Camera-Conditioned Stable Feature Generation for Isolated Camera Supervised Person Re-IDentification

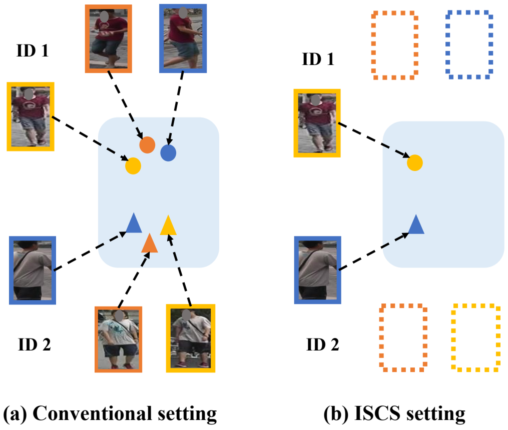
Chao Wu[1], Wenhang Ge[4], Ancong Wu[4], Xiaobin Chang[1,2,3*]
[1]School of Artificial Intelligence, Sun Yat-sen University, China
[2]Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou 510006, P.R.China
[3]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
[4]School of Computer Science and Engineering, Sun Yat-sen University, China
wuch76@mail2.sysu.edu.cn, gewh@mail2.sysu.edu.cn, wuanc@mail.sysu.edu.cn, changxb3@mail.sysu.edu.cn

Abstract

To learn camera-view invariant features for person Re-IDentification (Re-ID), the cross-camera image pairs of each person play an important role. However, such cross-view training samples could be unavailable under the ISolated Camera Supervised (ISCS) setting, e.g., a surveillance system deployed across distant scenes. To handle this challenging problem, a new pipeline is introduced by synthesizing the cross-camera samples in the feature space for model training. Specifically, the feature encoder and generator are end-to-end optimized under a novel method, Camera-Conditioned Stable Feature Generation (CCSFG). Its joint learning procedure raises concern on the stability of generative model training. Therefore, a new feature generator, σ-Regularized Conditional Variational Autoencoder (σ-Reg. CVAE), is proposed with theoretical and experimental analysis on its robustness. Extensive experiments on two ISCS person Re-ID datasets demonstrate the superiority of our CCSFG to the competitors. [1]

1. Introduction

Person re-identification (Re-ID) aims to retrieve the same person across different cameras in a surveillance network. Extracting the discriminative view-invariant features of person images play a central role for the Re-ID task. With the cross-camera images of each person available during training, existing methods have made great progress under different settings, e.g., the supervised [2, 38, 42, 50] and the unsupervised [5, 21, 23, 40]. The importance of cross-camera samples for model training is also demonstrated. However, such cross-camera person images are not guaranteed during training under some realistic scenarios. For example, a surveillance system is needed to re-identify a person across distant scenes, e.g., different cities, and each camera is isolated. It is too expensive to collect sufficient cross-camera person images for model training. A more applicable solution is exploiting the large amount of camera-specific images of different persons instead. As the cross-camera image pairs no longer exist during training, many existing methods [17, 38, 50, 57] fail to obtain the ideal performance on such data. This challenging person Re-ID setting, called ISolated Camera Supervised (ISCS), is first proposed by [52] as Single-Camera-Training (SCT). The comparison between different settings is shown in Fig. 1.



(a) Conventional setting (b) ISCS setting

Figure 1. Illustrations of the training samples under different person Re-ID settings. The light-blue areas indicate the feature space. Different shapes corresponding to identities. Different colors mean under different cameras. (a) Cross-camera person images are available under the conventional settings. (b) No cross-camera image pairs under the ISCS setting for model training.

To handle the challenging ISCS settings, existing methods [10, 52] explicitly align the feature distributions across cameras with new losses and network architectures. In this

*indicates corresponding author.
[1] https://github.com/ftd-Wuchao/CCSFG

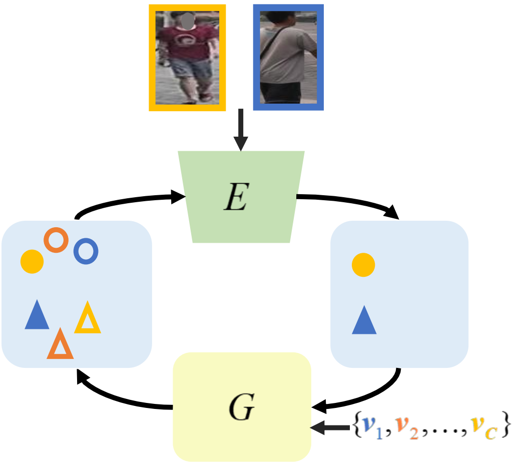


Figure 2. Our pipeline for the ISCS setting. The dots with different shapes represent the features of different IDs. Different colors mean different camera views $\{v_1, v_2, \ldots, v_C\}$. The feature samples conditioned on different camera views are generated by the generator G, shown as the non-solid dots. The cross-camera pairs are thus available for better training of the encoder E.

paper, we follow an alternative pipeline based on generation. The motivation behind this is rather straightforward: As the cross-camera samples play an important role in person Re-ID model training while such paired images do not exist under the ISCS setting, the missing camera view data can be compensated by the generated ones. Specifically, the cross-camera samples are generated in the feature space rather than as images with two considerations. Firstly, it takes great efforts of the generative model to capture details, e.g, backgrounds and illuminations, to improve the visual quality of images. The payoffs of such efforts may not directly reflect on Re-ID and the not ideal generated images can even harm the performance. On the contrary, the feature generation is not distracted by visual quality and is more concentrated on introducing camera-view information while preserving the discriminative power of the generated samples. With the camera-conditioned features of different persons generated, the cross-camera samples are recovered and can be used to train a better encoder of person images. To sum up, a new pipeline is introduced to handle the ISCS setting by synthesizing the cross-camera samples for better encoder training, as illustrated in Fig. 2.

To instantiate the pipeline above, a novel method, Camera-Conditioned Stable Feature Generation (CCSFG), is proposed. A common CNN backbone is used as the image encoder E and a camera conditioned variational autoencoder (CVAE) as the feature generator G. As the encoder E and the generator G are not ideal at first, they should be jointly optimized for iterative improvements. On the one hand, with the more reliable features conditioned on cameras generated by G, the person appearance features from E can be more discriminative and less variant across cameras. On the other hand, with the more discriminative features from E, the generator G can be more focused on capturing the camera information. However, this joint learning procedure forms an obstacle in training the generator G. The input of G is the output of encoder E that is still under training. Therefore, the enlarging dynamic variance of such input causes instability in training G and eventually leads to collapsed learning of the whole model. To handle this issue, a novel generative model, σ-Regularized CVAE (σ-Reg. CVAE), is proposed with a simple yet effective solution based on feature normalization and used as the generator G. More importantly, we provide the theoretical analysis and demonstrate it with experiments.

The main contributions of this paper are in three-fold. (1) To handle the challenging ISCS person Re-ID problem, a novel pipeline is proposed to explicitly generate the cross-view samples in the feature space for better encoder learning. (2) Following the pipeline above, a novel method, CCSFG, is instantiated. The encoder E and generator G are jointly optimized for iterative improvements. (3) To achieve stable joint learning in CCSFG, a novel generative model, σ-Reg. CVAE, is proposed with detailed analysis provided. The effectiveness of the proposed CCSFG is demonstrated by its state-of-the-art performance on two ISCS person Re-ID benchmark datasets.

2. Related Work

Person Re-ID Settings. To study the varied application scenarios of person re-id, different benchmark settings have been proposed for research. The person images in a dataset are usually assumed to be captured from a surveillance network with adjacent cameras but disjoint monitoring areas. Under the **supervised setting** [2,38,41,48,50], their identities are elaborated labeled and aligned across different cameras as supervision. The **unsupervised setting** [5,9,21,40,53] is more challenging than the supervised one by abandoning all the ID labels for model training. To help learning a model on the unsupervised target dataset, the extra source labeled data is available under the **unsupervised domain adaptation (UDA)** setting [13,16,36,49,51]. Moreover, the **intra-camera supervised (ICS)** setting [22, 29, 46, 47, 59] provides the camera-specific ID labels and without a global correspondence across cameras. All these settings are with cross-camera images of each person for model training. Their differences lie in the supervision extent and manners. The recent proposed **ISolated Camera Supervised (ISCS)** person Re-ID setting [10, 52] focuses on a distinctive scenario where no cross-camera person images are available for model training. Therefore, to learn the view-invariant models, existing methods [10,52] handle this challenging setting with the alignment losses on the data distributions rather than the sample pairs of different cameras.

In this paper, the alternative pipeline based on generation is proposed under a simple and sound motivation: to recover the crucial cross-camera samples and use them for enhancing the model training. To synthesis the person images under new camera views, existing generative methods, e.g., HHL [57], can be exploited. Our proposed method, CCSFG, is based on cross-camera feature generation instead. As a unified model, its image encoder E and feature generator G are end-to-end optimized for mutual improvement. To achieve stable joint learning, a novel feature generator, σ-Reg. CVAE, is proposed.

Generative Models. The Variational Autoencoders (VAE) [18] and Generative Adversarial Networks (GAN) [11] are two widely exploited generative methods for computer vision problems, such as medical image segmentation [1, 34], latent representations disentanglement [8, 12] and image background modeling [19, 33]. The GAN and VAE based methods also play important role in person Re-ID problems. Different GAN-based methods have been proposed to augment the training person images under the supervised setting [24, 30, 55, 58]. To bridge the domain gaps of different datasets under the UDA setting, the GAN-based methods [6, 7, 15, 25, 57] are proposed to transfer person image styles across domains. Existing VAE-based methods [28, 31] for person Re-ID mainly focus on disentangled representation learning rather than explicitly generating samples. To our best knowledge, the proposed σ-Reg. CVAE is the first VAE-based feature generator for the ISCS person Re-ID. The generator σ-Reg. CVAE and the encoder are unified under our CCSFG method for joint learning. However, the input for generator in CCSFG is the output of encoder that is still under training. Therefore, the huge dynamic variance of such inputs causes instability in training G and eventually leads to collapsed learning [4, 32, 39]. With theoretical and experimental analysis, a simple yet effective solution to this issue is proposed and incorporated by σ-Reg. CVAE.

3. Methodology

3.1. Isolated Camera Supervised Person Re-ID

The training set is denoted as $\mathcal{D} = \{(\boldsymbol{x}_n, y_n, c_n)\}_{n=1}^{|\mathcal{D}|}$, where each training sample is a triplet with the person image $\boldsymbol{x}_n$, its identity label $y_n \in \{p_1, \ldots, p_M\}$ and the camera label $c_n \in \{v_1, \ldots, v_C\}$. M and C denote the total numbers of different identities and camera views for training respectively. Under the ISCS setting, the cross-camera images of the same person do not exist in the training set, i.e., $\forall i, j \in \{1, ..., |\mathcal{D}|\}$, if $c_i \neq c_j$, then $y_i \neq y_j$. The testing protocol follows the regular routine. Given a query image of a pedestrian, a Re-ID model aims to retrieve the images of the same person from the gallery set.

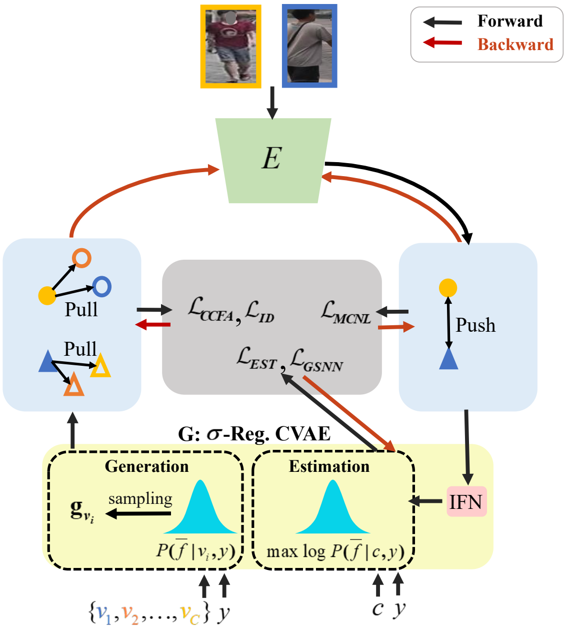


Figure 3. Overview of the Camera-Conditioned Stable Feature Generation (CCSFG) method. E denotes the encoder to extract features from pedestrian images. The features $\{\boldsymbol{g}_{v_i}\}_{i=1}^{C}$ are generated by G conditioning on the identity y and camera views $\{v_i\}_{i=1}^{C}$. IFN denotes the normalization module of generator G.

3.2. Camera-Conditioned Stable Feature Generation

We propose a new method, Camera-Conditioned Stable Feature Generation (CCSFG), for the ISCS person Re-ID problem. The model structure is illustrated in Fig. 3. It consists of two components, the person image encoder E and the feature generator G. The common person Re-ID backbones, e.g., Resnet-50, can be used as E and our novel σ-Reg. CVAE as generator G. During training, E and G are jointly optimized for mutual improvements. On the one hand, the generator G is trained with the feature extracted by E as input, which is encoded by the loss $\mathcal{L}_{G|E}$. On the other hand, the features across different cameras $\{v_1, \ldots, v_C\}$ are generated by G for the training of E. This is modeled as the loss $\mathcal{L}_{E|G}$. The overall training objective of CCSFG thus becomes,

$$\min_{G,E} \mathcal{L}_{(G,E)} = \mathcal{L}_{G|E} + \mathcal{L}_{E|G} \, , \tag{1}$$

where E and G are end-to-end optimized with $\mathcal{L}_{(G,E)}$.

To simplify the presentation of the training procedure, one training sample $(\boldsymbol{x}, y, c)$ is considered by default. The extension to mini-batch is straightforward. The appearance feature of person image $\boldsymbol{x}$ is extracted by encoder E,

$$\boldsymbol{f} = E(\boldsymbol{x}), \tag{2}$$

where $\boldsymbol{f} \in \mathbb{R}^d$.

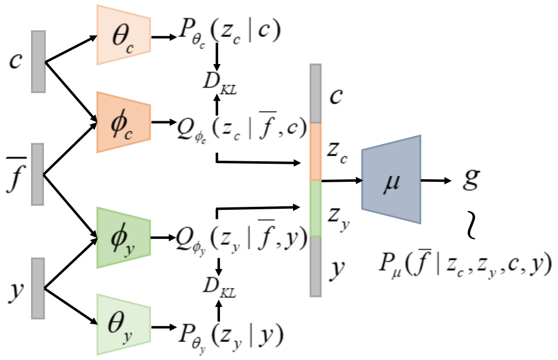

c
θ_c
$P_{\theta_c}(z_c \mid c)$
D_{KL}
ϕ_c
$Q_{\phi_c}(z_c \mid \bar{f}, c)$
$\bar{f}$
ϕ_y
$Q_{\phi_y}(z_y \mid \bar{f}, y)$
D_{KL}
y
θ_y
$P_{\theta_y}(z_y \mid y)$
c
z_c
z_y
y
μ
g
$\wr$
$P_\mu(\bar{f} \mid z_c, z_y, c, y)$

Figure 4. Illustration of the training (estimation) process of σ-Reg. CVAE. The generated feature g obeys the decoding distribution $P_\mu(\bar{\boldsymbol{f}}|z_c, z_y, c, y)$ given the feature $\bar{f}$, identity label y and camera label c.

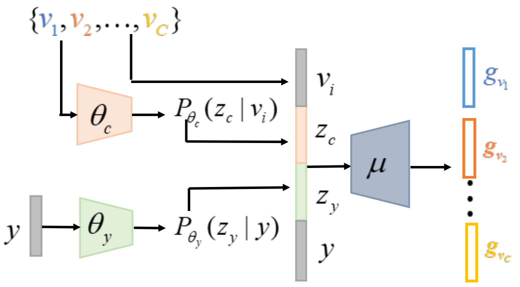

$\{v_1, v_2, \ldots, v_C\}$
v_i
θ_c
$P_{\theta_c}(z_c \mid v_i)$
z_c
μ
g_{v_1}
g_{v_2}
z_y
y
θ_y
$P_{\theta_y}(z_y \mid y)$
y
g_{v_C}

Figure 5. Illustration of the generation process of σ-Reg. CVAE. Conditioning on the identity label y and different camera views $\{v_i\}_{i=1}^{C}$ as inputs, the corresponding generated features $\{\boldsymbol{g}_{v_i}\}_{i=1}^{C}$ are obtained from the decoder network μ.

Given the image feature $\boldsymbol{f}$ under one camera, we expect the image features of the same person under other cameras to be generated. Therefore, the proposed generator G is built on the conditional variational autoencoder (CVAE) architecture for the convenience of introducing the side information, such as the identity y and camera view c.

Training (Estimation) of G. To learn the parameters of our generator σ-Reg. CVAE, the loss $\mathcal{L}_{G|E}$ is used. It corresponds to the Estimation phase of G, as shown in Fig. 3. Specifically, image feature $\boldsymbol{f}$ is the input of σ-Reg. CVAE and the normalized feature $\bar{\boldsymbol{f}}$ is obtained,

$$\bar{\boldsymbol{f}} = \mathrm{IFN}(\boldsymbol{f}), \tag{3}$$

where $\mathrm{IFN}(\cdot)$ is a normalization function in σ-Reg. CVAE. It plays the important role in the stable joint learning of CCSFG and will be discussed in greater details in Sec. 3.3.

The direct learning objective of σ-Reg. CVAE is maximizing the conditional log-likelihood $\log P(\bar{\boldsymbol{f}}|c, y)$, which is often intractable. Its variational lower bound is optimized instead by introducing latent variables [35]. Specially, $\boldsymbol{z}_y$ and $\boldsymbol{z}_c$ are the two latent variables introduced in the σ-Reg. CVAE and correspond to the given identity condition y and camera condition c respectively. Their prior distributions, $P_{\theta_y}(\boldsymbol{z}_y|y)$ and $P_{\theta_c}(\boldsymbol{z}_c|c)$, are modeled with two prior networks θ_y and θ_c. The two recognition networks ϕ_y and ϕ_c map $(\bar{\boldsymbol{f}}, y)$ and $(\bar{\boldsymbol{f}}, c)$ to their posterior distribution $Q_{\phi_y}(\boldsymbol{z}_y|\bar{\boldsymbol{f}}, y)$ and $Q_{\phi_c}(\boldsymbol{z}_c|\bar{\boldsymbol{f}}, c)$. Moreover, the decoding distribution $P_\mu(\bar{\boldsymbol{f}}|\boldsymbol{z}_c, \boldsymbol{z}_y, c, y)$ is modeled by the decoder network μ. Based on the previous sub-networks and distributions defined, an applicable learning objective of σ-Reg. CVAE is,

$$\begin{aligned}
&\mathcal{L}_{\text{EST}}(\bar{\boldsymbol{f}}, y, c \mid \theta_y, \theta_c, \phi_y, \phi_c, \mu) = \\
&\mathbb{E}_{Q_{\phi_y}(\boldsymbol{z}_y|\bar{\boldsymbol{f}},y)Q_{\phi_c}(\boldsymbol{z}_c|\bar{\boldsymbol{f}},c)}[-\log P_\mu(\bar{\boldsymbol{f}}|\boldsymbol{z}_c, \boldsymbol{z}_y, c, y)] \\
&+ D_{KL}(Q_{\phi_y}(\boldsymbol{z}_y|\bar{\boldsymbol{f}}, y)||P_{\theta_y}(\boldsymbol{z}_y|y)) \\
&+ D_{KL}(Q_{\phi_c}(\boldsymbol{z}_c|\bar{\boldsymbol{f}}, c)||P_{\theta_c}(\boldsymbol{z}_c|c)),
\end{aligned} \tag{4}$$

where D_{KL} denotes the Kullback-Leibler divergence. The construction of this loss is depicted in Fig. 4.

The latent variables $\boldsymbol{z}_y$ and $\boldsymbol{z}_c$ are sampled from different distributions across stages. During training $\boldsymbol{z}_c \sim Q_{\phi_c}(\boldsymbol{z}_c|\bar{\boldsymbol{f}}, c)$ and $\boldsymbol{z}_y \sim Q_{\phi_y}(\boldsymbol{z}_y|\bar{\boldsymbol{f}}, y)$ while testing $\boldsymbol{z}_c \sim P_{\theta_c}(\boldsymbol{z}_c|c)$ and $\boldsymbol{z}_y \sim P_{\theta_y}(\boldsymbol{z}_y|y)$. Such inconsistency can harm the quality of the generated feature samples. The Gaussian Stochastic Neural Network (GSNN) [35] method is exploited to alleviate this issue with a loss,

$$\begin{aligned}
&\mathcal{L}_{GSNN}(\bar{\boldsymbol{f}}, y, c \mid \theta_y, \theta_c) = \\
&\mathbb{E}_{P_{\theta_y}(\boldsymbol{z}_y|y)P_{\theta_c}(\boldsymbol{z}_c|c)}[-\log P_\mu(\bar{\boldsymbol{f}}|\boldsymbol{z}_c, \boldsymbol{z}_y, c, y)].
\end{aligned} \tag{5}$$

The overall objective for the estimation of the generator σ-Reg. CVAE is,

$$\min_{\theta_y, \theta_c, \phi_y, \phi_c, \mu} \mathcal{L}_{G|E} = \alpha \mathcal{L}_{\text{EST}} + (1 - \alpha)\mathcal{L}_{GSNN}, \tag{6}$$

with α as the balancing hyper-parameter.

Training of E. The cross-camera images of the same person play the central role in training image encoder E but not available under the ISCS setting. The feature samples of a person under different camera views are compensated from our generator G, σ-Reg. CVAE, for the training of encoder E. Therefore, $\mathcal{L}_{E|G}$ is used to indicate the overall training loss of E.

To obtain the synthesized features, the proposed σ-Reg. CVAE is functioned under the Generation mode, as illustrated in Fig. 5. With an input feature $\bar{\boldsymbol{f}}$, its person identity label y, the camera views v_i, $v_i \in \{v_1, ..., v_C\}$ and the latent variables $\boldsymbol{z}_y \sim P_{\theta_y}(\boldsymbol{z}_y|y)$ and $\boldsymbol{z}_c \sim P_{\theta_c}(\boldsymbol{z}_c|v_i)$ given, camera-conditioned features $\boldsymbol{g}$ can be generated from the decoder network μ of G,

$$\boldsymbol{g}_{v_i} = \mu(\boldsymbol{z}_c, \boldsymbol{z}_y, v_i, y). \tag{7}$$

Different $\{\boldsymbol{g}_{v_i}\}_{i=1}^{C}$ are generated by keeping the identity label y the same and traversing over cameras v_i. Therefore, $\bar{\boldsymbol{f}}$ and its corresponding generated features $\{\boldsymbol{g}_{v_i}\}_{i=1}^{C}$ form the cross-camera samples of the same person.

Different discriminative loss can then be applied on $\bar{\boldsymbol{f}}$ and $\{\boldsymbol{g}_{v_i}\}_{i=1}^{C}$ for the learning of encoder E. On the one hand, conditioning on the identity label y and camera views v_i, $i = 1, ..., C$, the generated features $\{\boldsymbol{g}_{v_i}\}_{i=1}^{C}$ from the generator σ-Reg. CVAE are ID discriminative and camera view specified. However, an ideal encoder E should extract the discriminative and view-invariant feature $\bar{\boldsymbol{f}}$ from a person image. To achieve this goal, the averaged distance between an image feature $\bar{\boldsymbol{f}}$ and the corresponding $\{\boldsymbol{g}_{v_i}\}_{i=1}^{C}$ should be minimized. By pulling $\bar{\boldsymbol{f}}$ towards different $\boldsymbol{g}_{v_i}$ can not only preserves their id distinctive information but also eliminating the camera-view dependent information in $\bar{\boldsymbol{f}}$. We propose a novel Cross-Camera Feature Align (CCFA) loss for the purpose described above,

$$\mathcal{L}_{CCFA}(\bar{\boldsymbol{f}} \mid G) = \frac{1}{C}\sum_{i=1}^{C} ||\bar{\boldsymbol{f}} - \boldsymbol{g}_{v_i}||^2, \tag{8}$$

where $||\cdot||$ denotes the feature norm. This loss is for the learning of encoder E only.

On the other hand, the cross-entropy loss $\mathcal{L}_{ID}$ is used,

$$\mathcal{L}_{ID}(y, \bar{\boldsymbol{f}} \mid G) = -\log(\boldsymbol{q}[y]), \tag{9}$$

where $\boldsymbol{q}[y]$ denotes the identity predictions of $\bar{\boldsymbol{f}}$ on the ground-truth y.

Moreover, the MCNL loss [52], denoted as $\mathcal{L}_{MCNL}$, is also exploited for the feature similarity learning on the extracted person image features by E. By aggregating the training losses for E, the overall loss $\mathcal{L}_{E|G}$ is,

$$\mathcal{L}_{E|G} = \lambda_1\mathcal{L}_{CCFA} + \lambda_2\mathcal{L}_{ID} + \lambda_3\mathcal{L}_{MCNL}, \tag{10}$$

where λ_1, λ_2 and λ_3 are balancing hype-parameters.

The encoder E and our generator G, σ-Reg. CVAE, are end-to-end optimized with the following loss,

$$\begin{aligned}\mathcal{L}_{(G,E)} &= \alpha\mathcal{L}_{\text{EST}} + (1-\alpha)\mathcal{L}_{GSNN} \\ &+ \lambda_1\mathcal{L}_{CCFA} + \lambda_2\mathcal{L}_{ID} + \lambda_3\mathcal{L}_{MCNL}.\end{aligned} \tag{11}$$

During testing, the person appearance feature are extracted by the learned encoder E for retrieving.

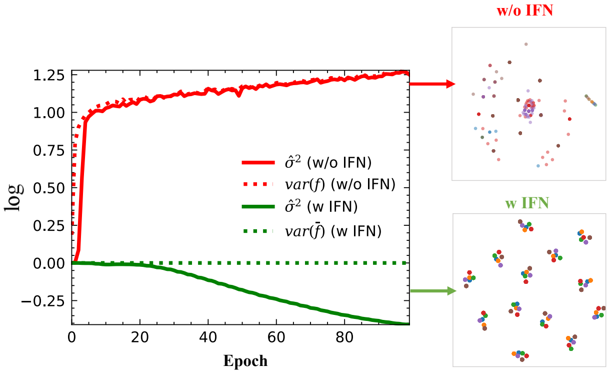


Figure 6. The curves on the left are the values of $\hat{\sigma}^2$ and $var(\boldsymbol{f})$ of the generative models with and without IFN during training on Market-SCT. The feature map, w/o IFN, on the top-right corresponds to the *failure* training cases. The feature map, w IFN, at the bottom-right corresponds to the *successful* joint learning. Each dot in a feature map is a generated feature of a person and the colors indicate the different conditioning camera views.

3.3. The stability in training G

In the proposed CCSFG, the encoder E and generator G are jointly learned. The image feature $\boldsymbol{f}$ is extracted by E and used as the input for training G. However, $\boldsymbol{f}$ will be drastically changed across training steps as its encoder E is also training. Based on such inputs, the optimization on generator G can easily fail and ruin the whole learning procedure. As illustrated in Fig. 6, the generated features from the failure training are compared with the successful one. The meaningless scattered features are produced by the failure case (top-right) while the more ideal features with the clear and meaningful clusters (IDs) are from the stable training (bottom-right).

To achieve the successful joint learning of G and E, the proposed σ-Reg. CVAE is exploited as generator G. Its $\text{IFN}(\cdot)$ module normalizes the image feature $\boldsymbol{f}$ to $\bar{\boldsymbol{f}}$ as input for feature generation and plays the key role in stabilizing the whole training process. Without $\text{IFN}(\cdot)$, the σ-Reg. CVAE degenerates to a conventional CVAE using the image feature $\boldsymbol{f}$ as input. Joint learning of image encoder E and the conventional CVAE as G will end up with posterior collapse. The explanations on them are provided as follows.

Considering the conventional CVAE (w/o IFN) is used as the generator G. Its learning objective follows Eq. (4) as $\mathcal{L}_{\text{EST}}(\boldsymbol{f}, y, c \mid \theta_y, \theta_c, \phi_y, \phi_c, \mu)$ with $\boldsymbol{f}$ rather than $\bar{\boldsymbol{f}}$ as input. Following [4, 32, 39], this loss can be rewritten as,

$$\begin{aligned}\mathcal{L}_{\text{EST}} =& \frac{d}{\sigma^2}||\boldsymbol{f} - \boldsymbol{g}||^2 + \frac{d}{2}\ln\sigma^2 \\ &+ D_{KL}(Q_{\phi_y}(\boldsymbol{z}_y|\boldsymbol{f}, y)||P_{\theta_y}(\boldsymbol{z}_y|y)) \\ &+ D_{KL}(Q_{\phi_c}(\boldsymbol{z}_c|\boldsymbol{f}, c)||P_{\theta_c}(\boldsymbol{z}_c|c)),\end{aligned} \tag{12}$$

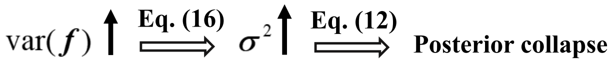


Figure 7. The impacts of huge $var(\boldsymbol{f})$ on the training of CVAE.

by replacing $\mu(\boldsymbol{z}_c, \boldsymbol{z}_y, c, y)$ with $\boldsymbol{g}$ as in Eq. (7) and assuming the decoding distribution $P_\mu(\boldsymbol{f}|\boldsymbol{z}_c, \boldsymbol{z}_y, c, y)$ as an Isotropic Gaussian distribution,

$$P_\mu(\boldsymbol{f}|\boldsymbol{z}_c, \boldsymbol{z}_y, c, y) = \mathcal{N}(\boldsymbol{g}, \sigma^2 I), \tag{13}$$

where d in Eq. (12) is the feature dimension of $\boldsymbol{f}$. Eq. (13) assumes the input feature $\boldsymbol{f}$ obeys the Isotropic Gaussian distribution with mean as the generated feature $\boldsymbol{g}$ and a variance value as σ^2. With the feature pairs $(\boldsymbol{f}, \boldsymbol{g})$ available, σ^2 is estimated via,

$$\hat{\sigma}^2 = \frac{1}{d}\mathbb{E}(||\boldsymbol{f} - \boldsymbol{g}||^2). \tag{14}$$

Moreover, the input feature $\boldsymbol{f}$ has its own variance $var(\boldsymbol{f})$ defined as,

$$var(f) = \frac{1}{d}\mathbb{E}(||\boldsymbol{f} - \mathbb{E}(\boldsymbol{f})||^2). \tag{15}$$

With the alignment loss $\mathcal{L}_{CCFA}$ (Eq. (8)) for the training of E the reconstruction term in $\mathcal{L}_{\text{EST}}$ (Eq. (12)) for the training of G, the intrinsic characteristics of input $\boldsymbol{f}$, e.g., $\mathbb{E}(\boldsymbol{f})$, can be captured by the decoder network μ (for $\boldsymbol{g}$ generation as in Eq. (7)) in CVAE, i.e., $\boldsymbol{g} \approx \mathbb{E}(\boldsymbol{f})$, and thus,

$$\sigma^2 \approx var(\boldsymbol{f}). \tag{16}$$

In the joint learning procedure, the training of encoder E leads to the rapid changes in image features and thus huge $var(\boldsymbol{f})$ occurs. Without normalization on the input $\boldsymbol{f}$, CVAE captures such variance and results in large σ^2 as in Eq. (16). A concrete example of the joint learning of E and CVAE on a ISCS dataset is shown in Fig. 6 (left), where σ^2 is approximated by $\hat{\sigma}^2$. As the red curves shown, the values of both $var(\boldsymbol{f})$ and $\hat{\sigma}^2$ rise drastically as expected. However, large value of σ^2 prevents the CVAE to learn from the its input as the weight $\frac{d}{\sigma^2}$ on reconstruction term $||\boldsymbol{f} - \boldsymbol{g}||^2$ in $\mathcal{L}_{\text{EST}}$ Eq. (12) becomes relatively small. This is known as the Posterior Collapse [4, 32, 39] in training VAEs. The analysis above is depicted in Fig. 7. The failure in training G also ruins the training of E.

The IFN$(\cdot)$ exploited by our σ-Reg. CVAE is a simple statistical standardization technique as,

$$\bar{\boldsymbol{f}} = \text{IFN}(\boldsymbol{f}) = \frac{\boldsymbol{f} - \mathbb{E}(\boldsymbol{f})}{\sqrt{var(\boldsymbol{f}) + \epsilon}}, \tag{17}$$

where ϵ is a small value. It puts a hard constraint $var(\bar{\boldsymbol{f}}) = 1$ on the input $\bar{\boldsymbol{f}}$ to CVAE and eliminates the impact of drastic changes on inputs to the generator. From Eq. (16), the value of σ^2 is thus regularized by the introduction of IFN. Therefore, our proposed generator is called σ-Reg. CVAE to highlight such a mechanism. The values of $var(\bar{\boldsymbol{f}})$ and $\hat{\sigma}^2$ with IFN are the green curves in Fig. 6. $var(\bar{\boldsymbol{f}})$ fixed at 1 because of IFN$(\cdot)$ applied. $\hat{\sigma}^2 \approx 1$ at first according to Eq. (16). However, $\hat{\sigma}^2$ is the estimation rather than σ^2 itself. As shown in Eq. (14), $\hat{\sigma}^2$ also reflects the reconstruction loss in $\mathcal{L}_{\text{EST}}$ Eq. (12), which is gradually decreasing during training, as the green solid curve behaves.

4. Experiments

Datasets. To evaluate and compare different methods under the ISCS person Re-ID settings, two benchmark datasets [10, 52], i.e., Market-SCT and MSMT-SCT, are exploited. Such datasets are built on the source ones, Market-1501 [54] and MSMT17 [44], via only keeping the images of each person from one single camera for training. With no cross-camera person images and fewer training samples, the datasets under the ISCS setting are much more challenging than the source ones. Note that MSMT17 is a challenging dataset with the person images collected from different time periods and across largely varied scenes. It contains much more camera views, 15, than its counterparts with 6 and 8 only. Therefore, the MSMT-SCT can better stimulates the ISCS person Re-ID scenario. With the testing data unchanged, conventional person Re-ID evaluation metrics, Cumulative Matching Characteristic (CMC) and Mean Average Precision (mAP), are reported.

Implementation Details. Our image feature encoder E is the ImageNet pre-trained Resnet-50, following existing work [10, 52] for fair comparison. We also adopt the architecture with local branches as in [10]. The mini-batch size is set to 128 with image data augmentation [14]. Adam optimizer is used with an initial learn rate 3.5×10^{-4}, which decays at the 100th and 000th epoch with a decay factor of 0.1, and a weight decay of 5×10^{-4}. The total number of training epochs is 500. The hyper-parameters α, λ_1, λ_2 and λ_3 are set to 0.2, 0.5, 4, 1, respectively. All experiments can be run on an NVIDIA 2080Ti GPU.

4.1. Results

The proposed CCSFG is compared with different state-of-the-art methods. Besides the existing methods (CCFP [10], MCNL [52]) for the ISCS setting, other methods, such as the image generation (HHL [57]), distribution alignment (MMD [27], CORAL [43]), self-supervised learning (SimSiam [3]), metric learning (Center Loss [45], A-Softmax [20], ArcFace [26]), and baselines (PCB [38], Suh's method [37], MGN-ibn [41], Bagtrick [17], AGW [50]) are included. The results are shown in Tab. 1.

The proposed CCSFG achieves superior results to all its competitors. Clear margins can be observed between our CCSFG and the second place method CCFP [10] which

Table 1. The performance of different methods under the ISCS person Re-ID setting. † denotes the re-ranking technique [56] is used.

Methods	MSMT-SCT				Market-SCT			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
PCB [38](ECCV'18)	-	-	-	-	43.5	-	-	23.5
Suh's method [37](ECCV'18)	-	-	-	-	48.0	-	-	27.3
MGN-ibn [41](ACMMM'18)	27.8	38.6	44.1	11.7	45.6	61.2	69.3	26.6
Bagtrick [17](CVPR'19)	20.4	31.0	37.2	9.8	54.0	71.3	78.4	34.0
AGW [50](TPAMI'21)	23.0	33.9	40.0	11.1	56.0	72.3	79.1	36.6
Center Loss [45](ECCV16)	-	-	-	-	40.3	-	-	18.5
A-Softmax [20](CVPR'17)	-	-	-	-	41.9	-	-	23.2
ArcFace [26](CVPR'19)	-	-	-	-	39.4	-	-	19.8
SimSiam [3](CVPR'21)	2.8	5.9	8.4	1.2	36.2	51.9	59.1	18.0
MMD [27](ICML'15)	42.2	55.8	61.4	18.2	67.7	83.1	88.2	44.0
CORAL [43](ECCV'16)	42.6	55.8	61.5	19.5	76.2	88.5	93.0	51.5
HHL [57](ECCV'18)	31.4	42.5	48.1	11.0	65.6	80.6	86.8	44.8
MCNL [52](AAAI'20)	26.6	40.0	46.4	10.0	67.0	82.8	87.9	41.6
CCFP [10](ACMMM'21)	50.1	63.3	68.8	22.2	82.4	92.6	95.4	63.9
CCFP† [10](ACMMM'21)	**54.9**	65.0	69.5	**33.6**	84.1	90.9	93.1	**78.2**
CCSFG(Ours)	54.6	**67.7**	**73.1**	24.6	**84.9**	**94.3**	**96.2**	67.7
CCSFG† (Ours)	**61.2**	**71.1**	**75.1**	**37.8**	**87.1**	**92.8**	**95.0**	**82.6**

is the state-of-the-art ISCS Re-ID model based on self-learning and feature alignment. Comparing with CCFP, CCSFG achieves 6.3% R-1 and 4.2% mAP improvements on MSMT-SCT. Such improvements on Market-SCT are 3.0% R-1 and 4.4% mAP. The ISCS setting of person Re-ID is challenging. Many existing methods fail to achieve the ideal performance on it. The image generation method HHL [57] can improve the baseline methods with the cross-camera images generated for training and achieve comparable performance to the ISCS method MCNL [52]. However, generating the person images with cross-camera view information captured is a challenging task. The distribution alignment methods, MMD [27] and CORAL [43], also achieve substantially good performance. They align the holistic feature distributions of different camera views. A feature alignment loss, $\mathcal{L}_{CCFA}$, is also used in CCSFG (Eq. (8)) to align the image feature and its generated features under different cameras.

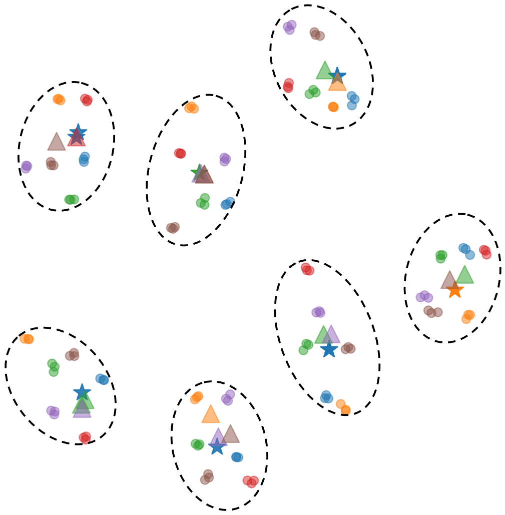


Figure 8. Visualization (t-SNE) of the training features produced by CCSFG on Market-1501. Dots are the generated features by G. Stars are the image features from E of the images retained in Market-SCT. Triangles are the image features from E of the deleted images of Market-1501 under the ISCS setting. Different colors indicate different cameras. The features in a oval belong to the same person.

4.2. Detailed Analysis

In this subsection, we conducted the detail analysis on our CCSFG from different perspectives.

Visualization. As shown in Fig. 8, meaningful features can be extracted and generated by the proposed CCSFG model. Firstly, obvious clusters are formed based on the person identities, which reflects that different features are discriminative. Secondly, the generated features $\boldsymbol{g}$s are id discriminative and view variant, as shown by the dense dots of different colors within a oval. Therefore, the generator G can handle both the identity and camera view information provided by y and c to generate meaningful features $\boldsymbol{g}$s (dots). Thirdly, the features $\boldsymbol{f}$ extracted by E on the training images under the ISCS setting are shown as stars. $\boldsymbol{f}$s are able to keep distances with the generated $\boldsymbol{g}$s under specified cameras. Moreover, the deleted cross-camera person images under the ISCS setting are fed into the trained encoder E and their features are shown as triangles. The

Table 2. The joint learning stability with different generators. CCSFG is with our σ-Reg. CVAE. Baseline is with encoder only.

Methods	MSMT-SCT		Market-SCT	
	Rank-1	mAP	Rank-1	mAP
Baseline	26.6	10.0	67.0	41.6
CVAE [35]	12.1	5.2	58.1	37.7
σ-CVAE [39]	11.2	4.1	59.3	38.6
CCSFG	**54.6**	**24.6**	**84.9**	**67.7**

star and triangles of the same person are highly overlapped which indicates the discriminative and view-invariant person image features can be extracted by encoder E. Last but not least, such an ideal feature map demonstrates that the stable and effective joint training is achieved by CCSFG.

The analysis of model stability. The qualitative results of analyzing the stability of training with different generators are presented in Sec. 3.3 along with the theoretical analysis. The quantitative results are provided here to further evaluate the impact on joint learning with different generator models, as shown in Tab. 2. Besides the comparisons with the vanilla CVAE [35] that without any regularization on σ^2, the proposed σ-Reg. CVAE is compared with the σ-CVAE [39] which attempts to optimize the σ^2 value during training. However, neither the CVAE nor the σ-CVAE can stabilize the training procedure. Jointly learning the image feature encoder with such generators can even harm the performance.

The impact of the conditional variables. To verify the necessity of conditional variables y and c for feature generation, we conduct the ablation study on them, as shown in Tab. 3. When both the identity label y and camera label c are not used in G for feature generation, our σ-Reg. CVAE degenerates into a VAE-based model. Its generated feature $\boldsymbol{g}$ is not conditioned on camera and ID information, the corresponding training objective will be reducing the distance between $\boldsymbol{g}$ and the input feature $\bar{\boldsymbol{f}}$ only. As shown in the first row of Tab. 3, such the generator harms the person Re-ID performance. Moreover, G can incorporate either y or c only for feature generation. Substantial improvements can be obtained by considering more conditions, especially the identity label y. Since conditioning on y can guarantee the discriminative power in the generated features and the encoder E can benefit from them in the joint learning. Conditioning on both y and c in our generator, σ-Reg CVAE, clearly boosts the performance. These results demonstrate the importance of both y and c for generating useful features in the joint learning of CCSFG.

Table 3. Analysis on the importance of condition variables to the generator. The condition variables y and c denote the identity label and camera label, respectively. G is the σ-Reg. CVAE.

Methods	MSMT-SCT		Market-SCT	
	Rank-1	mAP	Rank-1	mAP
G w/o y & c	12.4	6.3	43.6	24.8
G w/ c	22.6	12.3	49.3	31.6
G w/ y	36.2	15.3	77.1	54.3
G w/ y & c	**54.6**	**24.6**	**84.9**	**67.7**

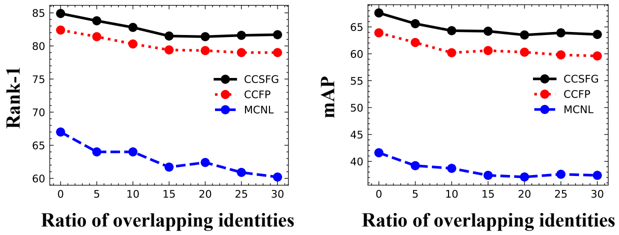


Figure 9. Analysis on the radios of overlapping identities on Market-1501.

Cross-camera identity overlap ratio. In the real-world surveillance application, fully non-overlapped persons across different cameras could be a strong assumption. Therefore, different ratios of cross-camera overlapping identities should be considered. With the more cross-camera images of more persons existing (indicated by the higher ratio of overlapping IDs), the more training samples of same person are given distinctive ID labels. The obtained models are thus worse, as shown in Fig. 9. However, Our CCSFG can withstand this challenge and stabilize at the SOTA level performance.

5. Conclusion

In this paper, we focus on handling the challenging ISolated Camera Supervised (ISIC) person Re-ID problem where the cross-camera image pairs are not available for model training. To compensate the missing cross-camera data pairs, a novel pipeline based on feature generation is introduced. Following this pipeline, we propose the camera-conditioned stable feature generation (CCSFG), the first method to synthesize the cross-camera feature samples and end up with the joint learning between image encoder E and feature generator G. A novel generative model, σ-Reg. CVAE, is then proposed as G to achieve stable joint learning. The effectiveness of CCSFG is demonstrated by theoretical analysis and experimental results. **Potential negative societal impact**: As a more advanced and robust feature learning technique for visual data, the proposed method might be abused for unauthorized monitoring.

Acknowledgement. This work was supported by the National Science Foundation for Young Scientists of China (62106289, 62106288), China National Postdoctoral Program for Innovative Talents (BX20200395), China Postdoctoral Science Foundation (2021M693616) and Zhuhai Industry-University-Research Cooperation Project (ZH22017001210010PWC).

References

[1] Christian F. Baumgartner, Kerem Can Tezcan, Krishna Chaitanya, Andreas M. Hötker, Urs J. Muehlematter, Khoschy Schawkat, Anton S. Becker, Olivio Donati, and Ender Konukoglu. Phiseg: Capturing uncertainty in medical image segmentation. In *22nd International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2019)*, volume 17765, pages 119–127, 2019. 3

[2] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2109–2118, 2018. 1, 2

[3] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 6, 7

[4] Bin Dai, Ziyu Wang, and David P. Wipf. The usual suspects? reassessing blame for VAE posterior collapse. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 2313–2322. PMLR, 2020. 3, 5, 6

[5] Yongxing Dai, Jun Liu, Yan Bai, Zekun Tong, and Ling-Yu Duan. Dual-refinement: Joint label and feature refinement for unsupervised domain adaptive person re-identification. *IEEE Transactions on Image Processing*, 30:7815–7829, 2021. 1, 2

[6] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 994–1003, 2018. 3

[7] Hao Ding, Siyuan Qiao, Alan L. Yuille, and Wei Shen. Deeply shape-guided cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8278–8288, 2021. 3

[8] Chris Donahue, Zachary C. Lipton, Akshay Balsubramani, and Julian J. McAuley. Semantically decomposing the latent spaces of generative adversarial networks. In *International Conference on Learning Representations*, 2018. 3

[9] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14(4):83, 2018. 2

[10] Wenhang Ge, Chunyan Pan, Ancong Wu, Hongwei Zheng, and Wei-Shi Zheng. Cross-camera feature prediction for intra-camera supervised person re-identification across distant scenes. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3644–3653, 2021. 1, 2, 6, 7

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 3

[12] Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taïga, Francesco Visin, David Vázquez, and Aaron C. Courville. Pixelvae: A latent variable model for natural images. In *ICLR (Poster)*, 2016. 3

[13] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7073–7082, 2018. 2

[14] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*, 2020. 6

[15] Yan Huang, Qiang Wu, Jingsong Xu, and Yi Zhong. Sbsgan: Suppression of inter-domain background shift for person re-identification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9527–9536, 2019. 3

[16] Takashi Isobe, Dong Li, Lu Tian, Weihua Chen, Yi Shan, and Shengjin Wang. Towards discriminative representation learning for unsupervised person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8526–8536, 2021. 2

[17] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 427–431, 2017. 1, 6, 7

[18] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2014. 3

[19] Bo Li, Zhengxing Sun, and Yuqi Guo. Supervae: Superpixelwise variational autoencoder for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33(1), pages 8569–8576, 2019. 3

[20] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu. Deep speaker: an end-to-end neural speaker embedding system. *arXiv preprint arXiv:1705.02304*, 2017. 6, 7

[21] Jianing Li and Shiliang Zhang. Joint visual and temporal consistency for unsupervised domain adaptive person re-identification. In *ECCV (24)*, pages 483–499, 2020. 1, 2

[22] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 772–788, 2018. 2

[23] Yuanyuan Li, Sixin Chen, Guanqiu Qi, Zhiqin Zhu, Matthew Haner, and Ruihua Cai. A gan-based self-training framework for unsupervised domain adaptive person re-identification. *Journal of Imaging*, 7(4):62, 2021. 1

[24] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4099–4108, 2018. 3

[25] Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Meng Wang. Adaptive transfer network for cross-domain person re-identification. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7202–7211, 2019. 3

[26] Weilun Liu, Jichao Jiao, Yaokai Mo, Jian Jiao, and Zhongliang Deng. Maaface: Multiplicative and additive angular margin loss for deep face recognition. In *International Conference on Image and Graphics*, pages 642–653, 2019. 6, 7

[27] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of The 32nd International Conference on Machine Learning*, volume 1, pages 97–105, 2015. 6, 7

[28] Nan Pu, Wei Chen, Yang Liu, Erwin M. Bakker, and Michael S. Lew. Dual gaussian-based variational subspace disentanglement for visible-infrared person re-identification. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 3

[29] Lei Qi, Lei Wang, Jing Huo, Yinghuan Shi, and Yang Gao. Progressive cross-camera soft-label learning for semi-supervised person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30:2815–2829, 2020. 2

[30] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yugang Jiang, and X. Xue. Pose-normalized image generation for person re-identification. *ArXiv*, abs/1712.02225, 2018. 3

[31] Jiawei Ren, Xiao Ma, Chen Xu, Haiyu Zhao, and Shuai Yi. Havana: Hierarchical and variation-normalized autoencoder for person re-identification. *ArXiv*, abs/2101.02568, 2021. 3

[32] Oleh Rybkin, Kostas Daniilidis, and Sergey Levine. Simple and effective vae training with calibrated decoders. In *ICML*, 2021. 3, 5, 6

[33] Sandro Schönborn, Bernhard Egger, Andreas Forster, and Thomas Vetter. Background modeling for generative image models. *Computer Vision and Image Understanding*, 136:117–127, 2015. 3

[34] Hoo Chang Shin, Neil A. Tenenholtz, Jameson K. Rogers, Christopher G. Schwarz, Matthew L. Senjem, Jeffrey L. Gunter, Katherine P. Andriole, and Mark Michalski. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *3rd International Workshop on Simulation and Synthesis in Medical Imaging, SASHIMI 2018 Held in Conjunction with 21st International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2018*, pages 1–11, 2018. 3

[35] Kihyuk Sohn, Xinchen Yan, and Honglak Lee. Learning structured output representation using deep conditional generative models. In *NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, volume 28, pages 3483–3491, 2015. 4, 8

[36] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. Unsupervised domain adaptive re-identification: Theory and practice. *Pattern Recognition*, 102:107173, 2020. 2

[37] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, volume 11218 of *Lecture Notes in Computer Science*, pages 418–437. Springer, 2018. 6, 7

[38] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 501–518, 2018. 1, 2, 6, 7

[39] Yuhta Takida, Wei-Hsiang Liao, Toshimitsu Uesaka, Shusuke Takahashi, and Yuki Mitsufuji. Preventing posterior collapse induced by oversmoothing in gaussian vae. *arXiv preprint arXiv:2102.08663*, 2021. 3, 5, 6, 8

[40] Binquan Wang, Muhammad Asim, Guoqi Ma, and Ming Zhu. Central feature learning for unsupervised person re-identification. *International Journal of Pattern Recognition and Artificial Intelligence*, 35(8):2151007, 2021. 1, 2

[41] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282, 2018. 2, 6, 7

[42] Hanxiao Wang, Xiatian Zhu, Shaogang Gong, and Tao Xiang. Person re-identification in identity regression space. *International Journal of Computer Vision*, 126(12):1288–1310, 2018. 1

[43] Zhi-Yong Wang and Dae-Ki Kang. P-norm attention deep coral: Extending correlation alignment using attention and the p-norm loss function. *Applied Sciences*, 11(5267):5267, 2021. 6, 7

[44] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[45] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515, 2016. 6, 7

[46] Guile Wu, Xiatian Zhu, and Shaogang Gong. Tracklet self-supervised learning for unsupervised person re-identification. In *AAAI*, 2020. 2

[47] Jinlin Wu, Hao Liu, Yang Yang, Zhen Lei, Shengcai Liao, and Stan Li. Unsupervised graph association for person re-identification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8321–8330, 2019. 2

[48] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1249–1258, 2016. 2

[49] Fengxiang Yang, Zhun Zhong, Zhiming Luo, Yuanzheng Cai, Yaojin Lin, Shaozi Li, and Nicu Sebe. Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4855–4864, 2021. 2

[50] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C.H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 2, 6, 7

[51] Yunpeng Zhai, Shijian Lu, Qixiang Ye, Xuebo Shan, Jie Chen, Rongrong Ji, and Yonghong Tian. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9021–9030, 2020. 2

[52] Tianyu Zhang, Lingxi Xie, Longhui Wei, Yongfei Zhang, Bo Li, and Qi Tian. Single camera training for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12878–12885, 2020. 1, 2, 5, 6, 7

[53] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593, 2013. 2

[54] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 6

[55] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2138–2147, 2019. 3

[56] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3652–3661, 2017. 7

[57] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero- and homogeneously. In *ECCV*, 2018. 1, 3, 6, 7

[58] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2018. 3

[59] Xiangping Zhu, Xiatian Zhu, Minxian Li, Pietro Morerio, Vittorio Murino, and Shaogang Gong. Intra-camera supervised person re-identification. *International Journal of Computer Vision*, 129(5):1580–1595, 2021. 2