IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 22, 2021, accepted July 16, 2021, date of publication July 27, 2021, date of current version August 26, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3100571

A Color/Illuminance Aware Data Augmentation and Style Adaptation Approach to Person Re-Identification

ZHOUCHI LIN, CHENYANG LIU, WENBO QI, AND S. C. CHAN, (Member, IEEE)
Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong
Corresponding author: S. C. Chan (scchan@eee.hku.hk)

ABSTRACT Person re-identification problems usually suffer from large subject appearance variations and limited training data. This paper proposes a novel physically motivated Color/Illuminance-Aware data-augmentation (CIADA) scheme and a style-adaptive fusion approach to address these issues. The CIADA scheme estimates the color/illuminance distribution from the training data via manifold learning and generates new samples under different color/illuminance perturbations to better capture objects' appearance for mitigating the small-sample-size and color variation problems. A Color/Illuminance Aware Feature Augmentation (CIAFA) approach, which is applicable to state-of-the-art features and metric learning algorithms, is then proposed to integrate the features generated by the augmented samples for metric learning. A new Color/Illuminance-Aware Style Fusion (CIASF) scheme, which allows the learning and matching process to be performed independently on each pair of datasets generated for estimating a set of 'local' distance functions, is also proposed. A canonical correlation analysis-based weighting scheme is developed to fuse these local distances to an overall distance for recognition. This reduces the memory requirement and complexity over the original CIAFA. Experiments on common datasets show that the proposed methodologies substantially improve the performance of state-of-the-art subspace learning algorithms. It is applicable to both small and large datasets with hand-craft and deep features.

INDEX TERMS Data augmentation, local metric learning, small sample size problem, person re-identification.

I. INTRODUCTION

Pedestrian recognition across multiple cameras, or, person re-identification (Person Re-id), has been extensively studied in the past decade due to the rapid deployment of large-scale video-based surveillance networks for social security and other applications. Despite the increasing amount of published literatures, it remains challenging due to large variations across camera views and limited availability of object data.

The problems in Person Re-id is three-fold. First, image samples captured by different cameras in different angles exhibit dramatic variations in body gesture, background and occlusions, as shown in Figure 1(a). Moreover, the large variation due to view-specific illumination/color (VSIC) halts the improvement of feature extraction techniques and recognizers. In addition, the sample size of commonly used datasets is usually small as compared with the feature dimension, which limits the usage of more complex models. Given that we need to distinguish each subject from all the others, it is a recognition problem with many classes but with few sample size [1].

State-of-the-art methods for the person re-id problem usually involve feature extraction and subsequent feature matching. Feature extraction aims to extract a robust and discriminative representation of a person across camera views by exploring: 1) color cues such as color histogram; 2) texture cues such as Histogram of Gradient (HoG) [2] and 3) learning-based representations such as Deep Part-aligned Representations [3]. Based on these components, many handcrafted features such as Ensemble of Local Features (ELF) [4], Symmetry-Driven Accumulation of Local Features (SDALF) [5] and Local Maximum Occurrence (LOMO) [6] have been proposed for tackling the re-id

The associate editor coordinating the review of this manuscript and approving it for publication was Vicente Alarcon-Aquino.

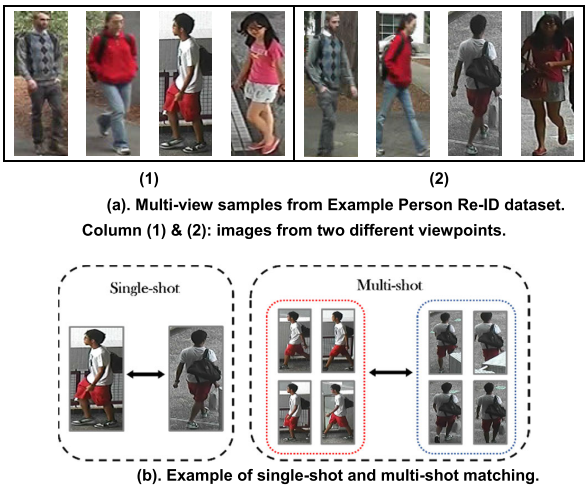


FIGURE 1. **(a). Multi-view samples from example person Re-ID dataset. Column (1) & (2): images from two different viewpoints. (b). Example of single-shot and multi-shot matching.**

problem. To address the view-specific pose variation, these feature extraction methods usually rely on over sampling from the image, and therefore result in redundant feature representations and large feature dimensions.

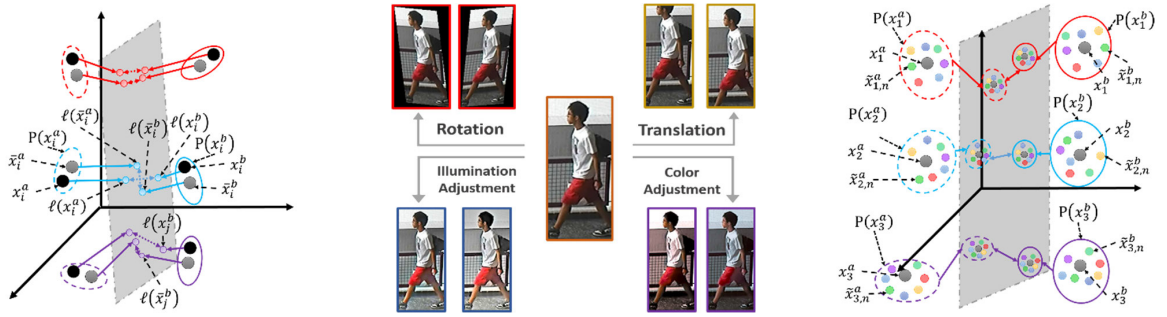
To facilitate feature matching of these high dimensional features, much focus has been given to subspace learning algorithms and distance metric learning methods, which usually relies on the Mahalanobis distance. The main objective of subspace learning [6], [7] is to find a discriminating subspace that best represents the essential information of the extracted features to reduce the matching complexity. On the other hand, distance metric learning methods [1], [8], [9] facilitate the matching process in these subspaces to differentiate different subjects. KISSME [1] and Cross-view Quadratic Discriminant Analysis (XQDA) [6] have proven to be fast and effective in solving the re-id problem. Null Foley-Sammon Transform (NFST) [7], which claims to optimally solve the small sample size (SSS) problem, also achieves high recognition rate comparing with state-of-the-art methods.

Recently, inspired by the success of deep learning in image classification. A series of CNNs that jointly learn two subnetworks, a feature extraction network for learning the object representation and a matching network for learning the similarity, have shown performance gain in Re-id tasks. Apart from global feature (output of the fully-connected layer [10], [11]) and combination of stripes /cell-like local feature [12], [13], Deep Part-aligned feature [3], which aligns human parts across camera views, has shown substantial performance improvement in common datasets. For the matching sub-network, it varies upon the formulation of inputs and feature types. It penalizes the misalignment of features caused by global feature or combination-of-strip/cell-based features. For part-aligned features, usually simple Euclidean distance is used since they are assumed to be well-aligned in feature dimension.

Previous studies [14]–[17] on larger datasets with multiple samples of a subject for each camera view (a.k.a. multi-shot matching) showed that the performance of state of-the-art re-id algorithms can be significantly enhanced with increased samples per camera view and the view-specific illuminance/color (VSIC) problem can be better solved with multi-shot samples. This can be explained from the illustration in Figure 1 (b) where the extracted feature from multi-shot samples better handles the color variation under different illumination conditions of the cameras and suppresses the noise and shading effect when representing the subject's appearance.

If we consider the cropped sample images of a subject, say in Figure 1(b), as samples from a given class, the problem of person re-id can be viewed as finding a discriminative subspace and measurement metric in the feature space such that, in this lower dimensional space, intra-class distance (i.e. distances between samples from the same identity/class but at different views) is smaller than inter-class distance (i.e. distances of samples of different identities/ class from arbitrary views). This is illustrated in Figure 2 (a) and (c) using a 3-dimensional feature space example. For the single-shot matching scenario shown in Figure 2(a), the captured samples are represented as the black dots. Due to the VSIC problem, the samples may deviate substantially from the center of the subject's appearance distribution (shown as grey dots in Figure 2(a)). Consequently, the low dimension subspace projection computed may separate the noisy feature samples but not the original subject's appearance distributions. This leads to a lower recognition rate. In contrast, multiple samples of the appearance distribution are available in the multi-shot scenarios at Figure 2(c). Due to the increased data samples, the subspace computed will be more robust to imperfections in the sampling process, leading to better generalization and increased recognition rate.

In real-world applications, multi-shot samples are not always available, or the sample number of each subject may be limited, which is commonly referred to as the small sample size (SSS) problem [1]. Thus, existing person re-id tasks still suffers substantially from the VSIC problem. To overcome the SSS problem, data augmentation (DA) such as horizontal reflections, translations and rotations of the original sample have been used to obtain an artificially enlarged dataset [18]. Fawzi *et. al* [19] suggested that incorporating these worst-case DA samples provides additional information for further improving the robustness and performance of the original classifier. Techniques using color jittering [20] and addressing the issues of location and scale variability [21] have also been proposed. In the person re-id problem, spatial transformation via translation, perspective transformation and subsampling of the original samples have been proposed [18] to increase the sample size [16] while perturbation-based approaches are yet to be explored. [22] proposed a multiple instance attention learning framework to address the weakly supervised video person re-identification task by utilizing the similarity between person identities and videos. A divided regions-based feature extraction method is proposed in [23] for face images. A Color to Gray Video



(a) Subspace matching in single-shot person re-identification.

(b) Approaches in data augmentation using image transformation and color adjustment.

(c) Subspace matching in multi-shot matching under same protocol of (a).

FIGURE 2. (a): Illustration of subspace matching process in single-shot person re-identification: grey dots refer to the mean appearances of objects, and black dots refer to the sampled appearance; dash circle and solid circle enclosed samples from the two views. Superscripts a and b denotes two camera views. Indices i and j denote sample numbers. x denotes feature vector and $\bar{x}$ denotes class mean. (b): Conventional approaches in data augmentation using image transformation and color adjustment. (c): Illustration of subspace matching in multi-shot matching under same protocol of (a): in (a), the single-shot sampled appearance (black dots) may deviate substantially from the mean appearance (grey dots) due to noise and background occluders and thus lower the recognition rate. In (c), multiple samples (color dots) are available and the robustness of the subspace learning and hence recognition accuracy are improved.

Person Re-identification (CGVPR) dataset was collected to address the problem of person re-id between grayscale and true-color videos [24] and an asymmetric within-video projection based Semi-coupled Dictionary Pair Learning (SDPL) approach was also proposed to address the problem. In [25], a set of new samples under different camera/dataset styles have been generated using CycleGAN [26] to enlarge the training set so as to improve the stability under complex backgrounds. However, the enlarged training set was not efficiently utilized in each mini-batch and epoch, and a long time is required to train the Cycle GAN and the final CNN.

In this paper, a novel Color/Illuminance-Aware data augmentation (CIADA) and Style Fusion (CIASF) approach for person re-identification is proposed. In particular, a new Color/Illuminance-Aware Data Augmentation (CIADA) scheme through illumination and color perturbation is proposed to obtain more physically meaningful data to address the important VSIC problem. Moreover, the 'physically motivated' perturbations/transformations in the CIADA is derived from the color/illumination characteristics of the training data, instead of through addition of random noise or sub-sampling. More precisely, the prior knowledge of color and illuminance distribution is firstly estimated by manifold learning. Then, a set of new samples based on this prior distribution on color and illuminance of the original pedestrian samples are generated to better capture the person's appearance by reducing background noise and mitigating camera-specific styles. With the multiple samples of a given subject generated, the original re-id problem becomes a *multi-sample* problem. It differs from traditional multi-shot scenario in that the Re-id task is solved through mitigating the VSIC difference between cameras with no additional information of the subjects is introduced. Moreover, the multi-shot samples are usually assumed to be independent and hence a single subspace/metric learning classifier is usually employed for recognition. However, the samples in the proposed augmented approach are generated through color/illuminate perturbation, and they can be correlated. This calls for a new approach to fuse the augmented data and hence the associated features for metric learning.

To this end, a style-adaptive fusion approach is proposed to address this important training issue. It is applicable to state-of-the-art features and metric learning algorithms such as LOMO and XQDA, respectively, and can also be extended to deep CNN approaches. Specifically, in traditional approaches, features from the augmented samples of the same identity can be stacked together for metric learning. Since the feature augmentation is derived from color/luminance perturbation, it is referred to as color/illuminance-aware feature augmentation (CIAFA). Due to the increased feature dimension of the augmented features, the memory/computational requirement is considerably increased. Therefore, an alternative called col-or/illuminance-aware style fusion (CIASF) is thus proposed and it allows the subspace/metric/deep learning to be performed independently on each pair of augmented data for estimating a set of 'local' distance functions. Since the augmented dataset are generated from the original samples, they are correlated rather than independent. Therefore, simply averaging them to form the overall distance is undesirable. Moreover, motivated by [27], we introduced a canonical correlation analysis (CCA)-based approach to quantify the correlation between the styles of generated samples for improving the fusion of distance function. This differs from the training in conventional multi-shot problem where one single classifier is used, as the samples are assumed to be independent. [28] studies the cross-modality person re-identification problem which is different from the present study. In [29], it is proposed to narrow down the color discrepancy between visual image and infrared image using the gray scale image as the bridge. It is reasonable that the gray scale visible images have similar style with infrared images. However, in our setting, only color images are employed instead of the infrared images and hence using the useful color information will be lost if gray scale images is used instead.

We now evaluate the performance of the proposed algorithm using several publicly available databases. Extensive experiments were carried out on benchmark datasets of various scales including VIPeR [30], CUHK01 [15] CUHK03 [16] and Market-1501 [39] to verify the effectiveness of the proposed methodology. We employed state-of-the-art features namely LOMO [6], KCCA [14], deep features such as Part-aligned Network (Partnet) [3] and ID-discriminative embedding (IDE) [31] and two metric learning algorithms XQDA [6] and NFST [7] for fair comparison. We compare the performance of the proposed DA and fusion approach with state-of-the-art methods in [3], [6], [7], [31] and experimental results show that the proposed framework achieve much improved rank-1 recognition accuracy against the benchmark in all tested features and similarity measurement techniques. The proposed approach also achieves comparable performance as more sophisticated GAN-based data augmentation method CamStyle [25], but with a much lower complexity (while GAN has the potential to generate sophisticated augmented data as demonstrated in a recent paper [31] after the submission of our first manuscript, it requires large dataset and a long training time). Moreover, it can work for both small and large training data. This demonstrates the generality and efficiency of the proposed approach in the person re-id problem.

This paper is organized as follows. In Section 2, the system overview and problem formulation will be described. Section 3 is dedicated to the proposed CIADA approach. In Section 4, the proposed CIAFA and CIASF approaches are introduced. Experimental results and discussions are presented in Section 5 with conclusions drawn in Section 6.

II. SYSTEM OVERVIEW AND PROBLEM FORMULATION

A. SYSTEM OVERVIEW

The key components of our proposed data augmentation framework are summarized in Figure 3. In the CIADA step, illumination and color styles of the original images from different camera views, A and B, shown on the left in Figure 3 are perturbed to generate more, which are outlined using different colors. In the feature extraction step, state-of-the-arts feature extraction methods such as LOMO can be used to extract features from different samples of the same subject/identity, which are denoted by color bars with the same color code as the outline of their corresponding image samples on the left of Figure 3. In the similarity measurement step, the features from all samples of the same subject/identity are then stacked together in the feature augmentation step for training and similarity measurement. To parallelize this process, in the style-adaptive fusion approach, the correlation between each illuminance/color pair generated is first estimated. A set of 'local' distance functions can be obtained independently by subspace/metric/deep learning. These 'local' pair-wise sample distance learned can be linearly fused with the help of their corresponding correlations to form the overall distance for recognition. The framework can readily be extended to include other state-of-the-art features and learning methods so as to leverage their merits.

The main contributions of this work include: 1) we introduce a CIAFA framework with CIADA to generate more informative samples so as to improve state-of-the-art person re-id methods. Our framework can be extended to existing feature extraction algorithms to enhance their robustness against environmental variations; 2) we model pair-wise correlation between samples generated using different illuminance/color settings to quantify their impact on subspace learning process and incorporate them in a style adaptive distance fusion framework for solving the re-id problem in parallel and reduced complexity.

B. PROBLEM FORMULATION

As mentioned above, the person re-identification problem aims to find the corresponding pedestrian in one view (gallery set) given its sample(s) from another view (probe set). Denote the samples from the probe and gallery sets respectively by $\boldsymbol{S}^a = \left\{\boldsymbol{s}_1^a, \boldsymbol{s}_2^a, \ldots, \boldsymbol{s}_{N_a}^a\right\}$ and $\boldsymbol{S}^b = \left\{\boldsymbol{s}_1^b, \boldsymbol{s}_2^b, \ldots, \boldsymbol{s}_{N_b}^b\right\}$, where s_i^φ is the input image of the *i-th* subject at view φ, and N_φ is the number of image samples at view φ.

Given $\boldsymbol{s}_i^a$ from the probe set, the searching of its corresponding sample in the gallery set is usually formulated as the following similarity ranking problem:

$$j^* = \arg\max_{j\in 1,2,\ldots,N_b} S\left(\boldsymbol{s}_i^a, \boldsymbol{s}_j^b\right) \tag{1}$$

where $\boldsymbol{s}_{j^*}^b$ is assumed to have same label with $\boldsymbol{s}_i^a$ and $S(\cdot,\cdot)$ is the similarity measurement between samples. Alternatively, we can define a distance function $d(\cdot,\cdot)$ to measure the dissimilarity between features from the corresponding samples. Let the feature extraction process be denoted by $\boldsymbol{x}_i^\varphi = \boldsymbol{f}\left(\boldsymbol{s}_i^\varphi\right)$, $\varphi = a, b$, $i = 1, \ldots, N_\varphi$ where $\boldsymbol{x}_i^a \in \Re^d$ and $\boldsymbol{x}_i^b \in \Re^d$, are respectively the features of the probe and gallery sets and $\boldsymbol{f}(\boldsymbol{s})$ is the feature extraction operator on the image samples. As the features are usually of high dimension, subspace learning algorithms are frequently employed to construct a distance function between the features in the reduced dimension subspace as follows:

$$\begin{aligned} d(i,j) &= d\left(\ell\left(\boldsymbol{x}_i^a\right), \ell\left(\boldsymbol{x}_j^b\right)\right) \\ &= \left(\boldsymbol{W}^T\left(\boldsymbol{x}_i^a - \boldsymbol{x}_j^b\right)\right)^T \boldsymbol{M}\boldsymbol{W}^T\left(\boldsymbol{x}_i^a - \boldsymbol{x}_j^b\right) \end{aligned} \tag{2}$$

where $\ell\left(\boldsymbol{x}_i^a\right) = \boldsymbol{W}^T\boldsymbol{x}_i^a$ is the subspace transformation $\boldsymbol{x}_i^a$, $\boldsymbol{W} \in \Re^{d\times r}$ is the learned discriminative subspace transformation, and $\boldsymbol{M} \in \Re^{r\times r}$ is the measurement metric which is positive semi-definite with rank r. Here, r is the size of the reduced feature space. Usually, $\boldsymbol{W}$ and $\boldsymbol{M}$ can be obtained by solving some kind of eigenvalue problem. For instance, it was shown in [33] that they can be solved by the following subspace/metric learning problem:

$$\begin{aligned} &\arg\min_{W,M} \sum_{l_i^a = l_j^b} d(i,j) \\ &s.t. \sum_{l_i^a \neq l_j^b} d(i,j) \geq 1, \boldsymbol{M} \succ 0 \end{aligned} \tag{3}$$

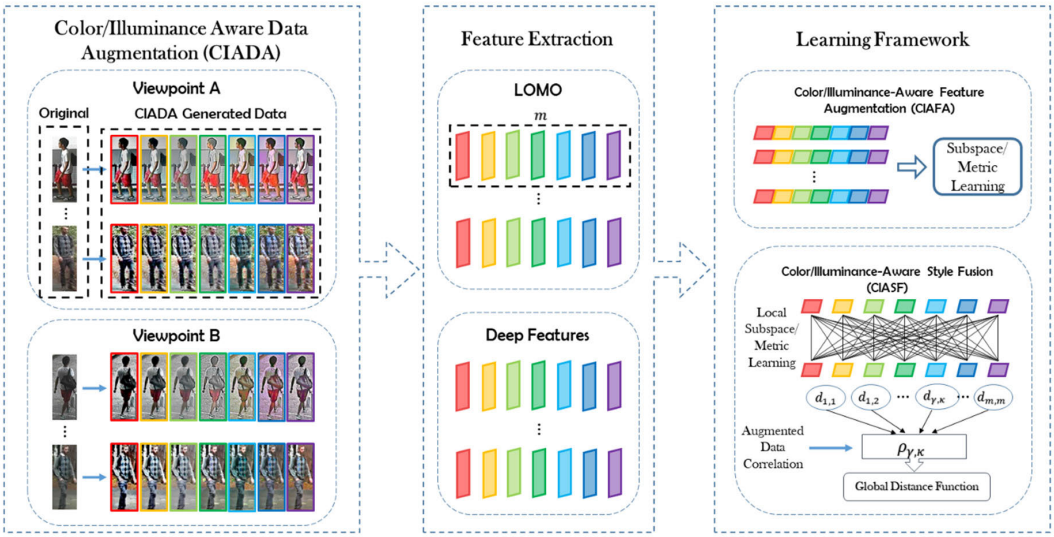


FIGURE 3. (a): Illustration of subspace matching process in single-shot person re-identification: grey dots refer to the mean appearances of objects, and black dots refer to the sampled appearance; dash circle and solid circle enclosed samples from the two views. Superscripts a and b denotes two camera views. Indices i and j denote sample numbers. x denotes feature vector and $\bar{x}$ denotes class mean. (b): Conventional approaches in data augmentation using image transformation and color adjustment. (c): Illustration of subspace matching in multi-shot matching under same protocol of (a): in (a), the single-shot sampled appearance (black dots) may deviate substantially from the mean appearance (grey dots) due to noise and background occluders and thus lower the recognition rate. In (c), multiple samples (color dots) are available and the robustness of the subspace learning and hence recognition accuracy are improved.

where l_i^a is the label of the *i-th* element, $\boldsymbol{s}_i^a$, from view a and l_j^b is the label of its corresponding element $\boldsymbol{s}_j^b$ in view b.

As mentioned earlier, color/illuminance can substantially affect the recognition rate in the person re-id problem. Therefore, data normalization is usually performed before feature extraction. This process can be statistical justified by viewing $\boldsymbol{x}_i^a$ as a random variable and the illuminance-adjusted version $\bar{\boldsymbol{x}}_i^a$ is likely to be closer to the class mean $\mathrm{E}\left(\boldsymbol{x}_i^a\right)$ of the underlying distribution, $\mathrm{P}\left(\boldsymbol{x}_i^a\right)$, thus reducing the variance of the estimated subspace. However, the spread of $\mathrm{P}\left(\boldsymbol{x}_i^a\right)$ is also an important factor which affects the estimated subspace to account for the much larger variations encountered in real-world applications. One possibility to acquire such information is to explore multiple samples of an object captured at different time instances or body angles via multi-shot. However, such multi-shot samples may be unavailable, tedious to obtain, or be very limited in number.

This motivated us to propose a physically motivated method to generate new samples from the limited training data to emulate the multi-shot scenario for improving the recognition rate. More specifically, given an input image $\boldsymbol{s}_i^a$, we utilize the proposed CIADA to generate a set of m new samples $\hat{S}_i^a = \left\{\tilde{s}_i^{a,1}, \tilde{s}_i^{a,2}, \ldots, \tilde{s}_i^{a,m}\right\}$ through certain prior human knowledge on contrast and color perturbation, where $\tilde{s}_i^{a,\gamma} = \boldsymbol{g}_\gamma\left(\boldsymbol{s}_i^a\right)$, and $\boldsymbol{g}_\gamma(\cdot)$ is the $\gamma - th$ augmentation transformation, $\gamma = 1, \ldots, m$. If we augment s_i^a with these physically motivated new samples to form a larger labelled training set $\left\{\boldsymbol{s}_i^a, \hat{S}_i^a\right\}$, more prior information such as geometrical transformation, color, and illuminance between the two cameras can be included to better separate the subjects.

We then augment these new samples $\hat{S}_i^\varphi$, $\varphi = a, b$, to the original set $\boldsymbol{s}_i^\varphi$ to form the augmented data set $\tilde{S}_i^\varphi = \left\{\boldsymbol{s}_i^\varphi, \hat{S}_i^\varphi\right\}$. Since we are now given multiple augmented samples per shot at each viewpoint, the original re-id problem is re-formulated as a special multi-sample problem with the extracted feature $\tilde{x}_i^a = \boldsymbol{f}\left(\tilde{S}_i^\varphi\right)$, which can be solved for the corresponding data augmented distance function $\tilde{d}(i,j) = \tilde{d}\left(\tilde{\ell}\left(\tilde{x}_i^a\right), \tilde{\ell}\left(\tilde{x}_j^b\right)\right)$. Through this approach, the recognition rate of both single-shot and multi-shot setting can be further improved. In the proposed approach, we also explore the correlations of these augmented samples and proposed a kernel XQDA to address the data non-linearity problem, which will be further elaborated in Sections 4 and 5.

III. THE PROPOSED CIADA SCHEME

The proposed CIADA scheme comprising of two label- preserving data augmentation techniques namely: illumination and color perturbation based on color/illuminance appearance prior learnt from the data using manifold learning. They aim to alter respectively the contrast level in intensity and RGB values of the original image to mimic different appearance variations of the original subject. By including these augmented samples as new training samples, an improved description of object appearance can be obtained. Experimental results show that the proposed method can effectively improve the recognition rate of person re-identification by enhancing the salient color/texture features of the subjects.

A. ILLUMINATION/CONTRAST ADJUSTMENT

As shown in Figure 4 (a), the appearance of an object can vary greatly under different illumination in the scene. To account

FIGURE 4. Example of our CIADA scheme. Column (a) is the original image. Column (b)-(d) are samples generated using our MSRCR with parameters [0.75, 0.9, 10], [1.5, 1, 0] and [3, 1.2, 10]. Column (e)-(g) are samples with random color perturbation.

for this variation and possible shading, we propose a new method called Multi-scale Retinex with Contrast Restoration (MSRCR), which is based on the Retinex algorithm [34]. It alters the contrast level of the intensity in the original samples to mimic different camera exposure levels for data augmentation. The Retinex algorithm is a popular image restoration method for local contrast enhancement based on the Retinex theory [35]. It explores human color perception to improve the image contrast of color images. As the single scale Retinex algorithm can only provide either dynamic range compression (small scale) or tonal rendition (large scale) [34], multi-scale version of the Retinex [34] is adopted in the proposed MSRCR algorithm.

More specifically, let the original image sample be $\boldsymbol{I}$, and $\boldsymbol{F}_n(x, y) = \Phi_n \exp\left[-\left(x^2 + y^2\right)/2\sigma_n^2\right]$ be the *n-th* scale normalized surrounding function at pixel location (x, y) with σ_n the corresponding scale parameter and Φ_n the normalization factor such that $\int\int \boldsymbol{F}_n(x, y)dxdy = 1$. The output of the proposed MSRCR algorithm $\boldsymbol{R}$ at location (x, y) is constructed by a weighted sum of N Retinex outputs, each estimates the local illumination levels at a scale σ_n:

$$\begin{aligned}\boldsymbol{R}(x, y) &= \sum_{n=1}^{N} \omega_n \boldsymbol{R}_n(x, y) \\ &= \sum_{n=1}^{N} \omega_n \left[\log(\boldsymbol{I}(x, y)) - \log\left(\boldsymbol{I}(x, y) * \boldsymbol{F}_n(x, y)\right)\right],\end{aligned} \tag{4}$$

where N is the number of scales, $\boldsymbol{R}_n$ is the Retinex output at the *n-th* scale, and ω_n is the importance weight of the *n-th* scale with $\sum_{n=1}^{N} \omega_n = 1$. "$*$" denotes the convolution operation of the original image $\boldsymbol{I}$ and the surrounding function $\boldsymbol{F}_n$. Usually, to achieve a balance between dynamic range compression and color rendition, a combination of small-scale parameter and a larger one is employed. We choose 5 and 60 as the scale parameters σ_n with equal weight in our MSRCR algorithm.

As the Retinex outputs $\boldsymbol{R}_n$, and hence the final output $\boldsymbol{R}$, are in logarithmic scale, they need to be projected back to a linear scale through quantization. However, it was found in [34] that direct quantization of the MSR output can generate unexpected color distortion if the original colors are desaturated. Therefore, we propose a new data-adaptive quantization scheme with a user controllable local contrast parameter, called 'Dynamic' factor δ. More specifically, to get the quantized output from $\boldsymbol{R}$, the dynamic scale is first estimated using the mean μ_R and standard deviation σ_R of $\boldsymbol{R}$. The dynamic factor δ is then introduced to control the spread in $\boldsymbol{R}$ via the effective range of $\boldsymbol{R}$, v_{max} and v_{min}, as follows:

$$v_{\min} = \mu_R - \delta \cdot \sigma_R, v_{\max} = \mu_R + \delta \cdot \sigma_R \tag{5}$$

The larger the dynamic factor, the lower the contrast. To further control the exposure condition, we incorporate the 'Gain' and 'Offset' concept in [34] into our quantized output. The 'Gain' G and 'Offset' b are incorporated over the given range through the following linear relationship:

$$\boldsymbol{I}_R = 255 \cdot G \cdot (\boldsymbol{R} - v_{\min})/(v_{\max} - v_{\min}) - b \tag{6}$$

$\boldsymbol{I}_R$ is truncated to a range of [0, 255] to obtain the final output.

We found that by adjusting the gain/offset parameters and the 'Dynamic' factor, we were able to control the contrast and exposure of the original image to generate various intensity variations of the original samples. Example images obtained by using different parameter settings of $[\delta, G, b]$ are shown in Figure 4 (b)-(d). We can see that the samples generated can mimic considerable intensity variations, which help to enhance the generality of the training samples.

B. MANIFOLD LEARNING-BASED COLOR PERTURBATION

Appearance-based algorithms, which rely on statistical models of gray-scale or color pixels, have attracted increasing attention in object recognition and pose estimation problems [36], [37]. However, if we consider only the color components of an image as a vector in a high-dimensional space spanned by the color value, the dimensionality is usually very large as it is proportional to the image resolution. Such representation is not robust under large illumination change or varying camera parameters and thus resulting in different color appearances.

It has long been recognized that the high dimensional image space can be simplified to a low-dimensional subspace in many applications to alleviate the dimensionality problem.

In fact, data with high dimensionality usually lie in a much lower dimensional subspace locally in form of a nonlinear manifold. In [30], the nonlinear manifold is represented by a mixture of local Gaussian densities with a global coordinate. Motivated by the effectiveness of nonlinear manifold representations, we model the color pixels of the image samples as a nonlinear manifold and aim to model it in forms of a mixture of representative color components. With this representation, we then mimic the color variations between different camera views as perturbation of these representative color components. This is motivated by the fact that different image samples may contain different representative colors. Adding uniform perturbation to the RGB color may not effectively capture such variations. Rather, it is important to extract salient or representative colors in the given samples and perform perturbation in each representative color group.

More specifically, for a set of training images, we per-form Arbitrary HyperPlane Vector Quantization (AHPVQ) [38] on all color pixels and separate them into K groups or clusters (A brief review and experimental results of the Arbitrary HyperPlane Vector Quantization algorithm will be given in the supplementary materials). The AHPVQ algorithm is a hierarchical clustering method that successively partitions the original data by a series of separating hyperplanes to obtain the desired clusters. It aims to minimize the total sum-of-squared errors (SSE) by splitting the cluster which will lead to the largest reduction in SSE. Due to its hierarchical nature, it can generate successive partitions with cluster number less than or equal to the target value and the complexity in increasing the cluster number is very low. Therefore, it can be used directly to estimate the representative color groups or serving as a stable initial guess for further refinement using other sophisticated clustering algorithms to reduce the overall computational requirement. Our approach aims to balance the color/illuminance variance in the datasets via data augmentation. Therefore, we need to estimate the color/illuminance variation of the specific dataset in order to guide the data augmentation. For instance, if the VIPER dataset has basically 3 kinds of color styles due to different weather or camera location, we then wish to augment the image with a specific color style to have a similar style with the other two levels. The RGB image is transformed to YCbCr color space and we measure the style by clustering the image color component in its low-dimensional manifold as the raw image pixel is of very high-dimension.

To obtain the local color manifold model, we first perform PCA on the associated data of *k-th* group to determine their *i-th* principal direction, $\boldsymbol{p}_i^{(k)}$, and associated standard deviation $\lambda_i^{(k)}$, $i = 1,2,3$. To generate each augmented data set via color perturbation, a zero mean random vector, $\boldsymbol{\alpha}^{(k)} \in \Re^3$, is generated for the *k-th* group $k = 1,\ldots, K$, with individual standard deviation δ_c. For the *k-th* group, a uniform perturbation is assigned on all colors in this group as:

$$\hat{\boldsymbol{C}}^{(k)} = \boldsymbol{P}^{(k)}[\alpha_1^{(k)}\lambda_1^{(k)}, \alpha_2^{(k)}\lambda_2^{(k)}, \alpha_3^{(k)}\lambda_3^{(k)}]^T \tag{7}$$

where $\boldsymbol{P}^{(k)} = [\boldsymbol{p}_1^{(k)}, \boldsymbol{p}_2^{(k)}, \boldsymbol{p}_3^{(k)}]$ and $\alpha_i^{(k)}$ is the *i-th* component of $\boldsymbol{\alpha}^{(k)}$. Given a color $\boldsymbol{C} = [C^R, C^G, C^B]^T$ belonging to the *k-th* group, its value after perturbation is $\boldsymbol{C}' = \boldsymbol{C} + \hat{\boldsymbol{C}}^{(k)}$. Thus, all the colors are perturbed by the random vector $\boldsymbol{\alpha}$ and their local color model.

In our experiment, we found that δ_C can be chosen between [0.1, 0.3] to obtain realistic augmented data. The number of cluster number K is set to 15 in our experiments. As we can see from Figure 4 (e-g), the generated samples emphasize the salient color components across camera views while preserving the important texture information for the re-identification process. With the augmented samples of a given object, the person re-id problem is now re-formulated as a specific multi-sample problem with the original shot of an object and its associated augmented samples via color/ illuminance perturbation. To cope with these physically intuitive and correlated data, a new Color/Illuminance-Aware Feature Augmentation (CIAFA) approach is proposed to construct an augmented feature for subsequent metric learning.

IV. THE PROPOSED FEATURE AUGMENTATION AND DISTANCE FUSION METHODS

In subsection 4.1, a global recognition framework called the Color/Illuminance-Aware Feature Augmentation for metric learning methods is first introduced. An alternative approach for locally pair-wise distance fusion called Color/Illuminance-Aware Style Fusion (CIASF) is then proposed in subsection 4.2 together with a pair-wise CCA-based correlation method for constructing the final distance metric. The latter allows the training problem to be solved in parallel with reduced memory complexity.

A. COLOR/ILLUMINANCE-AWARE FEATURE AUGMENTATION (CIAFA)

We have seen in Section IV that the proposed CIADA scheme utilizes two types of non-parametric label-preserving transformation for sample generation and the original data will be enlarged by m times as in a "**multi-sample problem**". Traditional algorithms for the multi-shot re-id problem usually treat the multi-shot samples as "independent samples" within the same class due to the unequal number of samples for each identity. On the contrary, in our CIAFA approach, as the number of generated data for each object is identical (m sets of augmented data and the original data) and each color/illuminance set are used to emulate a certain color/lighting condition. We therefore propose to treat the augmented features as an extension of the original sample's feature and concatenate them together to form a new feature vector. Specifically, the augmented feature $\tilde{\boldsymbol{x}}_i^{\varphi}$ is obtained by stacking the features extracted from the augmented data and the original data as follows:

$$\begin{aligned}\tilde{x}_i^{\varphi} &= \left[\boldsymbol{f}\left(\boldsymbol{g}_0\left(\boldsymbol{s}_i^{\varphi}\right)\right)^T, \boldsymbol{f}\left(\boldsymbol{g}_1\left(\boldsymbol{s}_i^{\varphi}\right)\right)^T, \ldots, \boldsymbol{f}\left(\boldsymbol{g}_m\left(\boldsymbol{s}_i^{\varphi}\right)\right)^T\right]^T \\ &= \left[\left(\tilde{x}_i^{\varphi,0}\right)^T, \left(\tilde{x}_i^{\varphi,1}\right)^T, \ldots, \left(\tilde{x}_i^{\varphi,m}\right)^T\right]^T \end{aligned} \tag{8}$$

where $\boldsymbol{f}\left(\boldsymbol{g}_{\gamma}\left(\boldsymbol{s}_{i}^{\varphi}\right)\right)$ is the ***local augmented feature*** extracted from the γ-*th* augmented transformation of the *i-th* sample $\tilde{s}_{i}^{\varphi,\gamma}$ where $\tilde{s}_{i}^{\varphi,\gamma} = \boldsymbol{g}_{\gamma}\left(\boldsymbol{s}_{i}^{\varphi}\right)$, $\gamma = 0,1,\ldots, m$. For notational convenient, $\gamma = 0$ is used to refer to the original dataset.

Consequently, we get a similar subspace/metric learning problem with a local distance function similar to (3) as follows:

$$\tilde{d}(i,j) = \left(\tilde{\boldsymbol{W}}^{T}\left(\tilde{\boldsymbol{x}}_{i}^{a} - \tilde{\boldsymbol{x}}_{j}^{b}\right)\right)^{T}\tilde{\boldsymbol{M}}\tilde{\boldsymbol{W}}^{T}\left(\tilde{\boldsymbol{x}}_{i}^{a} - \tilde{\boldsymbol{x}}_{j}^{b}\right) \tag{9}$$

where $\tilde{\boldsymbol{M}} \in \Re^{r\times r}$ is the local measurement metric and $\tilde{\boldsymbol{W}} \in \Re^{md\times r}$ is the local subspace transformation matrix on the augmented feature space. Both $\tilde{\boldsymbol{W}}$ and $\tilde{\boldsymbol{M}}$ can be solved by substituting $d(i,j)$ with $\tilde{d}(i,j)$ in (3).

On the other hand, as the number of available data is increased by m times, the computational requirement is substantially increased. We now introduce a local metric learning approach by fusing local distance function from each set of samples with the help of the pair-wise data correlation between the augmented datasets. This allows the classifier/metric to be learned separately.

B. COLOR/ILLUMINANCE-AWARE STYLE FUSION (CIASF)

In the proposed local fusion approach, the feature extraction and recognition steps are applied individually to *each pair of the augmented data set* and the overall distance function can be obtained by fusing the resultant local distances as shown in the bottom of Figure 3 (c). Since the m augmented data set are generated from a single sample, they are correlated statistically. It is therefore important to quantitatively measure the correlation between the features obtained from different pairs of data set so as to determine their contributions to the overall distance. Here, we introduce a CCA-based method to estimate the correlation between them for subsequent fusion.

More specifically, state-of-the-art view-generic metric learning algorithm is performed on the 'local' pair-wise feature $\{\tilde{\boldsymbol{x}}_{i}^{a,\gamma}\}$ and $\{\tilde{\boldsymbol{x}}_{i}^{b,\kappa}\}$ in (9) with $\gamma, \kappa \in \{0, 1, \ldots, m\}$. A 'local' dis-similarity measure is then estimated between the pair-wise sets of local augmented features from the γ-*th* and κ-*th* local transformed samples. Using the formulation in (4), one gets the following "*pair-wise local distance function*" for the γ-*th* and κ-*th* transformed data sets:

$$d_{\gamma,\kappa}(i,j) = (\boldsymbol{W}_{\gamma,\kappa}^{T}(\boldsymbol{x}_{i}^{a,\gamma} - \boldsymbol{x}_{i}^{b,\kappa}))^{T}\boldsymbol{M}_{\gamma,\kappa}\boldsymbol{W}_{\gamma,\kappa}^{T}(\boldsymbol{x}_{i}^{a,\gamma} - \boldsymbol{x}_{i}^{b,\kappa}) \tag{10}$$

where $\boldsymbol{W}_{\gamma K} \in \Re^{d\times r_{\gamma,\kappa}}$ and $\boldsymbol{M}_{\gamma K} \in \Re^{r_{\gamma,\kappa}\times r_{\gamma,\kappa}}$ are the locally learned discriminative subspaces and metric, respectively.

To aggregate these pair-wise local distance functions to form the overall distance, we consider the distances as statistical samples from an underlying distribution and hence they should be appropriately weighted to reflect the dependency between the samples. This is in contrast to traditional multi-shot matching problem where a uniform weighting is commonly used as the samples are assumed to be independent.

It is noticed in [27] that the cross-view correlation may vary due to view-specific settings and this correlation will affect the overall recognition rate. Thus, similar to [27], we measure the correlation between the augmented data using canonical correlation analysis (CCA). Specifically, consider the set of original and local augmented features, $\tilde{\boldsymbol{X}}^{\varphi,\gamma} = [\tilde{\boldsymbol{x}}_{1}^{\varphi,\gamma}, \ldots, \tilde{\boldsymbol{x}}_{N_{\varphi}}^{\varphi,\gamma}]^{T}$ from the probe view $\varphi = a$ and gallery view $\varphi = b$ with $\gamma, \kappa = 0,\ldots,m$. In linear subspace modeling of the data, they are assumed to lie on a linear subspace or Grassmann manifold with projection matrices $\boldsymbol{A}_{a,\gamma} \in \Re^{N_a\times\eta}$ and $\boldsymbol{B}_{b,\kappa} \in \Re^{N_b\times\eta}$. The corresponding projection scores of $\tilde{\boldsymbol{X}}^{\varphi,\gamma}$, $\varphi = a, b$, are denoted by $\tilde{\boldsymbol{G}}_{\varphi,\gamma} \in \Re^{N_{\varphi}\times\eta}$ and $\tilde{\boldsymbol{G}}_{\varphi,\kappa} \in \Re^{N_{\varphi}\times\eta}$ where η is the dimension of the subspace. The principal angles between the subspaces are related to the manifold geodesic distance with a good geometrical interpretation. Therefore, the similarity of these two sets of manifold points can be measured by the cosines of their principal angles θ_t, $(t = 1,\ldots, \eta)$. In fact, the cosines of these principal angles are also referred to as canonical correlation coefficients in CCA. Consequently, their similarity can be measured from the canonical correlation coefficients which can be obtained by maximizing the correlations of $\tilde{\boldsymbol{G}}^{a,\gamma}$ and $\tilde{\boldsymbol{G}}^{b,\kappa}$ with respect to their projections $\boldsymbol{A}_{a,\gamma}$ and $\boldsymbol{B}_{b,\kappa}$ as follows:

$$\boldsymbol{\rho} = \mathrm{Corr}\left(\boldsymbol{A}_{a,\gamma}^{T}\tilde{\boldsymbol{G}}_{a,\gamma}, \boldsymbol{B}_{b,\kappa}^{T}\tilde{\boldsymbol{G}}_{b,\kappa}\right) \tag{11}$$

One may then compute $\boldsymbol{\rho}$ from the Singular Value Decomposition (SVD) of the matrix:

$$(\tilde{\boldsymbol{G}}_{a,\gamma})^{T}\tilde{\boldsymbol{G}}_{b,\kappa} = \boldsymbol{A}_{a,\gamma}\boldsymbol{\Lambda}_{\gamma,\kappa}\boldsymbol{B}_{b,\boldsymbol{k}}^{T} \tag{12}$$

where $\boldsymbol{\Lambda}_{\gamma,\kappa}$ is a diagonal matrix containing the canonical correlation coefficients. The overall similarity can thus be measured from the sum of canonical correlation coefficients as:

$$\rho_{\gamma,\kappa} = \frac{1}{\eta - 1}Tr(\boldsymbol{\Lambda}_{\gamma,\kappa}) \tag{13}$$

where $Tr(\boldsymbol{Y})$ is the trace of matrix $\boldsymbol{Y}$. $\rho_{\gamma,\kappa}$ is also referred to as the augmented data commonness degree (ADCD) in [27]. In our experiments we empirically set η to 100.

Finally, we obtain the following overall distance function:

$$\bar{d}(i,j) = \sum_{\gamma=0}^{m}\sum_{\kappa=0}^{m}\rho_{\gamma,\kappa}d_{\gamma,\kappa}(i,j) \tag{14}$$

where $\rho_{\gamma,\kappa}$ is the ADCD in (13). As $d_{\gamma,\kappa}(i,j)$ can be computed separately from $\tilde{\boldsymbol{X}}^{a,\gamma}$ and $\tilde{\boldsymbol{X}}^{b,\kappa}$, the learning of the local distances can be done in parallel. Consequently, the above CIASF is more efficient to implement than the CIAFA formulation on large datasets through parallelization. Although the 'local' subspace learning process only relies on the pair-wise sample correlation to exchange information across dimensions, experimental results show that the CIASF can achieve similar performances as CIAFA in most datasets and consistently outperforms the benchmark single-data setting.

More importantly, for traditional metric learning algorithms such as XQDA and NFST, if the features from the original image and CIADA images are concatenated together as in CIAFA, the style information is implicitly added into the

TABLE 1. Comparison of using different data augmentation on viper and CUHK03. feature LOMO is used. Features are concatenated in CIAFA. ROT. refers to rotation, C.N. refers to color noise, IA and CP refers to our proposed CIADA approach.

VIPeR (m=3)

		Rank-1	Rank-5		Rank-1	Rank-5
Ori.	XQDA	35.04	63.96	NFST	36.41	65.72
Rot.		42.13	71.44		43.21	73.04
Crop		39.03	67.85		40.84	70.72
C.N.		41.52	70.57		40.99	70.71
IA		**43.11**	**73.41**		**43.36**	**73.99**
CP		**42.56**	**71.99**		**42.06**	**72.15**

CUHK03 (m=2)

		Rank-1	Rank-5		Rank-1	Rank-5
Ori.	XQDA	52.40	81.95	NFST	55.60	83.10
C.N.		63.80	88.10		69.00	91.80
Rot.		63.70	89.15		70.45	92.05
IA		**66.20**	**90.15**		**70.80**	**92.95**
CP		**65.25**	**89.40**		**69.90**	**92.10**

feature vector (features of each part are consistently extracted from a single style). Hence the metric learning approaches can implicitly adapt to the style information. However, deep-learning based methods usually adopt a single model and therefore the style discrepancy may have adverse effect on model generalization ability. Although CIASF increases the training cost, it can effectively address the style discrepancy problem. As we shall see later from the experimental results that the proposed CIASF is essential to apply our CIADA to deep-learning based model for performance improvement. Moreover, the training can be performed in parallel and hence accelerated using multi-GPUs.

V. EXPERIMENTS

A. DATASETS

The proposed methodology is evaluated using five popular datasets VIPeR [30], CUHK01 [15], CUHK03 [16], Market-1501 [39] and DukeMTMC-reID [40]. Comparisons with state-of-the-arts methods are also included.

VIPeR. VIPeR contains 632 identities with two outdoor views available under various angles. The 632 people are randomly divided into two equal halves for training and testing. Unless stated otherwise, we repeat the random division for 50 times and present the average performance. Meanwhile, 5 sets of augmented samples are used in our experiment (m=5) for all the datasets.

CUHK01. CUHK01 contains 971 identities, each of which has two images taken in a campus environment at two different viewpoints. All images are randomly split into two halves with 485 people for training and 486 for testing as in [41].

CUHK03. CUHK03 contains 1360 persons with over 13,000 images captured by six surveillance cameras. Each person was captured by two distinct views with roughly 5 samples each. The dataset consists of cropped images by labor and by machine, and will be referred to as 'labeled' and 'detected', respectively. We report our results on both sub-datasets. The 20 training/testing splits provided in [16] are used and the settings are the same as [16].

Market-1501. Market-1501 contains over 32000 labeled images from ∼1500 people in 6 camera views. The training set contains ∼13000 images from 751 identities and the testing set contains ∼19700 images from 750 identities. In testing, 3368 samples of 750 people are used as queries and single-query evaluation is used.

DukeMTMC-reID. DukeMTMC-reID is a large-scale re-id dataset with over 36000 labeled images from 8 different camera views and 1404 identities. The training set contains 16522 images from 702 identities and the testing set contains 2228 query images from the rest identities with 17661 database images.

Since CUHK03, Market 1501 and DukeMTMC-reID are large datasets, to save computation time, we only include 2 sets of augmented samples in our experiment (m=2).

B. FEATURES AND RECOGNITION ALGORITHMS

To verify the applicability of the proposed training methodology, we employed the state-of-the-art feature extraction methods namely, Local Maximal Occurrence (LOMO) [6] and the one proposed in [14], which is referred to as 'KCCA*' in subsequent discussion. The LOMO, KCCA, have 26960 and 5138 dimensions, respectively.

For handcrafted features, we choose XQDA [6], NFST [7] as metric learning methods for evaluation as they are state-of-the-art metric learning methods in person re-id problem. We also include results of other metric learning methods provided in their publication for comparison.

For experiments relating to deep features, we employ the framework of Part-aligned Network (Partnet) [3] and ID-discriminative embedding (IDE) [31] for fair comparison as they are the state-of-the-arts in Re-id task.

C. EVALUATION PROTOCOL

The Cumulated Matchin g Characteristics (CMC) is used to evaluate the recognition accuracy. Due to page limitation, we only report the cumulated matching ac-curacy at 4 selected levels (1st, 5th, 10th, and 20th).

D. COMPARISON WITH OTHER DATA AUGMENTATION AND POOLING METHODS

To present the effectiveness of our CIADA approach, we compare the proposed data augmentation approach, illumination adjustment (IA) and color perturbation (CP), with existing data perturbation methods such as rotation, cropping and random color noise described in [18]. In addition, to verify the performance of feature augmentation method in CIAFA, we also compare different pooling methods such as concatenation, max pooling and mean pooling.

Data augmentation via traditional transformation-based approaches are usually employed to avoid over-fitting while preserving useful information. In hand-crafted features, rotation is usually used to perturb the original data so that the human representation is less affected by the un-normalized detection box. Cropping and color noise addition, apart from mitigating over-fitting, usually erase useful information and

TABLE 2. Comparison of using different pooling methods on viper and CUHK03. Feature LOMO is used. CAT. refers to concatenation.

VIPeR (m=3)

XQDA						
	Cat.	Rank-1	Max	Rank-1	Mean	Rank-1
Rot.		42.13		39.35		40.33
Crop		39.03		39.18		38.32
C.N.		41.52		39.98		37.72
IA		**44.09**		**41.43**		**42.48**
CP		**42.56**		**39.60**		**41.12**
NFST						
Rot.		43.21		40.82		42.85
Crop		40.84		41.52		41.10
C.N.		40.99		40.75		39.68
IA		**43.36**		**44.23**		**45.53**
CP		**42.06**		**41.81**		**41.12**

CUHK03 (m=2)

XQDA						
	Cat.	Rank-1	Max	Rank-1	Mean	Rank-1
Rot.		63.70		55.00		46.30
IA		**66.20**		**56.05**		**50.25**
CP		**65.25**		**55.65**		**38.40**
NFST						
Rot.		70.45		64.95		66.70
IA		**70.80**		**61.70**		**65.50**
CP		**69.90**		**64.95**		**65.35**

introduce noise to the system. This can be verified by the experimental results in Table 1, where rotation achieves the best result amongst traditional DA methods. For the proposed CIADA approach, both illumination adjustment and color perturbation successfully improve the system performance. Similar trend can also be found in the CUHK03 dataset. We can also see that, as the number of available training sample increases, the performance gain of DA also increases.

The analysis of different pooling method is less than trivial. We can see from Table 2 that for XQDA, concatenation appears to be the best pooling method for all DA methods in both VIPeR and CUHK03. This can be explained by that the concatenation preserves all the information for recognition and XQDA enjoys the blessing of dimensionality. Mean pooling, comparing with max pooling, reduces noise and mitigates the style difference between augmented data and therefore behaves better. However, for NFST, as it looks for a discriminative null-space, faces the curse of dimensionality. In this case, the performance depends on the balance between size of training data and dimensionality. For VIPeR dataset, where the feature dimension is sufficiently larger than the data size, increasing the dimensionality may not guarentese the improve of performance. But for CUHK03 dataset, where the feature dimension is comparable to the data size, the improvement of the recognizer will rely on the increase of dimensionality.

To sum up, different DA methods and pooling techniques will have different effect on different datasets and recognizers. Based on the above analysis, we believe our proposed CIADA approach with concatenation pooling is effective in reducing over-fitting.

E. EFFECT OF NUMBER OF AUGMENTED DATA

In this subsection, we compare the Rank-1 recognition rate of CIAFA and CIASF under our CIADA approach with different numbers of augmented datasets m. We repeat each algorithm for 20 times. For fairness, we generated 7 augmented datasets and in each trial, we randomly pick m augmented data out of the 7 data sets generated for training and testing.

The experimental results on the selection of augmented data number m are shown in Figure 5. We can see that the proposed CIADA approach effectively improves the recognition accuracy of state-of-the-art metric learning algorithms. As m grows, the recognition rate increases but at a lower rate. This is reasonable as the data augmentation scheme may not provide substantial additional information of the background/subject via VSIC prior.

TABLE 3. Recognition rate on viper and CUHK01 using LO-MO & KCCA against state-of-the-arts.

		VIPeR			CUHK01		
	Rank	1	5	20	1	5	20
LOMO	kCCA [14]	30.16	62.69	86.80	56.30	80.66	93.00
	kLFDA [42]	38.58	69.15	89.15	54.63	80.45	92.02
	MFA [42]	38.67	69.18	89.02	54.79	80.08	92.72
	XQDA	35.04	63.96	88.90	55.24	79.64	92.56
	NFST	36.41	65.72	90.09	**65.43**	**86.05**	**95.64**
KCCA	XQDA	34.80	63.69	86.73	50.96	76.11	91.13
	NFST	**38.72**	**68.94**	**90.57**	57.03	80.38	93.09
CIAFA							
LOMO	XQDA	43.58	73.27	93.56	70.73	86.64	95.17
	NFST	42.46	72.18	93.46	**73.98**	**90.52**	**97.52**
KCCA	XQDA	40.22	68.77	89.78	59.98	80.46	91.70
	NFST	**44.30**	**74.45**	**93.46**	64.94	86.35	95.73
CIASF							
LOMO	XQDA	46.04	75.49	94.37	64.07	85.52	94.89
	NFST	**49.96**	**78.93**	**95.60**	**74.81**	**90.45**	**97.26**
KCCA	XQDA	43.18	72.01	91.28	62.22	83.43	94.50
	NFST	49.34	78.13	95.15	66.53	86.54	95.66

TABLE 4. Recognition rate on CUHK03 using LOMO & KCCA against state-of-the-arts.

		Manual			Detected		
	Rank	1	5	20	1	5	20
LOMO	DeepReID [16]	20.65	51.50	80.00	19.89	50.00	78.50
	Improved Deep [13]	54.74	86.50	98.10	44.96	76.01	93.15
	Metric Ensemble[43]	62.10	89.10	97.80	-	-	-
	XQDA	52.40	81.95	96.65	46.15	77.95	94.00
	NFST	**55.60**	**83.10**	**96.00**	**50.10**	**79.75**	**93.25**
KCCA	XQDA	39.30	70.90	90.35	37.00	66.90	88.65
	NFST	35.35	59.25	81.55	30.80	54.85	76.95
CIAFA							
LOMO	XQDA	66.80	91.30	98.25	59.10	85.40	97.00
	NFST	**70.10**	**92.40**	**98.45**	**67.10**	**88.70**	**97.20**
KCCA	XQDA	54.65	81.50	94.85	49.80	78.35	93.60
	NFST	46.80	72.75	88.80	47.00	72.05	88.70
CIASF							
LOMO	XQDA	59.40	86.40	97.55	49.25	78.75	94.65
	NFST	**70.95**	**91.80**	**97.90**	**63.90**	**87.25**	**96.15**
KCCA	XQDA	45.75	75.90	92.90	42.90	71.40	89.80
	NFST	47.45	71.60	87.50	42.00	64.50	83.90

** The results of the 'Detected' set under Metric Ensemble using LOMO are not available in the paper [43].*

F. COMPARISON WITH STATE-OF-THE-ARTS

We verify the effectiveness of the proposed CIADA approach under both CIAFA and CIASF training approaches with other competitive methods.

Results on smaller datasets. The results for the VIPeR and CUHK01 datasets are showed in Table 3. For VI-PeR, the CIASF further boosts the performance of CIAFA by

TABLE 5. Recognition rate on market-1505 using LOMO against state-of-the-arts.

	Rank-1	Rank-5	Rank-10
WARCA [44]	45.20	68.20	76.00
TMA [45]	47.90	-	-
SCSP [46]	51.90	72.00	79.00
XQDA	41.95	65.05	74.35
NFST	54.81	75.83	82.57
CIAFA			
XQDA	**53.62**	**74.76**	**81.56**
NFST	**57.42**	**78.24**	**84.86**
CIASF			
XQDA	**43.47**	**65.71**	**74.88**
NFST	**57.84**	**77.26**	**84.47**

TABLE 6. Performance comparison of style-adaptive distance fusion on deep feature networks on CUHK03 and market-1501 dataset.

CUHK03 (m=2)	Rank-1	Rank-5	Rank-10
Deep Metric[41]	61.3	88.5	96.0
PersonNet [47]	64.8	89.4	94.9
DCSL [48]	80.2	97.7	99.2
Partnet [3]	85.4	97.6	99.4
Partnet+IA+SADF	88.71	98.42	99.47
Partnet+CP+SADF	88.32	98.42	99.45
Partnet+IA+CP+SADF	**89.42**	**98.70**	**99.61**

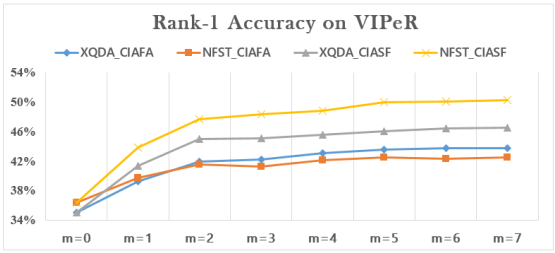
Market-1501 (m=2)	Rank-1	Rank-5	Rank-10
PIE[49]	78.65	90.26	93.59
PEI[49]+KISSME[1]	79.33	90.76	94.41
Partnet [3]	81.0	92.0	94.7
Partnet+IA+CIASF	84.98	93.56	96.02
Partnet+CP+CIASF	83.58	93.47	95.87
Partnet+IA+CP+CIASF	**86.46**	**94.09**	**96.62**
IDE [31]	84.98	93.50	96.35
IDE+IA+CIASF	87.08	94.89	96.59
IDE+CP+CIASF	85.96	93.94	96.38
IDE+IA+CP+CIASF	**87.20**	**94.63**	**96.35**

TABLE 7. Performance comparison of CIADA and CAM-STYLE on market-1501 and DUKEMTMC-reid dataset. re: random erasing.

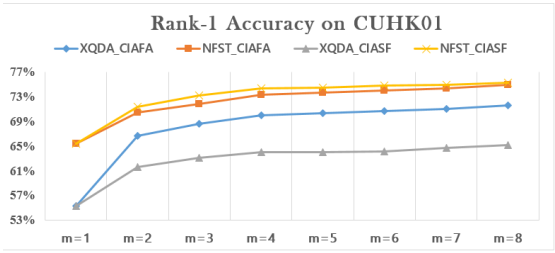
	Rank-1	
	DukeMTMC-reID	Market-1501
IDE [31]	72.31	85.66
IDE+CamSylte	75.27	88.12
IDE+CamSylte+RE [50]	**78.32**	**89.49**
IDE+ IA_CP +NDA	71.49	83.87
IDE+IA_CP+ CIASF	75.00	87.20
IDE+IA_CP+RE+ CIASF	**78.41**	**89.40**

another 3∼5% on both metric learning methods. Similar trends appear on the CUHK01 dataset except for the LOMO feature under XQDA. By comoparing with the state-of-the-arts, we can see that our proposed CIADA approach successfully improves the original method and out-performs state-of-the-arts.

Results on larger datasets. CUHK03 and Market-1501 are two much larger datasets and their results are shown in Tables 4 and 5, respectively. The performances of CIAFA and CIASF are more consistent on these two datasets. Both the global feature augmentation and local style manifold frameworks are effective in boosting the performance of metric learning algorithms. For NFST, both global and local training share similar performances while for XQDA, global training is preferred. This may be due to the increased information available in the augmented features during training, which is also observed in Figure 5(b) where the performance increases substantially with the number of dataset generated.



(a) Rank-1 Accuracy on VIPeR dataset with different number sets of augmented data m.



(b) Rank-1 Accuracy on CUHK01 dataset with different number sets of augmented data m.

FIGURE 5. (a) Rank-1 Accuracy on VIPeR dataset with different number sets of augmented data m. (b) Rank-1 Accuracy on CUHK01 dataset with different number sets of augmented data m.

G. EXTENSION TO DEEP NEURAL NETWORK APPROACHES

We further extend our CIADA approach and style-adaptive fusion framework to deep neural network approaches. To fairly justify our effectiveness, we choose two state-of-the-art methods: Partnet [3] and IDE [31]. We follow the standard procedures as in the original publication to train the original training samples and use the augmented data for fine tuning in each specific set of generated data. At the end, for each method, m different networks are trained for m pairs of augmented data and the local distance function were fused by CIASF, we also conduct the comparison with normal data augmentation (NDA) method which trains a single network from raw data and augmented data.

The experimental results are shown in Table 6. We can see that the CIADA is effective in boosting the overall performance of the original DNN approaches. It also verifies that DA can still play an important role in avoid over-fitting in DNNs.

We further include the comparison of our CIADA with state-of-the-art GAN-based data augmentation method CamStyle [25]. We follow the same experimental procedure in [25] and the results are shown in Table 7. It can be seen that our CIADA shares comparable performance with CamStyle in terms of boosting the recognition rate by reducing over-fitting on large datasets. In addition, the proposed CIADA requires very little training overhead and can be easily adapted to other datasets. Furthermore, the CIASF is essential to apply our CIADA to deep-learning based model for performance improvement. In fact, training a single network with original and augmented data together shows inferior performance than without using data augmentation. This can be explained by the fact that our IA and CP

effectively capture the illuminance/color characteristic of the entire dataset and a single model is difficult to fit such images characteristics with different styles. In [51], a novel unsupervised embedding learning approach which learns the augmentation invariant feature representation was proposed. Similar to the metric learning, the network is learnt to minimize the intra-instance distance while maximizing the inter-category distance. As it is self-supervised, the framework is different from the metric learning approach studied in this paper. The main focus of the present manuscript is the novel data augmentation technique as well as the fusion strategy under the supervised learning setting. As the work in [51] is very promising, it may be possible to integrate the two approaches together, which is left for future work.

VI. CONCLUSION

An effective CIADA and local metric learning approach to person re-identification problem has been presented. The CIADA scheme generates new samples via different color/illumination perturbation and manifold learning. Such perturbation of the original data was found to mitigate considerably the SSS and VSIC problems. A CIAFA approach, which is applicable to conventional and deep features and metric learning algorithms, is developed to integrate the features generated. A color/illumination-aware style fusion scheme is also proposed for learning the 'local' distance functions from each pair of datasets generated to reduce memory requirement in CIAFA. The overall distance measure is obtained by fusing these local distances using a CCA-based scheme. The CIAFA framework shows consistently superior performance over traditional state-of-the-art person re-id methods, at the expense of higher memory and computational requirement, while the CIASF can be applied to much larger dataset. Experiments on commonly used small and large datasets show that the proposed methodologies can improve the performance of state-of-the-arts subspace learning algorithms significantly from 5 to 10%. It shares comparable performance with a GAN-based DA method, CamStyle, but requiring a much lower complexity and is applicable to small dataset.

ACKNOWLEDGMENT

(Zhouchi Lin and Chenyang Liu contributed equally to this work.)

REFERENCES

[1] L.-F. Chen, H.-Y.-M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," *Pattern Recognit.*, vol. 33, no. 10, pp. 1713–1726, Oct. 2000.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[3] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3239–3248.

[4] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2008, pp. 262–275.

[5] L. Bazzani, M. Cristani, and V. Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," *Comput. Vis. Image Understand.*, vol. 117, no. 2, pp. 130–144, Feb. 2013.

[6] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2197–2206.

[7] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1239–1248.

[8] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2666–2672.

[9] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian face recognition," *Pattern Recognit.*, vol. 33, no. 11, pp. 1771–1782, Nov. 2000.

[10] S.-Z. Chen, C.-C. Guo, and J.-H. Lai, "Deep ranking for person re-identification via joint representation learning," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2353–2367, May 2016.

[11] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "End-to-end deep learning for person search," 2016, *arXiv:1604.01850*. [Online]. Available: https://arxiv.org/abs/1604.01850

[12] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1335–1344.

[13] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3908–3916.

[14] G. Lisanti, I. Masi, and A. Del Bimbo, "Matching people across camera views using kernel canonical correlation analysis," in *Proc. Int. Conf. Distrib. Smart Cameras (ICDSC)*, Nov. 2014, pp. 1–6.

[15] W. Li and X. Wang, "Locally aligned feature transforms across views," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3594–3601.

[16] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.

[17] Y.-J. Cho and K.-J. Yoon, "Improving person re-identification via pose-aware multi-shot matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1354–1362.

[18] N. McLaughlin, J. M. Del Rincon, and P. Miller, "Data-augmentation for reducing dataset bias in person re-identification," in *Proc. 12th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2015, pp. 1–6.

[19] A. Fawzi, H. Samulowitz, D. Turaga, and P. Frossard, "Adaptive data augmentation for image classification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3688–3692.

[20] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," 2014, *arXiv:1405.3531*. [Online]. Available: http://arxiv.org/abs/1405.3531

[21] N. Karianakis, J. Dong, and S. Soatto, "An empirical evaluation of current convolutional architectures' ability to manage nuisance location and scale variability," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4442–4451.

[22] X. Wang, M. Liu, D. S. Raychaudhuri, S. Paul, Y. Wang, and A. K. Roy-Chowdhury, "Learning person re-identification models from videos with weak supervision," *IEEE Trans. Image Process.*, vol. 30, pp. 3017–3028, 2021.

[23] M. A. Hossain and B. Assiri, "Facial emotion verification by infrared image," in *Proc. Int. Conf. Emerg. Smart Comput. Informat. (ESCI)*, Mar. 2020, pp. 330–335.

[24] F. Ma, X.-Y. Jing, X. Zhu, Z. Tang, and Z. Peng, "True-color and grayscale video person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 115–129, 2020.

[25] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5157–5166.

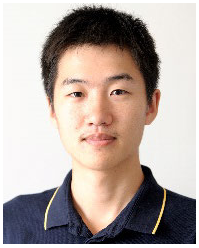
[26] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2242–2251.

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

[28] M. Ye, J. Shen, and L. Shao, "Visible-infrared person re-identification via homogeneous augmented tri-modal learning," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 728–739, 2020.

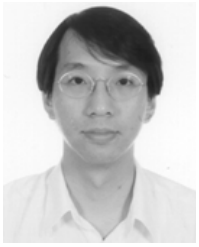
[29] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 26, 2021, doi: 10.1109/TPAMI.2021.3054775.
[30] J. Verbeek, "Learning nonlinear image manifolds by global alignment of local linear models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 8, pp. 1236–1250, Aug. 2006.
[31] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3346–3355.
[32] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2133–2142.
[33] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image Vis. Comput.*, vol. 32, no. 4, pp. 270–286, Apr. 2014.
[34] D. J. Jobson, Z.-U. Rahman, and G. A. Woodell, "A multiscale Retinex for bridging the gap between color images and the human observation of scenes," *IEEE Trans. Image Process.*, vol. 6, no. 7, pp. 965–976, Jul. 1997.
[35] E. H. Land, "The Retinex theory of color vision," *Sci. Amer.*, vol. 237, no. 6, pp. 108–129, 1977.
[36] A. Leonardis and H. Bischof, "Kernel and subspace methods for computer vision," *Pattern Recognit.*, vol. 36, no. 9, pp. 1925–1927, Sep. 2003.
[37] J. Metzler, "Appearance-based re-identification of humans in low-resolution videos using means of covariance descriptors," in *Proc. IEEE 9th Int. Conf. Adv. Video Signal-Based Surveill.*, Sep. 2012, pp. 191–196.
[38] S. C. Chan, C. W. Kok, and S. W. Chau, "Codebook generation and search algorithm for vector quantization using arbitrary hyperplanes," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 1992, pp. 1885–1888.
[39] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.
[40] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3774–3782.
[41] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li, "Embedding deep metric for person re-identification: A study against large variations," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 732–748.
[42] F. Xiong, M. Gou, O. I. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 1–16.
[43] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1846–1855.
[44] C. Jose and F. Fleuret, "Scalable metric learning via weighted approximate rank component analysis," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 875–890.
[45] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury, "Temporal model adaptation for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 858–877.
[46] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1268–1277.
[47] L. Wu, C. Shen, and A. van den Hengel, "PersonNet: Person re-identification with deep convolutional neural networks," 2016, *arXiv:1601.07255*. [Online]. Available: http://arxiv.org/abs/1601.07255
[48] Y. Zhang, X. Li, L. Zhao, and Z. Zhang, "Semantics-aware deep correspondence structure learning for robust person re-identification," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 3545–3551.
[49] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose-invariant embedding for deep person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4500–4509, Sep. 2019.
[50] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 13001–13003.
[51] M. Ye, J. Shen, X. Zhang, P. C. Yuen, and S.-F. Chang, "Augmentation invariant and instance spreading feature for softmax embedding," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Aug. 3, 2020, doi: 10.1109/TPAMI.2020.3013379.

ZHOUCHI LIN received the B.Eng. degree from Sun Yat-sen University, in 2012, and the M.Eng. degree from Cornell University, in 2013. He is currently pursuing the Ph.D. degree with the Department of Electrical and Electronic Engineering, The University of Hong Kong, under the supervision of Prof. S. C. Chan. His research interests include image processing, computer vision, and pattern recognition.



CHENYANG LIU received the B.Eng. degree from Harbin Institute of Technology, in 2016. He is currently pursuing the Ph.D. degree with the Department of Electrical and Electronic Engineering, The University of Hong Kong, under the supervision of Prof. S. C. Chan. His research interests include computer vision, machine learning, and image retrieval.

WENBO QI received the B.Eng. degree from the University of Science and Technology of China, in 2019. He is currently pursuing the Ph.D. degree with the Department of Electrical and Electronic Engineering, The University of Hong Kong, under the supervision of Prof. S. C. Chan. His research interests include computer vision, machine learning, and image processing.



S. C. CHAN (Member, IEEE) received the B.Sc. (Eng.) and Ph.D. degrees from The University of Hong Kong, in 1986 and 1992, respectively. Since 1994, he has been with the Department of Electrical and Electronic Engineering, The University of Hong Kong, where he is currently a Professor. His research interests include fast transform algorithms, filter design and realization, multirate and biomedical signal processing, communications and array signal processing, high-speed A/D converter architecture, bioinformatics, smart grid, and image-based rendering. He is currently a member of the Digital Signal Processing Technical Committee, IEEE Circuits and Systems Society, and an Associate Editor of *Journal of Signal Processing Systems*, *Digital Signal Processing*, and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: EXPRESS BRIEFS. He was the Chair of the IEEE Hong Kong Chapter of Signal Processing, from 2000 to 2002, an Organizing Committee Member of the 2003 IEEE ICASSP and the 2010 IEEE ICIP, and an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS, from 2008 to 2009.