognition, image processing, and pattern recognition. He is a member of the Chinese Association of Automation.

HANGBIN YU was born in Pingdingshan, Henan, China, in 1993. He received the bachelor's degree in automation science from Henan Polytechnic University, China, in 2016. He is currently pursuing the master's degree with the North China University of Technology. His research interests include machine learning and pattern recognition.