Soft-attention based person re-identification in real-world settings using variational autoencoders

Emna Ben Baoues, Imen Jegham, Mounim El Yacoubi, Anouar Ben Khalifa

Soft-Attention Based Person Re-Identification in Real-world Settings using Variational AutoEncoders

Emna BEN BAOUES*, Imen JEGHAM †, Mounim El Yacoubi ‡, Anouar BEN KHALIFA¶‖
*‡§Université de Sousse, Ecole Nationale d'Ingénieurs de Sousse,
LATIS- Laboratory of Advanced Technology and Intelligent Systems, 4023, Sousse, Tunisie;
‖Université de Jendouba, Institut National des Technologies et des Sciences du Kef, 7100, Le Kef, Tunisie;
†Samovar, CNRS, Télécom SudParis, Institut Polytechnique de Paris 19 Place Marguerite Perey, 91120 Palaiseau, France;
Email: *emna.benbaoues@eniso.u-sousse.tn, ¶‖anouar.benkhalifa@eniso.u-sousse.tn,

***Abstract*—Person re-identification is still an open challenging task in various fields due to numerous factors, including illumination changes, background clutter, pose state variations and cloth changes. Several approaches have been suggested to address this problem in the context of deep learning. Generative models, particularly Variational Autoencoders (VAEs), have emerged as promising tools to address these challenges by learning discriminative feature representations of individual images. In this paper, we present Soft-Attention based Person Re-Identification (SAPRI), a novel approach that combines VAEs with a supervised ReID method to enhance the resilience and efficacy of ReID systems. The proposed approach focuses on data reconstruction based on soft attention. Variational autoencoders encode principally person data, while ignoring irrelevant information. By incorporating supervised ReID, the model learns to appropriately classify persons in real world environments. Our SAPRI proposed method has been evaluated on well-known benchmarks, DukeMTMC-reID and CUHK03, demonstrating superior performance compared to existing state-of-the-art techniques in terms of the mean Average Precision evaluation metric (mAP). Additionally, qualitative results show the effectiveness of the VAE in generating discriminative representations of person images.**

***Index Terms*—Person ReID, Soft attention, Generative models, Variational autoencoders, Data reconstruction, Latent space, Surveillance systems.**

I. Introduction

Person Re-identification (ReID) has gained prominence in computer vision, particularly in security & video surveillance systems domains [1]. In recent times, within the increasing need for efficient and accurate techniques for person matching across different camera views and environments, person ReID has gained attention from numerous researchers [1]–[6]. However, person ReID remains challenging with the pose variations in uncontrollable environments. Other issues include different illumination conditions, camera view points, cloth changes, background clutter and occlusion [7]. Owing of these problems, a ReID system remains incapable of re-identifying the right person with good precision.

Deep generative models, have been proposed as a promising solution for feature learning and representation, including data enhancement and data augmentation, learning better feature representations of the data distribution or enhancing the diversity and richness of the training dataset. These models are useful for capturing complex changes in person appearances across different camera angles. Nevertheless, the quality of generated images for person ReID depends on various factors, including image size, variability, resolution and GAN's generalization capabilities. Among these deep models, the literature highlights two prominent types, deep autoencoders (AE) [8] and generative adversarial networks (GAN) [9]. GAN stands out as a quintessential unsupervised learning framework, comprising both generative and discriminative components. On the other side, AE is a specialized model designed for high-dimensional data, featuring two neural networks namely, encoder and decoder. The encoder analyzes input images to extract underlying characteristics, while the decoder attempts to recreate the original image.

In particular, variational autoencoder (VAE) [10], a specific type of auto-encoder, learns to encode and decode data, but with the added capability of learning a latent space that has useful properties such as smoothness and continuity. In a VAE, the encoder depicts the input as a probability distribution within a compressed, reduced-dimensional representation named the latent space, instead of directly mapping it to a single point. Subsequently, the decoder utilizes this distribution to sample and produce new data points.

Recently, VAEs and their variants [11]–[15] have emerged as promising tools in addressing the multifaceted challenges of ReID and enhancing generated images' quality, for better matching of individuals across camera views. These kind of deep generative models are trained to not only reconstruct input data but also to regularize the latent feature space to follow a prior distribution. This regularization stimulates the model to acquire a structured and continuous latent space, which can be advantageous for ReID tasks. Several studies have explored the VAE framework to learn better latent representations of complex data, including images of individuals captured from various camera angles and lighting conditions.

In this study, we explore the soft attention capability to effectively recognize individuals [16]. Attention mechanisms enable the model to concentrate on pertinent aspects of the inputs, improving its capacity to capture crucial features and spatial relationships within the original data, consequently

resulting in more discriminant latent representations. Notably, we aim to offer promising avenues and innovative solutions for addressing numerous complexities inherent in person ReID in challenging environments and enabling more effective person matching and identification. We outline our contributions in the following manner:

- We propose SAPRI (Soft-Attention based Person Re-Identification), a Re-ID scheme that reliably classifies people under real world environments.
- We introduce a novel soft attention mechanism that transforms input data into a latent space while simultaneously learning to reconstruct the real-data and handle the complexity inherent in ReID datasets.
- Our experimental results on two popular benchmarks illustrate that our approach ameliorates the effectiveness and resilience of the ReID system when compared to state-of-the-art approaches.

In Section 2, we summarize the various research works based on VAEs, explore their various extensions, and examine their applications to person ReID. Our proposed SAPRI is detailed in Section 3. In Section 4, we detail our extensive experiments on different benchmark datasets to showcase the assess SAPRI w.r.t existing methods. Finally, we present a summary of our findings and discuss the implications of our work in the concluding section.

II. VAEs for Person-ReID

In the domain of person ReID, scholars have extensively explored various methodologies to address the challenges of identifying persons across different multi-camera views. Traditional approaches often rely on handcrafted features or deep learning techniques. In [17], the researchers propose an approach that integrates deep learning techniques with graph-based modeling to effectively handle occlusions and extract high-level features. Nevertheless, the complexity of graph-based modeling can make the approach computationally expensive. Zhong *et al.* [2] have combined metric learning with cross-view feature aggregation and re-ranking strategies, facilitating the learning of discriminative representations across different camera views. This method can also lead to additional computational costs due to the re-ranking strategies and can be sensitive to variations in video quality.

To overcome these limitations, numerous works on person ReID have turned to generative deep models to tackle the challenges inherent to cross-camera person matching. These models, including GANs and VAEs, offer sophisticated techniques for discriminative feature learning and representation, essential in capturing the complex variations in person appearances across diverse camera views. The discriminative power of GANs can be leveraged to generate realistic person images, aiding in data augmentation and enriching the diversity of training datasets. By integrating such generative deep models into person ReID pipelines, researchers strive to enhance the robustness, scalability, and generalization capabilities of existing systems, thus pushing forward the forefront of person matching in intricate surveillance settings. Notable studies in this domain include the work by Baoues *et al.* [5], who proposed a GF2PReID framework for generating person images which reconstructed facial features with remarkable accuracy. This method is helpful in restoring missing or damaged facial regions in images, and has demonstrated prowess in synthesizing high-quality person images. On the other hand, Jiang *et al.*, proposed a GAN-based method for generating diverse image variations with better quality for data augmentation, which can improve the precision of person re-identification systems.

In the recent years, VAEs have gained significant attention in ReID tasks. This surge in interest is attributed to the VAEs' capability to acquire discriminative latent representations of data, capturing important features and characteristics, while keeping the GAN's advantage of generating synthetic examples to augment the training set. Nowadays, several research works showcase the diverse applications of VAEs instead of GANs in person ReID tasks, including feature learning, image generation, attention mechanisms, and graph embedding. Zheng *et al.* (2017) [3] proposed a VAE-based framework which integrates both body and latent part features for feature learning-based person ReID. By capturing both global and local contextual information, this approach effectively enhances the discriminative power of learned representations. Kim *et al.* (2019) introduced the variational discriminator bottleneck framework [18], which combines VAEs with adversarial learning to improve generative models and representation learning tasks. Similarly, Ma *et al.* (2019) focused on image generation in person ReID, leveraging VAEs with a progressive pose-attention transfer mechanism [19]. This work emphasizes the importance of VAEs in capturing pose-related information, further enhancing the discriminative capabilities of ReID systems. Moreover, Liu *et al.* explored the use of adversarially regularized graph autoencoders based on VAEs for graph embedding [20], showcasing the flexibility and potential of VAEs in learning discriminative latent representations.

Furthermore, in [21], Wu *et al.* introduce a novel approach to video-based person ReID, aiming to match individuals across camera views in unaligned video footage with limited labeled data. The proposed method leverages variational recurrent neural networks trained adversarially to generate discriminative view-invariant latent variables capturing temporal dependencies. Other authors integrate attention mechanisms with VAEs for person ReID tasks, enabling the models to capture context-aware features and enhance the precision of person matching in re-identification scenarios. For example, Wei *et al.* (2022) propose, in a recent work, a reciprocal bidirectional framework for infrared person ReID [22], attaining state-of-the-art performance through unifying modalities and enhancing discriminative feature learning. Their model employs bidirectional image translation sub-networks and attention mechanism-based feature embedding networks. Our proposed approach differs from the mentioned works by using an explicit cost function to facilitate stable training and empowers our VAE-model to grasp the intrinsic structure of the initial data. Table I highlights diverse VAE-based approaches, by illustrating the various innovations and modifications made

to VAE architectures.

TABLE I
SUMMARY OF VAE VARIANTS

Category	VAE Variants	Year
Regularization	Information Maximizing VAE [11]	2017
Normalization Techniques	Group Normalization VAE [14]	2018
	Batch Normalization VAE [23]	2020
Add Noise to Inputs	Denoising Variational Autoencoder [24]	2020
	Mixture of gaussian vae [12]	2020
Self-Attention	Self-Attention VAE [15]	2019
	Recurrent Self-Attention VAE [13]	2019

VAEs have showed their ability to learn compact and informative representations of person images while simultaneously reconstructing the original data. This reconstruction process imposes a constraint on the learned latent space, encouraging the extraction of robust and discriminative features to correctly re-identify persons.

III. PROPOSED SAPRI APPROACH

In this section, we detail our novel approach utilizing a mixed architecture that combines VAE and a CNN model for person ReID. Our SAPRI aims to reliably re-identify people while addressing the challenges faced by traditional person ReID systems including lighting variations, pose changes, and cluttered backgrounds.

As shown in Figure 1, the first component of our approach is the VAE, which is trained on an extensive dataset of person images. The VAE effectively learns to encode the fundamental features that differentiate one individual from another by capturing both their appearance and structural information. By modeling the distribution of person images in a latent space, the VAE provides a compact and meaningful representation in reconstructed images where relevant data is highlighted. The second element of our approach is the CNN model, which leverages the VAE output for person recognition. It relies on a Residual Neural Network (ResNet-18) baseline, initially trained on ImageNet, and later fine-tuned specifically for person ReID.

A. Data Reconstruction

VAEs belong to a class of generative based models that, once trained to model a distribution, can reconstruct new samples that match the distribution, even if the VAE has not seen those specific samples before. Introduced by Kingma *et al.* [10] in 2013, the architecture of a VAE is relatively straightforward, comprising two components: the encoder and decoder. Taking an image denoted I as input, the encoder produces feature embeddings. This is commonly accomplished utilizing a neural network, which translates the input data into mean (μ) & variance (σ^2) constants of a Gaussian distribution within the latent feature space. Mathematically, the encoder process follows Bayesian inference to describe the data in the latent space. This can be expressed in the following manner:

$$\mu, \log \sigma^2 = \text{Encoder}(I) \qquad z \sim \mathcal{N}(\mu, \sigma^2)$$

Within this framework, the notation $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian distribution characterized by mean μ and variance σ^2. These parameters are then utilized to generate a latent feature z. This process is mathematically represented by the following equation:

$$z = \mu + \sigma \odot \epsilon$$

Here ϵ is obtained from a Gaussian distribution.

On the other hand, the decoder receives the embedding space z from the encoder and reconstructs the original image I. This is accomplished by employing a neural-network, which leverages its internal layers and parameters to decode the latent representation and generate the reconstructed data. Mathematically, the decoder process can be represented as follows:

$$\hat{I} = \text{Decoder}(z)$$

By inputting the latent representation into the decoder, the network creates the reconstructed image $\hat{I}$.

The encoder & decoder components of a VAE are characterized by their probabilistic nature, represented by $q_\phi(z|x)$ for the approximate posterior and $p_\theta(x|z)$ for the likeliness of data x given the latent feature z:

$$q_\phi(z|x) = \mathcal{N}(z; \mu_{z|x}, \sigma^2_{z|x} I)$$

Here, the neural network computes the values of $\mu_{z|x}$ and $\sigma_{z|x}$ based on the input data x using the variational parameter ϕ.

Similarly, $p_\theta(x|z)$ is modeled as a multivariate Gaussian distribution derived from the embedding space z:

$$p_\theta(x|z) = \mathcal{N}(x; \mu_{x|z}, \sigma^2_{x|z})$$

Here, $\mu_{x|z}$ and $\sigma_{x|z}$ are calculated from the data x throughout a neural network using variational parameter θ.

Throughout the training process, the VAE strives to reduce the reconstruction error measuring the disparity between the real data and its corresponding reconstructed version. This reconstruction loss encourages the VAE to acquire meaningful representations in the latent feature space, facilitating accurate reconstruction of the original data. Furthermore, VAEs include a regularization term associated with Kullback-Leibler (KL) divergence, denoted as $D_{KL}(q_\phi(z|x)||p_\theta(z|x))$. The purpose of this term is to encourage a closer approximation between the learned distribution in the latent space $q_\phi(z|x)$, and a target distribution $p_\theta(z|x)$. By minimizing the KL divergence, the VAE aims to align the learned distribution with the desired target distribution, promoting more meaningful and structured latent representations.

Since direct computation of this KL divergence is not feasible, our approach involves maximizing the cumulative variational lower bound on the marginal likelihood for each individual data point x_i, where i ranges from 1 to n.

Hence, the VAE cost function can be expressed as follows:

$$L_{\text{VAE}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} q\phi(z|x_i)[\log p_\theta(x_i|z)] - \frac{1}{N} \sum_{i=1}^{N} D_{KL}(q_\phi(z|x_i)||p_\theta(z))$$

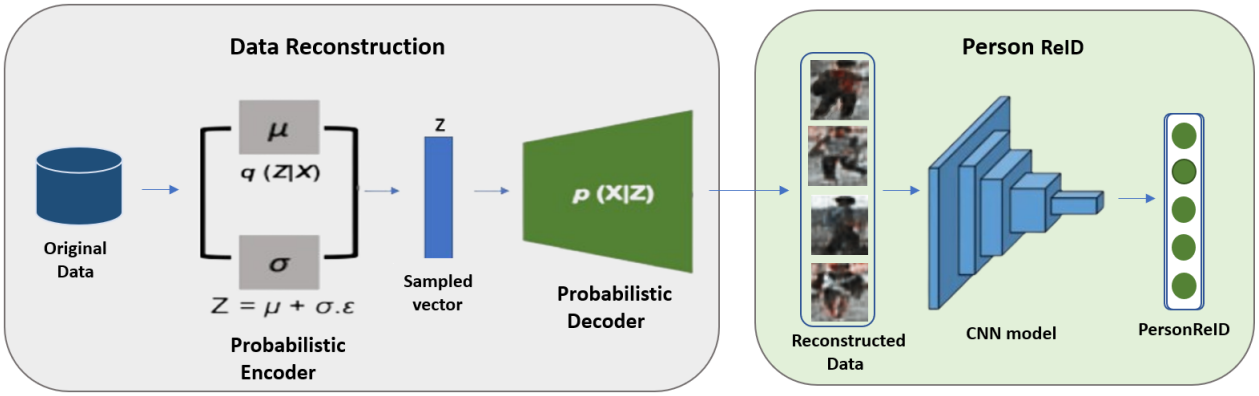

Fig. 1. The general layout of our SAPRI framework involves the encoder producing two latent variables , representing the mean and standard deviation (μ & σ) parameters of the distribution learned during training. These parameters allow us to sample a latent vector Z. Furthermore, we employ ResNet-18 to extract discriminative representations from the generated images. These features are subsequently utilized to compare and match individuals across various images.

Here, the summation is performed across the training samples $\{x_i\}_{i=1}^{N}$. The expression on the right-hand side of this equation can be interpreted as a metric quantifying the reconstruction error, compelling the VAE to reconstruct the inputs accurately. Meanwhile, the second component serves as a regularization term (KL), which is both analytically and differentiably defined. A more detailed expression of this component is as follows:

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{J}\left(1+\log\left((\sigma_{(i)j})^2\right)-(\mu_{(i)j})^2-(\sigma_{(i)j})^2\right)$$

B. Person ReID

In this section, we describe our image recognition method based on a residual neural network. After training the VAE, new images were reconstructed and fed into the CNN model, which further refines the features and performs the final identification by effectively extracting discriminative features from the generated images, enabling accurate matching and identification of individuals across different camera views in different illumination variations. The CNN model in our study is a ResNet-18 containing 18 deep residual connections layers, which help in addressing the problem of vanishing gradients during training. This model is pre-trained using the ImageNet large-scale dataset, and subsequently fine-tuned for individual classification. During the fine-tuning process, an adjustment is made to ResNet-18, by substituting the last original fully-connected layer with a novel linear layer. This new linear layer is specifically designed to handle the number of person identities, as depicted in Figure 2.

We employ the ResNet-18 model to initially extract high-level features from input images of individuals and then to classify persons. ResNet-18 is known for its ability to capture rich and discriminative features, making it suitable for person reID tasks. The VAE learns a robust latent representation, while the CNN Person ReID model utilizes this representation to perform accurate matching and re-identification.

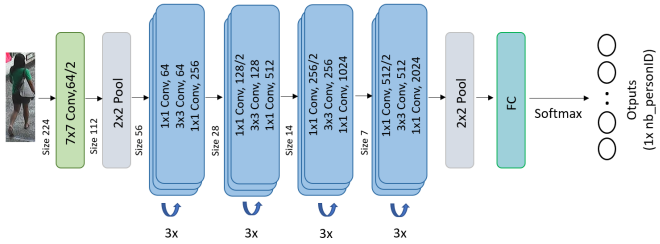

Fig. 2. ResNet-18 Architecture

IV. Experiments

This section describes our experiments carried out on two widely public datasets, CUHK03 and DukeMTMC-reID, with the objective of assessing the effectiveness of VAE leveraging a soft attention mechanism to efficiently re-identify individuals in realistic settings.

A. Experimental Setup

All experiments were carried out using a Dell G15 laptop, powered by an Intel Core-i7 processor, and outfitted with an NVIDIA GeForce RTX 3060 graphics card using 8 GB of DDR6 RAM. The experiments were conducted on the Windows 10 operating system. The codebase was implemented in Python 3.10.9 within Jupyter Notebook environment, relying on PyTorch libraries for machine learning tasks. Each dataset used in our study was divided into two parts, following an evaluation protocol where 80% of the data were designated for training, while the remaining 20% were allocated for testing.

B. Datasets

We used two public large-scale ReID datasets to evaluate and accurate our person tracking system across different camera views and environments.

DukeMTMC-reID The Duke Multi-Tracking Multi-Camera ReID dataset [25], stands out as one of the most

extensive collections of pedestrian images captured by 8 distinct cameras, This dataset comprises a total of 16,522 training images, representing 702 distinct individuals. Additionally, the dataset includes 2,228 query images and 17,661 gallery images.

CUHK03 The CUHK03 dataset [26] consists of a total of 14,097 images corresponding to 1,467 unique individuals. The images were captured using six campus cameras, with each individual being recorded by two different cameras. This dataset provides annotations in the form of bounding boxes, which are available in two variations. The first type includes manually labeled bounding boxes, while the second type consists of bounding boxes generated through an automatic detector. CUHK03 offers 20 randomized train/test divisions, each split entails selecting 100 identities for testing and the remainder for training.

C. Implementation details

Our SAPRI approach comprises two main components: the VAE autoencoder and the ResNet-18 CNN model fine-tuned for person ReID.

During the first part of the process, a VAE is trained on two separate datasets. The encoder, which consists of multiple convolutional layers, is tasked with transforming the input images into a latent space representation of lower dimensionality. The decoder component of the VAE reconstructs then the original images from this latent space representation.

As a result, the trained VAE is able to generate new images. Subsequently, in the second phase, these generated images are resized and provided as input to the ResNet-18 model. The ResNet-18 model is able to extract meaningful features from the input images, allowing for effective representation learning. This facilitates the discrimination of individuals based on their appearance, which is crucial in our person ReID task.

The training parameters for both the VAE and ResNet-18 are detailed in Table II:

TABLE II
TRAINING PARAMETERS

Parameters	**VAE**	**ResNet18**
Learning Rate	0.001	0.0001
Batch Size	32	32
Number of Epochs	50	30
Optimizer	Adam	Adam

D. Experimental results

In this section, we showcase the experiments to validate the robustness of our ReID system. Our SAPRI approach is extensively evaluated on two widely used benchmarks. To assess the effectiveness of our proposed approach, we adopt the mean Average Precision (mAP) score as performance metric.

Qualitative Results. To showcase the efficacy of the VAE in image reconstruction, as well as its applicability to the person ReID task, we employed visualization techniques to examine the final output of randomly selected images generated by the model. Figure 3 illustrates the soft attention mechanism obtained after the VAE process and the influence of the latent space on generating data using the CUHK03 dataset. By

Fig. 3. Individual image generation using CUHK03 dataset.

leveraging the learned latent space, this observation proves that our VAE-model can effectively encode discriminative representations relevant to the ReID task, facilitating accurate matching and retrieval of individuals across different camera views and conditions.

Quantitative Results. To authenticate the performance of our ReID system, we conducted comparative analysis against several state-of-the-art methods. The evaluation was performed on the 2 benchmarks: DukeMTMC-ReID & CUHK03.

TABLE III
SAPRI VS STATE-OF-THE-ART USING DUKEMTMC-REID & CUHK03 DATASETS

DukeMTMC-reID		CUHK03	
Method	mAP (%)	Method	mAP (%)
PAN [27]	51.51	PAN [27]	35.03
PT [4]	56.91	Basel (R)+XQ+Re [2]	37.4
TriNet+REDA* [28]	62.44	PT [4]	38.7
FD-GAN [29]	64.5	TriNet+REDA* [28]	50.74
PGFA [30]	65.5	LOMO + XQDA [31]	51.5
Ours	**69.50**	**Ours**	**65.71**

The comparative results indicate that our SAPRI approach, leveraging VAEs, exceeds the performance of the state-of-the-art methods, as it achieves the highest mAP scores of 69.50% on DukeMTMC-reID and 65.71% on CUHK03, showcasing notable improvements in person ReID performance. The robustness of our SAPRI approach is due to its reliance on soft attention, enabling the model to selectively concentrate on specific weights assigned to different sections of the input data.

Discussions. The utilization of VAEs extends beyond their ability to reconstruct data distributions. It enables the generation of intermediate images by smoothly transitioning between two latent representations. This property showcases the smoothness and continuity of the learned latent space. In our ReID task, VAEs serve as valuable tools for data reconstruction across various conditions specially to address real-world issues. VAEs play a pivotal role to eliminate noise while preserving essential features and fill in missing parts of images, facilitating image restoration. Moreover, fine-tuning hyper-parameters including network architecture, learning rate

and batch size have a considerable impact on the model's reconstruction quality and generalization capacity. .

V. Conclusion

Person ReID is becoming a crucial task for surveillance systems and security. However, it remains challenging due to numerous factors including illumination changes, background clutter, pose state variations and cloth changes. Within this research work, we introduce a novel method that integrates VAEs with a supervised deep learning based model to tackle the challenges related to person ReID tasks, particularly in security and surveillance applications. Thanks to its soft attention mechanism, our SAPRI achieves improves the effectiveness of ReID systems. Through extensive experimentation on two widely used benchmarks, DukeMTMC-reID & CUHK03, our SAPRI outperforms existing approaches, by achieving the highest mAP scores of up to 69.5% and 65.71%, respectively.

References

[1] M. I. Khedher, M. A. E. Yacoubi, and B. Dorizzi, "Multi-shot surf-based person re-identification via sparse representation," in *2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*. IEEE, Aug. 2013. [Online]. Available: http://dx.doi.org/10.1109/AVSS.2013.6636633

[2] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1318–1327.

[3] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 384–393.

[4] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4099–4108.

[5] E. B. Baoues, I. Jegham, S. Ameur, and A. B. Khalifa, "Gf2preid: A novel framework for person re-identification using generative networks," in *2023 International Conference on Cyberworlds (CW)*. IEEE, 2023, pp. 102–109.

[6] M. Ibn Khedher, M. A. El-Yacoubi, and B. Dorizzi, "Fusion of appearance and motion-based sparse representations for multi-shot person re-identification," *Neurocomputing*, vol. 248, p. 94–104, Jul. 2017. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2016.11.073

[7] N. M. Bhiri, S. Ameur, I. Alouani, M. A. Mahjoub, and A. B. Khalifa, "Hand gesture recognition with focus on leap motion: An overview, real world challenges and future directions," *Expert Systems with Applications*, p. 120125, 2023.

[8] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," *Machine learning for data science handbook: data mining and knowledge discovery handbook*, pp. 353–374, 2023.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[10] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[11] S. Zhao, J. Song, and S. Ermon, "Infovae: Information maximizing variational autoencoders," *arXiv preprint arXiv:1706.02262*, 2017.

[12] D. B. Lee, D. Min, S. Lee, and S. J. Hwang, "Meta-gmvae: Mixture of gaussian vae for unsupervised meta-learning," in *International Conference on Learning Representations*, 2020.

[13] M. Jang, S. Seo, and P. Kang, "Recurrent neural network-based semantic variational autoencoder for sequence-to-sequence learning," *Information Sciences*, vol. 490, pp. 59–73, 2019.

[14] D. Bouchacourt, R. Tomioka, and S. Nowozin, "Multi-level variational autoencoder: Learning disentangled representations from grouped observations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[15] J.-T. Chien and C.-W. Wang, "Self attention in variational sequential learning for summarization." in *INTERSPEECH*, 2019, pp. 1318–1322.

[16] I. Jegham, A. B. Khalifa, I. Alouani, and M. A. Mahjoub, "Soft spatial attention-based multimodal driver action recognition using deep learning," *IEEE Sensors Journal*, vol. 21, no. 2, pp. 1918–1925, 2021.

[17] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun, "High-order information matters: Learning relation and topology for occluded person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6449–6458.

[18] X. B. Peng, A. Kanazawa, S. Toyer, P. Abbeel, and S. Levine, "Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow," *arXiv preprint arXiv:1810.00821*, 2018.

[19] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai, "Progressive pose attention transfer for person image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2347–2356.

[20] S. Pan, R. Hu, G. Long, J. Jiang, L. Yao, and C. Zhang, "Adversarially regularized graph autoencoder for graph embedding," *arXiv preprint arXiv:1802.04407*, 2018.

[21] L. Wu, Y. Wang, H. Yin, M. Wang, and L. Shao, "Few-shot deep adversarial learning for video-based person re-identification," *IEEE Transactions on Image Processing*, vol. 29, pp. 1233–1245, 2019.

[22] Z. Wei, X. Yang, N. Wang, and X. Gao, "Rbdf: Reciprocal bidirectional framework for visible infrared person reidentification," *IEEE Transactions on Cybernetics*, vol. 52, no. 10, pp. 10 988–10 998, 2022.

[23] Q. Zhu, J. Su, W. Bi, X. Liu, X. Ma, X. Li, and D. Wu, "A batch normalized inference network keeps the kl vanishing away," *arXiv preprint arXiv:2004.12585*, 2020.

[24] B. Biswas, S. K. Ghosh, and A. Ghosh, "Dvae: deep variational auto-encoders for denoising retinal fundus image," *Hybrid Machine Intelligence for Medical Image Analysis*, pp. 257–273, 2020.

[25] M. Gou, S. Karanam, W. Liu, O. Camps, and R. J. Radke, "Dukemtmc4reid: A large-scale multi-camera person re-identification dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 10–19.

[26] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 152–159.

[27] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 10, pp. 3037–3045, 2018.

[28] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13 001–13 008.

[29] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang *et al.*, "Fd-gan: Pose-guided feature distilling gan for robust person re-identification," *Advances in neural information processing systems*, vol. 31, 2018.

[30] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 542–551.

[31] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2197–2206.