2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Attention-based Model with Attribute Classification for Cross-domain Person Re-identification

Simin Xu
School of Aeronautics and Astronautics
Shanghai Jiao Tong University
Shanghai, China
Email: siminxu0613@sjtu.edu.cn

Lingkun Luo
School of Aeronautics and Astronautics
Shanghai Jiao Tong University
Shanghai, China
Email: lolinkun@gmail.com

Shiqiang Hu
School of Aeronautics and Astronautics
Shanghai Jiao Tong University
Shanghai, China
Email: sqhu@sjtu.edu.cn

***Abstract*—Person re-identification (re-ID) which aims to recognize a pedestrian observed by non-overlapping cameras is a challenging task due to high variance between images from different viewpoints. Although remarkable progresses on research of re-ID had been obtained via leveraging the merits of deep learning framework through sufficient quantity training on a large amount of well labeled data, whereas, in real scenarios, re-ID generally suffers from lacking of well labeled training data. In this paper, we propose an attention-based model with attribute classification (AMAC) to facilitate a well trained model transferring across different data domains, which further enables an efficient cross-domain video-based person re-ID. Specifically, an attention-based sub-network is proposed for deep insight into the quality variations of local parts, hence, different local parts are cooperated with different weights to avoid the heavy occlusions or the cluttered background in datasets. Moreover, we introduce a transferred attribute classification sub-network to extract attribute-semantic features of any new target datasets without the requirement for new training attribute labels which are costly to annotate. Attribute-semantic features can be considered as valuable complementary information for person re-identification since they are robust to illumination varieties and different viewpoints across cameras. Due to the large gap between different datasets, we finetune each sub-network with pseudo labels on the target datasets respectively to strengthen the original model trained on other labeled datasets. Extensive comparable evaluations demonstrate the superiority of our AMAC in solving cross-domain person re-ID task on two benchmarks including PRID-2011 and iLIDS-VID.**

I. Introduction

Person re-identification has been attached great importance by researchers due to its broad applications and significant research sense. The accuracy of person re-ID has been significantly improved via borrowing the advantages of convolutional neural network (CNN) [1], [2], [3], [4], [5], however, they depend on a large amount of training samples which are costly and sometimes unpractical to obtain.

To address this issue, more researchers begin to exploit cross-domain person re-ID methods which transfer the knowledge from labeled source domain to the unlabeled target domain. Existing works follows three main directions. **(1) GAN-based re-ID:** PTGAN [6] and SPGAN [7] use CycleGAN [8] to translate the source images to the target camera style and then train a model on the translated images. However, due to the large gap between translated images and real images, style transfer methods merely care about the marginal distribution alignment while significantly ignoring the conditional distribution approximation, therefore, the performance is still unsatisfactory. **(2) Clustering-based re-ID:** Clustering-based methods also contribute significantly to unsupervised person re-ID by predicting pseudo labels of unlabeled target data. [9], [10] propose an iterative training scheme to finetune the original model trained on the source dataset. They apply the k-means clustering on the feature representations of the target dataset to assign pseudo labels. Despite it lifts the accuracy in solving person re-ID, the rationale of its pseudo label-based classification strategy still requires discussed. **(3) Auxiliary information-based re-ID:** Beyond the above two directions, some approaches focus on the auxiliary information as an assistant to help predicting the identity labels. [11] combines a visual classifier trained from the source dataset and the pedestrians spatio-temporal patterns in the target domain. [12] simultaneously learns an attribute-semantic and identity-discriminative feature representation space. It is worth noting that, the training process in [12] requires quite amount attribute labels of pedestrians, which fails to replicate on new datasets due to the insufficient human labeling.

Besides, most existing research merely leverages knowledge from still images in research of cross-domain person re-ID, whereas, which neglects the spatial-temporal information between image sequences. However, video-based person re-ID is more practical since our surveillance system always capture videos of pedestrians which contain more information than still images. Many supervised video-based person re-ID methods aggregate the information from a sequence of images to represent a pedestrian and achieve high accuracies. An intuitive way is to assume that all the training images are equally important for aggregating an identity-discriminal feature of a video [13], but the re-identification process will be misled by some bad samples containing undesired noise. To handle this challenge, [14] has proposed a quality aware network which aggregates all the images according to their quality. However, this network predicts the quality scores of all samples from the integral view and ignores which part causes noise. Thus, some recent works design a sub-network to predict quality scores for regional parts of images.

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9413309

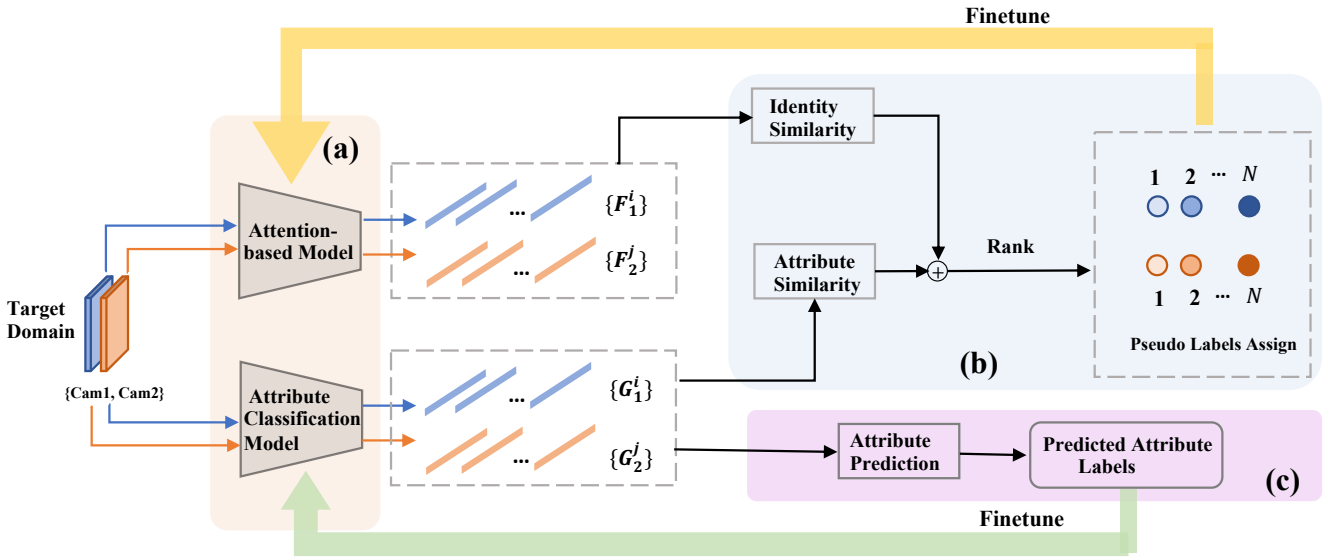

Fig. 1. An overview of our proposed framework. We use blue and orange lines to distinguish samples from Camera1 and Camera2 of the target dataset, respectively. (a) illustrates the two transferred models trained on the source re-ID dataset and the attribute dataset. F_1^i, F_2^i are identity-discriminative features of every pedestrians videos capturedy by the two cameras and G_1^i, G_2^i are attribute-semantic features. (b) and (c) are pseudo labels generation parts: (b) is responsible for identity labels according to both identity similarity and attribute similarity while (c) predicts the attribute labels merely depends on the attribute features extracted by the attribute classification model. After pseudo labels generation, we finetune the two models in part (a) to optimize their performance on the target dataset.

For instance, [15] proposes a region-based quality estimation network consisting of regional feature generation part and region-based quality prediction part. [16] proposes a spatial-temporal attention-aware learning method in which a joint spatial-temporal attention model is used for evaluating the quality scores of discriminative salient body parts. All these mentioned works have achieved remarkable results through effective aggregation method to fully exploit the information in pedestrians videos, which proves that solving person re-identification task using video-based setting is much more valuable than image-based setting.

Considering the observation above, this paper proposes an attention-based model with attribute classification for cross-domain person re-identification. The whole network consists two main parts: **an attention-based model** jointly learns local features of images and their corresponding quality scores, which supposes to obtain a more robust feature representation of video clips; **an attribute classification model** learns attribute-semantic features as further discriminative information to restrain the search range in the re-ID task. We pretrain the attention-based model and attribute classification model on the labeled source dataset and PETA attribute dataset [17] respectively, then use the pretrained models to extract identity features and attribute features of images in the target dataset. After that, we label the videos of the target dataset from Camera1 in sequence and compare the similarities between videos from Camera1 and Camera2. According to the similarities, we assign the videos from Camera2 the same label with the most similar videos from Camera1. Using the pseudo labels, we finetune the two pretrained models on the target dataset and repeat the label assigning and finetuning to obtain our final results.

To sum up, the main contributions of this paper are summarized as follows.

- So far as we know, we are the first to apply the attention model to cross-domain person re-ID problem. Additionally, in Table.IV and Table.V, we highlight its contributions in reducing the effects of heavy occlusions or the cluttered background especially in cross-domain experiment settings.
- We propose a transfer learning method to predict the attributes of the target dataset as a complementary information for identity classification. Thanks to our specifically deigned transfer learning framework, **AMAC** enjoys the merits of attribute-semantic features while ignoring the tedious manual annotation process.
- We conduct extensive experiments on the iLIDS-VID and PRID-2011 datasets to show that our method achieves competitive results in solving the person re-ID task.

II. RELATED WORK

A. Attention Model

The concept of attention model imitates the human perception scheme in which we tend to concentrate on discriminative local parts. We can divide the attention learning into hard attention and soft attention. The former one aims to decide the coarse position of discriminative regions while the latter one which can be further divided into soft spatial attention and channel attention is dedicated to train attention scores for every pixel. [4] integrates a temporal attention model to select

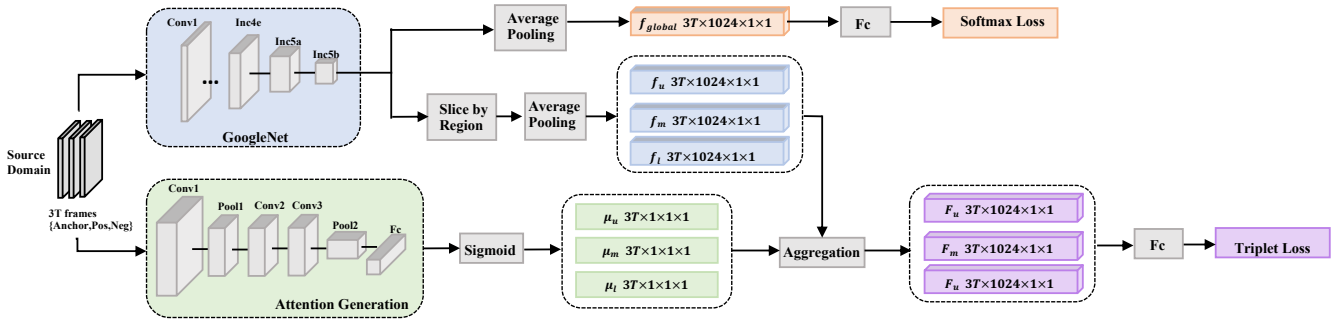

Fig. 2. The detailed architecture of attention-based model. The input of this sub-network is a triple of 3T frames from the source re-ID dataset. Each frame generates middle representation through GoogleNet before being sliced by region. Then the local features f_u, f_m, f_l will be aggregated according to the quality scores μ_u, μ_m, μ_l predicted by the attention generation part to form the final representation F_u, F_m, F_l.

discriminative frames and a spatial recurrent model to exploit the contextual information when measuring the similarity. [18] designs a triplet recurrent neural network to dynamically generate attention scores for person re-ID. And [19] designs a harmonious attention module to joint learn hard attention and soft attention in an end-to-end fashion. Different from the above works, our method learns an attention map to evaluate the qualities of three local parts of pedestrians images.

B. Attribute Learning

Attribute learning has received much attention in person re-ID due to its high reputation in identifying pedestrians. [20] proposes a semi-supervised attribute learning framework and boost the accuracy of attribute classification with triplet loss. [21] provides the attribute labels of the Market-1501 and DukeMTMC-reID datasets and prove that predicting multiple attribute labels along with the identity labels improves the re-ID performance. However, it is costly and sometimes impossible to manually label all the attributes of pedestrians, which shrinks the application of attribute learning. With the development of transfer learning methods, this paper proposes to transfer a CNN model pretrained on the already existing attribute dataset to the target dataset. Although the transferred model cannot obtain accurate results, the attribute-semantic features it extracts can be considered as a complement of the main identity-discriminative features, thereby improving the accuracy of person re-ID.

III. METHODOLOGY

Fig.1 illustrates an overview of our network, where the input target video clips firstly go through the attention-based model pretrained on the source dataset and the attribute classification model pretrained on the PETA dataset simultaneously. Then we assign the pseudo labels to the target dataset according to feature similarities between all the images from two cameras. Finally, we finetune the two pretrained models according to the pseudo labels.

A. Attention-based Module

Due to occlusions and cluttered background, the qualities of different regions and frames of pedestrian videos vary dramatically. The function of attention-based model is to divide the videos into several spatial-temporal units and predict the quality scores of these units. We aggregate the features of all the units in a video with their corresponding scores to obtain the final representation of this video.

Given a pedestrian video $V = \{I_t\}_{\{t=1:T\}}$, where T is the number of frames and I^t denotes the t-th frame. In our experiments, we set $T = 8$. As Fig.2 shows, when training this module on the source dataset, we input a triplet $V = \{V^a(I_t), V^+(I_t), V^-(I_t)\}$ into two parts: regional feature generation part and attention estimation part. We define $V^a(I_t)$ as anchor set, $V^+(I_t)$ as positive set and $V^-(I_t)$ as negative set. In regional feature generation part, we use GoogleNet with batch normalization [22] to obtain the middle representation of all the input images and slice every frame into three parts according to the local ratio 3:2:2 in height direction. This ratio setting follows as [15] which detects the key points of human body for confirming a proper ratio to divide the middle representation. We denote $f^V = \{f^a(I_t), f^+(I_t), f^-(I_t)\}$ as the middle representation of the triplet, where $f^a(I_t) = \{f^a_u(I_t), f^a_m(I_t), f^a_l(I_t)\}$ and so as $f^+(I_t)$ and $f^-(I_t)$.

The attention estimation part is a simply convolutional network. Fig. 2 shows the detailed architecture which consists a 7×7 convolution layer, two 3×3 convolution layers, a 7×7 max pooling layer, a fully connected layer and a sigmoid layer. The final output is a $T \times 3$ score map for each image set in the triple and the scores are scaled to [0,1] by the sigmoid layer. Take the anchor set as an example, we denote $\mu^a = \{\mu^a_u(I_t), \mu^a_m(I_t), \mu^a_l(I_t)\}$ as the attention map, then the final representation of the anchor set can be denoted as $F^a = \{F^a_u, F^a_m, F^a_l\}$. We generate the final representation of each local part by the formulations below:

$$F^a_{part} = \sum_{t=1}^{T} \mu^a_{part}(I_t) f^a_{part}(I_t) \tag{1}$$

where $\mu^a_{part}(I_t)$ and $f^a_{part}(I_t)$ represent different parts (upper, middle, lower) scores and features, respectively.

For the training stage of the attention-based sub-network, the overall loss consists the softmax loss $L_{softmax}$ and the

triplet loss L_t. We use the global feature f_{global} obtained by the output of $pool5$ layer to compute the softmax loss, which reflects the identity classification accuracy for every frame in a video. And the triplet loss can be formulated as below:

$$L_t = max\left\{0, d\left(F^a, F^+\right) - d\left(F^a, F^-\right)\right\} \tag{2}$$

where the $d(\cdot)$ is the L_2-norm distances.

To reduce the triplest loss, we can make the samples from the same class compact while the samples from different classes far away. Then the total loss is:

$$L = L_{softmax} + L_t \tag{3}$$

B. Attribute Classification Module

Attribute learning provides useful information to help identify whether two images belong to the same person. Since there are no available video-based datasets containing attribute labels, our proposed method pretrained a simple CNN model on the PETA attribute dataset and then use this model to extract attribute-semantic features of the target dataset. PETA dataset is organized by 10 publicly available small-scale datasets, covering more than 60 attributes on 19000 images of different pedestrians. Fig.3 illustrates the architecture of this model and we choose ResNet-50 as the backbone network. We divide

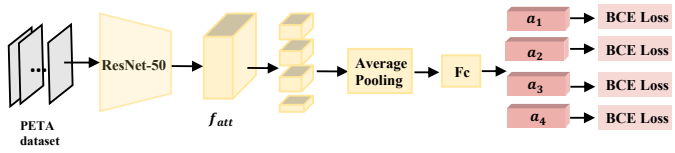

Fig. 3. The overall architecture of attribute classification model

the output f_{att} of $Conv5x$ block in ResNet-50 into four parts according to the ratio 2:2:2:1 as $f_{att}^1, f_{att}^2, f_{att}^3, f_{att}^4$, which correspond to the head, the upper body, the lower body and the foot. For the attributes in the PETA dataset, we select 38 attributes which are also divided into 4 groups according to different locations and the attribute labels of each group are predicted by the attribute feature $a_i, (i = 1, 2, 3, 4)$ which are acquired from f_{att}^i by applying a global average pooling layer and a fully connected layer. Tab. 1 lists the details of attribute groups.

TABLE I
THE PARTITIONED ATTRIBUTE GROUPS

Group Name	Attribute Names
Head	Age, Gender, Haircolor
UpperBody	Color, Casual, Jacket, Logo, LongSleeve, ShortSleeve, NoSleeve, Plaid, ThinStripes, Suit, Sweater, ThickStripes,Tshirt,Other,VNeck
LowerBody	Color, Casual, Jean, Logo, Shorts, LongSkirt, ShortSkirt Suits, Trousers, Capri, HotPants, Plaid, ThinStripes
Foot	Boots, LeatherShoes, Sandals, Sneaker, Stocking, Luggagecase, BabyBuggy

We use Binary Cross-Entropy (BCE) loss to train this attribute classification model and the attribute loss in the i-th group is:

$$L_i = -\sum_{k}^{M_i}[l_k^i log p_k + (1 - l_k^i) log(1 - p_k)] \tag{4}$$

where M_i is the number of attributes in each group, l_k^i is the binary-value label of the k-th attribute in the i-th group and p_k is the probability of having this attribute predicted by the model.

Then the total loss all the groups can be formulated as:

$$L_{att} = \sum_{i=1}^{4} L_i \tag{5}$$

C. Pseudo Labels Assign and Finetuning

After obtaining the two pretrained models above, the next step is to assign pseudo labels to the target dataset according to the similarities between the features extracted by the pretrained models.

Although there are no identity labels in the target dataset, we can annotate all the videos captured by one camera sequentially from 1 to N, where N is the number of pedestrians in the dataset. For the videos from another camera, we compare the feature similarities of them to all the labeled videos from the first camera and find the most similar one. We denote the aggregation features of videos from the two camera extracted by the attention-based model as $\left\{F_1^i\right\}$, $\left\{F_2^i\right\}$ and the labels of videos are $\left\{y_1^i\right\}$, $\left\{y_2^i\right\}$ $(i = 1, 2, ..., N)$ respectively. Notice that $y_1^i = i$. Also, we apply the pretrained attribute classification model to the target dataset to obtain the attribute features $\left\{G_1^i\right\}$ and $\left\{G_2^i\right\}$. Then we calculate the similarity of the identity feature and attribute feature between the i-th person from Cam1 and the j-th person from Cam2:

$$S(i, j) = S_1(F_1^i, F_2^j) + \lambda S_2(G_1^i, G_2^j) \tag{6}$$

where S_1, S_2 can be determined simply by the dot product of the two feature vectors and λ is a parameter that balances the two losses. The determination of λ will be discussed in Section IV.

As mentioned before, we rank the similarities between the j-th pedestrian from Cam2 and all the pedestrians from Cam1 and decide the pseudo label of this pedestrians as below:

$$y_2^j = \arg\max_i S(i, j) \quad i = 1, 2, ..., N \tag{7}$$

For the attribute labels, we use the pretrained attribute classification model to extract the predicted probabilities of a certain attribute for all the frames of a person. If the averaged predicted probability of the m-th attribute for the i-th person is beyond 0.5, we set its pseudo label $l_k^i = 1$, else $l_k^i = 0$.

Finally, we can finetune the attention-based identity classfication model and the attribute classification model with pseudo identity labels and attribute labels, respectively.

IV. EXPERIMENTAL RESULT

A. Datasets

We evaluate our proposed method on two video-based person re-ID datasets: iLID-VID [23] and PRID-2011 [24].

1) iLID-VID: contains 600 image sequences of 300 identities. Each image sequence has a length of 23 to 192 frames captured by two non-overlapping cameras in an airport terminal. Due to its cluttered background and extremely heavy occlusion, this dataset is very challenging.

2) PRID-2011: includes 200 pedestrians and each has 2 image sequences under two different cameras. The length of each image sequence varies from 5 to 675 frames, with an average of 100. In our experiments, only sequences with more than 27 frames will be used. Since the images of this dataset are captured in an open area whose background is rather clean and has little occlusion, it is less challenging.

B. Implement Details

1) Evaluation Metrics: To conduct our experiments, we split each dataset into equal-sized training and testing sets and repeats experiments 10 times to calculate the average accuracy. We use the Rank-1, Rank-5, Rank-10, Rank-20 scores of the Cumulative Matching Characteristic (CMC) curve to evaluate the performance. The Rank scores denote the probability whether one or more correctly matched image sequences appear in top-1, top-5, top-10, top-20 respectively.

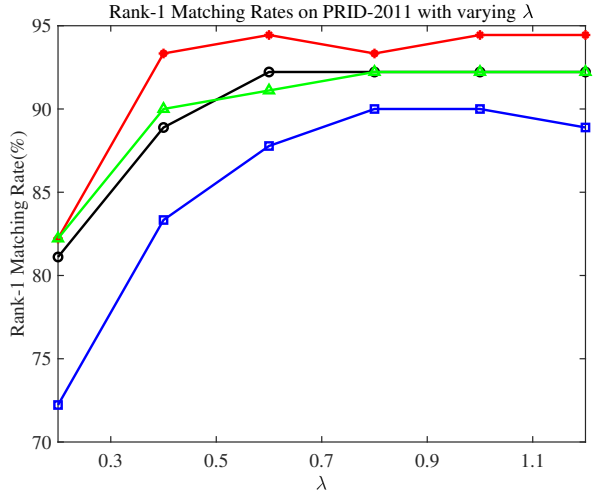

Fig. 4. Rank-1 matching rates on the PRID-2011 dataset with varying λ. We set $\lambda = 0.2, 0.4, 0.6, 0.8, 1.0, 1.2$.

2) Parameter Validation: In this section, we evaluate the influence of λ in equation (6) on CMC curves. λ is an important parameter which balances the contribution of the identity classification loss and the attribute classification loss. We test our model on four groups of training/testing data from the PRID-2011 dataset under different values of λ and the Rank-1 scores of CMC curves are illustrated in Fig. 4. The values of λ ranges from 0.2 to 1.2 and the step size is 0.2. Notice that we directly apply the model trained on the source re-ID dataset and the attribute dataset to extract features of the target dataset and test the re-ID accuracy during the parameter validation.

From the result we obtain that the optimal value is $\lambda = 0.8$ on the PRID-2011 dataset. Other values like $\lambda = 0.7$ or $\lambda = 0.9$ can achieve similar performance. We use $\lambda = 0.8$ for the four datasets in our experiments.

C. Comparison with the State-of-the-art Methods

We compare our **AMAC** with the state-of-the-arts methods and report the rank-1, rank-5, rank-10, rank-20 accuracy in TABLE II and TABLE III, respectively. Since our work is the first attempt we know to perform cross-domain methods on the video-based re-ID datasets, we compare our method with the existing supervised video-based methods. Although there are no labels in the target dataset, our method achieve comparable results, which shows its effectiveness and flexibility when applied to any new target datasets.

TABLE II
COMPARISON RESULTS OF DIFFERENT METHODS ON THE PRID-2011 DATASET

Source → Target	**iLIDS-VID → PRID-2011**			
Methods	**Rank-1**	**Rank-5**	**Rank-10**	**Rank-20**
DVDL[25]	40.6	69.7	77.8	85.6
SDALF[26]	31.6	58.0	-	85.3
DVR[23]	48.3	74.9	87.3	94.4
RFA-Net[27]	68.3	81.1	-	96.8
WSTF[28]	70.3	86.9	-	96.5
CNN+XQDA[2]	72.2	93.1	96.7	99.1
QAN[14]	90.3	98.2	99.3	100
RQEN[15]	91.8	98.4	99.3	99.8
AMAC	**94.7**	**99.3**	**99.8**	**100**

TABLE III
COMPARISON RESULTS OF DIFFERENT METHODS ON THE iLIDS-VID DATASET

Source → Target	**PRID2011 → iLIDS-VID**			
Methods	**Rank-1**	**Rank-5**	**Rank-10**	**Rank-20**
DVDL[25]	25.9	48.2	57.3	68.9
SDALF[26]	26.7	49.3	-	71.6
DVR[23]	41.3	63.5	72.7	83.1
RFA-Net[27]	39.6	65.8	-	85.3
WSTF[28]	41.5	70.5	-	87.4
CNN+XQDA[2]	54.1	80.7	90.0	95.4
QAN[14]	68.0	86.6	95.4	97.4
RQEN[15]	77.1	93.2	97.7	98.8
AMAC	**47.7**	**69.0**	**78.7**	**90.3**

TABLE II and TABLE III compare the results of our cross-domain method with other methods under supervised setting on the target dataset including DVDL[25], SDALF[26], DVR[23], RFA-Net[27], WSTF[28], CNN+XQDA[2], QAN[14] and RQEN[15]. The first three methods use the hand-crafted features while the latter five methods apply deep learning methods. From the results we can see that our proposed method achieved the Rank-1 accuracy of 94.7% on the PRID-2011 dataset, which is superior to other methods listed in II. On the iLIDS-VID dataset, the Rank-1 accuracy of the proposed method is 47.7%. Since this dataset is

more challenging due to its heavy occlusion, the accuracy of our method is inferior to CNN+XQDA[2], QAN[14] and RQEN[15]. However, although the target dataset is unlabeled in our training stage, our method is still comparable to most of the state-of-arts.

D. Evaluation of Different Branches

In order to evaluate the impact of all branches of our proposed framework, we conduct extensive experiments under four settings to show their effectiveness. TABLE IV and TABLE V list the results of all the settings. We denote B1 (GoogleNet) as the baseline and diretly apply the model pretrained on the source dataset to the target dataset. The second setting consists of GoogleNet and attention-based sub-network and the third setting adds the attribute classification sub-network to the second setting. The last setting is the proposed whole network which generates the pseudo labels and finetune the pretrained models.

In TABLE IV and TABLE V, B1 baseline yields only 31.3% and 9.3% Rank-1 accuracy on the PRID-2011 dataset and the iLIDS-VID dataset, respectively. When attention scores are added to aggregate features, B2 obtains the better results compared with B1 on both two datasets. For the PRID-2011 dataset, B3 achieves great improvements of 32.2% and 58.9 % in Rank-1 accuracy compared with B2 and B1 with the help of attribute classificaiton. Such improvements show that the attribute classification provides effective information for person re-identification. Finally, we can see that our framework gets a 4.5%, 36.7%, 63.4% improvement compared to B3, B2, B1 on PRID-2011 and 3.0%, 16.1%, 37.6% on iLIDS-VID, which verifies the effectiveness of all the sub-networks in our poposed method.

TABLE IV
ABLATION STUDY ON EACH MODULE OF OUR PROPOSED METHOD ON PRID-2011

Source → Target	**iLIDS-VID → PRID-2011**			
Settings	**Rank-1**	**Rank-5**	**Rank-10**	**Rank-20**
B1[14]	31.3	58.1	73.6	89.8
B2 (B1+Attention)	58.0	87.2	94.3	97.8
B3 (B2+Attribute	90.2	98.2	99.8	99.8
AMAC (B3+Finetune)	**94.7**	**99.3**	**99.8**	**100**

TABLE V
ABLATION STUDY ON EACH MODULE OF OUR PROPOSED METHOD ON iLIDS-VID DATASET

Source → Target	**PRID-2011 → iLIDS-VID**			
Settings	**Rank-1**	**Rank-5**	**Rank-10**	**Rank-20**
B1[14]	9.3	24.0	30.0	40.7
B2 (B1+Attention)	23.6	44.5	56.5	68.5
B3 (B2+Attribute	44.7	67.7	77.6	86.6
AMAC (B3+Finetune)	**47.7**	**79.0**	**78.7**	**89.3**

V. CONCLUSION

In this work, we develop an attention-based model with attribute classification (**AMAC**) for joint learning identity-discriminative features and attribute-semantic features under an unsupersived setting in order to alleviate the limitation of existing methods in real-word large-scale person re-identification. In contrast to most cross-domain re-ID methods considering all the images as independent, our method exploits the relationship between all the frames belonging to the same person and aggregate them according to their qualities to obtain a more representative feature of the person. Furthermore, a transferred attribute classification model is proposed to predict attribute labels for the target dataset without labeled training data. Extensive experimental results demonstrate that our proposed framework achieves very competitive re-ID accuracy to the state-of-the-art approaches.

Our future work intends to extend the proposed **AMAC** to adapt the transferred model in jointly optimizing the distribution divergence among the source and the target domains for better performance.

REFERENCES

[1] L. Zheng, L. Shen, L. Tian, S. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *IEEE International Conference on Computer Vision*, 2015.

[2] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," pp. 868–884, 2016.

[3] F. Wang, W. Zuo, L. Liang, D. Zhang, and Z. Lei, "Joint learning of single-image and cross-image representations for person re-identification," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[4] Z. Zhen, H. Yan, W. Wei, W. Liang, and T. Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[5] Zheng, Zhedong, Liang, Yang, and Yi, "A discriminatively learned cnn embedding for person reidentification," *ACM transactions on multimedia computing communications and applications*, 2018.

[6] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," pp. 79–88, 2018.

[7] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[8] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," pp. 2242–2251, 2017.

[9] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, and X. Wang, "Unsupervised domain adaptive re-identification: Theory and practice.," *arXiv: Computer Vision and Pattern Recognition*, 2018.

[10] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, no. 4, p. 83, 2018.

[11] J. Lv, W. Chen, Q. Li, and C. Yang, "Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns," pp. 7948–7956, 2018.

[12] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[13] S. Karanam, L. Yang, and R. J. Radke, "Sparse re-id: Block sparsity for person re-identification," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015.

[14] Y. Liu, J. Yan, and W. Ouyang, "Quality aware network for set to set recognition," 2017.

[15] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai, "Region-based quality estimation network for large-scale person re-identification," 2017.

[16] G. Chen, J. Lu, M. Yang, and J. Zhou, "Spatial-temporal attention-aware learning for video-based person re-identification," *IEEE Transactions on Image Processing*, pp. 1–1, 2019.
[17] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," pp. 789–792, 2014.
[18] 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*, 2017.
[19] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
[20] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," 2016.
[21] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, vol. 95, pp. 151–161, 2019.
[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015.
[23] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *European Conference on Computer Vision*, 2014.
[24] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Scandinavian Conference on Image Analysis (SCIA)*, 2011.
[25] B. Gao, M. Zeng, S. Xu, F. Sun, and J. Guo, "Person re-identification with discriminatively trained viewpoint invariant orthogonal dictionaries," *Electronics Letters*, vol. 52, no. 23, pp. 1914–1916, 2016.
[26] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc IEEECONFERENCE on Computer Vision Patternrecognition*, 2010.
[27] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang, "Person re-identification via recurrent feature aggregation," 2017.
[28] D. Zhang, R. Chen, Z. Qiu, W. Zhang, and Q. Wang, "Person re-identification with weighted spatial-temporal features," in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018.