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

Chao Wu¹, Wenhang Ge⁴, Ancong Wu⁴, Xiaobin Chang¹,²,³ *
¹School of Artificial Intelligence, Sun Yat-sen University, China
²Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou 510006, P.R.China
³Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
⁴School 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-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. 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.

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
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, $\sigma$-Regularized CVAE ($\sigma$-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, $\sigma$-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, their identities are elaborately labeled and aligned across different cameras as supervision. The unsupervised setting 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. Moreover, the intra-camera supervised (ICS) setting 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 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 handle this challenging setting with the alignment losses on the data distributions rather than the sample pairs of different cameras.

20207
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 synthesize 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, $\sigma$-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 $\sigma$-Reg. CVAE is the first VAE-based feature generator for the ISCS person Re-ID. The generator $\sigma$-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 $\sigma$-Reg. CVAE.

3. **Methodology**

3.1. **Isolated Camera Supervised Person Re-ID**

The training set is denoted as $D = \{(x_n, y_n, c_n)\}_{n=1}^{|D|}$, where each training sample is a triplet with the person image $x_n$, 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., $\forall i,j \in \{1, \ldots, |D|\}$, 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.
Given the image feature \( f \) 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 \( \sigma \)-Reg. CVAE, the loss \( \mathcal{L}_{EST} \) is used. It corresponds to the Estimation phase of \( G \), as shown in Fig. 3. Specifically, image feature \( f \) is the input of \( \sigma \)-Reg. CVAE and the normalized feature \( \tilde{f} \) is obtained,

\[
\tilde{f} = \text{IFN}(f),
\]

where \( \text{IFN}(\cdot) \) is a normalization function in \( \sigma \)-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 \( \sigma \)-Reg. CVAE is maximizing the conditional log-likelihood \( \log P(\tilde{f}|c, y) \), which is often intractable. Its variational lower bound is optimized instead by introducing latent variables [35]. Specially, \( z_y \) and \( z_c \) are the two latent variables introduced in the \( \sigma \)-Reg. CVAE and correspond to the given identity condition \( y \) and camera condition \( c \) respectively. Their prior distributions, \( P_\phi(y|z_y) \) and \( P_\phi(z_c|c) \), are modeled with two prior networks \( \phi_y \) and \( \phi_c \). The two recognition networks \( \phi_y(z_y|f, y) \) and \( \phi_c(z_c|f, c) \) map \( (f, y) \) and \( (f, c) \) to their posterior distribution \( Q_{\phi_y}(z_y|f, y) \) and \( Q_{\phi_c}(z_c|f, c) \). Moreover, the decoding distribution \( P_\mu(f|z_y, z_c, c, y) \) is modeled by the decoder network \( \mu \). Based on the previous sub-networks and distributions defined, an applicable learning objective of \( \sigma \)-Reg. CVAE is,

\[
\mathcal{L}_{EST}(f, y, c | \theta_y, \theta_c, \phi_y, \phi_c, \mu) =
E_{Q_{\phi_y}(z_y|f, y)}Q_{\phi_c}(z_c|f, c) \left[ -\log P_\mu(f|z_y, z_c, c, y) \right]
+ DKL(Q_{\phi_y}(z_y|f, y) \| P_\theta(y|z_y|y))
+ DKL(Q_{\phi_c}(z_c|f, c) \| P_\theta(z_c|c)),
\]

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

The latent variables \( z_y \) and \( z_c \) are sampled from different distributions across stages. During training \( z_c \sim Q_{\phi_c}(z_c|f, c) \) and \( z_y \sim Q_{\phi_y}(z_y|f, y) \) while testing \( z_c \sim P_\theta(z_c|c) \) and \( z_y \sim P_\theta(y|z_y|y) \). Such inconsistency may 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,

\[
\mathcal{L}_{GSNN}(f, y, c | \theta_y, \theta_c) =
E_{Q_{\phi_y}(z_y|f, y)} \left[ -\log P_\mu(f|z_y, c, y) \right],
\]

The overall objective for the estimation of the generator \( \sigma \)-Reg. CVAE is,

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

with \( \alpha \) 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 \( G \), \( \sigma \)-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 \( \sigma \)-Reg. CVAE is functioned under the Generation mode, as illustrated in Fig. 5. With an input feature \( f \), its person
identity label \( y \), the camera views \( v_i, v_i \in \{ v_1, ..., v_C \} \) and the latent variables \( z_y \sim P_{\theta_y}(z_y|y) \) and \( z_c \sim P_{\theta_c}(z_c|v_i) \) given, camera-conditioned features \( g \) can be generated from the decoder network \( \mu \) of \( G \),

\[
g_{v_i} = \mu(z_c, z_y, v_i, y).
\]

(7)

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

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

\[
L_{CCFA}(\bar{f} | G) = \frac{1}{C} \sum_{i=1}^{C} \| \bar{f} - g_{v_i} \|^2,
\]

(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 \( L_{ID} \) is used,

\[
L_{ID}(y, \bar{f} | G) = -\log(q[y]),
\]

(9)

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

Moreover, the MCNL loss [52], denoted as \( L_{MCNL} \), 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 \( L_{E|G} \) is,

\[
L_{E|G} = \lambda_1 L_{CCFA} + \lambda_2 L_{ID} + \lambda_3 L_{MCNL},
\]

(10)

where \( \lambda_1, \lambda_2 \) and \( \lambda_3 \) are balancing hype-parameters.

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

\[
L_{E,G} = \alpha L_{EST} + (1-\alpha) L_{GSNN} + \lambda_1 L_{CCFA} + \lambda_2 L_{ID} + \lambda_3 L_{MCNL}.
\]

(11)

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

3.3. The stability in training \( G \)

In the proposed CCSFG, the encoder \( E \) and generator \( G \) are jointly learned. The image feature \( f \) is extracted by \( E \) and used as the input for training \( G \). However, \( f \) 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 \( \sigma \)-Reg. CVAE is exploited as generator \( G \). Its IFN(\( \cdot \)) module normalizes the image feature \( f \) to \( \bar{f} \) as input for feature generation and plays the key role in stabilizing the whole training process. Without \( \text{IFN}(\cdot) \), the \( \sigma \)-Reg. CVAE degenerates to a conventional CVAE using the image feature \( f \) 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 \( L_{ EST}(f, y, c | \theta_y, \theta_c, \phi_y, \phi_c, \mu) \) with \( f \) rather than \( \bar{f} \) as input. Following [4, 32, 39], this loss can be rewritten as,

\[
L_{EST} = \frac{d}{\sigma^2} \| f - g \|^2 + \frac{d}{2} \ln \sigma^2 + \mathcal{D}_{KL}(Q_{\phi_y}(z_y|f, y)||P_{\theta_y}(z_y|y)) + \mathcal{D}_{KL}(Q_{\phi_c}(z_c|f,c)||P_{\theta_c}(z_c|c)),
\]

(12)
Figure 7. The impacts of huge $\text{var}(f)$ on the training of CVAE. 

by replacing $\mu(z_c, z_y, c, y)$ with $g$ as in Eq. (7) and assuming the decoding distribution $P_\mu(f|z_c, z_y, c, y)$ as an isotropic Gaussian distribution,

$$P_\mu(f|z_c, z_y, c, y) = \mathcal{N}(g, \sigma^2 I),$$

(13)

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

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

(14)

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

$$\text{var}(f) = \frac{1}{d} \mathbb{E}((f - \mathbb{E}(f))^2).$$

(15)

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

$$\sigma^2 \approx \text{var}(f).$$

(16)

In the joint learning procedure, the training of encoder $E$ leads to the rapid changes in image features and thus huge $\text{var}(f)$ occurs. Without normalization on the input $f$, CVAE captures such variance and results in large $\sigma^2$ as in Eq. (16). A concrete example of the joint learning of $E$ and $G$ on the ISCS dataset is shown in Fig. 6 (left), where $\sigma^2$ is approximated by $\hat{\sigma}^2$. As the red curves shown, the values of both $\text{var}(f)$ and $\hat{\sigma}^2$ rise drastically as expected. However, large value of $\sigma^2$ prevents the CVAE to learn from the its input as the weight $\frac{d}{\sigma^2}$ on reconstruction term $||f - g||^2$ in $L_{EST}$ 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(.) exploited by our $\sigma$-Reg. CVAE is a simple statistical standardization technique as,

$$\tilde{f} = \text{IFN}(f) = \frac{f - \mathbb{E}(f)}{\sqrt{\text{var}(f)} + \epsilon},$$

(17)

where $\epsilon$ is a small value. It puts a hard constraint $\text{var}(\tilde{f}) = 1$ on the input $\tilde{f}$ to CVAE and eliminates the impact of drastic changes on inputs to the generator. From Eq. (16), the value of $\sigma^2$ is thus regularized by the introduction of IFN. Therefore, our proposed generator is called $\sigma$-Reg. CVAE to highlight such a mechanism. The values of $\text{var}(\tilde{f})$ and $\hat{\sigma}^2$ with IFN are the green curves in Fig. 6. $\text{var}(\tilde{f})$ fixed at 1 because of IFN(.) applied. $\hat{\sigma}^2 \approx 1$ at first according to Eq. (16). However, $\hat{\sigma}^2$ is the estimation rather than $\sigma^2$ itself. As shown in Eq. (14), $\hat{\sigma}^2$ also reflects the reconstruction loss in $L_{EST}$ 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 $3.5 \times 10^{-4}$, which decays at the 100th and 000th epoch with a decay factor of 0.1, and a weight decay of $5 \times 10^{-4}$. The total number of training epochs is 500. The hyper-parameters $\alpha$, $\lambda_1$, $\lambda_2$ and $\lambda_3$ 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     | 12.8   | 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, $L_{CCFA}$, is also used in CCSFG (Eq. (8)) to align the image feature and its generated features under different cameras.

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 $g_s$ 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 $g_s$ (dots). Thirdly, the features $f$ extracted by $E$ on the training images under the ISCS setting are shown as stars. $f_s$ are able to keep distances with the generated $g_s$ 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
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 $\sigma^2$, the proposed $\sigma$-Reg. CVAE is compared with the $\sigma$-CVAE [39] which attempts to optimize the $\sigma^2$ value during training. However, neither the CVAE nor the $\sigma$-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 $\sigma$-Reg. CVAE degenerates into a VAE-based model. Its generated feature $g$ is not conditioned on camera and ID information, the corresponding training objective will be reducing the distance between $g$ and the input feature $f$ 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, $\sigma$-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.

**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, $\sigma$-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 Zuhai Industry-University-Research Cooperation Project (ZH22017001210010PWC).
References

[1] Christian F. Baumgartner, Kerem Can Tezcan, Krishna Chaitanya, Andreas M. Hölter, Urs J. Muehlematter, Koshy 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. 

[2] Xiaolin 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. 

[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. 

[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, 13-17 July 2020, Virtual Event, volume 119 of Proceedings of Machine Learning Research, pages 2313–2322. PMLR, 2020. 

[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. 

[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. 

[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. 

[8] Chris Donahue, Zachary C. Lipton, Akshay Balsubramani, Hao Ding, Siyuan Qiao, Alan L. Yuille, and Wei Shen. The usual suspects? reassessing blame for VAE posterior collapse. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-17 July 2020, Virtual Event, volume 119 of Proceedings of Machine Learning Research, pages 2313–2322. PMLR, 2020. 

[9] Yi Yang, and Jianbin Jiao. Image-image domain adaptation for unsupervised domain adaptive person re-identification. IEEE Transactions on Multimedia, 23(22), pages 119–127, 2019. 

[10] Lingxiao He, Jian Liang, Haqing 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. 

[11] 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. 

[12] 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. 

[13] 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. 

[14] 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. 

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

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

[17] 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. 

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

[19] 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. 

[20] 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. 

[21] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In Proceedings of the 29th ACM International Conference on Multimedia, volume 33(1), pages 8569–8576, 2019. 

[22] Jianing Li and Shiliang Zhang. Joint visual and temporal consistency for unsupervised domain adaptive person re-identification. In Proceedings of the European Conference on Computer Vision (ECCV), pages 772–788, 2018. 

[23] 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. 

[24] Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Hongwei Zheng. Cross-camera feature prediction for unsupervised domain adaptive person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 7202–7211, 2019. 

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

[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.

[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.

[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.

[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.

[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.

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

[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.

[59] Xiangping Zhu, Xiatain 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.