HAVANA:
Hierarchical and Variation-Normalized Autoencoder for Person Re-identification

Jiawei Ren*[1], Xiao Ma*[2], Chen Xu*[2], Haiyu Zhao†[1], and Shuai Yi[1]

[1] SenseTime Research, `renjiawei@sensetime.com`
[2] National University of Singapore, `xiao-ma@comp.nus.edu.sg`

Abstract. Person Re-Identification (Re-ID) is of great importance to the many video surveillance systems. Learning discriminative features for Re-ID remains a challenge due to the large variations in the image space, e.g., continuously changing human poses, illuminations and point of views. In this paper, we propose HAVANA, a novel extensible, lightweight HierArchical and VAriation-Normalized Autoencoder that learns features robust to intra-class variations. In contrast to existing generative approaches that prune the variations with heavy extra supervised signals, HAVANA suppresses the intra-class variations with a Variation-Normalized Autoencoder trained with no additional supervision. We also introduce a novel Jensen-Shannon triplet loss for contrastive distribution learning in Re-ID. In addition, we present Hierarchical Variation Distiller, a hierarchical VAE to factorize the latent representation and explicitly model the variations. To the best of our knowledge, HAVANA is the first VAE-based framework for person ReID.

Keywords: Person Re-Identification, Variational Autoencoder, Representation Learning

1 Introduction

Pedestrian Re-Identification (Re-ID) aims to associate pedestrian images across different cameras and time periods. An efficient robust Re-ID system is of fundamental importance to various surveillance applications. For example, finding a lost child in the mall, searching for a suspect at the airport and etc [47].

As a well-studied problem, Re-ID, however, remains challenging and an unsolved task. One major challenge is the large variations in the image space. It is common that a camera does not always capture the human face and the Re-ID model has to be able to match the back view of a pedestrian with the front view. The constantly changing human poses, illumination conditions and point of views also requires a robust Re-ID model against the variations.

---

[*] Equal contribution
[†] Corresponding author

Most state-of-the-art Re-ID models often aim to learn invariant features to be insensitive against the noise. In particular, with the recent advances of the deep generative models, GANs have been widely used to generate clean representations from the noisy observations. However, these models normally rely on additional supervised training signals, *e.g.*, keypoints/poses, camera-styles, *etc.*, which not only requires heavy data labeling efforts, but also burdens the training process.

In this paper, we propose HAVANA, a novel extensible, light-weight HierArchical and VAriation-Normalized Autoencoder framework for Re-ID with explicit variation modeling and no extra data labeling. In contrast to the prior works that focus on filtering the variations, HAVANA adopts a *Variational Autoencoder* (VAEs) [22] based structure, that explicitly models the feature variation with a learned latent variable. HAVANA consists of three main components. First, we introduce *Variation-Normalized Autoencoder* (VNAE) framework, which is the base of HAVANA, that applies standard supervised Re-ID training signals, classification loss and triplet loss, to regularize the variance of the latent distribution in VAE. Specifically, we backpropagate the supervised gradients to the mean of the latent distribution, ignoring the noisy variations, which automatically learns the shared clean feature of an identity. Second, we introduce a novel variance aware triplet loss function, *Jensen-Shannon Triplet Loss* (JS triplet loss), that computes the contrastive loss of two distributions. JS triplet loss encourages the VNAE to consider distributional distance and improves the robustness of VNAE against the noise. Third, *Hierarchical Variation Distiller* (HVD) is proposed to further improve variation filtering and disentangled representation learning. HVD factorizes the latent variations into a hierarchical structure with two additional variables to model high-level variations.

We experiment on three large scale datasets: Market-1501, DukeMTMC-reID and MSMT17. We train HAVANA by only the unsupervised ELBO loss, and the combination of Jensen-Shannon triplet loss and cross-entropy loss as a supervised signal for Re-ID, without any extra supervision. Our results significantly outperform the state-of-the-art (SOTA) generative approaches that extensively rely on other labels. Comparing with all existing Re-ID methods, we achieve SOTA results on two datasets and strong performance on the rest.

We summarize our contributions as follows: 1) we propose HAVANA, a VAE-based Re-ID framework with minimal data labeling requirements that achieves the SOTA performance on 3 commonly used Re-ID benchmarks; 2) we introduce Jensen-Shannon triplet loss, a robust variance aware contrastive feature learning loss function for Re-ID; 3) we design a hierarchical variation distiller module to for a factorized variation modeling.

2   Background

2.1   Variational Autoencoders

Variational Autoencoder (VAE) [22] is an unsupervised representation learning algorithm that embeds high-dimensional observations into a low-dimensional

space such that the learned embeddings, *i.e.*, meta-prior [4], can reconstruct to the original observations. A VAE has an encoder-decoder structure: we encode observation $\mathbf{x}$ into a latent variable $\mathbf{z}$ and decode into $\mathbf{x}$ again. Specifically, we define a family of prior distribution $p_\theta(\mathbf{z})$ over the latent variable $\mathbf{z}$ and decoder $p_\theta(\mathbf{x} \mid \mathbf{z})$ over observations $\mathbf{x}$. A VAE learns to maximize the observation likelihood $\sum_{n=1}^{N} \log p_\theta(\mathbf{x}^n)$, of a given dataset $\{\mathbf{x}^n\}_{n=1}^{N}$. The observation likelihood can further be factorized as $\int p_\theta(\mathbf{x} \mid \mathbf{z})p_\theta(\mathbf{z})dz$. VAE proposes to tackle the intractable log marginal likelihood problem by importance sampling with a encoder $q_\phi(\mathbf{z} \mid \mathbf{x})$, *i.e.*, a conditional distribution of $\mathbf{z}$ given $\mathbf{x}$, and maximizes the Evidence Lower Bound (ELBO) of the observation likelihood,

$$\text{ELBO} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{x} \mid \mathbf{z})p_\theta(\mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})} \right] \tag{1}$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x} \mid \mathbf{z}) \right] - D_{\text{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z})) \tag{2}$$

where D_{KL} is the Kullback–Leibler divergence (KL Divergence).

The intuition behind VAE is that we sample $\mathbf{z}$ from encoder $q_\phi(\mathbf{z} \mid \mathbf{x})$ and reconstruct x with decoder $p_\theta(\mathbf{x} \mid \mathbf{z})$. We want to minimize the distance between the encoder $q_\phi(\mathbf{z} \mid \mathbf{x})$ and the prior $p_\theta(\mathbf{z})$, as well as maximize the observation likelihood. In implementations, the distributions are assumed to be Gaussians and $p_\theta(\mathbf{z})$ is usually assumed as a diagonal Normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ for simplicity.

2.2 Related Works

Deep Person Re-ID Person Re-ID studies the person retrieval problem under a multi-camera setup [12]. Great progress has been achieved in the field lately [45,17,16,9,42,28,24,37,19,5,31,54,34,43,1,46,50,47]. Many strong feature extractors has been designed, including deep neural networks that are attention guided, body-part aligned, keypoint aligned, semantic aligned, *etc*. A recent survey [44] reviews the existing methods in detail.

Among many approaches, multiple prior works attempt to utilise deep generative models to help improve the feature embedding. In particular, [51] initiated the trend by using GAN to generate unlabeled pseudo images to supplement limited image-per-identity in Re-ID datasets. Later, [25] proposed a generative network to transfer pose variations from a large-scale dataset to target dataset for robust feaure learning. Similarly, [33] fused the feature embeddings from the original image and a GAN-synthesized image with a different pose and thus normalised the human-poses. [11] designed FD-GAN to distill a pose-unrelated feature by pruning the pose embedding learned by GAN. Meanwhile, [53] employed CycleGAN to adapt labeled images to different camera styles and hence augment the training set. So far, all preceding methods treat the generative model as a standalone module. [50] for the first time proposed a pipeline to jointly train a shared encoder that is supervised by both discriminative loss and generative loss under the generative-adversarial framework. However, one major issue of GAN-based methods is that training GAN has been notoriously difficult and unstable [14,3].
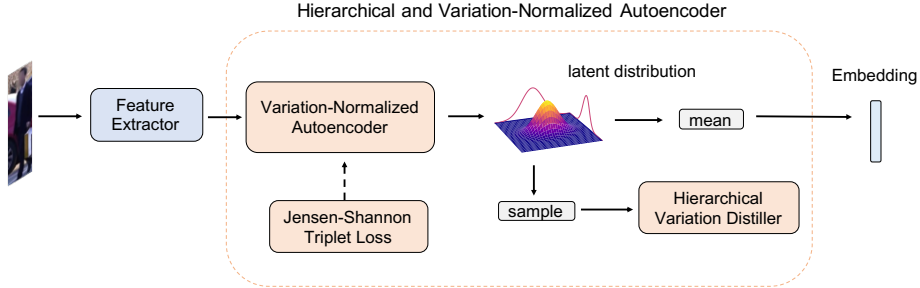
Hierarchical and Variation-Normalized Autoencoder

Fig. 1: Overview of our proposed HAVANA framework.

Disentangled Representation Learning for VAEs [39] presents recent progresses in the field. The characteristics of the aforementioned meta-priors varies for downstream applications, for example, image-to-image translation[13], manipulating attributes on images[23]. One group of meta-prior [18,20,6,48,26] encourages disentanglement, which factorizes variations in observations like viewpoint, lighting condition into independent dimensions of the meta-prior. In the same time, another group of works [15,36,21,8,32] proposed hierarchical VAEs that attempt to describe the observations with hierarchical attributes.

Our framework HAVANA is a combination of the above-mentioned approaches: we propose a hierarchical β-VAE approach that improves the disentangled latent representation learning. To the best of our knowledge, HAVANA is the first work to perform feature learning on the Person Re-ID task with VAEs. It also eases the training process compared with prior GAN-based generative approaches.

3   Proposed Method

3.1   Overview

In this section, we present our VAE-based, joint generative and discriminative learning framework, Hierarchical and Variation-Normalized Autoencoder. The general Person Re-ID pipeline uses person identity labels to supervise a feature extractor that embeds the cropped person images into multi-dimensional feature vectors. Treating the feature vectors as observations, we propose a Variation-Normalized Autoencoder that normalizes the intra-class variations in the embedding space. To facilitate the above variational modeling process, we introduce a novel metric learning loss, Jensen-Shannon triplet loss, to assimilate the latent distributions of positive pairs and separate the latent distributions of negative pairs. Additionally, we distill the features by learning a structural representation of features with Hierarchical Feature Distiller and meanwhile explicitly model the intra-class variation in an unsupervised manner. Lastly, we wrap up all proposed modules/losses in a holistic framework, and introduce a novel constraint to practically improve the framework's performance on multiple Person Re-ID datasets. Fig. 1 demonstrates an overview of our method.
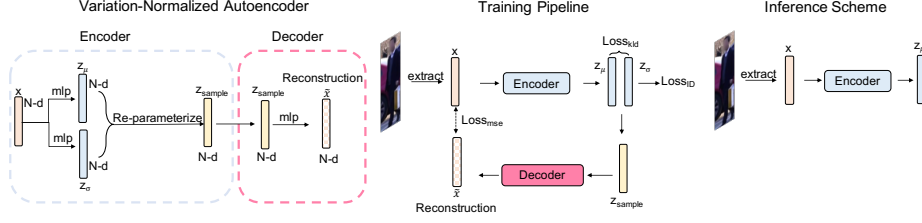
Fig. 2: Network design, training/inference scheme of VNAE. VNAE consists of an encoder and a decoder. We first use a trained feature extractor to embedd an image to a feature vector $\mathbf{x}$. Then $\mathbf{x}$ is encoded to latent distribution $\mathbf{z}$. $\mathbf{z}$ is supervised by both ID loss and KL divergence. A sample of $\mathbf{z}$ is decoded to $\tilde{\mathbf{x}}$ to compute reconstruction loss with input $\mathbf{x}$. During inference, VNAE directly outputs the encoded mean vector $\mathbf{z}_\mu$ as the new embedding vector.

3.2   Variation-Normalized Autoencoder for Re-identification

Prior works designed powerful feature extractors to learn discriminative features from person images with supervision of person identity labels. However, since the image-per-identity in most Re-ID datasets is limited [51], the intra-class variations, e.g., poses and illumination changes, cannot be successfully filtered. To alleviate the issue, we introduce Variation-Normalized Autoencoder (VNAE) to learn a variation-insensitive representation from the embedding features.

We first employ a feature extracting network, which can be selected from any existing Person Re-ID models and not restricted to a specific type, to embed a cropped person image into a N-dimensional feature vector $\mathbf{x} \in \mathbb{R}^N$. We then build our VNAE over the feature x, using VAE described in Sect. 2.1. Specifically, we first encode $\mathbf{x}$ with encoder $q_\phi(\mathbf{z} \mid \mathbf{x})$ and sample $\mathbf{z}$ from it:

$$\mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{x}), \quad q_\phi(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(\mathbf{z}_\mu, \mathbf{z}_\sigma^2) \tag{3}$$

$$\mathbf{z}_\mu = f_\mu(\mathbf{x}), \quad \mathbf{z}_\sigma = f_\sigma(\mathbf{x}) \tag{4}$$

where $\mathbf{z}_\mu$ and $\mathbf{z}_\sigma$ are the corresponding mean and standard deviation of the Gaussian distribution $q_\phi(\mathbf{z} \mid \mathbf{x})$, learned by two parameterized functions f_μ and f_σ with $\mathbf{x}$ as input. Normally, f_μ and f_σ are simple MLPs with non-linear activations. To achieve differentiable sampling, we apply the reparameterization trick [22]: we first sample $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and set $z = \mathbf{z}_\mu + \epsilon * \mathbf{z}_\sigma$. Lastly, we decode to observation $\mathbf{x}$ with decoder $p_\theta(\mathbf{x} \mid \mathbf{z})$. Fig. 2 shows VNAE in detail.

VNAE differs from VAE in terms of the variation normalization during training and inference. Specifically, Re-ID task requires distinctive features for different identities. To disentangle the identities, we supervise $\mathbf{z}_\mu$ with a cross-entropy Re-ID classification loss and triplet loss, where the classification loss learns to predict the correct identity of $\mathbf{x}$ using $\mathbf{z}_\mu$, and the triplet loss learns to separate the embeddings $\mathbf{z}$ of different identities in the latent space. The gradient optimizes the mean of the random variable $\mathbf{z}$ to represent the common features of an identity and filters the variance of the distribution, thus regularizes the

variations. A detailed definition can be found in the following sections. As a result, VNAE is optimized by

$$\mathcal{L}_{\text{VNAE}} = \mathcal{L}_{\text{cls}} + \lambda\mathcal{L}_{\text{triplet}} - \alpha\text{ELBO}_{\text{VNAE}} \tag{5}$$

$$\text{ELBO}_{\text{VNAE}} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right] - \beta D_{\text{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z})) \tag{6}$$

where λ and α are hyper-parameters for balancing the losses, and $\text{ELBO}_{\text{VNAE}}$ follows the β-VAE [18] setup that encourages to learn disentangled representation with β parameter. Together with the discriminative losses $\mathcal{L}_{\text{triplet}}$ and $\mathcal{L}_{\text{cls}}$, VNAE would be able to learn a disentangled latent representation, *i.e.*, distinctive features, for Re-ID.

Comparing with methods that utilise GAN to generate pseudo-images with additional annotations, and thus augment the training set to remove a specific variation from the feature embedding, VNAE has following merits: 1) VNAE models variation in an unsupervised manner and requires no additional annotations; 2) VNAE is cost-efficient with a unified generative and discriminative model; 3) VNAE is decoupled with the feature extraction methods and hence extensible to work with SOTA feature extractors.

3.3   Jensen-Shannon Triplet Loss

Triplet loss is a contrastive loss that optimizes toward letting the distance within positive pairs greater than the distance within negative pairs by a margin. Formally, for an anchor sample z_i, same-identity sample z_j and different-identity sample z_k, the triplet loss $\mathcal{L}_{\text{triplet}}$ is defined as:

$$\mathcal{L}_{\text{triplet}} = \max(d(z_i, z_j) - d(z_i, z_k) + \gamma, 0) \tag{7}$$

where $d(\cdot, \cdot)$ denotes Euclidean distance for a pair of samples and γ is a pre-defined margin. When applying to distribution means, the triplet loss can pull the centers of positive distribution pairs together and push the centers of negative distributions pairs away. However, conventional triplet loss ignores the distribution assumption of embeddings and might be susceptible to variations.

To alleviate the issue, we propose a novel metric learning loss for distribution learning, Jensen-Shannon triplet loss (JS triplet loss). We measure the pairwise distance with the Jensen-Shannon divergence (JS divergence)

$$D_{\text{JS}}(\mathbf{p} \parallel \mathbf{q}) = \frac{1}{2}D_{\text{KL}}(\mathbf{p} \parallel \mathbf{m}) + \frac{1}{2}D_{\text{KL}}(\mathbf{q} \parallel \mathbf{m}) \tag{8}$$

where $\mathbf{m} = \frac{1}{2}(\mathbf{p} + \mathbf{q})$.

The JS divergence can adequately reflect the similarity between two distributions. We compute the JS triplet loss following the scheme of computing the triplet loss between two embedding vectors. For an anchor distribution $\mathbf{z}_i$, a same-identity distribution $\mathbf{z}_j$ and a different-identity distribution $\mathbf{z}_k$, the JS triplet loss is:

$$\mathcal{L}_{\text{jst}} = \max(D_{\text{JS}}(\mathbf{z}_i \parallel \mathbf{z}_j) - D_{\text{JS}}(\mathbf{z}_i \parallel \mathbf{z}_k) + \gamma, 0) \tag{9}$$
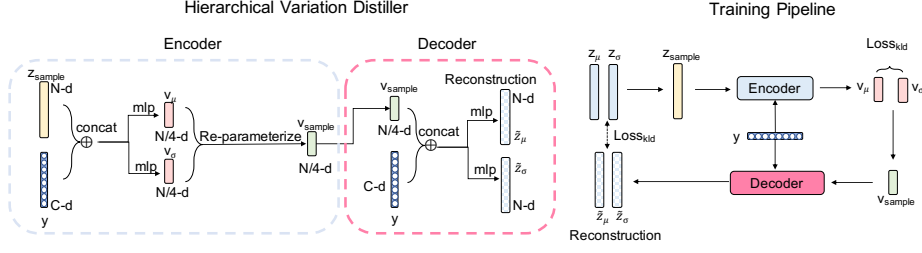
Fig. 3: Network design and training scheme of HVD. HVD is of encoder-decoder structure. A sample of latent distribution $\mathbf{z}$ is concatenated with the one-hot label $\mathbf{y}$ to encode variation distribution $\mathbf{v}$. A sample from $\mathbf{v}$ is again concatenated with the one-hot vector $\mathbf{y}$ to decode a reconstructed distribution $\tilde{\mathbf{z}}$. Two unsupervised loss are computed.

where γ is the margin parameter.

JS triplet loss has two major advantages: 1) JS triplet loss is variance aware. It considers the sample variance when computing the loss, which identifies noisy samples with large variance and output a smaller loss that improves the robustness of the network; 2) as JS divergence is symmetric and bounded in $[0, 1]$, $\mathcal{L}_{\mathrm{jst}}$ is bounded in $[0, 1 + \gamma]$. This gives a robust loss function against outliers and a numerically stable training process.

3.4   Hierarchical Variation Distiller

We introduce the Hierarchical Variation Distiller (HVD) module that further improves VNAE variation modeling with a hierarchical conditional VAE.

HVD is a higher level variational autoencoder following the VNAE framework. Besides random variable $\mathbf{x}$ and $\mathbf{z}$ in VNAE, we introduce two additional random variables, variation variable $\mathbf{v}$ and identity variable $\mathbf{y}$. We model a joint distribution of $p(\mathbf{x}, \mathbf{z}, \mathbf{v}, \mathbf{y})$, factorized as

$$p(\mathbf{x}, \mathbf{z}, \mathbf{v}, \mathbf{y}) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z} \mid \mathbf{v}, \mathbf{y})p(\mathbf{v})p(\mathbf{y}) \quad (10)$$

where its graphical model representation is given by Fig. 4. The sub-model, $p(\mathbf{z} \mid \mathbf{v}, \mathbf{y})p(\mathbf{v})p(\mathbf{y})$, forms the HVD part. The intuition behind is: we hope that a higher level random variable $\mathbf{v}$ is able to capture the variations over the embedding of the same identity, e.g., pose variations and occlusions. An identity variable $\mathbf{y}$ is assumed to be independent of $\mathbf{v}$, i.e., variation should be universally applied to all

Fig. 4: Hierarchical variation distillation. HVD distills the variance of embedding variable $\mathbf{z}$ of observation $\mathbf{x}$ with variation variable $\mathbf{v}$ and identity variable $\mathbf{y}$.

identities. Specifically, we assume variation variable has a prior of $p_\theta(\mathbf{v}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and identity variable follows a categorical distribution corresponding to the one-hot identity label $\mathbf{y} \sim \mathrm{Cat}(\mathbf{y})$.
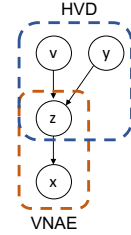
To efficiently sample $\mathbf{v}$, we follow the VAE setup and sample $\mathbf{v}$ from a proposal distribution, *i.e.*, encoder, $q_\phi(\mathbf{v} \mid \mathbf{z}, \mathbf{y})$. The encoder generates variation variable samples with higher likelihood to produce samples $\mathbf{z}$ and $\mathbf{y}$. As a result, this leads to a new ELBO for unsupervised learning:

$$
\begin{aligned}
\mathrm{ELBO}_{\mathrm{HVD}} = {} & \log \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})q_\phi(\mathbf{v}|\mathbf{z},\mathbf{y})} \left[p_\theta(\mathbf{x} \mid \mathbf{z}) \right] + \log p_\theta(\mathbf{y}) \\
& - D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z} \mid \mathbf{v}, \mathbf{y})) - D_{\mathrm{KL}}(q_\phi(\mathbf{v} \mid \mathbf{z}, \mathbf{y}) \parallel p_\theta(\mathbf{v})) \\
\propto {} & \log \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})q_\phi(\mathbf{v}|\mathbf{z},\mathbf{y})} \left[p_\theta(\mathbf{x} \mid \mathbf{z}) \right] \\
& - D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z} \mid \mathbf{v}, \mathbf{y})) - D_{\mathrm{KL}}(q_\phi(\mathbf{v} \mid \mathbf{z}, \mathbf{y}) \parallel p_\theta(\mathbf{v}))
\end{aligned}
\tag{11}
$$

where $\log p_\theta(\mathbf{y})$ is a constant and can be ignored during optimization. The condition $\mathbf{y}$ encourages the network to learn a disentangled representation for each identity. The derivation of the $\mathrm{ELBO}_{\mathrm{HVD}}$ can be found in the supplementary material.

Essentially, HVD is an auxiliary module to the VNAE. We illustrate details of HVD in Fig. 4. HVD improves the disentangled representation learning during training. It can be removed during inference and the remaining VNAE network keeps to be a light-weight Re-ID module.

Algorithm 1: HAVANA

Input: $\mathbf{x}_{\mathrm{img}}$: cropped person image, $\mathbf{y}$: person identity label, **extractor**($\cdot$): trained feature extractor network

while *not converged* **do**

$\quad$ $\mathbf{x} \leftarrow$ **extractor**($\mathbf{x}_{\mathrm{img}}$) $\qquad\qquad$ (extract feature $\mathbf{x}$ from $\mathbf{x}_{\mathrm{img}}$)

$\quad$ $\mathbf{z}_\mu, \mathbf{z}_\sigma \leftarrow$ VNAE.encode($\mathbf{x}$) $\qquad$ (map $\mathbf{x}$ to latent distribution $\mathbf{z}$)

$\quad$ $\mathbf{z}_{\mathrm{sample}} \leftarrow$ reparameterize($\mathbf{z}_\mu, \mathbf{z}_\sigma$) $\qquad\qquad$ (sample from $\mathbf{z}$)

$\quad$ $\tilde{\mathbf{x}} \leftarrow$ VNAE.decode($\mathbf{z}_{\mathrm{sample}}$) $\qquad\qquad$ (reconstruct to $\mathbf{x}$)

$\quad$ $\mathbf{v}_\mu, \mathbf{v}_\sigma \leftarrow$ HVD.encode($\mathbf{z}_{\mathrm{sample}}, \mathbf{y}$) $\quad$ (map $\mathbf{z}$ to variation distribution $\mathbf{v}$)

$\quad$ $\mathbf{v}_{\mathrm{sample}} \leftarrow$ reparameterize($\mathbf{v}_\mu, \mathbf{v}_\sigma$) $\qquad\qquad$ (sample from $\mathbf{v}$)

$\quad$ $\tilde{\mathbf{z}}_\mu, \tilde{\mathbf{z}}_\sigma \leftarrow$ HVD.decode($\mathbf{v}_{\mathrm{sample}}, \mathbf{y}$) $\qquad$ (reconstruct to distribution $\mathbf{z}$)

$\quad$ $\mathcal{L}_{\mathrm{cls}} \leftarrow$ CrossEntropy(classifier($\mathbf{z}_\mu$), $\mathbf{y}$)

$\quad$ $\mathcal{L}_{\mathrm{jst}} \leftarrow$ JSTripletLoss($\mathcal{N}(\mathbf{z}_\mu, \mathbf{z}_\sigma^2)$, $\mathbf{y}$)

$\quad$ $\mathcal{L}_{\mathrm{reconx}} \leftarrow$ MeanSquareError($\mathbf{x}, \tilde{\mathbf{x}}$)

$\quad$ $\mathcal{L}_{\mathrm{klz}} \leftarrow -D_{\mathrm{KL}}(\mathcal{N}(\mathbf{z}_\mu, \mathbf{z}_\sigma^2) \parallel N(\tilde{\mathbf{z}}_\mu, \tilde{\mathbf{z}}_\sigma^2))$

$\quad$ $\mathcal{L}_{\mathrm{klv}} \leftarrow -D_{\mathrm{KL}}(\mathcal{N}(\tilde{\mathbf{v}}_\mu, \mathbf{v}_\sigma^2) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I}))$

$\quad$ $\mathcal{L}_{\mathrm{total}} \leftarrow \mathcal{L}_{\mathrm{cls}} + \mathcal{L}_{\mathrm{jst}} + \alpha(\mathcal{L}_{\mathrm{reconx}} + \beta(\mathcal{L}_{\mathrm{klz}} + \mathcal{L}_{\mathrm{klv}}))$

$\quad$ Backpropagate($\mathcal{L}_{\mathrm{total}}$)

end

Output: $\mathbf{z}_\mu$: the output embedding vector

3.5   HAVANA

Combining the two network modules, VNAE and HVD, and one new metric learning loss, JS triplet loss, we present the holistic framework as Hierarchical and Variation-Normalized Autoencoder (HAVANA). HAVANA is a hybrid system
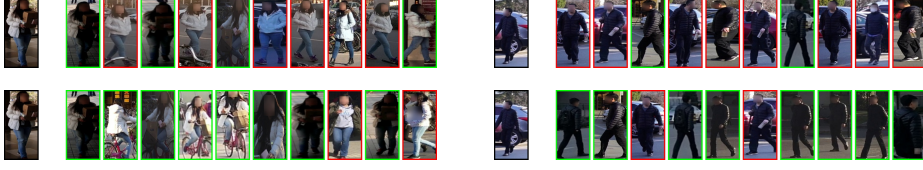
Fig. 5: Comparison of top-10 retrieval results. Upper row: Baseline results. Lower row: HAVANA results. Images in green boxes are positive results and images in red boxes are negative results.

with both unsupervised and supervised training signals. The overall loss is

$$\mathcal{L}_{\text{HAVANA}} = \mathcal{L}_{\text{cls}} + \lambda\mathcal{L}_{\text{jst}} - \alpha\text{ELBO}_{\text{HVD}} \tag{12}$$

$$\text{ELBO}_{\text{HVD}} = \log \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})q_\phi(\mathbf{v}|\mathbf{z},\mathbf{y})}\left[p_\theta(\mathbf{x} \mid \mathbf{z})\right]$$
$$- \beta\left(D_{\text{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z} \mid \mathbf{v},\mathbf{y})) + D_{\text{KL}}(q_\phi(\mathbf{v} \mid \mathbf{z},\mathbf{y}) \parallel p_\theta(\mathbf{v}))\right) \tag{13}$$
$$= \mathcal{L}_{\text{reconx}} + \beta(\mathcal{L}_{\text{klz}} + \mathcal{L}_{\text{klv}}) \tag{14}$$

where λ, α and β are hyper-parameters for loss ratio.

HAVANA benefits from the ELBO_{HVD} for unsupervised disentangled representation learning with hierarchical variance regularization; simple extra supervised signals, including classification loss $\mathcal{L}_{\text{cls}}$ and powerful Jensen-Shannon triplet loss $\mathcal{L}_{\text{jst}}$ help to optimize the latent representation to generate robust and accurate Re-ID results. We summarize the overall pipeline in Alg. 1.

In practice, we found that imposing a covariance constraint to variable $\mathbf{z}$ can significantly improve the performance. Specifically, we enforce $\mathbf{z}_\sigma = \mathbf{I}$. The intuition is that the limited images-per-identity [51] might form a biased empirical distribution; covariance constraint softens the biased distribution and tries to recover the invariant identity distribution. We empirically validate the influence of covariance constraint in our ablation studies.

4   Experiments

In this section, we evaluate our approach on three popular large-scale datasets: Market-1501, DukeMTMC-reID and MSMT17. We also perform extensive ablation studies to understand the influence of each proposed component.

4.1   Experiment Setup and Implementation Details

Evaluation is performed on Market-1501[49], DukeMTMC-reID [35] and MSMT17 [41]. The detailed dataset statistics can be found in Table 1. Cumulative matching characteristics (CMC) Rank-1 accuracy and mAP are used as evaluation metrics. We adopt ResNet-50 as our backbone. Our implementation is based on an open-source code-base [30,29]. We follow their training schemes: the last stride of ResNet-50 is set from 2 to 1; the training and testing images are cropped into a size of 256x128; the dimension of output embedding is 2048;

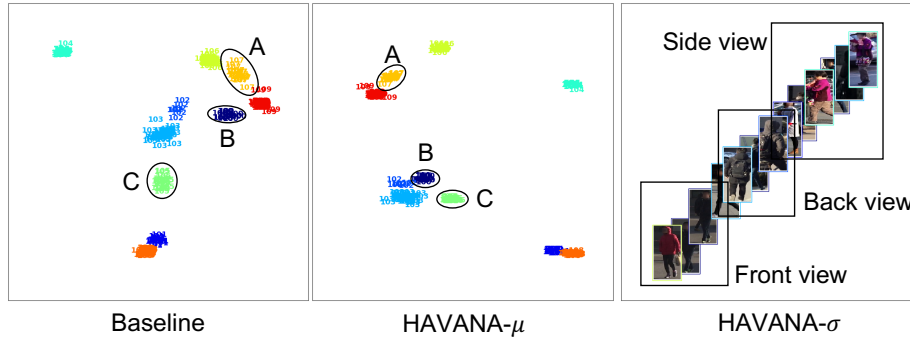Baseline   HAVANA-μ HAVANA-σ

Fig. 6: t-SNE visualization results. 10 random identities are sampled and represented in different colors. Baseline is the visualization of baseline embedding. HAVANA-μ is the visualization of HAVANA embedding, $i.e.$, $\mathbf{z}_\mu$. HAVANA-σ is the visualization of the standard deviation vector of the learned distribution, $i.e.$, $\mathbf{z}_\sigma$. Regions A, B and C contains the embedding from same IDs across methods. Comparing with baseline, HAVANA achieves a more compact identity group in the embedding space. In HAVANA-σ, we observe that pose variation has been unsupervisingly clustered in the $\mathbf{z}_\sigma$ vector, which strongly advocates that HAVANA is capable of variation modeling.

learning rate warm-up, label smoothing [38], random flip, Random Erasing [52] and BNNeck are employed; batch size is set to 64 with 4 images of 16 identities sampled each batch. We set the learning rate and weight decay to 1e-5 and 5e-4 respectively, and train the network for 300 epochs without

Dataset	# IDs (T-Q-G)	# images (T-Q-G)
Market-1501	751-750-751	12936-3368-15913
DukeMTMC	702-702-1110	16522-2228-17661
MSMT17	1041-3060-3060	30248-11659-82161

Table 1: Dataset statistics. T: Train. Q: Query. G: Gallery.

learning rate decay. We have no test-time augmentation unless specified. We take ResNet-50 with the cross-entropy loss and the triplet loss as baseline, and train the baseline under above described setup.

4.2   Experiment Results

We compare our model with state-of-the-art methods on Person Re-ID. We report our results on all three aforementioned datasets in Table 2. Specifically, we compare HAVANA with the SOTA methods, including both non-generative Re-ID approaches and generative approaches. With a baseline model with ResNet-50 backbone as the feature extractor, we show that our approach clearly outperforms all prior generative model based Person Re-ID methods. Replacing our feature extractor from the baseline model to a Multiple Granularities Network (MGN) [40] model with ResNet-50 backbone, HAVANA achieves state-of-the-art results on both Market-1501 dataset and DukeMTMC dataset, meanwhile a strong result

Method	Market-1501		DukeMTMC		MSMT17	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
MGN [40] (2018 ACMM)	86.9	**95.7**	78.4	88.7	-	-
BoT [30] (2019 CVPRW)	85.9	94.5	76.4	86.4	-	-
MHN [5] (2019 ICCV)	85.0	95.1	77.2	89.1	-	-
BDB [9] (2019 ICCV)	86.7	95.3	76.6	89.0	-	-
P^2-Net [16] (2019 ICCV)	85.6	95.2	73.1	86.5	-	-
SFT [28](2019 ICCV)	82.7	93.4	73.2	86.9	47.6	73.6
CAR [55](2019 ICCV)	84.7	96.1	73.1	86.3	-	-
BAT-net [10] (2019 ICCV)	87.4	95.1	77.3	87.7	56.8	79.5
SONA [42] (2019 ICCV)	**88.8**	95.6	78.3	**89.4**	-	-
ABD-Net [7] (2019 ICCV)	88.3	95.6	**78.6**	89.0	**60.8**	**82.30**
Generative Approaches						
USG-GAN[51] (2017 ICCV)	56.2	78.1	-	-	-	-
CSA-GAN[53] (2018 CVPR)	71.6	89.5	57.6	78.3	-	-
PN-GAN [33](2018 ECCV)	72.6	89.4	53.2	73.6	-	-
FD-GAN [11] (2018 NIPS)	77.7	90.5	64.5	80.0	-	-
DGNet [50] (2019 CVPR)	86.0	94.8	74.8	86.6	52.3	77.2
HAVANA	86.9	94.4	77.4	87.7	52.3	76.8
HAVANA[†]	87.4	94.5	77.9	87.9	53.3	76.9
HAVANA(MGN)	88.9	95.2	80.4	89.1	55.3	77.5
HAVANA(MGN)[†]	**89.2**	**95.3**	**80.8**	**89.4**	**56.0**	**77.9**

Table 2: Experimental comparison with state-of-the-art. Red denotes our performance and blue denotes the best performance achieved by existing methods. † denotes that the result is obtained with flip-test.

on MSMT17 dataset. The results concretely validate HAVANA's extensibility to be combined with other SOTA methods. Qualitative results are shown in Fig. 5 and Fig. 6.

4.3   Component Analysis

Variation-Normalized Autoencoder Firstly, we experiment on Variation-Normalized Autoencoder, which our later proposed methods developed from. In the experiment, the hyper-parameter α and β are set to 0.2 and 1.0 for Market-1501 dataset, 0.2 and 2.0 for MSMT17 dataset. The result in Table 3 shows that VNAE outperforms the baseline model by an observable margin, especially on the larger datasets MSMT17 where the intra-class variations signify. The experiment results indicate that our variation normalization is effective, particularly on large-scale datasets.

J	γ	mAP	Rank-1	Rank-5
-		50.1	74.3	85.9
✓	0.3	50.1	74.4	86.1
✓	0.4	50.3	74.2	86.2
✓	**0.5**	**50.6**	**75.0**	**86.3**
✓	0.6	50.2	74.3	86.1
✓	0.7	50.0	74.3	85.9

Table 4: Analyses on Jensen-Shannon triplet loss and margin value γ. J: JS triplet loss.

V	J	H	C	Dataset	mAP	Rank-1	Rank-5	Rank-10
				MSMT17	48.2	72.2	83.9	87.7
✓				MSMT17	50.1 (+1.9)	74.3	85.9	89.6
✓	✓			MSMT17	50.6 (+2.4)	75.0	86.3	89.8
✓	✓	✓		MSMT17	52.0 (+3.8)	75.9	87.2	90.3
✓	✓	✓	✓	MSMT17	52.3 (+4.1)	76.7	87.2	90.6
				Market-1501	85.8	94.2	98.4	99.0
✓				Market-1501	86.5 (+0.7)	94.2	98.3	98.9
✓	✓			Market-1501	86.5 (+0.7)	94.2	98.3	98.9
✓	✓	✓		Market-1501	86.5 (+0.7)	93.9	98.2	98.9
✓	✓	✓	✓	Market-1501	86.9 (+1.1)	94.4	98.3	99.0

Table 3: Component analysis. V: Variation-Normalized Autoencoder. J: Jensen-Shannon Triplet Loss. H: Hierarchical Variation Distiller. C: Covariance constraint.

Jensen-Shannon Triplet Loss To optimize the training of VNAE, we replace the conventional triplet loss with Jensen Shannon triplet loss. We set the loss ratio λ for JS triplet to 1.0. In Table 4, we analyse margin γ's effect to the performance of JS triplet loss. The experiments show that setting the γ to 0.5 results in best performance on MSMT17 dataset, which improves 0.5 mAP from the previous VNAE's performance. The success of combining latent distribution learning and the Jensen-Shannon triplet loss not only supports the novel metric loss's potency, but also implies that our intra-class variation modelling is meaningful.

Hierarchical Variation Distiller To enhance the representation quality, we employ Hierarchical Variation Distiller module to complete the HAVANA framework. The hyper-parameter α and β are set to 0.2 and 1.0 for Market-1501 dataset, 0.2 and 2.0 for MSMT17 dataset. Table 3 shows a comparison before and after attaching the HVD module. The results show that a hierarchical representation can be beneficial for the Person Re-ID task.

Covariance Constraint We put a covariance constraint on the latent representation by rewriting the original loss term $\mathcal{L}_{\text{klz}}^{\text{ori}} = \mathcal{L}_{\text{reconz}} + \mathcal{L}_{\text{klz}}$, where $\mathcal{L}_{\text{reconz}} = \text{MSE}(\mathbf{z}_\mu, \tilde{\mathbf{z}}_\mu)$ and $\mathcal{L}_{\text{klz}} = D_{\text{KL}}(\mathcal{N}(\mathbf{z}_\mu, \mathbf{z}_\sigma^2) \| \mathcal{N}(\mathbf{0}, \mathbf{I}))$. We show the comparison of training loss curves in Fig. 7. Without covariance constraint, the generative loss L_{klv} dissipates due to a well-known phenomenon, posterior collapse [27], which leads the representation to be less meaningful. Covariance constraint balances the generative loss L_{klz} and the reconstruction loss L_{reconz} and improves representation learning.
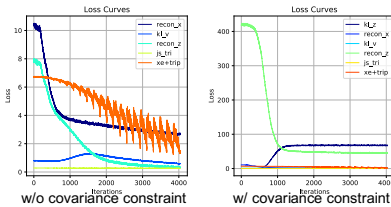

Fig. 7: Comparison of loss curves before and after imposing the covariance constraint (CC). Without CC, the loss curve of the generative loss $\mathcal{L}_{\text{klv}}$ dissipates in an early stage of training. With CC, the generative loss $\mathcal{L}_{\text{klz}}$ and the reconstruction loss $\mathcal{L}_{\text{reconz}}$ are maintained at a stable level.

Method	Market-1501				MSMT17			
	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
baseline	85.8	94.2	98.4	99.0	48.2	72.2	83.9	87.7
$+\mathcal{L}_{\mathrm{reconx}}$	85.7	93.8	98.0	98.9	47.9	74.1	85.7	89.3
$+\mathcal{L}_{\mathrm{klz}}$	86.5	94.2	98.3	98.9	50.1	74.3	85.9	89.6
$+\mathcal{L}_{\mathrm{jst}}$	86.5	94.3	98.2	98.9	50.6	75.0	86.3	89.8
$+\mathcal{L}_{\mathrm{reconz}}$	86.6	94.5	98.2	99.0	52.3	76.5	87.2	90.4
$+\mathcal{L}_{\mathrm{klv}}$	86.9	94.4	98.3	99.0	52.3	76.7	87.2	90.6

Table 5: Component analysis on all loss proposed in the HAVANA framework. We show that each loss contributes to final performance.
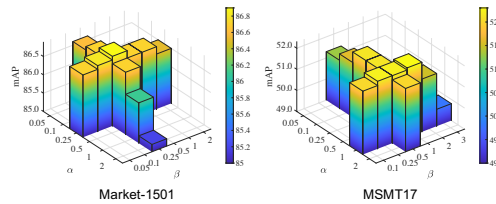

Fig. 8: 3-D bar chart visualisation of hyper-parameter analysis. HAVANA is insensitive to α and β values.

Dataset	Method	mAP	Rank-1
Market-1501	Baseline	85.8	94.2
Market-1501	Ours w/o RE	86.2	94.3
Market-1501	Ours w/ RE	86.9	94.4
MSMT17	Baseline	48.2	72.2
MSMT17	Ours w/o RE	50.8	75.9
MSMT17	Ours w/ RE	52.3	76.7

Table 6: Ablation study on artificial variation, Random Erasing.

Table 3 shows that our proposed HAVANA performs better when trained with the covariance constraint, which not only justifies our intuition, but also suggests that the bias stemmed from unseen data can be alleviated in such a way.

4.4   Experiment Results

Overall Loss Analysis In this section, we use various experiments to analyse the contribution of each loss proposed in this paper, as shown in Table 5. The analysis is conducted following the final scheme, HAVANA with covariance constraint. The results show that all losses proposed in the paper are important to the final performance.

Analyses on Hyper-parameters α and β In Fig.8, we use 3-D bar chart to visualize hyper-parameters' effect on performance: two of the axis are α and β, and the other axis is the performance metric. The experiment results show that the optimal β in a larger dataset, MSMT17, is higher than the optimal β in a smaller dataset, Market-1501, which aligns with the fact that more intra-class variations need to be accommodated in a larger dataset.

Effectiveness of Capturing Variations in Feature Space Random Erasing (RE) [52] is an augmentation method commonly seen in Person Re-ID works. RE randomly erases part of input images to imitate occlusions in real world. Our feature extractor is trained with the RE technique, which improves the extracted features' robustness to this artificial variation. To test whether HAVANA is capable of capturing such a subtle variation in the feature space, we train our model both with and without RE. The result in Table 6 shows that our HAVANA performs better when with RE, which suggests that the artificial variation leaking through the feature extractor can be further filtered out by our model.

Computational Cost In HAVANA, only the two-layer MLP that encodes from 2048-dimensional features to 2048-dimensional mean vectors is required test-time, whose parameters are $(2048 + 1) \times 2048 \times 2$. In Fig. 9, we show a comparison of the performance increment against the number of parameters increment. We firstly experiment on baseline with various ResNet-like backbones. In the figure, ResNet-18 baseline's result, which achieves 78.1 mAP on Market-1501 and 39.2 mAP on MSMT17 using 11M parameters, is set as the origin of the coordinate plane. We then experiment on HAVANA using baseline feature extractors previously obtained. We compare with a SOTA method, ABD-Net [7], as well. The number of parameters in ABD-Net is recorded from the official implementation. Our method outperforms the baseline models consistently. HAVANA also beats the SOTA model with fewer parameters on Market-1501. We thus show that HAVANA is cost-efficient.
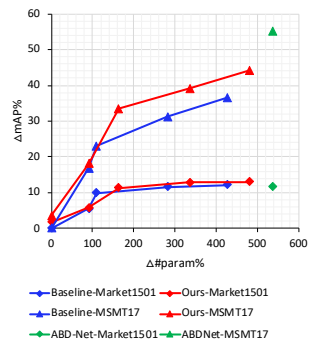

Fig. 9: Computational cost comparison. Our model (Red) shows consistent advantage over the baseline model (Blue) on the accuracy-parameter curve. Our model also outperforms SOTA model (Green) with less parameters on Market1501.

5 Conclusion

We present HAVANA, an extensible, light-weight hierarchical and variation-normalized autoencoder for the Person Re-ID task. We conducted extensive experiments and ablation studies on Market-1501, DukeMTMC-reID and MSMT17 datasets. We show that HAVANA significantly outperforms existing GAN-based methods on all datasets, and outperforms the SOTA Re-ID models on 2 datasets while achieving comparable results on the rest, with minimum supervision. We demonstrate VAE is naturally suitable to model variation in the Re-ID task.

However, our current HAVANA framework considers only basic VAE models. More complex VAE models, e.g., hyper-prior approaches [2], could be considered to improve disentangled representation learning.

References

1. Alemu, L.T., Pelillo, M., Shah, M.: Deep constrained dominant sets for person re-identification. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
2. Ansari, A.F., Soh, H.: Hyperprior induced unsupervised disentanglement of latent representations. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 3175–3182 (2019)
3. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan. ArXiv **abs/1701.07875** (2017)
4. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(8), 1798–1828 (Aug 2013). https://doi.org/10.1109/TPAMI.2013.50
5. Chen, B., Deng, W., Hu, J.: Mixed high-order attention network for person re-identification. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
6. Chen, T.Q., Li, X., Grosse, R.B., Duvenaud, D.: Isolating sources of disentanglement in variational autoencoders. In: NeurIPS (2018)
7. Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z., Wang, Z.: Abd-net: Attentive but diverse person re-identification. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
8. Chen, X., Kingma, D.P., Salimans, T., Duan, Y., Dhariwal, P., Schulman, J., Sutskever, I., Abbeel, P.: Variational lossy autoencoder. ArXiv **abs/1611.02731** (2016)
9. Dai, Z., Chen, M., Gu, X., Zhu, S., Tan, P.: Batch dropblock network for person re-identification and beyond. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
10. Fang, P., Zhou, J., Roy, S.K., Petersson, L., Harandi, M.: Bilinear attention networks for person retrieval. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
11. Ge, Y., Li, Z., Zhao, H., Yin, G., Yi, S., Wang, X., Li, h.: Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31, pp. 1222–1233. Curran Associates, Inc. (2018), `http://papers.nips.cc/paper/7398-fd-gan-pose-guided-feature-distilling-gan-for-robust-person-re-identification.pdf`
12. Gong, S., Cristani, M., Yan, S., Loy, C.C.: Person Re-Identification (10 2014)
13. Gonzalez-Garcia, A., van de Weijer, J., Bengio, Y.: Image-to-image translation for cross-domain disentanglement. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31, pp. 1287–1298. Curran Associates, Inc. (2018), `http://papers.nips.cc/paper/7404-image-to-image-translation-for-cross-domain-disentanglement.pdf`
14. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 5767–5777. Curran Associates, Inc. (2017), `http://papers.nips.cc/paper/7159-improved-training-of-wasserstein-gans.pdf`

15. Gulrajani, I., Kumar, K., Ahmed, F., Taïga, A.A., Visin, F., Vázquez, D., Courville, A.C.: Pixelvae: A latent variable model for natural images. ArXiv **abs/1611.05013** (2016)

16. Guo, J., Yuan, Y., Huang, L., Zhang, C., Yao, J.G., Han, K.: Beyond human parts: Dual part-aligned representations for person re-identification. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)

17. He, L., Wang, Y., Liu, W., Zhao, H., Sun, Z., Feng, J.: Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)

18. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. Iclr **2**(5), 6 (2017)

19. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X.: Interaction-and-aggregation network for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)

20. Kim, H., Mnih, A.: Disentangling by factorising. In: ICML (2018)

21. Kingma, D.P., Mohamed, S., Rezende, D.J., Welling, M.: Semi-supervised learning with deep generative models. ArXiv **abs/1406.5298** (2014)

22. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

23. Lample, G., Zeghidour, N., Usunier, N., Bordes, A., DENOYER, L., Ranzato, M.A.: Fader networks:manipulating images by sliding attributes. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 5967–5976. Curran Associates, Inc. (2017), `http://papers.nips.cc/paper/7178-fader-networksmanipulating-images-by-sliding-attributes.pdf`

24. Liu, F., Zhang, L.: View confusion feature learning for person re-identification. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)

25. Liu, J., Ni, B., Yan, Y., Zhou, P., Cheng, S., Hu, J.: Pose transferrable person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)

26. Louizos, C., Swersky, K., Li, Y., Welling, M., Zemel, R.S.: The variational fair autoencoder. CoRR **abs/1511.00830** (2015)

27. Lucas, J., Tucker, G., Grosse, R.B., Norouzi, M.: Understanding posterior collapse in generative latent variable models. In: DGS@ICLR (2019)

28. Luo, C., Chen, Y., Wang, N., Zhang, Z.: Spectral feature transformation for person re-identification. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)

29. Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S., Gu, J.: A strong baseline and batch normalization neck for deep person re-identification. IEEE Transactions on Multimedia pp. 1–1 (2019). https://doi.org/10.1109/TMM.2019.2958756

30. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019)

31. Miao, J., Wu, Y., Liu, P., Ding, Y., Yang, Y.: Pose-guided feature alignment for occluded person re-identification. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)

32. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning. In: NIPS (2017)

33. Qian, X., Fu, Y., Xiang, T., Wang, W., Qiu, J., Wu, Y., Jiang, Y.G., Xue, X.: Pose-normalized image generation for person re-identification. In: The European Conference on Computer Vision (ECCV) (September 2018)
34. Quan, R., Dong, X., Wu, Y., Zhu, L., Yang, Y.: Auto-reid: Searching for a part-aware convnet for person re-identification. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
35. Ristani, E., Solera, F., Zou, R.S., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. CoRR **abs/1609.01775** (2016)
36. Sønderby, C.K., Raiko, T., Maaløe, L., Sønderby, S.K., Winther, O.: Ladder variational autoencoders. In: NIPS (2016)
37. Sun, Y., Xu, Q., Li, Y., Zhang, C., Li, Y., Wang, S., Sun, J.: Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
38. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2818–2826 (2015)
39. Tschannen, M., Bachem, O., Lucic, M.: Recent advances in autoencoder-based representation learning. ArXiv **abs/1812.05069** (2018)
40. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: MM '18 (2018)
41. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
42. Xia, B.N., Gong, Y., Zhang, Y., Poellabauer, C.: Second-order non-local attention networks for person re-identification. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
43. Yang, W., Huang, H., Zhang, Z., Chen, X., Huang, K., Zhang, S.: Towards rich feature discovery with class activation maps augmentation for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
44. Ye, M., Shen, J., jie Lin, G., Xiang, T., Shao, L., Hoi, S.C.H.: Deep learning for person re-identification: A survey and outlook. ArXiv **abs/2001.04193** (2020)
45. Yu, T., Li, D., Yang, Y., Hospedales, T.M., Xiang, T.: Robust person re-identification by modelling feature uncertainty. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
46. Zhang, Z., Lan, C., Zeng, W., Chen, Z.: Densely semantically aligned person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
47. Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., Tang, X.: Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1077–1085 (2017)
48. Zhao, S., Song, J., Ermon, S.: Infovae: Information maximizing variational autoencoders. ArXiv **abs/1706.02262** (2017)
49. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. 2015 IEEE International Conference on Computer Vision (ICCV) pp. 1116–1124 (2015)

50. Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., Kautz, J.: Joint discriminative and generative learning for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
51. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
52. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. ArXiv **abs/1708.04896** (2017)
53. Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y.: Camera style adaptation for person re-identification. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 5157–5166 (2017)
54. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
55. Zhou, S., Wang, F., Huang, Z., Wang, J.: Discriminative feature learning with consistent attention regularization for person re-identification. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)