Adversarial View Confusion Feature Learning for Person Re-identification

Lei Zhang, Fangyi Liu, David Zhang

***Abstract*—The performances of person re-identification tasks can be seriously degraded because of variations caused by view changes. In recent years, there are many methods focusing on how to solve cross view challenges which can be roughly divided into two categories: 1) learning view-invariant features without the help of view information. 2) combining view-wise features with the guide of view information. However, these methods are neither perfect enough. Methods of the first category are not roust enough for different kinds of view-invariants while methods of the other category can not generalize well in real-world applications. In this paper, we aim to learn view-invariant features with the help of view information. We proposed an end-to-end trainable framework, called View Confusion Feature Learning (VCFL), to learn view-invariant features by getting rid of view specific information. To the best of our knowledge, VCFL is originally proposed to learn view-invariant identity-wise features, and it is a kind of combination of view-generic and view-specific methods. The whole view confusion learning mechanism consists of three parts: 1) adversarial learning between feature extractor and the view classifier; 2) drawing the features with the same ID close to centers; 3) the guidance of SIFT, for seamlessly integration of hand-crafted features and deep features. In order to make the whole confusion mechanism work better, we further propose a VCFL+ model, which improves the fusion process in the feature map level through the thoughts of attention mechanism. Experiments on three benchmark datasets including Market1501, CUHK03, and DukeMTMC prove the superiority of our method over state-of-the-art approaches.**

***Index Terms*—Person re-identification, View variability, view-invariant features.**

I. Introduction

In recent years, as the surveillance cameras become widespread over many situations, and it's time-consuming and hard to manually deal with massive data. Person re-identification (ReID) is thus widely applied in many situations such as long-term multi-camera tracking and forensic search. However, person ReID tasks are still challenging for many reasons. It's rather difficult to re-identify pedestrians using appearance features and analyze their activities across cameras in time and space cues because pedestrian images are from different environments. Inter-similarity under the same camera becomes more significant than intra-similarity under different cameras because of the variability of camera view. Just as is shown in Figure 1, pedestrian images are captured from different camera views and the goal of this work is to solve cross-view problems through the view confusion mechanism.

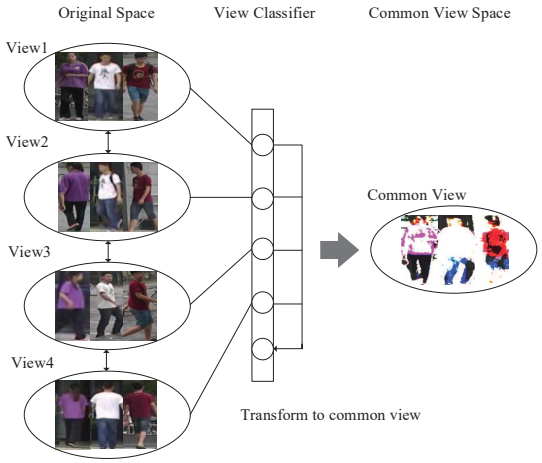


Fig. 1: We take four camera views as example. Dissimilarity is caused by view changes, and our goal is to achieve view confusion. View confusion is achieved by drawing the distance of each view close and expecting the view classifier to identify each view to common view. View confusion mechanism is mainly based on the classifier structure. Note that the showed common view is obtained by the average of the 4 views.

Recent state-of-the-art methods can be divided into two groups: learning robust and discriminative representations [1], [2], [3], [4], [5], or robust similarity match metrics [6], [7], [2], [8], [9], [10] in a supervised manner. Recently, deep learning has gained much attention for learning deep features and metrics in an end-to-end network, and achieves promising results in re-id tasks. Powerful image features should be invariant to variations in illumination, image quality, and especially viewpoint. Many hand-crafted feature types have been used for re-identification, e.g. color, textures, edges and shape, but the discrimination is unsatisfactory. Although deeply learned features have been proved to be powerful in Re-ID tasks, the deep representation is still easily wrapped with view changes. Therefore, a essential part in person ReID is to mitigate the influence caused by view variability. We mainly focus on methods solving cross view changes in deep neural network, which are usually achieved by either designing view-generic models or designing view-specific models with camera view information. View-generic models [11] aim to learn view-invariant features without taking the view information (e.g. pose labels) into account. However, they may still suffer

This work was supported by the National Science Fund of China under Grants (61771079), Chongqing Natural Science Fund (No. cstc2018jcyjAX0250) and Chongqing Youth Talent Program.

Lei Zhang and Fangyi Liu are with the School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China (E-mail: leizhang@cqu.edu.cn, fangyiliu@cqu.edu.cn).

David Zhang is with School of Science and Engineering, Chinese University of Hong Kong (Shenzhen), Shenzhen, China. (E-mail: csdzhang@comp.polyu.edu.hk).

from feature distortion caused by camera view variations. The reason is that different views have different impacts on feature extraction, and we can not use only one model to extract features invariant to all views. View-specific models [12] aim to utilize camera view information to help cross-view data adaptation and learn view-specific features. However, these features are often limited compared with view-shared features because they are only suitable in specific views, and it clearly can not satisfy the retrieval demands in real-word applications. Considering the drawbacks of aforementioned models, we propose to learn view-invariant features by combining view-generic and view-specific models, such that our method is robust to feature distortion caused by camera view changes.

Motivation. In this work, we argue that the key to learn an effective re-id model lies in learning view-invariant features with the help of view information, and we claim that view information is also beneficial for view-invariant feature learning. The reasons can be listed as follows. (1) The traditional methods which learn view-invariant features are all paying attention to the identity information, ignoring the internal view information of features which may deteriorate the performance. (2) The features that are insensitive to views is recognized to be view-invariant and view-independent. Therefore, it may be impossible to eliminate this influence without the help of view information. Motivated by the above reasons, we propose the concept of view confusion, which is similar to "domain confusion" in some aspect. Actually, the view confusion can be understood as view-agnostic, so that the features of a subject from different views can be view-agnostic. With the effect of view confusion, the performance of person re-identification is greatly improved over others. To achieve view confusion, we consider to implement our method from three aspects: classifier, center clustering and sift guidance, the purpose of each part are described below.

Ideas. View bias is a well known problem in person ReID field, and it can be solved with transfer learning methods when each view is treated as an independent domain. Taking advantages of adversarial thoughts, we propose that features can be confused to be view-invariant via the iterative training between feature extractor and view classifier. Similar to "domain confusion", a view classifier is proposed as the discriminator to eventually enable each view confused. Specifically, we can try to classify each view in the discriminator step, then we learn features that can not be differentiated by the discriminator in the adversarial generator step. In benchmark person ReID datasets, pedestrians with the same identity vary with view information, thus it is reasonable to make features with the same label close to feature centers. In order to make the view information confused, we aim to achieve this by drawing features of pedestrians with different view information to the same pattern in each identity. Center loss [13], which simultaneously learns a center for deep features of each identity and penalizes the distances between the deep features and their corresponding identity centers, can make up for the our purpose. Considering the good interpretability of hand-crafted features (e.g. sift), we propose to take the scale-invariant feature transform (SIFT) [14] into account for guiding the learning of the deep feature network. SIFT has the following advantages: first, it is a local and view-invariant feature descriptor; second, it is suitable for rapid and accurate matching due to its good distinctiveness; the last but not least, its extensibility can be easily combined with other forms of feature vectors. There are many methods based on SIFT feature, which mostly rely on the bag of word (BOW) methods, before deep learning was sprung out. In this paper, we pay more attention to the view-independence of SIFT features, and propose SIFT guided feature learning for better improving the robustness of deep network.

Extension. This article is extended version of our conference work [5]. As is shown in Figure 3, the new contributions can be described as follows. 1) We change the view information. In the pervious conference version, we use the view information predicted by the RAP dataset pre-trained model. Although the view prediction accuracy is not our focus in VCFL, prediction error still exists and it is better to use camera labels as the view information for convenience and better accuracy. The view classifier varies as the number of camera IDs of the dataset change, and it is more applicable in real worlds. 2) We change the implementation details of the view confusion part. As mentioned in [15], the gradient reverse layer (GRL) is used to make the feature exactor to learn features that can not be classified by the view discriminator. 3) The conference work was conducted in feature aspect, which ignores the distribution of feature map level. Since the view distribution has an impact on the feature map, we propose the attention mechanism and enable the distribution of feature maps to be aligned. Specifically, a feature map alignment (FMA) module is proposed, in which a feature map regularization (FMR) loss is proposed. We propose that the view-invariant features can be learnt through the view confusion process on both feature level and feature map level. For convenience, we name our new contribution as VCFL+, which achieves a new state of art in person re-identification. Note that, different from VCFL, in VCFL+, the proposed feature map alignment (FMA) with two-stream structure aims to alleviate the conflict between identity classifier and view classifier, such that both identity discrimination and view independence of features can be guaranteed.

In summary, the contributions are as follows.

- We propose a VCFL approach for learning view-invariant features by using the view confusion learning mechanism. It is different from other cross-camera methods that it learns view-invariant features by taking view information into account. Domain adaptation idea is combined with person re-id methods to solve cross-view problem, which can provide new solution for solving this problem.
- In VCFL, we integrate the SIFT guidance strategy for further improving the view independence of the deep features. The proposed SIFT guidance is also an important step to explore the interpretability of deep network in feature representation by designing the network structure following the hand-crafted descriptor.
- We extensively propose a VCFL+ model, in which an attentive feature map alignment with a feature map regularization loss is presented, such that the feature learning robustness is improved by integrating the identity

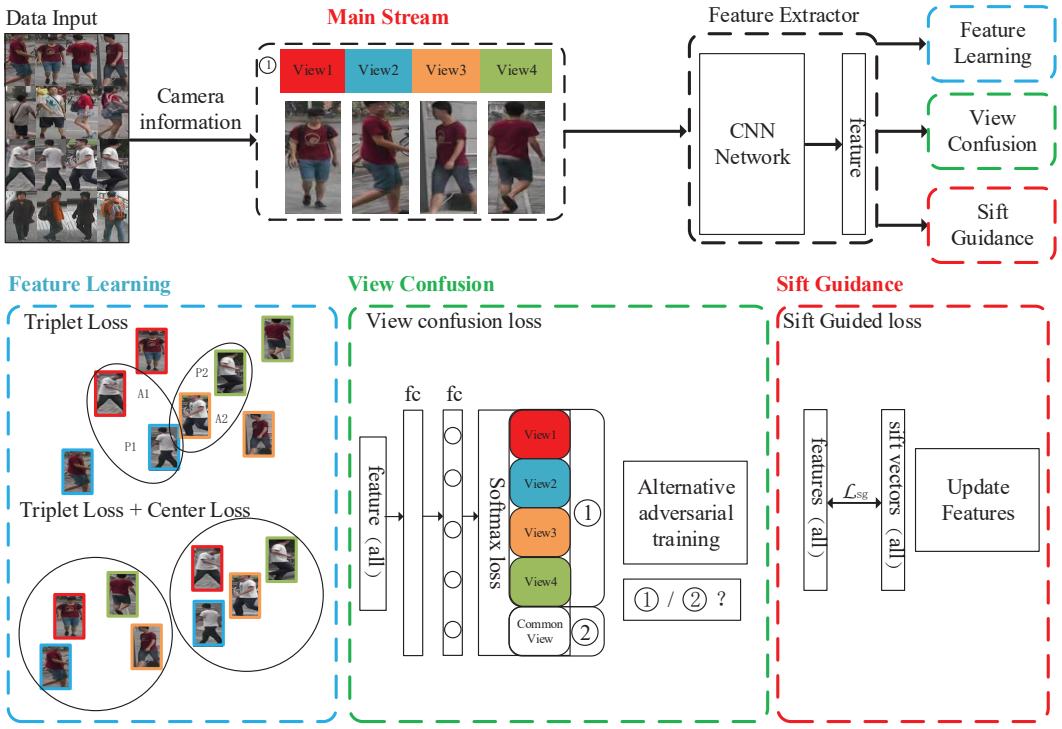


Fig. 2: Illustration of the proposed VCFL. We use the camera ID as the view information and we take four views as example. Our goal is to find the view-agnostic features from multiple views. The model consists of View Confusion, Feature Learning and Sift Guidance. The View Confusion is achieved by adversarial learning between the feature learning and camera ID guided view classifier, which ensures that the feature distributions over different views are made view indistinguishable, i.e. view-agnostic features. The Feature Learning can be supervised by triplet loss, center loss, and identity classification loss. The Sift guidance is a smooth combination of deep and hand-crafted features, aiming to improve the interpretability and view-invariance.

discrimination and view-independence of features.

- Extensive experiments and performance analysis are conducted on a number of benchmark datasets to show that our models outperform state-of-the-art deep re-id models.

II. Related Work

A. Person Re-identification

Person re-id is a task to find the same person across cameras. The challenge is that images of the same person under different cameras may be dissimilar to the images of different persons under the same camera. Recently, most of methods treat re-id as a ranking problem, which means that the distance between the images of the same identity should be smaller than that of different identities. There are two main challenges: (1) Learning discriminative features which are robust to illumination, poses, view variation, etc. (2) Learning similarity metrics which are used to predict whether two images describe the same person.

However, in order to address the challenges, existing methods try to learn robust features in different ways. In order to solve the problem of pose changes and various human spatial distributions in the person bounding box, Zhao *et al.* [16] proposed a simple yet effective human part-aligned representation for handling the body part misalignment problem. Zhang *et al.* [17] proposed a novel method called Aligned Re-ID that extracts global feature jointly with local features. In order to take advantages of body structure, Zhao *et al.* [18] proposed Spindle Net, in which the human body region guided multi-stage feature decomposition and tree-structured competitive feature fusion are used to facilitate feature learning. Li [19] design a Multi-Scale Context-Aware Network (MSCAN) to learn powerful features over full body and body parts, which uses attention methods to learn meaningful body parts rather than uses predefined parts.

For learning similarity metrics, most of previous methods propose to solve person re-id problems as ranking problems. Hermans *et al.* [7] proposed to use a variant of the triplet loss to perform end-to-end deep metric learning, providing guidance for triplet loss training. Chen *et al.* [20] designed a quadruplet loss, which can lead to the model output with a larger inter-class variation and a smaller intra-class variation compared to the triplet loss. With the improvements of triplet loss, many existing end-to-end frameworks can obtain good performance gain.

B. Cross-View Methods

Cross-view problems widely exist in the field of computer vision and they can be solved from different aspects. For example, the coupled projection learning methods proposed by Ben et al. [21], [22], [23] have provided a significant insight into the cross-view gait recognition and the CPL models

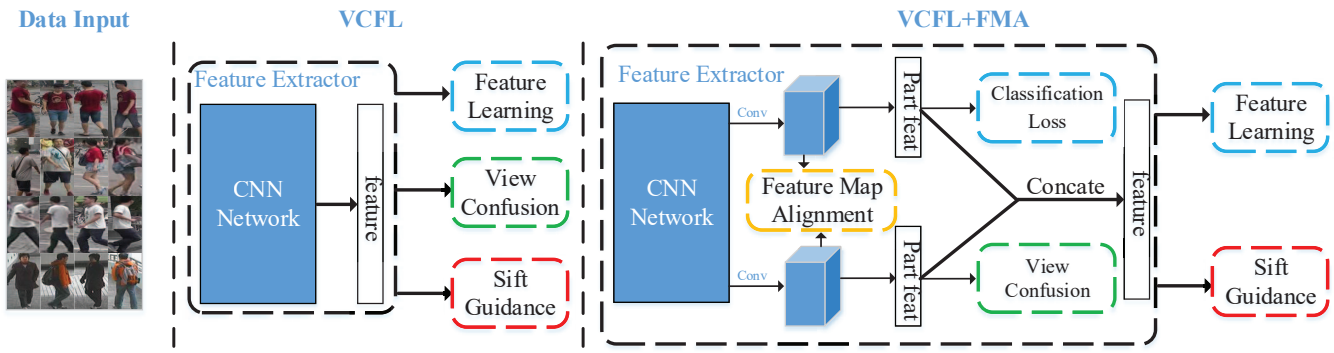


Fig. 3: Illustration of our proposed VCFL and VCFL+ models. Comparing to VCFL, the plus version introduces the feature map alignment (FMA) module with attention mechanism and the feature map regularization (FMR) term designed by two branches. Additionally, we introduce the cross-entropy classification loss in feature learning part to further improve the performance.

have achieved good performances in cross-view scenarios. In general human recognition problems, learning robust features for human appearance has been studied by exploiting different strategies such as [24], [25], [26]. It is also vital to learn discriminative features which are robust to view variation for person re-id problems. Most of methods for solving cross view challenge can be roughly divided into view-generic methods and view-specific methods. For example, Yu *et al.* [11] ignored the view information and tried to find a shared space where view-specific bias is alleviated. Feng *et al.* [12] proposed a deep neural network-based framework, which, in the feature extraction stage, utilizes view information to learn a view-specific network for each camera view. A cross-view Euclidean constraint (CV-EC) and a cross-view center loss (CV-CL) were designed. Similar to our method, Wu *et al.* [27] also takes advantage of adversarial learning, which, however, pays attention to the asymmetric mappings followed by an explicit view adaptation through two-class discriminator. Our approach aims at learning features that are robust to view changes. Specifically, our proposed VCFL method is a kind of combination of view-generic and view-specific methods, in which a novel view confusion mechanism is presented for cross-view person re-id application by removing the impact caused by view variation.

C. Sift Based Re-ID Methods

In non-deep learning era, traditional person re-identification methods usually extract low-level features using hand-crafted visual feature descriptors (e.g. Sift, HOG, etc.). Then the visual retrieval community has witnessed the prominence of the bag-of-words (BoW) [28] model over a decade, during which many algorithms were proposed. The SIFT-based methods for image classification mostly rely on the BoVW model [29]. Tao et al. proposed a multi-feature model [30]. In this paper, we take advantage of the view independence of Sift features to find the view invariant regions to guide the learning of deep model. We propose to integrate the handcrafted features with deep features in a smooth way considering the interpretability of Sift descriptor.

D. Domain Adaptation Based Re-ID Methods

Many person re-identification methods attempt to solve cross view problems using domain adaptation methods, because each view can be regarded as an independent domain. Zhong *et al.* [31] solved the cross-view problem by learning a camera-invariant descriptor subspace, which is a kind of camera-style adaptation. Deng *et al.* [32] used domain adaptation methods to achieve image translation while maintaining discriminative cues contained in the ID label. There is no doubt that domain adaptation methods are beneficial to solve distribution difference in person re-id. There are also many transfer learning methods trying to turn other view information into one specific view. However, it is not hard to find that the transformation matrix may not be suitable to all views, Take four views (*front*, *back*, *left*, *right*) for example, the transformation matrix for *back* to *front* may not be suitable for *left* to *front*. The same problem also exists in view-generic models that try to learn view-invariant features through only one model. We assume that there must be some common parts (average images) between images of these four views, and we call the common parts as common view in our work. There is no doubt that it will be more suitable to transform all views into the common view, and the extracted features in common view must be view-invariant. Our approach takes advantage of domain adaptation methods and proposes the concept of "view confusion", by removing the influence of specific views in feature aspect. In this paper, the work [15] inspires us to propose a new way to solve view variance, for which we can learn discriminative features for person re-identification task on the main domain and learn invariant features with respect to the shift between different views. To be specific, "view confusion" aims to confuse each view so that features can be regarded as view-agnostic. There is an assumption that the view and identity are coherent in features, and we aim to separate the view from the identity, which can also be understood as a kind of disentangling process.

E. Attention Mechanism Based Re-ID Methods

Attention mechanism has been widely used in various tasks such as semantic segmentation, object detection, person re-

identification, etc., which mimics human's vision in understanding images. In person re-id task, attention mechanism is usually used to capture human body part and align part features. Xu *et al.* [33] used the pose information to guide the attention mechanism to handle pose variations and background clutters. Chen *et al.* [34] added regularization on the attention network to increase the correlation and diversity of the attentive features. Xia *et al.* [35] used the second order feature statistics to generalize the part-based models. Chen *et al.* [36] designed a high-order attention mechanism by exploiting the complex and high-order statistics information for better enriching the attention knowledge. These attention methods are all creative in designing suitable attention structure or learning adaptive regions. Different from above attention methods, we try to alleviate the differences between feature maps with the same identity but different poses by taking advantages of the attentive thoughts.

Among these existing approaches, our model is different from them in the following four main aspects:

(1) Our goal is different from other cross-view feature learning methods, and we aim to learn features without impact of view information. It is quite novel to learn view-invariant features by considering the internal influence of view information in feature and feature map aspect. The view and identity are coherent, and the proposed model can disentangle the impacts and remove view impact.

(2) Traditional methods using hand-crafted features. We propose to learn deep feature with the guidance of hand-crafted features, which, to some extent, can improve the interpretability of deep models.

(3) Taking each view as an independent domain, transfer learning are integrated to solve cross-view problems via view confusion rather than unsupervised or semi-supervised ReID tasks. The fusion of the parts of confusion mechanism is not only determined by loss weights in feature aspect but also determined by the regularization in feature map.

(4) Previous methods only use attention mechanisms to extract part-based spatial patterns from person images, and we propose to use the simplest implementation of attention as the regularization of feature map.

III. The Proposed VCFL

The proposed model consists of three parts: Feature Learning, View Confusion and Sift Guidance. Further, we assume that the three parts may not promote each other when they work together, thus it is reasonable to put regularization on feature maps when combining the three parts. In this section, we will describe the feature learning network, the view confusion mechanism based on adversarial idea, and the SIFT guided feature loss. First, we describe the feature learning part including the triplet loss and the center loss. Second, we propose the concept of view confusion, which takes advantage of adversarial learning, aiming to learn view-invariant features through the gaming between feature extractor and view classifier. Third, we focus on the novelty of exploring the interpretability of deep learning by designing the network structure following the hand-crafted SIFT descriptor.

A. Feature Learning

Feature leaning has always been an important part in solving person re-id problems, which is beneficial to the feature matching. Person re-identification tasks are similar to image retrieval in some aspects, for which many methods treat re-id tasks as ranking problems. Our goal is to learn a network which maps images of the same id to similar features and map those with different id to different features. Thus, we propose to use triplet loss as [7]. The basic architecture is ResNet [37].

Triplet Loss. Triplet loss is proposed to improve the intra-person similarity and inter-person dissimilarity and it is the main loss of our basic network for similarity learning. According to the hard examples mining strategy in [7], we form the training set into a set of triplets, $\gamma = (I_i, I_j, I_k)$, where (I_i, I_j) is a positive pair of images with the same identity and (I_i, I_k) is a negative pair of images with different identity. Then, the triplet loss can be formulated:

$$\mathcal{L}_{tri} = [d(h(I_i), h(I_j)) - d(h(I_i), h(I_k)) + m]_+, \quad (1)$$

where $(I_i, I_j, I_k) \in \gamma$, m is the margin by which the distance between a negative pair of images is ensured to be larger than that between a positive pair of images, $h(I)$ represents the feature representation for image I, $d(\cdot)$ represents the distance function, and $[\cdot]_+$ means the positive operator.

Center Loss. In order to make the extracted features to be more view-invariant, we try to make features of the same person with different views as similar as possible. The most straightforward way is to use the center loss [13] which forces features to be close to the corresponding feature centers. The center loss intends to learn discriminative features by drawing intra-class distances smaller while inter-class distances larger. In [12], View Information is also considered when applying center loss to re-id to further improve the performance, however, its goal is to make the feature to be view specific but approaching the whole center simultaneously. Our method aims to achieve view confusion in feature aspect which means that the view specific centers should also be close to the whole center. The following center loss can achieve this without introducing any extra computation:

$$\mathcal{L}_{cen} = \frac{1}{2}\sum_{i=1}^{N} \parallel h(I_i) - h(C_{y_i}) \parallel_2, \quad (2)$$

where $h(I)$ represents the visual features, C_{y_i} represents the center (average feature) of identity y, and N is the number of samples. The network parameter θ and the center C_{y_i} can be updated as follows.

$$\begin{aligned} \frac{\partial \mathcal{L}_{cen}}{\partial h(I_i)} &= h(I_i) - h(C_{y_i}) \\ \frac{\partial \mathcal{L}_{cen}}{\partial h(C_{y_i})} &= h(C_{y_i}) - h(I_i) \\ \theta^{t+1} &= \theta^t - \mu \frac{\partial \mathcal{L}_{cen}}{\partial h(I_i)} \frac{\partial h(I_i)}{\partial \theta} \\ C_{y_i}^{t+1} &= C_{y_i}^t - \alpha \frac{\partial \mathcal{L}_{cen}}{\partial h(C_{y_i})}. \end{aligned} \quad (3)$$

where μ and α represent the learning rate for updating the network parameter and the center, respectively. t denotes the

iteration index. Therefore, the loss for feature learning is formulated as follows.

$$\mathcal{L}_f = \mathcal{L}_{tri} + \lambda_{cen}\mathcal{L}_{cen}, \tag{4}$$

where λ_{cen} is the trade-off parameter.

B. View Confusion

The concept of *view confusion* can be also understood as view-agnostic, that is, the features of a subject from different views can be view-agnostic. It means that the impact of view-specific information can be alleviated and obtain view-invariant features. As is shown in Figure 2, a view classifier θ_d and a feature extractor θ_f are included. Motivated by the mechanism of transfer learning, the confusion is achieved through the adversarial learning between the feature exactor and the view classifier. The feature extractor aims to learn powerful features that are robust to view changes, while the view classifier aims to identify the view of the extracted features. We recognize the features to be view-invariant if the view classifier can not distinguish the specific views of the samples. The view classifier is trained by using the camera label of each image and used to perform view classification task. For a particular feature extractor θ_f, we evaluate its view invariance by learning the best view classifier based on this feature extractor. This can be learned by optimizing the following objective:

$$\mathcal{L}_d(x, \theta_f; \theta_d) = -\sum_{i=1}^{N} y_c^i \log p_c^i, \tag{5}$$

where y_c^i and p_c^i represent view (camera) label and the softmax probabilities of the *i*-th image, respectively, and N is the number of person images.

For a particular view classifier θ_d, we can propose to 'maximally confuse' the specific views by computing the cross entropy between the output predicted view labels and a balanced distribution over view labels:

$$\mathcal{L}_{vc}(x, \theta_d; \theta_f) = -\sum_{i=1}^{N} \frac{1}{C} \log p_c^i, \tag{6}$$

where C is the number of cameras. This view confusion loss seeks to learn view invariance by computing a feature extractor for which the best view classifier is poorly performed.

Clearly, the above two losses stand in direct opposition to each other. That is, learning a completely view-invariant feature extractor means the view classifier is poor, and learning an effective view classifier means that the feature extractor is not view invariant. Then, we form the view confusion losses in the adversarial formulation:

$$\min_{\theta_d} \mathcal{L}_d(x, \theta_f; \theta_d) \tag{7}$$

$$\min_{\theta_f} \mathcal{L}_{vc}(x, \theta_d; \theta_f) \tag{8}$$

Optimization. Similar to GAN [38], we can perform iterative updates for the above two objectives given the fixed parameters from the previous iteration. At training time, in order to obtain view-invariant features, we seek the feature exactor θ_f that minimizes the loss of view confusion by making the feature distributions of different views as similar as possible. Simultaneously, we solve the parameters θ_d of the view classifier by minimizing the loss of the view classifier. Based on this idea, we propose to solve the parameters $\hat{\theta}_f$ and $\hat{\theta}_d$ of the feature extractor network and the view classifier in an adversarial manner. In practice, we implement this model with two steps: 1) Train the view classifier using camera labels by seeking the best θ_d in Eq. (7). 2) Fix the parameter θ_d, and train the feature extractor by seeking the best θ_f in Eq. (8). However, it is not stable enough as it is conducted in our conference version. Inspired by [15], we use the gradient reverse layer (GRL) for optimization. In the view classifier, two view classifier losses with different views are included. Different views can be regarded as different domains. Minimization of the view classifier loss (L_d) results in a better domain discrimination, while the view classifier loss (L_{vc}) is minimized when the domains are confused. Stochastic updates for θ_d and θ_f are then defined as:

$$\begin{aligned} \theta_d &\leftarrow \theta_d - \mu \cdot \frac{\partial \mathcal{L}_d}{\partial \theta_d} \\ \theta_f &\leftarrow \theta_f - \mu \cdot \frac{\partial \mathcal{L}_{vc}}{\partial \theta_f} \end{aligned} \tag{9}$$

To be specific, this kind of training is achieved through gradient reverse layer, which is theoretically supported by the formulation listed above. The discriminator tries to correctly classify multiple camera views, and the updating backward gradients are computed. For the purpose that the discriminator can not classify the views, we reverse the computed gradients, which means we propagate the minus gradients backward. It has the same effect by labeling each view with the same probability (i.e., $\frac{1}{C}$), and the overall View Confusion mechanism is outlined in Algorithm 1.

Algorithm 1 View Confusion (VC)

Input: A probe person image x, camera label, identity label
Output: View-invariant features

1: Extract global features $f(x)$ supervised by identity label;
2: Train the view classifier with camera label according to the Eq. (7) and (8);
3: Update parameters of view classifier as in Eq. (9) and compute the reversed gradients;
4: Back-propagate the reversed gradients and update parameters of feature extractor according to Eq. (9) .

C. Sift Guidance

We assume that there exists a kind of view confusion which can be achieved by the adaptive combination of deep features and hand-crafted features. We know that deep learning becomes popular due to its powerful representation, however, the interpretability of deep neural network is always a flaw. Scale invariant feature transformation (SIFT) has been a typical hand-crafted feature descriptor in low-level vision. SIFT features can provide local gradient description, and it is

meaningful to integrate the hand-crafted SIFT features into the network training losses, which may help deep features to be more robust to view changes. Specifically, for each image x_i in the set $\{x^i\}_{i=1}^n$, we extract SIFT features and then turn them into vectors using the BOW model, and we call these vectors as sift-bow vectors. Given the assumption that SIFT features are view-independent, the more the deep features are similar to sift-bow vectors, the more view-independent the deep features are. In other words, we use sift-bow vectors as a supervision to help the feature learning of network, then we propose the sift-guided loss formulated as follows:

$$\mathcal{L}_{sg} = \sum_{i=1}^{n} \parallel f(x_i) - g(x_i) \parallel_2, \tag{10}$$

where $f(x_i)$ and $g(x_i)$ denote deep feature of image x_i and sift-bow vector, respectively, and n is the number of images. We only update $f(x_i)$ in the back-propagation since the sift vectors are pre-computed as supervision. Then, the gradient can be computed as follows:

$$\frac{\partial \mathcal{L}_{sg}}{\partial f(x_i)} = 2 \sum_{i=1}^{n} f(x_i). \tag{11}$$

D. Overall Model

In our VCFL model, three parts are included: Feature Learning, View Confusion, and Sift Guidance. For feature learning, the triplet loss and center loss are included into $\mathcal{L}_f$ to keep the feature discrimination and view-invariance. For View Confusion, the view confusion loss $\mathcal{L}_{vc}$ is achieved by the adversarial learning between the feature extractor and view classifier. Then, by revisiting the Eq. (4), Eq. (7), Eq. (8) and Eq. (10), the whole loss of VCFL can be formulated:

$$\mathcal{L}_{VCFL} = \lambda_f \mathcal{L}_f + \lambda_{vc} \mathcal{L}_{vc} + \lambda_{sg} \mathcal{L}_{sg}, \tag{12}$$

where λ_f, λ_{vc} and λ_{sg} control the weights of losses. Note that the view discriminator loss $\mathcal{L}_d$ in Eq. (7) is co-trained alternatively with the above loss $\mathcal{L}_{VCFL}$.

IV. The Proposed VCFL+

Based on the VCFL model, we further propose a plus version, VCFL+, by introducing a new feature map alignment (FMA) module coupled with a feature map regularization loss and a cross-entropy based identity classification loss. The proposed VCFL+ model with an extra FMA and a classification loss is shown in Figure 3 (right).

A. Feature Map Alignment

When the three parts in VCFL are combined together, the interaction effect of the three parts are unknown. It is possible that the effect counteracts with each other when they are simply conducted on the same feature embedding. In this section, we propose to impose regularization on feature map, such that these losses are enabled to promote each other.

Specifically, we propose to achieve this with the attention thought, which is a kind of regularization on the attention mask of the feature map. We use convolution and sigmoid activation to get the final attention mask of the feature map. We expect that the attention mask obtained from the feature map of different parts can be similar to each other. The attention mask can be regarded as a selective map with the same size of feature map. Constraining the selective map to be same can also have impact on the original feature map without losing feature discrimination. In other words, what we aim to do is just regularizing the most discriminative parts of each part to be as similar as possible. This is because, as mentioned before, the view classifier and identity classifier may affect each other when they are conducted on the same feature simultaneously. Thus, we propose the two-stream structure deployed with different losses, i.e. identity classification loss and view confusion loss. The basic architecture of the proposed FMA module in Figure 3 (right) is specified in Figure 4, in which a feature map regularization (FMR) loss is further formulated to enable the two branches with similar attention map, so that the feature can be posed with both identity discrimination and view independence without conflict between the identity classification loss and the view confusion loss of VCFL. Technically, we use 1*1 convolution to get two feature maps with the same size but half of channels comparing to the original map. We then flatten the sigmoid map to get vectors. The two branches are supervised by the VCFL losses, the FMR loss and the classification loss.

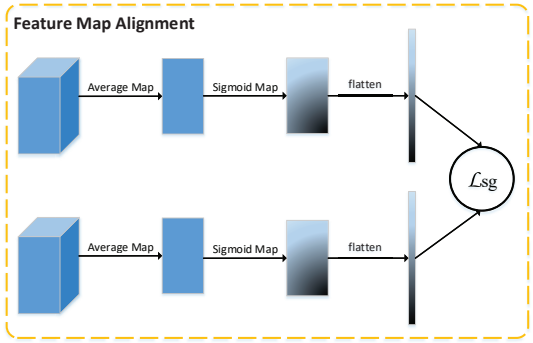


Fig. 4: Illustration of the Feature Map Alignment Part. In order to mitigate the negative effect when combing the three confusion parts together, we impose regularization on feature map level.

Feature Map Regularization. The proposed FMR loss can be formulated as:

$$\mathcal{L}_{fmr} = \sum_{i=1}^{n} \parallel \mathcal{T}(mask1) - \mathcal{T}(mask2) \parallel_2, \tag{13}$$

where $mask1 = \sigma(conv1(M))$, $mask2 = \sigma(conv2(M))$, M represents the feature map, $\mathcal{T}$ is the flatten operator, $conv1(\cdot)$ and $conv2(\cdot)$ represent the convolution operator, σ is the sigmoid function and n is the number of images.

Classification Loss. Many recent works combine metric leaning and classification learning together. In order to learn powerful discriminative features for person appearance representations, we add the identity guided classification loss in VCFL+, which learn to differentiate the different pedestrian identities from different scale features. Specifically, we use the softmax with N output neurons. Given an image, we denote

y as ground-truth ID label and p_i as the ID prediction logits of class i (the softmax outputs). We use label smoothing [39] (soft label) to prevent over-fitting for a classification task. The classification loss with soft label is formulated as follows:

$$\mathcal{L}_{cls} = \sum_{i=1}^{N} -q_i \log p_i, \tag{14}$$

where N denotes the number of pedestrian identities and q_i is the soft label with smoothing function represented as follows.

$$q_i = \begin{cases} 1 - \frac{N-1}{N}\varepsilon, if\ i = y \\ \frac{\varepsilon}{N}, otherwise \end{cases} \tag{15}$$

where ε is a small positive constant within $[0, 1]$.

B. Overall Model

Based on the VCFL model, the plus version is formulated with an extra feature map regularization and an identity guided soft classification loss. Specifically, by revisiting the Eq. (12), Eq. (13) and Eq. (14), the whole loss of the plus version VCFL+ can be formulated as follows:

$$\mathcal{L}_{VCFL_+} = \mathcal{L}_{VCFL} + \lambda_{fmr}\mathcal{L}_{fmr} + \mathcal{L}_{cls}. \tag{16}$$

where λ_{fmr} is the trade-off parameter.

V. Experiments

A. Datasets and Setting

Datasets: Our experiments are conducted on three benchmark datasets including CUHK03 [40], Market1501 [41] and DukeMTMC-reID [42], presented as follows.

- **Market1501** [41] contains 32,668 images of 1,501 labeled persons sampled from six camera views. There are 751 identities with 12936 images in the training set and 750 identities with 19732 images in the testing set, following the standard splits in [41].
- **CUHK03** [40] contains manually labeled 14,096 images and DPM detected 14,097 images of 1,467 identities captured by two camera views. The labeled dataset contains 7,368 training images, 5,328 gallery, and 1,400 query images for testing, while the detected dataset contains 7,365 images for training, 5,332 gallery, and 1,400 query images for testing. We report the results on the detected dataset following two kinds of protocols: 1) The standard splits in [40] based on GoogLeNet. 2) The hard new protocol [39] (CUHK03-NP) based on ResNet: 767 identities are used for training and 700 identities for testing.
- **DukeMTMC-reID** [42] is a subset of the DukeMTMC [43] dataset. It contains 1404 identities captured by eight camera views. Following the Market1501-like evaluation protocol in [42], 702 identities are used as the training set and the remaining 702 identities are used as the testing set. The testing set contains 17,661 gallery images and 2,228 query images.

Evaluation metrics: We adopt the widely-used evaluation protocol [40], [44]. In the matching process, we calculate the similarities between each query and all the gallery images, and then return the ranking list according to the similarities. All the experiments are conducted under the single query setting. The performances are evaluated by using the cumulated matching characteristics (CMC) curves, which is an estimate of the expectation of finding the correct match in the top n matches. We also report the mean average precision (mAP) score [41] over CUHK03, DukeMTMC-reID and Market1501.

B. Implementation Details

In this section, we describe the implementation details of baseline network and the proposed VCFL.

ResNet. We mainly use this net as baseline network, and many experiments are carried out for comparison with state-of-the-arts, ablation study and discussion.

- **Structure.** For baseline net, we use the ResNet-50 architecture and the initial weights are provided by He *etal.* [37]. The final output of 2048 channels is obtained by global average pooling. The initial parameters and the training strategy follow [7].
- **Training.** The ResNet network is implemented on Pytorch. The initial learning rate is 0.0003, which is fixed in the first 151 epochs and then decayed following exponentially decaying training schedule. For the view classifier, the initial learning rate is 0.001. The momentum for gradient update is 0.9 and the updating strategy is the same. The parameters of the feature extractor and view classifier are updated alternatively, which thus increases the training difficulty. The input image size is 256×128.

View Confusion. The implementation of view confusion is the most tricky part in the whole model. In the conference version [5], we use the predicted pose label, and the training method of adversarial feature learning is to alternatively train the feature extractor and view classifier. Different from GAN [38], we do not need to input noise variables because our goal is to generate more discriminative features instead of synthetic images. The training of feature extractor uses the stochastic gradient descent algorithm (SGD) while the training of view classifier uses the adaptive moment estimation algorithm (ADAM). In detail, we first train the feature extractor to get the initial features, which are then feed into the view classifier. Then we fix θ_f and begin to update θ_d. L_{d-} is given the common information while L_{d+} is given the specific view information predicted by view branch. In the experiments of our conference work, we find that the alternating optimization with pose label predicted by pre-trained view prediction model may not work well, so we propose to use the gradient reverse layer (GRL) in [15] with camera ID label for stable training. Specifically, for training the view classifier, the task specific camera ID labels are utilized and the number of views is also dataset dependent. For adversarial training, the gradients are reversed in the back propagation to update the parameters of feature extractor, such that the view classifier can not discriminate the camera views.

Center Loss. In order to make the features of the same identity but different camera view to be more similar, we propose to increase the intra-class compactness by minimizing the center loss. The principle of this part is described in

the former section. However, the center loss dees not usually converge well, and we set the loss weight as 0.0001 to make it balanced with other losses.

Sift Guidance. This part mainly guide the learning of deep features with hand-crafted features (i.e., Sift-BOVW vectors) in the form of Euclidean loss. The main process is the extraction of the Sift-BOVW vectors: 1) We use the sift detector to extract 128 key points with descriptors. 2) We use the K-means clustering algorithm to cluster the descriptors to certain groups, and we set the cluster number as 2048 which is identical with the expected deep feature dimension. 3) Compute the histogram of each image with respect to the cluster centers and get the final BOVW vectors. In order to guide the learning of deep features, we leverage the smooth Euclidean distance norm constraint to bridge the Sift features and deep features. We have to mention that we only update the deep features during training stage since the clustering can not back-propagate the gradients.

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

In order to verify the superiority of our method, we compare with the state-of-the-art methods on several popular ReID datasets. We compare the performance of the proposed VCFL models with SOTA methods and the comparisons on Market-1501 are presented in Table I. The compared methods can be categorized into three groups, i.e., hand-crafted methods, deep learning methods with global feature and deep learning methods with part features. Most of the recent methods take advantage of the part features such as PCB, Aligned-ReID, MGN to improve the performance. It is not appropriate to compare our method with them since we only use the global features. We can see that our method can outperform the most recent methods in MAP and our method mainly aim to provide specific ways for solving cross-view problems.

We conduct experiments on the detected CUHK03 and summarize the comparisons (new training/testing protocol) in Table II. In the compared methods, the MAP of our method can reach state-of-art by using the re-ranking strategy [39]. We have to mention the influence of identity classification loss, because the performance seems to drop much when it is used. It is assumed that the training process seems to be over-fitting, thus dropout and early stopping are applied in training.

We further evaluate our approach on DukeMTMC-reID as is shown in Table III. Similar to Market-1501, the benchmark is challenging because this dataset have more occlusions and complex backgrounds. Similar to market1501, our method are competitive with methods that only use global features since our motivation is to solve cross-view problems.

In this paper, from the reported results in Table I, Table II, Table III, we observe that our models achieve competitive accuracy on the three datasets comparing to other methods that only use global features. All the results are achieved under the single-query model. The re-ranking strategy can further boost the performance, especially the mAP. With re-ranking, our VCFL+ model can reach 90.80%, 70.45%, 83.12% of mAP for Market-1501, CUHK03 and DukeMTMC-reID, respectively. Additionally, we also see that the improved VCFL+ model outperforms the VCFL, which validates the effectiveness of the proposed feature map alignment module in the plus version.

TABLE I: The comparison with other state-of-art methods over Market1501. “tri.” denotes training with triplet loss. “cls.+tri.” denotes training with the combination of both losses.

Method	top1	top5	mAP
BoW+kissme[41]	44.4	63.9	20.8
CCAFA[45]	71.8	-	45.5
Consistent-Aware[46]	73.84	-	47.11
re-ranking[39]	77.11	-	63.63
GAN[42]	78.06	-	56.23
DLPAR[16]	81.0	92.0	63.4
SVDNet[47]	82.3	-	62.1
PAN[48]	82.8	-	63.4
HAP2S-P [49]	84.59	-	69.43
MultiScale[50]	88.9	-	73.1
MLFN [51]	90.00	-	74.30
HA-CNN [52]	91.20	-	75.70
DuATM [53]	91.42	-	76.62
Deep-Person [54]	92.31	-	79.62
Aligned-ReID [17]	92.62	-	82.31
PCB(UP)[55]	92.3	97.2	77.4
PCB(RPP)[55]	93.8	97.5	81.6
SGGNN [56]	92.30	-	82.80
Mancs [57]	93.10	-	82.30
MuDeep [58]	95.34	-	84.66
MGN [59]	95.70	-	86.90
baseline	86.58	95.10	70.91
VCFL	89.25	95.61	74.48
VCFL (re-rank) [39]	90.91	95.10	86.67
VCFL+	91.86	96.70	76.97
VCFL+ (re-rank) [39]	94.00	96.73	**90.80**

TABLE II: The comparison with other state-of-art methods over CUHK03(Detected). “tri.” denotes training with triplet loss. “cls.+tri.” denotes training with the combination of both losses.

Method	top1	top5	mAP
BoW[41]	6.36	-	6.39
LOMO[60]	12.8	-	11.5
Resnet50+XQDA[39]	31.1	-	28.2
Resnet50+XQDA+re-rank[39]	34.7	-	37.4
SVDNet[47]	41.5	-	37.3
MultiScale[50]	40.7	-	37.0
TriNet+Era[61]	55.5	-	50.7
SVDNet+Era[61]	48.7	-	43.5
PCB(UP)[55]	61.3	-	54.2
PCB(RPP)[55]	63.7	-	57.5
HA-CNN [52]	41.70	-	38.60
MLFN [51]	52.80	-	47.80
MGN [59]	66.80	-	66.00
Mancs [57]	65.50	-	60.50
MuDeep [58]	71.93	-	67.21
baseline	58.36	78.71	53.71
VCFL	61.43	79.71	55.61
VCFL (re-rank)[39]	70.36	81.14	**70.44**
VCFL+	61.29	78.71	54.26
VCFL+ (re-rank)[39]	69.93	81.43	**70.45**

D. Analysis of The Proposed Model

Our experiments are carried out on ResNet to validate our methods. The experiments are conducted on CUHK03(Detected), Market1501 and DukeMTMC-reID, and the performance and visualization of each part in the proposed model is shown. Further, the retrieval performance are shown to prove our superiority.

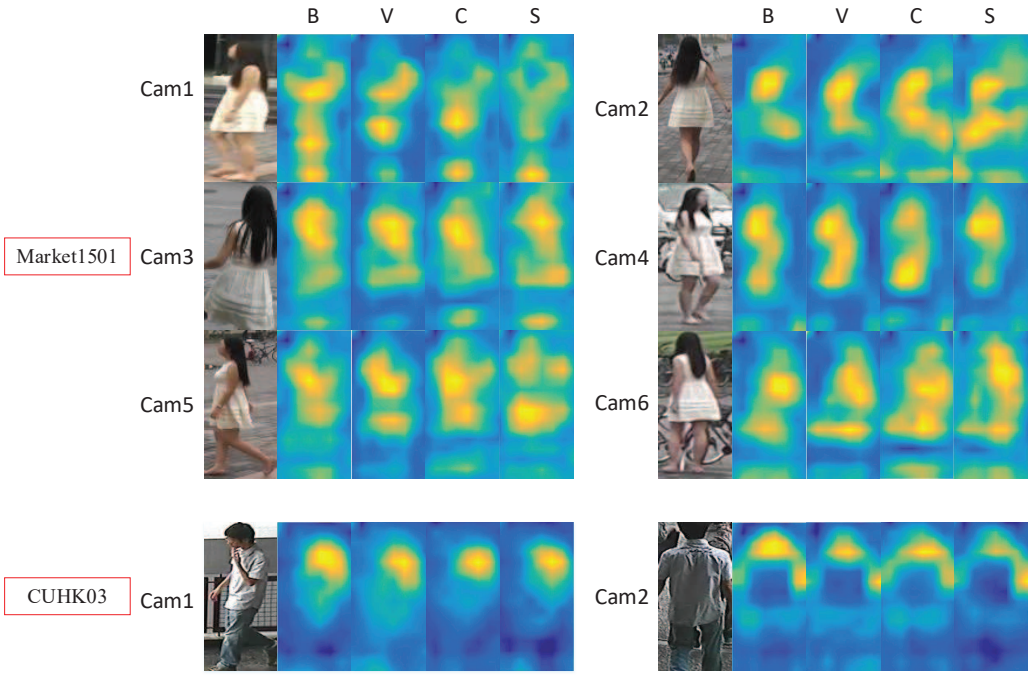


Fig. 5: View confusion visualization on Market1501 and CUHk03. In order to validate the performance of part confusions, we visualise the feature map of three part confusions to prove the interaction effect between them may not be promotive. As can be seen from the picture, different losses may have different promoting influence in feature maps. We have to mention that the feature maps are obtained using the baseline (B) and the single loss of L_{vc} (V), L_{cen} (C), L_{sg} (S), respectively, which are denoted as B, V, C, S in the figure.

TABLE III: The comparison with other state-of-art methods over DukeMTMC-reID. "tri." denotes training with triplet loss. "cls.+tri." denotes training with the combination of both losses.

Method	top1	top5	mAP
BoW+kissme[41]	25.13	-	12.17
SVDNet[47]	76.70	86.4	56.80
LOMO+XQDA [60]	30.80	-	17.00
ResNet50 [37]	65.20	-	45.00
PN-reID [62]	73.58	-	53.20
HAP2S-P [49]	75.94	-	60.64
CSA [31]	78.32	-	57.61
MultiScale[50]	79.20	-	60.60
HA-CNN [52]	80.50	-	63.80
MLFN [51]	81.00	-	62.80
Deep-Person [54]	80.91	-	64.83
SGGNN [56]	81.10	88.40	68.20
DuATM [53]	81.82	90.20	64.58
PCB(UP)[55]	82.6	-	68.8
PCB(RPP)[55]	84.5	-	71.5
Mancs [57]	84.90	-	71.80
MGN [59]	88.70	-	78.40
MuDeep [58]	88.19	-	75.63
baseline	79.22	90.62	64.33
VCFL	80.30	90.22	63.41
VCFL (re-rank)[39]	84.83	91.16	**80.41**
VCFL+	82.68	91.25	65.68
VCFL+ (re-rank)[39]	87.21	92.55	**83.12**

View confusion: The core experiments are focused on the performance of view confusion mechanism. We analyze the performance of each part and the their combinations and observe the visualization of the feature maps. In the next section, we empirically study the influence of different parts and the whole view confusion mechanism in benchmark datasets. As is shown in Table II, the contribution of each part has been validated. Specifically, we analyze the performance of the whole the view confusion mechanism and make the detailed ablation study on the view confusion parts. We further consider the cross-interaction effect inside the model and analyze the influence of the feature map regularization, which can also be seen through the visualization of feature maps.

- **Baseline:** Similar to [37], the main framework is based on ResNet50, and pre-trained by weights provided in [37]. We use the triplet loss as the supervision for baseline, and we further add cross-entropy loss as identity classification loss to improve the performance.
- **View confusion.** As is described in the former section, the implementation of this part has two kinds of ways. 1) We follow the implementation of GAN [38] and use the views predicted by a pose predictor, which are adopted in our conference version [5]. 2) We use the Gradient Reverse Layer and the true camera view label for convenience and simplifying the implementation difficulty. For the first way, the network training is not stable enough which is similar to the training of GAN [38]. The adversarial learning of the view classifier and the feature extractor is not stable in our training stage which means this kind of confusion has a great impact on the final performance. In this paper, we adopt the second way for improving the alternative training, which are proved to be better. As is shown in the third and fourth row of Table IV, the Classifier Based Confusion is proved

TABLE IV: Evaluation of the confusion with ResNet backbone. "tri." denotes training with triplet loss. "cls.+tri." denotes training with the combination of both losses. "VC" denotes View Confusion. "CL" denotes Center Loss. "SG" denotes Sift Guidance. "FMR" denotes Feature Map Regularization. "P" denotes predicted view information. "C" denotes camera view information. Market-1501, cuhk03(Detected) and DukeMTMC-reID are tested.

Methods	Market1501				Cuhk03				Duke			
	top1	top5	top10	mAP	top1	top5	top10	mAP	top1	top5	top10	mAP
baseline(tri.)	86.58	95.10	96.67	70.91	58.93	76.43	83.93	53.05	79.22	90.62	92.50	64.33
VC(tri.)(P)	85.18	94.27	96.32	69.04	-	-	-	-	-	-	-	-
VC(tri.)(C)	88.27	95.67	97.27	74.19	61.29	77.43	83.14	54.56	80.52	89.99	93.00	64.50
CL(tri.)	88.57	95.19	96.94	73.36	59.79	77.93	85.21	54.01	80.39	90.44	93.09	65.21
SG(tri.)	88.57	95.64	97.24	74.30	**61.43**	79.14	85.71	55.22	80.34	90.53	93.27	65.01
VC(C) + CL(tri.)	87.23	95.37	96.85	73.41	60.57	**79.29**	**85.86**	**55.27**	79.89	90.93	93.36	64.09
VC(C) + SG(tri.)	88.09	95.61	97.12	73.73	58.43	76.14	83.14	52.92	79.94	90.26	92.37	64.29
CL + SG(tri.)	88.45	95.25	96.70	73.91	58.50	76.93	84.36	53.15	79.80	90.80	93.27	64.40
VC(C)+CL+SG(tri.)	88.54	95.40	97.03	73.39	61.00	78.29	85.07	54.82	80.30	90.22	92.73	63.41
CL + SG(cls.+tri.)	90.23	96.26	97.77	77.08	59.36	77.29	84.93	53.20	82.32	91.07	94.12	**66.97**
VC(C) + CL(cls.+tri.)	89.13	95.87	97.48	75.73	56.14	75.21	82.00	51.17	82.27	91.07	93.49	65.77
VC(C) + SG(cls.+tri.)	89.34	95.75	97.42	75.35	60.21	76.43	82.64	54.37	82.36	91.11	93.36	65.75
VC(C)+CL+SG(cls.+tri.)	90.02	95.52	97.00	76.42	59.14	76.71	83.93	53.43	81.96	91.34	93.67	65.63
VC(C)+CL+SG+FMR(cls.+tri.)	**91.86**	**96.70**	**97.92**	**76.97**	61.29	78.71	85.14	54.26	**82.68**	**91.25**	**94.25**	65.68

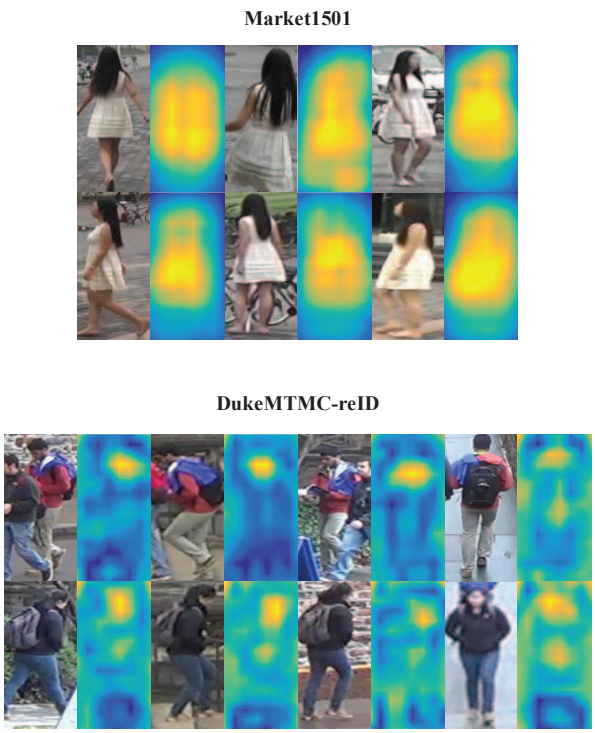
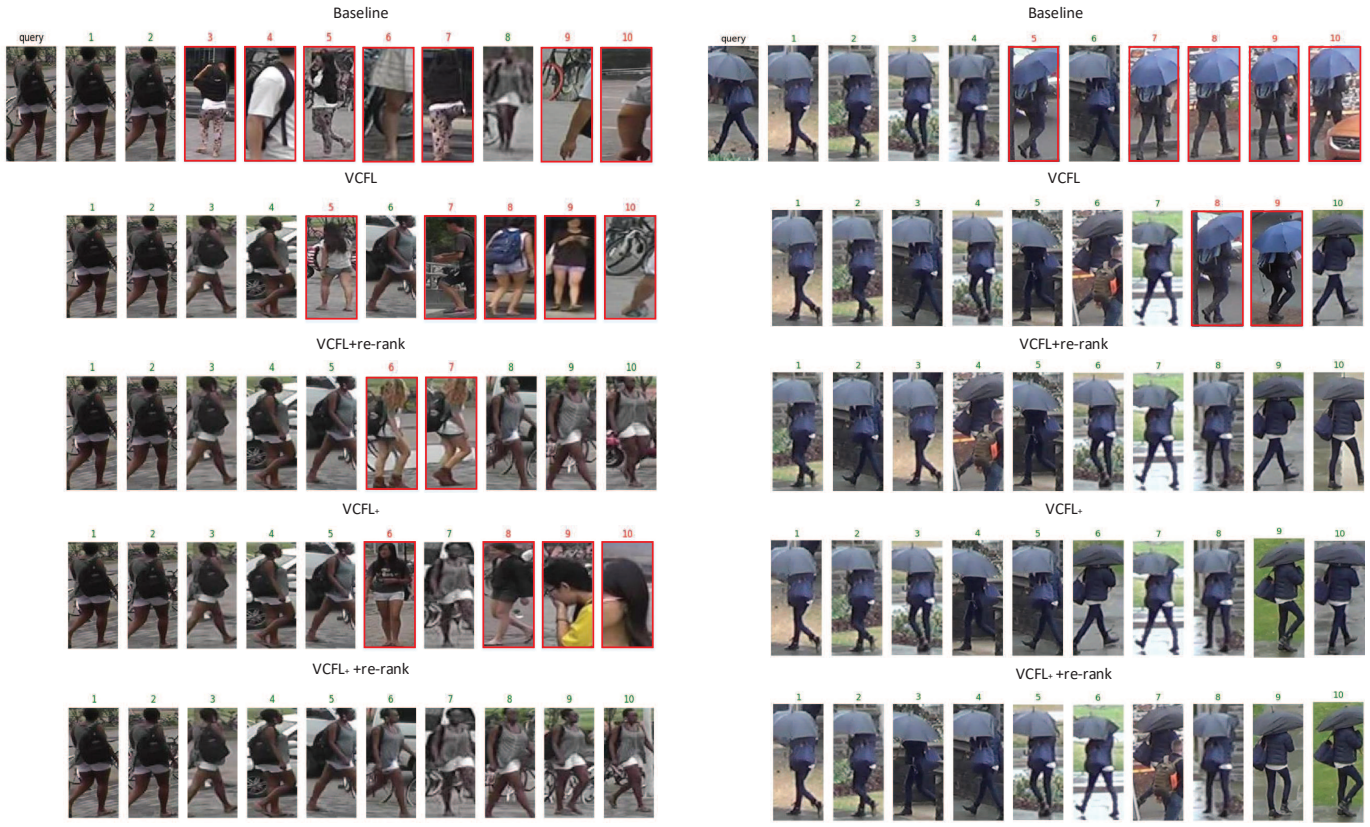


Fig. 6: In order to validate the performance of view confusion, we compare the feature map of same id among different camera information to show the view invariance of our method on ResNet, and it clear shows that feature maps tend to have similar distribution after view confusion mechanism. It means that our model can eliminate the view influence in person images, thus we can finally get view-agnostic features.

to be effective. Besides, it can provide us a new way to solve the cross-view person re-id problems using domain adaptation inspired methods.

- **Center loss.** It is not hard to admit the effectiveness of feature confusion loss from Table IV, which improves the performance much by drawing the features of different views close. In the training stage, we set $\lambda_{fc} = 10^{-4}$ to make the whole loss converge well. Compared with [12], the latter aims to specify all the views and it may suffer from the influence caused by the predicted error of view predictor. We adopt the camera ID as view labels.
- **Sift guidance.** It is meaningful when integrating sift features into deep features since it can make up for the deep features. In our opinion, sift features have with good local quality and view-independence, which may help deep feature to be view invariant. This part improves the performance much as well as explore the interpretability of deep learning by designing the network structure following the hand-crafted descriptors.
- **The combination of each part.** In order to have an insight on the mutual interference among the three part confusions, it is vital to explore the combination of part confusions. We conduct several experiments to explore multiple combinations of part confusions, aiming to explore the essence of each part in our model. The results are presented in Table IV, in which the performance based on triplet loss only or triplet loss plus classification loss are both showed. We can draw a conclusion that the Feature Based Confusion and Sift Based Confusion can work well on global features. We think the identity information and camera information are coherent and influenced each other, so it is possible that the view classifier and identity classifier can affect each other. For this reason, we propose the multi-branch structure as is shown in Figure 3, and propose the regularization on feature map. From the results in IV, we can see that each part has different contribution to the final model.
- **Whole model.** The final performance is illustrated in the Table IV, proving that the experimental results can be improved much through our view confusion mechanism. We have to mention that loss weights should be adjusted to achieve better performance. We analyze the difference of three parts as is shown in Figure 5 and we also



(a) (b)

Fig. 7: Illustration of the retrieval results on Market 1501 (a) and DukeMTMC-reID (b). The green title represents a true positive, and the red rectangle represents a negative positive. For each sample, the first row shows the results for the baseline network representation, the second and the third row show the results of our VCFL without/with re-ranking, and the fourth and fifth row show the results of our VCFL+ without/with re-ranking. Through this figure, the effectiveness of our approach can be clearly seen.

show the feature maps of whole model in Figure 6. The Market1501 and CHUK03 are performed. We have to mention that Market1501 is captured by 6 cameras while CUHK03 is captured by 2 cameras, and we show the feature map of individual confusion part with each camera view. As is shown in Figure 5, in Market1501, the feature maps of Cam1, Cam2, Cam5 and Cam6 seem to focus on different regions, and what we want to do is to reduce the conflict between different losses in feature learning. It means that each part of the whole confusion may not promote features in the same direction, thus further regularization on feature maps is needed. In this part, we also report our retrieval process and ranking result in Figure 7, which clearly shows that our approach has a good person re-id performance. In Figure 7, we also show the feature map comparisons of our model with different losses, including baseline (tri.), our approach (tri.), our approach (tri. + re-rank), our approach (tri.+cls.), our approach (tri.+cls.+re-rank).

- **The regularization on feature map.** In order to promote the model, we further regularize our confusion mechanism not only in feature aspect but also in the feature map aspect. We present the results in the final row of the Table IV. We can see that the proposed feature map regularization can improve the performance. Notably, for CUHK03 data, the identity classification loss shows negative impact which is alleviated by adding FMR.

VI. Conclusion

This paper aims to solve view changes in person Re-ID, and prevents the Re-ID system from dropping dramatically due to large variations of camera views and human poses. To improve the performance as well as solve view change problems, we present an adversarial view confusion feature learning (VCFL) framework for view-invariant person re-identification. Our VCFL is achieved from three aspects: 1) Adversarial learning between feature extractor and the view classifier. 2) Center loss for discriminative feature learning. 3) Sift guidance for improving the view invariance and interpretability of deep features. Thus, the view-invariant identity-wise features can be learned. Further, an improved VCFL+ model is proposed by introducing attentive feature map alignment. Experiments validate the superiority of our models. In our future work, solving cross-view problems guided by transfer learning and exploring the interpretability of deep features guided by hand-crafted low-level descriptors are worthy studied.

REFERENCES

[1] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *CVPR*, 2012.

[2] S. Liao and S. Z. Li, "Efficient psd constrained asymmetric metric learning for person re-identification," in *ICCV*, 2015.

[3] S. Zhou, F. Wang, Z. Huang, and J. Wang, "Discriminative feature learning with consistent attention regularization for person re-identification," in *ICCV*, October 2019.

[4] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *ICCV*, October 2019.

[5] F. Liu and L. Zhang, "View confusion feature learning for person re-identification," in *ICCV*, October 2019.

[6] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *CVPR*, 2016.

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

[8] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang, "Person re-identification with correspondence structure learning," in *ICCV*, 2015.

[9] X. Yang, M. Wang, and D. Tao, "Person re-identification with metric learning using privileged information," *IEEE TIP*, vol. 27, no. 2, pp. 791–805, 2018.

[10] X. Yang, P. Zhou, and M. Wang, "Person reidentification via structural deep metric learning," *IEEE TNNLS*, no. 99, pp. 1–12, 2018.

[11] H.-X. Yu, A. Wu, and W.-S. Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," *arXiv*, 2017.

[12] Z. Feng, J. Lai, and X. Xie, "Learning view-specific deep networks for person re-identification," *IEEE TIP*, 2018.

[13] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*, 2016, pp. 499–515.

[14] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, Nov 2004.

[15] Y. Ganin and V. S. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2015.

[16] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *ICCV*, 2017.

[17] X. Zhang, H. L. Xing, F. Weilai, X. Yixiao, S. Q. Xiao, W. Jiang, C. Zhang, and J. Sun, "Alignedreid: Surpassing human-level performance in person re-identification," *arXiv*, 2017.

[18] 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*, 2018.

[19] 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.

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

[21] X. Ben, C. Gong, P. Zhang, X. Jia, Q. Wu, and W. Meng, "Coupled patch alignment for matching cross-view gaits," *IEEE TIP*, vol. 28, no. 6, pp. 3142–3157, 2019.

[22] X. Ben, C. Gong, P. Zhang, R. Yan, Q. Wu, and W. Meng, "Coupled bilinear discriminant projection for cross-view gait recognition," *IEEE TCSVT*, vol. 30, no. 3, pp. 734–747, 2020.

[23] X. Ben, P. Zhang, Z. Lai, R. Yan, X. Zhai, and W. Meng, "A general tensor representation framework for cross-view gait recognition," *Pattern Recognition*, vol. 90, pp. 87–98, 2019.

[24] G. Gao, Y. Yu, M. Yang, H. Chang, P. Huang, and D. Yue, "Cross-resolution face recognition with pose variations via multilayer locality-constrained structural orthogonal procrustes regression," *Information Sciences*, vol. 506, pp. 19–36, 2020.

[25] C. Zhao, K. Chen, D. Zang, Z. Zhang, W. Zuo, and D. Miao, "Uncertainty-optimized deep learning model for small-scale person re-identification," *Sci China Inf Sci*, vol. 62, no. 12, pp. 1–13, 2019.

[26] C. Zhao, X. Wang, W. Zuo, F. Shen, L. Shao, and D. Miao, "Similarity learning with joint transfer constraints for person re-identification," *Pattern Recognition*, vol. 97, pp. 1–10, 2020.

[27] L. Wu, R. Hong, Y. Wang, and M. Wang, "Cross-entropy adversarial view adaptation for person re-identification," *IEEE TCSVT*, 2019.

[28] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *ICCV*, 2003.

[29] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *ECCV W*, 2004.

[30] D. Tao, Y. Guo, B. Yu, J. Pang, and Z. Yu, "Deep multi-view feature learning for person re-identification," *IEEE TCSVT*, 2017.

[31] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *CVPR*, 2018.

[32] 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 *CVPR*, 2018.

[33] X. Jing, Z. Rui, Z. Feng, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *CVPR*, 2018.

[34] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang, "Abd-net: Attentive but diverse person re-identification," in *ICCV*, October 2019.

[35] B. N. Xia, Y. Gong, Y. Zhang, and C. Poellabauer, "Second-order non-local attention networks for person re-identification," in *ICCV*, 2019.

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

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[38] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.

[39] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *CVPR*, 2017, pp. 3652–3661.

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

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

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

[43] R. S. Z. R. C. Ergys Ristani, Francesco Solera and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *ECCVw*, 2016.

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

[45] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai, "Person re-identification by camera correlation aware feature augmentation," *IEEE TPAMI*, vol. 40, no. 2, pp. 392–408, 2018.

[46] J. Lin, L. Ren, J. Lu, J. Feng, and J. Zhou, "Consistent-aware deep learning for person re-identification in a camera network," in *CVPR*, 2017.

[47] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," in *ICCV*, 2017, pp. 3800–3808.

[48] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE TCSVT*, 2018.

[49] R. Yu, Z. Dou, S. Bai, Z. Zhang, Y. Xu, and X. Bai, "Hard-aware point-to-set deep metric for person re-identification," in *ECCV*, 2018, pp. 188–204.

[50] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *ICCV*, 2017, pp. 2590–2600.

[51] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *CVPR*, June 2018.

[52] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *CVPR*, June 2018.

[53] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," in *CVPR*, June 2018.

[54] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, and Y. Xu, "Deep-person: Learning discriminative deep features for person re-identification," *Pattern Recognition*, vol. 98, p. 107036, 2020.

[55] 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, pp. 480–496.

[56] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, "Person reidentification with deep similarity-guided graph neural network," in *ECCV*, 2018.

[57] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," in *ECCV*, 2018.

[58] T. X. Y.-G. J. X. X. Xuelin Qiany, Yanwei Fuy, "Leader-based multi-scale attention deep architecture for person re-identification," *IEEE TPAMI*, 2019.

[59] G. Wang, Y. Yuan, C. Xiong, J. Li, and Z. Xi, "Learning discriminative features with multiple granularities for person re-identification," in *ACM MM*, 2018.

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

[61] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *arXiv*, 2017.

[62] 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.