

VSPeR: Self-Attention Based Hybrid Approach for Person Re-Identification in Realistic Environments

Emna Ben Baoues¹, Taher Slimi², Imen Jegham³, Anouar Ben Khalifa^{4*}

^{1,3,4}*Laboratory of Advanced Technology and Intelligent Systems, Ecole Nationale d'Ingénieurs de Sousse, Université de Sousse, Sousse, Tunisie*

^{1,2,4}*Institut National des Technologies et des Sciences du Kef, Université de Jendouba, Le Kef, Tunisie*

Abstract – Person Re-Identification (ReID) remains a pivotal challenge in intelligent surveillance, constantly tested by dramatic intra-class variations, pose changes, lighting shifts, occlusions and distracting backgrounds. This paper proposes a self-attention based hybrid framework (VSPeR) designed to improve data diversity and visual quality for person re-identification in real-world scenarios. At the heart of VSPeR lies a powerful synergy: the robust, identity preserving feature extraction of Variational Autoencoders (VAEs) meets the precise, context-aware focus of Self-Attention Generative Adversarial Networks (GANs). This hybrid architecture captures rich, semantically meaningful latent representations while zeroing in on the most discriminative visual cues necessary for accurate cross-view matching. The outcome is high-fidelity, identity-consistent synthetic images that significantly amplify the training dataset enabling smarter, more effective learning. Put to the test across four leading ReID benchmarks, VSPeR achieves competitive performance compared to existing approaches in terms of mean Average Precision (mAP) and Rank-1 accuracy.

Keywords – Computer vision, deep learning, generative models, person ReID, self-attention.

I. INTRODUCTION

With the rapid expansion of smart surveillance systems, person re-identification (ReID) has become a critical task in computer vision. It consists of recognising and matching individuals across multiple non-overlapping camera views and typically involves three stages: detection, tracking, and re-identification [1]. Despite substantial progress, ReID remains challenging due to strong variations in camera viewpoints, human poses, illumination, image resolution, and clothing appearance [2], further exacerbated by background clutter, dense crowds and occlusions in real-world environments [3], [4].

Early ReID approaches relied on handcrafted features such as colour histograms, texture descriptors, and shape-based representations, combined with metric-learning strategies to match query and gallery images [5]. Although effective under controlled conditions, these methods lack scalability and robustness in large-scale, unconstrained scenarios. Deep learning, particularly Convolutional Neural Networks (CNNs), has considerably improved ReID through hierarchical and

discriminative feature representation [6]. Yet, CNN-based models remain highly dependent on the diversity and quality of training data, while traditional augmentation techniques (e.g., rotations, flips, contrast changes) provide limited variability and fail to reflect the complexity of real surveillance imagery.

Generative models have therefore gained increasing interest for enriching and enhancing ReID datasets. They enable both data augmentation by synthesizing novel and diverse samples and data enhancement by refining noisy or low-resolution inputs, thus improving visual quality and model generalization [7]–[10]. Unlike conventional augmentation pipelines, generative models can simulate realistic and identity-preserving samples to address class imbalance, limited variability, and degraded image quality [11]–[13].

Generative approaches for ReID primarily fall into two categories: Autoencoder-based and GAN-based methods. Autoencoders (AEs) [7] learn compact latent representations suitable for noise reduction and feature extraction, but may fail to preserve fine-grained identity details. Variational Autoencoders (VAEs) [8] further model data distributions and generate diverse, identity-consistent samples [12], but often produce blurry outputs due to their likelihood-based objectives. Conversely, Generative Adversarial Networks (GANs) [9] synthesize sharper and more realistic images, improving domain and appearance diversity [11], but suffer from training instability, mode collapse, and sensitivity to hyperparameters. Hybrid VAE-GAN models [10] attempt to leverage the strengths of both paradigms; however, their convolution-based decoders often struggle to capture long-range spatial dependencies essential for structural coherence and identity preservation.

Recent advances have introduced self-attention (SA) mechanisms into generative architectures to capture global contextual relationships and emphasise identity-relevant regions, thereby enhancing synthesis fidelity and stability [14], [15]. Although promising, most existing works embed SA only within GAN generators or as separate refinement modules. They rarely integrate SA within a unified VAE-GAN pipeline optimised specifically for ReID, resulting in partial handling of long-range dependencies and suboptimal identity consistency.

* Corresponding author's e-mail: anouar.benkhailifa@enisou.u-sousse.tn
Article received 2026-04-04; accepted 2026-06-29

Motivated by these limitations, this work proposes VSPeR, a hybrid VAE-SAGAN framework explicitly designed for both data augmentation and enhancement in person ReID. Unlike conventional hybrid generative methods, VSPeR embeds self-attention directly inside a shared VAE-GAN decoder, allowing reconstruction and adversarial learning to jointly refine the same generative pathway while modelling long-range spatial dependencies crucial for identity preservation. Furthermore, the encoder is fine-tuned as an identity-discriminative feature extractor, aligning the generative process with downstream ReID requirements. This unified formulation enables the production of realistic, structurally coherent, and identity-consistent synthetic samples that significantly strengthen ReID model robustness.

The main contributions of this work are summarised as follows:

- We provide a structured analysis of generative models designed to enhance person ReID performance.
- We introduce a VAE-SAGAN hybrid framework incorporating self-attention to jointly perform data augmentation and enhancement for ReID.
- We conduct extensive evaluations on four benchmark datasets: Market-1501, DukeMTMC-reID, CUHK03, and IUST_PersonReId demonstrating the robustness and adaptability of the proposed approach.

The remainder of this paper is structured as follows: Section II reviews related work in generative approaches for person ReID. Section III describes the architecture and methodology of the proposed VSPeR model. Section IV presents experimental settings, results and analysis. Section V provides a detailed discussion and interpretation of the results. Section VI concludes the paper and outlines future directions.

II. REID REDEFINED: THE RISE OF GENERATIVE APPROACHES

Person ReID has garnered considerable attention due to its critical role in multi-camera surveillance applications. Early approaches relied on hand-crafted features such as local maximal occurrence descriptors, local binary patterns, and colour histograms, combined with similarity metrics for feature comparison [5]. Although these methods provided initial progress, they struggled in real-world conditions involving occlusions, pose variations, low illumination, and complex backgrounds.

To overcome these limitations, deep learning-based feature extractors, particularly convolutional neural networks (CNNs), were introduced [16]. Subsequent advancements focused on refining spatial alignment and discriminative capability. Body-part alignment methods, such as those proposed by Su et al. [17] and Zhao et al. [18], explicitly decompose person images into body regions to reduce sensitivity to pose changes and partial occlusions. Semantic alignment approaches, exemplified by Li et al. [19], incorporate semantic cues to improve robustness in the presence of background clutter and viewpoint variations. Attention-based models, including Liu et al.'s comparative attention network [20], selectively emphasise informative regions, enhancing feature discrimination under challenging visual conditions.

A comprehensive survey by Ye et al. [21] reviewed these developments and highlighted their remaining limitations. Part-based methods continue to struggle under severe occlusion; semantic alignment requires extensive annotations; and attention mechanisms may be misled by noisy data. These challenges underscore the difficulty of extracting reliable identity-preserving features from low-quality or highly variable images.

Generative models have emerged as a complementary direction to augment and enhance ReID training data. By learning underlying data distributions, generative models support both data augmentation and distribution alignment. VAEs generate diverse identity-preserving samples that improve generalization across environments [22], while GANs synthesize high-fidelity pseudo-images to simulate appearance variations across cameras and scenes [23]. Examples include PT-GAN [23] for cross-camera adaptation, CSA-GAN [24] for style normalization, FD-GAN [25] for pose-guided identity preservation, and SAPRI [12] for soft-attention-based synthesis. Despite their effectiveness, GANs are prone to mode collapse and training instability, while VAEs often produce lower-quality or blurry images. Both models struggle when dealing with low-quality, occluded, or highly cluttered inputs.

To mitigate these drawbacks, hybrid approaches such as VAE-GANs combine the stable latent modelling of VAEs with the visual fidelity of GANs, improving sample quality and diversity. VAEs provide structured latent representations that preserve identity-relevant information, whereas GANs enhance realism through adversarial learning. However, existing VAE-GAN hybrids still face challenges, including computational overhead, non-trivial joint optimization, and difficulty in modelling long-range spatial dependency factors that affect both synthesis quality and downstream feature embeddings.

Recent work has therefore explored integrating self-attention (SA) into generative architectures to capture global spatial dependencies [14]. SA enables generators to model long-range interactions, allowing them to focus more effectively on identity-critical regions such as clothing attributes and discriminative visual patterns. This enhances structural coherence and improves the ability to handle complex backgrounds or occlusions.

Rather than functioning as a simple aggregation of independent components, VSPeR employs a tightly integrated architecture in which structured latent modelling, adversarial synthesis, and self-attention interact within a shared generative pathway. This unified design enables the model to simultaneously enforce identity-preserving latent structure, high-fidelity synthesis, and long-range spatial consistency capabilities that conventional VAE, GAN, or hybrid models cannot achieve in isolation. As a result, VSPeR yields diverse, detailed, and identity-consistent images even under limited data, noisy inputs, or high visual variability, thereby providing more robust support for modern ReID systems.

III. PROPOSED VSPeR APPROACH

GANs are effective at synthesizing realistic person images, while VAEs provide a principled way to learn structured latent representations. Yet, VAEs often produce blurry outputs and GANs suffer from instability and difficulty preserving fine-grained identity cues. To address these limitations, VSPeR introduces a hybrid generative design that differs from conventional VAE-GAN models by integrating self-attention directly within the shared decoder-generator, enabling the

capture of long-range spatial dependencies essential for identity consistency. In addition, the decoder is jointly optimised by reconstruction and adversarial losses, ensuring that realism and latent consistency reinforce one another rather than compete. Finally, the encoder is fine-tuned as an identity-discriminative feature extractor, aligning generation with downstream ReID objectives. This unified formulation allows VSPeR (see Fig. 1) to generate more coherent, realistic, and identity-preserving images than prior hybrid approaches.

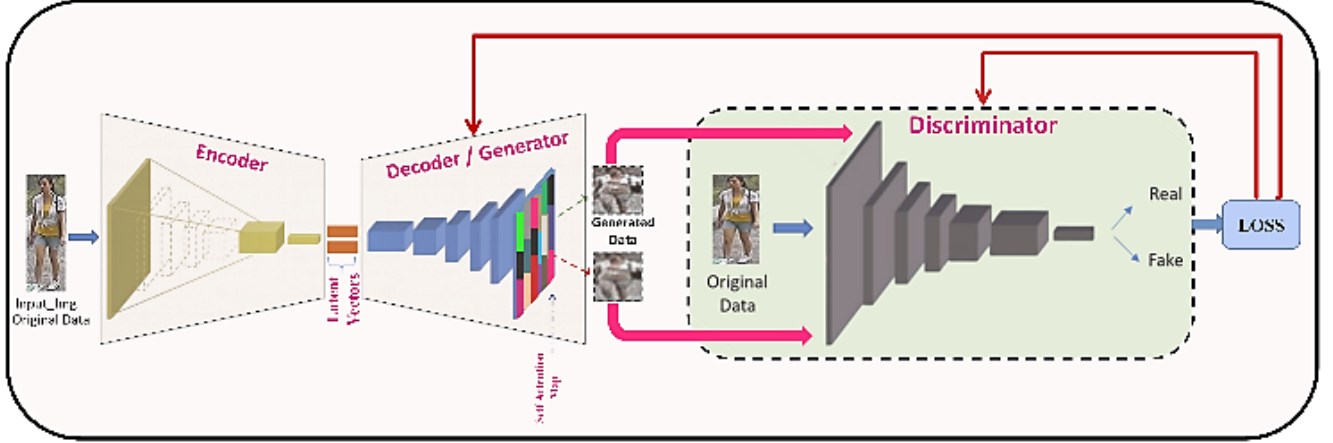


Fig. 1. VSPeR: hybrid VAE-SAGAN architecture with self-attention, enabling latent representation learning and ensuring identity preservation.

A. VSPeR for Improved Image Variety Generation

As shown in Fig. 1, VSPeR extends a VAE framework where a ResNet-50 based encoder maps input images to a Gaussian latent space, and a shared decoder reconstructs them. This component captures global semantics, ensuring faithful reconstruction and smooth latent interpolation.

The GAN module, enhanced with self-attention (Fig. 2), improves image realism by modelling long-range dependencies and focusing on identity-relevant regions. An adversarial objective partially replaces the VAE reconstruction loss, encouraging the generation of high-fidelity, realistic images.

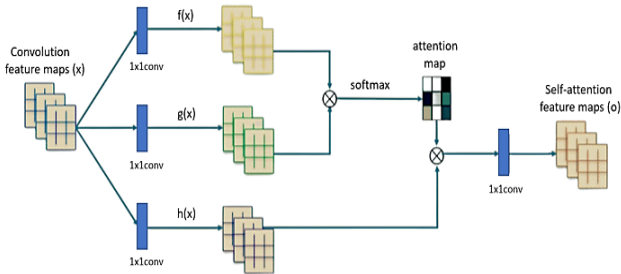


Fig. 2. The Self-Attention Module integrated into the decoder/generator of our proposed VSPeR.

The discriminator's role is to classify incoming samples as real or generated, guiding the generator toward producing realistic images. Simultaneously, the encoder is fine-tuned as an instance-level feature extractor, which benefits downstream

person ReID tasks. The loss function combines VAE and GAN objectives, incorporating reconstruction, regularization, and adversarial components, as given in (1).

$$\mathcal{L}_{VAE-GAN} = \lambda_1 \mathcal{L}_{recon} + \lambda_2 \mathcal{L}_{KL} + \lambda_3 \mathcal{L}_{GAN}, \quad (1)$$

where $\mathcal{L}_{recon}$ corresponds to the reconstruction loss, $\mathcal{L}_{KL}$ represents the Kullback-Leibler divergence enforcing a Gaussian latent distribution, and $\mathcal{L}_{GAN}$ denotes the adversarial loss. The weighting coefficients λ_1 , λ_2 and λ_3 control the contribution of each component.

In our implementation, $\lambda_1 = 1$, $\lambda_2 = 0.01$, and $\lambda_3 = 1$, selected empirically to balance reconstruction fidelity, latent regularization, and image realism.

The reconstruction loss is defined as:

$$\mathcal{L}_{recon} = \frac{1}{N} \sum_{i=1}^N \|x_i - \hat{x}_i\|^2. \quad (2)$$

The KL divergence is defined as:

$$\mathcal{L}_{KL} = -\frac{1}{2} \sum_{i=1}^d (1 + \log(\sigma_i^2) - \mu_i^2 - \sigma_i^2). \quad (3)$$

The adversarial loss is defined as:

$$\mathcal{L}_{GAN} = -\frac{1}{N} \sum_{i=1}^N [\log(D(x_i)) + \log(1 - D(G(z_i)))]. \quad (4)$$

In these equations, D denotes the discriminator and G the generator. The input image x_i is reconstructed as $\hat{x}_i$ from a latent vector z_i sampled from a Gaussian distribution $z_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, where μ_i and σ_i^2 parameterize the learned latent distribution, N denotes the number of samples, and d corresponds to the dimensionality of the latent space.

In our implementation, the latent space dimensionality d is set to 128, selected empirically to provide a balance between representation capacity and computational efficiency.

This formulation enables the model to jointly learn structured latent representations and generate high-fidelity synthetic images. The generated image features from the previous hidden layer ($x \in \mathbb{R}^{C \times N}$) are first mapped into two separate feature spaces $f(x) = W_f(x)$ and $g(x) = W_g x$ for the calculation of attention scores, as can be seen in Fig. 2. The attention map β_{ij} is computed using the softmax operation on the similarity scores $s_{ij} = f(x_i)^T g(x_j)$. The resulting attention layer output refines the feature representation by selectively focusing on the most relevant parts of the image. The final output of the attention layer is shown below in (5):

$$o_j = v \left(\sum_{i=1}^N \beta_{j,i} h(x_i) \right), \quad (5)$$

where $h(x_i) = W_h x_i$ and $v(x_i) = W_v x_i$, are combined with the input feature map using a learnable scalar γ , as depicted in (6), to generate the final output:

$$y_i = \gamma o_i + x_i. \quad (6)$$

The SA module enables our proposed method to focus on local neighbourhood cues before gradually learning to assign greater importance to non-local features, enhancing its ability to capture complex spatial relationships.

B. Person Classification

Person ReID aims to determine whether two images depict the same individual. ResNet-50 is widely adopted in ReID applications due to its residual learning architecture, which enables stable gradient flow and reliable extraction of discriminative features under challenging visual conditions. This makes it suitable for distinguishing individuals across variations in pose, lighting, and background.

In our implementation, a customized ResNet-50 model (Fig. 3) is fine-tuned for person ReID using both real and VSPeR-generated images.

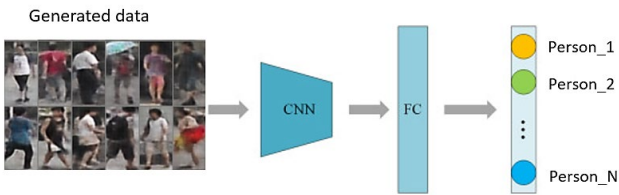


Fig. 3. Person ReID pipeline using ResNet-50 trained with augmented data generated by VSPeR.

The final classification layer is adapted to the number of training identities. Image preprocessing includes resizing, normalization, and tensor conversion. Datasets are loaded into DataLoader objects to support batch processing.

Feature extraction is performed by forwarding images through the network to obtain embeddings for both query and gallery sets. Person matching relies on deep metric learning, typically using Euclidean distance.

The Euclidean distance between two feature vectors f_i and f_j is calculated as depicted in (7):

$$d(f_i, f_j) = \sum_{k=1}^n (f_{ik} - f_{jk})^2, \quad (7)$$

where n is the dimension of the feature vectors. Additionally, performance is assessed using precision-based metrics, most notably the mAP score, which is pivotal for evaluating the model's retrieval capabilities. To compute the mAP score, each query image is compared with the gallery images by ranking them according to their computed distances, and binary labels indicating true matches are used to construct a precision-recall curve. Precision at rank k for a query image is defined as in (8):

$$Precision(k) = \frac{1}{k} \sum_{i=1}^k rel(i), \quad (8)$$

where $rel(i)$ is a binary indicator function equal to 1 if the i -th retrieved image matches the query identity, and 0 otherwise. The AP for a query is then computed as in (9):

$$AP = \frac{1}{R} \sum_{k=1}^N Precision(k) \times rel(k), \quad (9)$$

where R is the total number of relevant documents (gallery images sharing the same identity as the query), N is the total number of retrieved documents, and $rel(k)$ denotes the relevance at rank k .

Finally, the mAP score is obtained by averaging the AP scores across all queries, as defined in (10):

$$mAP = \frac{1}{Q} \sum_{i=1}^Q AP_i, \quad (10)$$

where Q represents the number of query images and AP_i is the Average Precision corresponding to the i -th query. These evaluation criteria collectively assess how effectively the ReID model distinguishes identities, ensuring accurate retrieval and proper ranking of correct matches in the gallery set.

Unlike many existing approaches that treat generative and attention mechanisms as independent modules, VSPeR jointly integrates VAE-based latent modelling, adversarial learning, and self-attention within a shared decoder, thereby improving feature representation and identity preservation for person ReID. This unified design supports simultaneous data augmentation and image enhancement, leading to more robust

performance under challenging real-world conditions, including occlusions, pose variations, and complex backgrounds.

IV. EXPERIMENTS

This section presents the experiments conducted on four challenging benchmarks to assess the effectiveness of the VSPeR approach in improving image quality and enhancing person ReID performance.

A. Setup

All experiments were conducted on a Dell G15 workstation equipped with an Intel Core-i7 processor and an NVIDIA GeForce RTX 3060 GPU (8 GB). The implementation was developed in Python 3.10.9 using PyTorch and executed in a Jupyter Notebook environment on Windows 10. Due to hardware constraints, the input resolution was fixed to 128×64, which is a commonly adopted setting in person ReID benchmarks and ensures fair comparison with existing methods.

To ensure stable training of the hybrid VAE-GAN architecture, convergence behaviour was monitored through preliminary experiments. Both reconstruction and adversarial losses stabilized after approximately 250–300 epochs; therefore, 300 epochs were used. The model is trained using the Adam optimizer ($\beta_1 = 0.5$, $\beta_2 = 0.999$) with an initial learning rate of 0.0002.

Data augmentation techniques, including random horizontal flipping, cropping, colour jittering, and normalization, were applied to improve generalization while preserving identity-specific features.

In terms of computational efficiency, the integration of self-attention increases the training cost compared to standard VAE-GAN models due to global dependency modelling. Training requires approximately 300 epochs on an RTX 3060 GPU, representing a trade-off between performance and efficiency. The total number of trainable parameters is approximately 32 million, dominated by the ResNet-50 backbone (≈ 25 M), with additional VAE-GAN components (≈ 5 – 6 M) and self-attention modules (< 1 M).

All experiments follow standard evaluation protocols with predefined splits.

B. Datasets

We evaluate the proposed VSPeR framework using four widely adopted person ReID benchmarks: Market-1501, DukeMTMC-reID, CUHK03, and IUST_PersonReID. The diversity of these datasets enables a comprehensive assessment of the model's robustness under different visual conditions, camera setups, and identity variations.

Market-1501 [26] contains 32,668 pedestrian images of 1501 identities captured from six cameras, including five high-definition and one low-resolution device. Its variety of viewpoints and backgrounds makes it one of the most widely used ReID benchmarks.

DukeMTMC-reID [27] is a large-scale multi-camera dataset with 36 441 images of 2702 pedestrians recorded from eight

synchronized outdoor cameras. It includes challenging conditions such as occlusions, illumination changes, and dynamic backgrounds.

CUHK03 [28] consists of 14 096 images of 1467 individuals collected from six surveillance cameras. It provides both manually labelled and detector-generated bounding boxes, with substantial variation in pose, illumination, and occlusion.

IUST_PersonReID [29] comprises images of 1847 individuals collected from 19 different cameras across various real-world settings such as markets, university campuses, mosques, and public gatherings in Islamic countries like Iran and Iraq. The dataset is organised into a training partition containing 72 393 images representing 1193 identities, while the testing partition consists of 1428 query images and 45 062 gallery images covering 654 identities.

These benchmarks are currently the largest and most representative datasets for ReID, featuring diverse conditions including multi-season and multi-time settings, a mix of high- and low-definition cameras, and a variety of scenes and backgrounds. They also incorporate complex lighting changes and are particularly challenging due to extensive intra-class variability, occlusions, and different camera viewpoints. Notably, IUST_PersonReID is a novel dataset introduced in 2024, presenting even greater challenges due to its collection in culturally diverse environments with modest clothing, frequent occlusions and high visual similarity among individuals, providing a valuable resource for evaluating ReID models under real-world, complex conditions.

Performance is measured using the mean Average Precision (mAP) metric, which evaluates both retrieval accuracy and ranking quality by sorting gallery images according to their similarity to each query, as well as the Rank-1 accuracy, which reflects the model's ability to correctly identify the top-most matching identity.

C. Implementation Details

The proposed VSPeR architecture consists of several components, including an encoder, a shared decoder/generator, a discriminator, a self-attention mechanism, and a CNN-based ReID model. The encoder, based on a pre-trained ResNet-50 backbone, extracts discriminative features from input images and maps them into a latent space represented by mean and variance vectors. The shared decoder reconstructs images from these latent representations, enforcing reconstruction consistency while enabling smooth latent interpolation.

The adversarial component is implemented through a generator-discriminator framework, where the generator, enhanced with self-attention layers, produces realistic images, and the discriminator distinguishes between real and generated samples. During training, the model jointly optimises reconstruction and adversarial objectives to balance latent structure preservation and visual fidelity.

The overall loss combines reconstruction, KL divergence, and adversarial terms, weighted by coefficients λ_1 , λ_2 , and λ_3 , respectively. In our implementation, $\lambda_1 = 1$, $\lambda_2 = 0.01$, and $\lambda_3 = 1$. The latent space dimensionality is set to $d = 128$,

providing a balance between representation capacity and computational efficiency.

Training is performed using the Adam optimiser, and the model is updated iteratively over multiple epochs. Reconstruction and adversarial losses are monitored throughout training to ensure convergence and stability.

For the ReID task, a CNN-based model using a modified ResNet-50 architecture is adopted. The network is fine-tuned with dropout regularization and trained using cross-entropy loss defined as: $L_{CE} = -\sum_{i=1}^C y_i \log(\hat{y}_i)$, where y_i and $\hat{y}_i$ denote the ground-truth and predicted probabilities, respectively. Feature embeddings are extracted for query and gallery images, and matching is performed using distance-based ranking to compute mAP and Rank-1 accuracy.

This implementation ensures a consistent integration between generative modelling and discriminative learning, enabling improved feature representation and robustness for person ReID tasks. Details of the hyperparameters used for both VSPeR and the CNN-based ReID model are summarised in Table I.

TABLE I

HYPER-PARAMETERS FOR VSPeR AND CNN-PERSONREID TRAINING

Hyperparameter	VSPeR	CNN-PersonReID
Image Size	(128, 64)	(128, 64)
Batch Size	32	32
Learning Rate	0.0002	0.0002
Number of Epochs	300	40
Optimizer	Adam	Adam
β_1 / β_2	0.5 / 0.999	0.9 / 0.999
Latent Dimension (d)	128	—
Loss Weights ($\lambda_1, \lambda_2, \lambda_3$)	1 / 0.01 / 1	—

D. Experimental Results

This section showcases the qualitative and quantitative outcomes from evaluating our proposed VSPeR approach across multiple datasets. These results highlight the performance and efficiency of our model in diverse scenarios, validating the effectiveness of combining VAE and GAN architectures with SA for enhanced person ReID.

Qualitative evaluations. The hypothesis for this experiment is that the synthesized images (generated data) should exhibit realistic textures and maintain structural coherence with their corresponding ground-truth (original data) within their complex environment.

Figure 4 illustrates the examples of synthesized images produced by our proposed approach over the Market-1501 dataset. This visualisation confirms our hypothesis and our model's competence and effectiveness, particularly highlighting how the attention mechanism enhances the learning of discriminative features, resulting in an excellent visual representation of human attributes. Furthermore, most images, including both foreground and background, appear realistic.

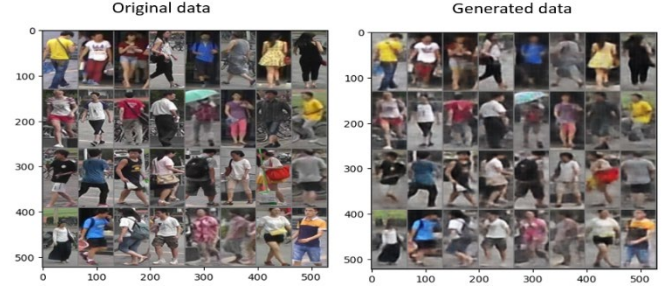


Fig. 4. Comparison between real and generated images for person ReID using the Market-1501 dataset.

Our proposed model, extracted by a shallow network, encapsulates the person positional and geometric information of the whole input image. Consequently, the interpolation between the low-level features maintains semantic smoothness and consistency which is not achieved by existing generative methods as depicted in Fig. 5. This success can be attributed to the significant roles played by VAEs, GANs, and the SA mechanism in our approach.



Fig. 5. Comparison of generated images on Market-1501 showing improved realism and structure with the proposed VSPeR model.

To validate the increased diversity achieved by our approach, we randomly selected augmented views generated from the Market-1501 dataset during each training iteration. These augmented views, obtained using our proposed VAE-SAGAN model based on the self-supervised learning framework [15], were used to assess the contribution of each sample to model performance. Importantly, the original (unaltered) view of each person was always retained in the training data.

Figure 6 illustrates VSPeR's ability to generate multiple realistic images of a given individual in various poses. The method successfully produces identity-preserving synthetic samples in complex real-world environments, maintaining fine-grained visual details. Moreover, the diversity introduced by these samples enriches the training dataset, effectively mitigating the limitations imposed by real-world data scarcity and enhancing the overall robustness of the ReID model.



Fig. 6. Generated samples showing increased diversity from VSPeR for improved person ReID training.

Figure 7 illustrates the retrieval results for CUHK03 dataset, where the gallery contains only a single image corresponding to the query identity. Despite the challenge posed by significant appearance changes and partial occlusions, as shown in the first row of Fig. 7, our VSPeR method consistently retrieves the correct individual. This demonstrates the model's robustness in handling various ReID challenges simultaneously, including drastic changes in appearance, occlusion, low-illumination, pose state changes, and low-camera resolution. Thus, we can conclude that even when the query identity undergoes significant changes in appearance, our approach demonstrates effective performance in matching and identifying the correct person from the gallery.



Fig. 7. Qualitative evaluation on CUHK03 dataset showing the query image and top-5 retrieved results from the gallery.

Quantitative Evaluations. The qualitative observations presented above are further supported by quantitative evaluations. To assess the realism and diversity of the generated images on the Market-1501 dataset, we used two metrics: FID and SSIM. FID measures how closely the distribution of generated images resembles that of real images. It is particularly sensitive to visual artefacts and thus serves as a strong indicator of image realism. SSIM, on the other hand, evaluates the intra-class similarity of identity-conditioned generations, providing insight into the structural fidelity and diversity of the generated images.

We present a direct performance comparison with prior work on the Market-1501 dataset in Table II, where VSPeR demonstrates superior results in terms of FID and SSIM. This confirms the effectiveness of our approach in producing identity-preserving, diverse and realistic images, which are essential for improving ReID training via data augmentation. For DukeMTMC-reID, CUHK03 and IUST_PersonReID datasets, due to the unavailability of comparable FID and SSIM metrics from previous studies, we report the performance of VSPeR alongside real data baselines in Table III, establishing reference points for future evaluations.

As observed in Table III, the FID scores of VSPeR-generated images remain within close proximity to those of the real samples, indicating strong perceptual quality and realism, an expected outcome, given that real data naturally represents the optimal case. More notably, the SSIM values of the synthetic

outputs consistently exceed those of the real images across all datasets. This outcome reflects VSPeR's ability to introduce rich structural diversity while preserving the core identity features of each individual, thereby enhancing intra-class variation.

TABLE II

PERFORMANCE COMPARISON ON THE MARKET-1501 DATASET. LOWER FID INDICATES BETTER IMAGE QUALITY, WHILE HIGHER SSIM INDICATES GREATER DIVERSITY AMONG GENERATED IMAGES

Method	FID ↓	SSIM ↑
GCL [30]	59.86	0.367
PN-GAN [31]	54.23	0.335
FD-GAN [25]	257.00	0.247
DPTN [32]	48.3663	0.0689
PATN [32]	–	0.30135
XING [32]	–	0.13606
CTFN [33]	27.14	0.342
DG-GAN [34]	18.24	0.360
StyleGAN [35]	12.67	0.3870
Proposed VSPeR approach	12.04	0.39

↓ Lower value is better. ↑ Higher value is better.

TABLE III

FID AND SSIM COMPARISON BETWEEN REAL AND VSPeR-GENERATED IMAGES ON SEVERAL DATASETS. LOWER FID INDICATES BETTER FIDELITY; HIGHER SSIM REFLECTS IMPROVED STRUCTURAL SIMILARITY

Dataset	Real FID ↓	VSPeR FID ↓	Real SSIM ↑	VSPeR SSIM ↑
Market-1501	7.14	12.04	0.35	0.39
CUHK03	8.40	11.86	0.38	0.42
DukeMTMC-reID	9.21	11.04	0.43	0.63
IUST_PersonReID	10	12.49	0.31	0.47

↓ Lower value is better. ↑ Higher value is better.

These findings highlight VSPeR's potential for generating high-quality and semantically diverse synthetic person images, making it a valuable tool for augmenting training data in person ReID tasks where annotated samples are limited or imbalanced. In addition to generative quality metrics, we also report retrieval performance using mAP and Rank-1 accuracy. The mAP metric provides a comprehensive measure by accounting for both retrieval correctness and ranking quality across the entire gallery. Rank-1 accuracy, in contrast, specifically reflects the ability of the system to correctly retrieve the most relevant identity at the top rank a critical criterion in real-world ReID scenarios. As illustrated in Tables IV–VII, VSPeR achieves superior performance across all the used benchmark datasets, showing competitive performance compared to recent methods. The consistently high Rank-1 accuracy obtained by our model further demonstrates its robustness and practical utility in accurately identifying individuals under varying conditions. This strong performance is attributed to the synergy between generative modelling and attention mechanisms, which together enhance feature representation and generalization.

Based on the performance comparisons above, we demonstrate that employing a deep convolutional ResNet-50

model as the feature extractor for our generated data, our VSPeR method consistently achieves strong performance compared to recent approaches across diverse benchmarks, indicating its robust and reliable performance in person ReID tasks.

For the Market-1501 dataset, our method achieved the highest mAP score of 71.63 %, surpassing notable methods including DASK (61.2 %) and GF2PReID (67.16 %). Similarly, for the DukeMTMC-reID dataset, our method attained a mAP of 75.97 %, exceeding the next best method, AMD+ResNet-50 (75.54 %). In the CUHK03 dataset, our method again led with a mAP of 75.86 %, outperforming GF2PReID (74.3 %) and SAPRI (65.71 %). For the IUST_PersonReID dataset, we additionally report comparative results with recent methods. SOLIDER [29] achieved a mAP score of 42.35 % and Rank-1 of 51.08 %, while CLIP-ReID [29] attained a mAP of 51.60 % and Rank-1 of 61.40 %. Our proposed VSPeR framework achieved a significantly higher mAP score of 82.41 % and Rank-1 of 89.64 %, further demonstrating its effectiveness on challenging, culturally diverse scenarios under real-world conditions.

TABLE IV
PERFORMANCE COMPARISON ON MARKET-1501 DATASET

Method	mAP (%)	Rank-1 (%)
ACAN [36]	50.6	73.3
PAT [37]	51.7	75.2
MMCL [38]	60.4	84.4
DASK [39]	61.2	82.3
CycAsasy [40]	64.8	84.8
ASG+ResNet-50 [41]	66.6	–
GCL [30]	66.8	87.3
GF2PReID [11]	67.16	–
PCSL [42]	69.4	87.0
2D-SNet-169 [43]	–	89.0
Proposed VSPeR	71.63	91.6

In addition, we include comparisons with recent transformer-based models such as Per-Former [44] and MHSA-Net [49]. As shown in Table V, the proposed VSPeR achieves competitive or superior performance, demonstrating that enhancing training data through generative modelling can effectively complement transformer-based feature extraction. These models have demonstrated strong performance through improved feature representation. In contrast, our approach focuses on enhancing training data quality rather than modifying backbone architecture, highlighting the complementary nature of the proposed framework.

To further validate the robustness and generalization capabilities of our approach, we also evaluated cumulative matching characteristics (CMC) at Rank-1, Rank-5, and Rank-10. The results, visualised in Fig. 8, clearly illustrate the superiority of VSPeR in accurately identifying individuals at varying rank levels across four datasets. The curves show the cumulative matching rates from Rank-1 to Rank-10. Key

performance values at R-1, R-5, and R-10 are: (91.6 %, 95.3 %, 96.42 %) for Market-1501, (92.3 %, 96.45 %, 98.2 %) for DukeMTMC-reID, (92.8 %, 96.71 %, 98.33 %) for CUHK03 and (89.64 %, 93.18 %, 96.5 %) for IUST_PersonReID. This additional evaluation highlights the model's effectiveness in both top-1 identification accuracy and broader retrieval scenarios.

TABLE V
PERFORMANCE COMPARISON ON DUKEMTMC-REID DATASET

Method	mAP (%)	Rank-1 (%)
ACAN [36]	45.1	67.6
PAT [37]	56.5	71.8
MMCL [38]	51.4	72.4
DASK [39]	58.5	74.6
CycAsasy [40]	60.1	77.9
ASG+ResNet-50 [41]	58.1	–
PCSL [42]	53.5	71.7
Per-former [44]	61.6	81.1
CORE [45]	64.1	–
PIRID [46]	65.4	79.6
IER [47]	66.4	–
SAPRI [12]	69.5	–
AMD+ResNet-34 [48]	71.71	–
MHSA-Net [49]	73.1	–
SAN [50]	75.5	87.9
AMD+ResNet-50 [48]	75.54	–
Proposed VSPeR	75.97	92.3

TABLE VI
PERFORMANCE COMPARISON ON CUHK03 DATASET

Method	mAP (%)	Rank-1 (%)
PAT [37]	31.5	31.1
DASK [39]	46.2	48.1
CycAsasy [40]	41.0	47.4
2D-SNet-218 [43]	53.8	55.8
2D-SNet-152 [43]	54.3	56.9
2D-SNet-169 [43]	57.3	58.6
SAPRI [12]	65.71	–
AGW [1]	62.0	–
SAN-basic [50]	73.7	77.9
GF2PReID [11]	74.3	–
Proposed VSPeR	75.86	92.8

TABLE VII
PERFORMANCE COMPARISON ON IUST_PERSONREID DATASET

Method	mAP (%)	Rank-1 (%)
SOLIDER [29]	42.35	51.08
CLIP-ReID [29]	51.60	61.40
Proposed VSPeR	82.41	89.64

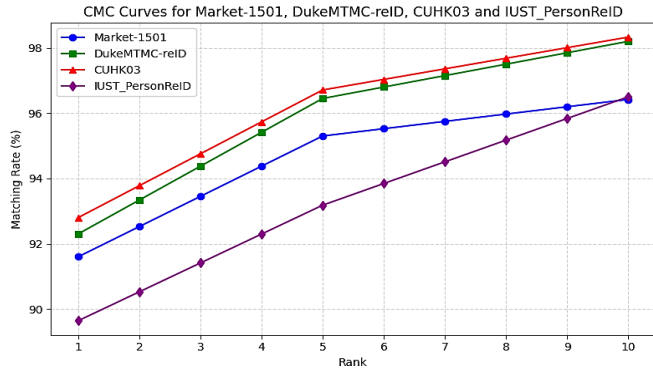


Fig. 8. CMC curves comparing four person ReID benchmarks: Market-1501, DukeMTMC-reID, CUHK03 and IUST_PersonReID.

Overall, these results underscore the effectiveness of VSPeR in combining the strengths of both VAE and GAN models, leading to enhanced data augmentation, robust generalization, and superior feature representation. Additionally, the incorporation of SA proves instrumental in enhancing the learning process. By enabling the model to dynamically and contextually prioritize different regions of the input data, SA facilitates a more nuanced understanding of complex patterns and dependencies in the data.

E. Ablation Study

To systematically analyse the contribution of each component within the proposed VSPeR framework, we perform a component-wise ablation study by removing one component at a time from the full architecture while keeping all other settings fixed. All ablation variants follow the same training and evaluation protocol as the complete VSPeR model, to ensure fair comparison.

Table VIII summarises the ablation configurations and their relative impact on mAP across four benchmark datasets: Market-1501, DukeMTMC-reID, CUHK03, and IUST_PersonReID.

Definition of ablation variants. (i) Full VSPeR: the complete model integrating both VAE-based latent representation

learning and SAGAN-based adversarial generation with self-attention. (ii) w/o VAE: the VAE component is removed from VSPeR, and the model relies only on adversarial generation with self-attention. (iii) w/o SAGAN: the adversarial self-attention module is removed from VSPeR, and the model relies only on latent reconstruction through the VAE. (iv) w/o generative modules: both VAE and SAGAN components are removed, reducing the model to a standard ResNet-50 backbone trained using classification loss. (v) VAE + GAN (without self-attention): both VAE and GAN components are retained, but the self-attention mechanism is removed, allowing the evaluation of adversarial learning without spatial attention. (vi) VAE + SAGAN (without shared decoder): both VAE and SAGAN components are preserved, but the shared decoder is removed, disabling the unified generative pathway and allowing the components to operate independently.

Across all datasets, removing both generative components (ResNet-50 baseline) results in the lowest performance, highlighting the limitations of purely discriminative approaches in complex ReID scenarios. When only one component is retained, the ablation variants generally improve performance compared to the purely discriminative baseline, confirming the contribution of both generative modules. Additional variants further confirm that removing self-attention or the shared decoder leads to a consistent performance drop, highlighting their importance in modelling long-range dependencies and enabling effective interaction between latent representation learning and adversarial synthesis.

However, the full VSPeR model consistently achieves the best performance across all datasets. The removal of the VAE component leads to a decrease in performance due to reduced diversity and structure in the latent representations, while removing the SAGAN module limits the model's ability to capture long-range dependencies and focus on discriminative visual cues through attention mechanisms.

These results confirm that each component contributes significantly to the overall framework and that their integration within a unified architecture is essential to achieve robust and generalizable ReID performance.

TABLE VIII

COMPONENT-WISE ABLATION STUDY OF THE VSPeR FRAMEWORK ACROSS FOUR BENCHMARK DATASETS. RESULTS ARE REPORTED IN TERMS OF mAP (%)

Configuration	VAE	GAN	SA	Shared Decoder	Market-1501	DukeMTMC-reID	CUHK03	IUST_PersonReID
Full VSPeR	✓	✓	✓	✓	71.63	75.97	75.86	82.41
w/o VAE	✗	✓	✓	✓	58.43	69.50	65.71	67.14
w/o SAGAN	✓	✗	✗	✓	67.25	76.82	79.13	80.04
w/o Generative (ResNet-50 baseline)	✗	✗	✗	✗	54.81	65.04	54.62	53.12
VAE + GAN (without self-attention)	✓	✓	✗	✓	68.91	73.06	73.82	79.74
VAE + SAGAN (without shared decoder)	✓	✓	✓	✗	71.20	75.54	75.58	81.93

To further evaluate the stability of the proposed approach, we report multi-run experimental results in Table IX. Each run corresponds to an independent training process performed using identical hyperparameters but different random seed initializations (42, 123, and 2024 for Runs 1, 2, and 3, respectively). The Mean $\pm$ Std (mAP %) reports the average performance across the three runs and the corresponding standard deviation, providing a quantitative measure of performance stability. The results show only minor variations across runs, confirming that the performance of VSPeR is consistent and not sensitive to stochastic training factors.

TABLE IX

MULTI-RUN STATISTICAL EVALUATION OF VSPeR ACROSS FOUR BENCHMARK DATASETS. RESULTS ARE REPORTED IN TERMS OF MAP (%) OVER THREE INDEPENDENT RUNS

Datasets	Run 1	Run 2	Run 3	Mean $\pm$ Std
Market-1501	71.63	71.28	71.31	71.41 $\pm$ 0.22
DukeMTMC-reID	75.97	75.42	75.81	75.73 $\pm$ 0.29
CUHK03	75.86	75.42	75.64	75.64 $\pm$ 0.26
IUST_PersonReId	82.41	81.97	82.16	82.18 $\pm$ 0.31

V. DISCUSSION

The comprehensive evaluation of our proposed VSPeR framework across Market-1501, DukeMTMC-reID, CUHK03 and IUST_PersonReId datasets confirms the effectiveness of combining VAE-GAN with SA mechanisms for person ReID. The strong performance in both quantitative (mAP, Rank-1 accuracy) and qualitative (FID, SSIM) metrics highlights several important insights.

First, the improved mAP and Rank-1 scores across all benchmarks clearly demonstrate that the hybrid model effectively captures both identity-preserving features and visual diversity. These gains validate the importance of using generative models not only for data augmentation but also for enhancing the visual quality and discriminative capability of training samples. The use of ResNet-50 as a feature extractor on synthetic images proved effective in maintaining robust feature embeddings, which positively impacted retrieval performance. Second, the visual quality of the generated images, as shown by the lower FID and higher SSIM scores, reinforces the role of SA in preserving structural integrity and modelling long-range dependencies. Unlike traditional GAN-based models that often suffer from mode collapse or unrealistic textures, the integration of SA allows the model to focus selectively on meaningful regions within an image, improving synthesis consistency and fidelity. Moreover, the performance of VSPeR across multiple datasets shows its strong generalization capabilities. It consistently outperforms competitive models despite the diverse and challenging nature of each benchmark. This suggests that the hybrid architecture is well-suited for real-world surveillance conditions involving occlusions, pose variations, low-resolution imagery, and diverse lighting environments. Finally, the ablation study further underscores the unique contributions of each component. The VAE component enhances latent space representation and regularizes training, the GAN component

improves realism through adversarial learning, and the SA module contributes to spatial discrimination and structural refinement. Their combination in VSPeR leads to a more powerful and balanced system.

However, some limitations were observed. While the generated images are realistic, occasional artefacts still appear in complex backgrounds or under extreme occlusion. Additionally, the inclusion of SA layers, while beneficial for performance, introduces computational overhead, which may limit deployment on low-resource platforms. Overall, the results indicate that the VSPeR framework not only demonstrates competitive performance across multiple benchmarks ReID models but also offers a robust, scalable and identity-consistent approach for training data enhancement and augmentation. These findings suggest promising directions for future work, including dynamic attention strategies, lightweight attention modules for edge deployment and integration with temporal or multi-frame ReID systems.

VI. CONCLUSION

In this paper, we proposed VSPeR, a generative framework that integrates VAE and SAGAN to enhance person ReID through both data reconstruction and data augmentation. By combining the structured representation learning of VAE with the high-quality image generation of SAGAN, VSPeR generates realistic and identity preserving synthetic images that improve training diversity and overall ReID performance. Experimental results on standard benchmarks demonstrate that VSPeR achieves competitive performance in terms of higher image quality and retrieval accuracy. The model shows strong generalization capabilities across varied poses, lighting conditions, and occlusions. However, challenges remain, particularly with occasional visual artefacts and increased computational demands due to the SA mechanism. Future work will focus on refining image quality, optimising computational efficiency, and extending the framework to broader surveillance scenarios.

REFERENCES

- [1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S.C. Hoi, "Deep learning for person reidentification: A survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2872–2893, Jan. 2021. <https://doi.org/10.1109/TPAMI.2021.3054775>
- [2] Y. Chen, K. Wang, H. Ye, L. Tao, and Z. Tie, "Person re-identification in special scenes based on deep learning: A comprehensive survey," *Mathematics*, vol. 12, no. 16, Aug. 2024, Art. no. 2495. <https://doi.org/10.3390/math12162495>
- [3] E. Ben Baoues, T. Slimi, I. Jegham, A.B. Khalifa, "PerDiff: Identity-preserving diffusion reconstruction for Person re-identification," *Engineering Research Express*, vol. 8, no. 12, 2026, Art. no. 125202. <https://doi.org/10.1088/2631-8695/ae7935>
- [4] J. Yan, Y. Wang, X. Luo, and Y. W. Tai, "PS-ReID: Advancing Person re-identification and precise segmentation with multimodal retrieval," *arXiv:2503.21595*, Aug. 2025. <https://arxiv.org/html/2503.21595v2>
- [5] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *2010 IEEE computer society conference on computer vision and pattern recognition*, San Francisco, CA, USA, June 2010, pp. 2360–2367. <https://doi.org/10.1109/CVPR.2010.5539926>
- [6] K. O'shea and R. Nash, "An introduction to convolutional neural networks," *arXiv:151108458*, Dec. 2015. <https://arxiv.org/pdf/1511.08458>

- [7] N. Deng, L. Pu, L.Y. Tao, L.L. Guo, R.T. Yang, and J.F. Lin, "Application of GAN deblurring algorithm in face recognition," *Traitement du Signal*, vol. 42, no. 5, pp. 2669–2678, Feb. 2025. <https://doi.org/10.18280/ts.420519>
- [8] D.P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv:1312.6114*, Dec. 2013. <https://doi.org/10.48550/arXiv.1312.6114>
- [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. https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf
- [10] A.B.L. Larsen, S.K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *International Conference on Machine Learning*, 2016, pp. 1558–1566. <https://backend.orbit.dtu.dk/ws/portalfiles/portal/127745553/larsen16.pdf>
- [11] E. Ben 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)*, Sousse, Tunisia, Oct. 2023, pp. 102–109. <https://doi.org/10.1109/CW58918.2023.00024>
- [12] E. Ben Baoues, I. Jegham, S. Ameur, M. El Yacoubi, and A.B. Khalifa, "Soft-attention based Person re-identification in real-world settings using variational AutoEncoders," in *2024 International Conference on Human System Interaction (HSI)*, Paris, France, July 2024, pp. 102–109. <https://doi.org/10.1109/HSI61632.2024.10613536>
- [13] E. Ben Baoues, I. Jegham, S. Ameur, and A.B. Khalifa, "Hard attention based EfficientNet for Person re-identification," in *2024 10th International Conference on Control, Decision and Information Technologies (CoDIT)*, Vallette, Malta, July 2024, pp. 976–981. <https://doi.org/10.1109/CoDIT62066.2024.10708438>
- [14] A. Hatamizadeh, J. Song, G. Liu, J. Kautz, and A. Vahdat, "DiffiT: Diffusion vision transformers for image generation," in *European Conference on Computer Vision*, Oct. 2024, pp. 37–55. https://doi.org/10.1007/978-3-031-73242-3_3
- [15] Zhang H, Goodfellow I, Metaxas D, Odena A. Self-attention generative adversarial networks. In: International conference on machine learning. PMLR; 2019. p. 7354–63.
- [16] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, and Y. Xu, "Deep-Person: Learning discriminative deep features for person re-identification," *Pattern Recognition*, vol. 98, Feb. 2020, Art. no. 107036. <https://doi.org/10.1016/j.patcog.2019.107036>
- [17] C.W. Su, S. Yang, G. Zhang, J. Gao, and S. Sclaroff, "Pose-driven deep convolutional model for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3960–3969.
- [18] R. Zhao, W. Ouyang, H. Li, and X. Wang, "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 (CVPR)*, Honolulu, HI, USA, July 2017, pp. 1077–1085. <https://doi.org/10.1109/CVPR.2017.103>
- [19] 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 (CVPR)*, Honolulu, HI, USA, July 2017, pp. 384–393. <https://doi.org/10.1109/CVPR.2017.782>
- [20] Y. Liu, Y. Yan, P. Liu, and L. Shao, "End-to-end comparative attention networks for person re-identification," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3492–3506, May 2018. <https://doi.org/10.1109/TIP.2017.2700762>
- [21] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S.C. Hoi, "A deep learning perspective on person re-identification: Survey and beyond," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–34, 2021.
- [22] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," in *Proceedings of the IEEE international conference on computer vision*, Venice, Italy, Oct. 2017, pp. 3754–3762. <https://doi.org/10.1109/ICCV.2017.405>
- [23] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Salt Lake City, USA, June 2018, pp. 79–88. <https://doi.org/10.1109/CVPR.2018.00016>
- [24] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Salt Lake City, USA, June 2018, pp. 5157–5166. <https://doi.org/10.1109/CVPR.2018.00541>
- [25] Y. Ge *et al.*, "Fd-GAN: Pose-guided feature distilling GAN for robust person re-identification," *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [26] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person reidentification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, Santiago, Chile, Dec. 2015, pp. 1116–1124. <https://doi.org/10.1109/ICCV.2015.133>
- [27] M. Gou, S. Karanam, W. Liu, O. Camps, and R.J. Radke, "DukeMTMC4ReID: A largescale multi-camera person re-identification dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, Honolulu, HI, USA, July 2017, pp. 10–19. <https://doi.org/10.1109/CVPRW.2017.185>
- [28] 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*, Columbus, OH, USA, June 2014, pp. 152–159. <https://doi.org/10.1109/CVPR.2014.27>
- [29] M.R. Mohammadi *et al.*, "IUST_PersonReID: A new domain in person re-identification datasets," *arXiv:2412.18874v1*, Dec. 2024. <https://arxiv.org/html/2412.18874v1>
- [30] H. Chen, Y. Wang, B. Lagadec, A. Dantcheva, and F. Bremond, "Joint generative and contrastive learning for unsupervised person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Nashville, TN, USA, June 2021, pp. 2004–2013. <https://doi.org/10.1109/CVPR46437.2021.00204>
- [31] X. Qian *et al.*, "Pose-normalized image generation for person re-identification," in *Proceedings of the European conference on computer vision (ECCV)*, Oct. 2018, pp. 650–667. https://doi.org/10.1007/978-3-030-01240-3_40
- [32] B. Kaur and S. Singh, "Performance analysis of state-of-the-art models for pose-guided person image generation," *International Journal of Computer Engineering and Technology*, vol. 16, no. 1, pp. 3747–3768, Jan.–Feb. 2025. <https://doi.org/10.34218/IJCET.16.01.258>
- [33] T. Zhang, X. Sun, X. Li, and Z. Yi, "Image generation and constrained two-stage feature fusion for person re-identification," *Applied Intelligence*, vol. 51, pp. 7679–7689, Mar. 2021. <https://doi.org/10.1007/s10489-021-02271-z>
- [34] Z. Zheng, L. Zheng, and Y. Yang, "Joint discriminative and generative learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, June 2019, pp. 2138–2147. <https://doi.org/10.1109/CVPR.2019.00224>
- [35] S.H. Salem Hussin and R. Yildirim, "StyleGAN-LSRO method for person re-identification," *IEEE Access*, vol. 9, pp. 13857–13869, Jan. 2021. <https://doi.org/10.1109/ACCESS.2021.3051723>
- [36] L. Qi, L. Wang, J. Huo, Y. Shi, X. Geng, and Y. Gao, "Adversarial camera alignment network for unsupervised cross-camera person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2921–2936, July 2022. <https://doi.org/10.1109/TCSVT.2021.3096668>
- [37] H. Ni, Y. Li, L. Gao, H.T. Shen, and J. Song, "Part-aware transformer for generalizable person re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, Paris, France, Oct. 2023, pp. 11280–11289. <https://doi.org/10.1109/ICCV51070.2023.01036>
- [38] D. Wang and S. Zhang, "Unsupervised person re-identification via multi-label classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Seattle, WA, USA, June 2020, pp. 10981–10990. <https://doi.org/10.1109/CVPR42600.2020.01099>
- [39] K. Xu, C. Jiang, P. Xiong, Y. Peng, and J. Zhou, "DASK: Distribution rehearsing via adaptive style kernel learning for exemplar-free lifelong person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 9, pp. 8915–8923, Apr. 2025. <https://doi.org/10.1609/aaai.v39i9.32964>
- [40] Z. Wang, J. Zhang, L. Zheng, Y. Liu, Y. Sun, Y. Li, and S. Wang, "CycAs: Self-supervised cycle association for learning re-identifiable descriptions," in *Computer Vision – ECCV 2020: 16th European Conference*, Glasgow, UK, Aug. 2020, pp. 72–88. https://doi.org/10.1007/978-3-030-58621-8_5
- [41] P. Hong, A. Wu, and W.S. Zheng, "Semi-supervised person re-identification by attribute similarity guidance," in *2020 25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy, Jan. 2021, pp. 6471–6477. <https://doi.org/10.1109/ICPR48806.2021.9412561>

- [42] L. Qi, L. Wang, J. Huo, Y. Shi, and Y. Gao, "Progressive cross-camera soft-label learning for semi-supervised person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 9, pp. 2815–2829, Mar. 2020. <https://doi.org/10.1109/TCSVT.2020.2983600>
- [43] W. Li, S. Shao, Z. Qiu, Z. Zhu, and A. Song, "2D-SNet: A lightweight network for person re-identification on the small data regime," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 6, no. 1, pp. 68–78, Jan. 2024. <https://doi.org/10.1109/TBIOM.2023.3332285>
- [44] N. Pervaiz, M.M. Fraz, and M. Shahzad, "Per-former: rethinking person re-identification using transformer augmented with self-attention and contextual mapping," *Visual Computer*, vol. 39, pp. 4087–4102, Sep. 2023. <https://doi.org/10.1007/s00371-022-02577-0>
- [45] M. Ye *et al.*, "Collaborative refining for person re-identification with label noise," *IEEE Transactions on Image Processing*, vol. 31, pp. 379–391, Dec. 2022. <https://doi.org/10.1109/TIP.2021.3131937>
- [46] L. Wang, W. Zhang, D. Wu, P. Hong, and B. Li, "Prototype-based inter-camera learning for person re-identification," in *ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 2022, pp. 4778–4782. <https://doi.org/10.1109/ICASSP43922.2022.9746640>
- [47] G. Zhang, Z. Luo, Y. Chen, Y. Zheng, and W. Lin, "Illumination unification for person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6766–6777, Oct. 2022. <https://doi.org/10.1109/TCSVT.2022.3169422>
- [48] X. Chen *et al.*, "Explainable person reidentification with attribute-guided metric distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, QC, Canada, Oct. 2021, pp. 11813–11822. <https://doi.org/10.1109/ICCV48922.2021.01160>
- [49] H. Tan, X. Liu, B. Yin, and X. Li, "MHSA-Net: Multihead self-attention network for occluded person re-identification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 8210–8224, Nov. 2023. <https://doi.org/10.1109/TNNLS.2022.3144163>
- [50] X. Jin, C. Lan, W. Zeng, G. Wei, and Z. Chen, "Semantics-aligned representation learning for person re-identification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 11173–11180, Apr. 2020. <https://doi.org/10.1609/aaai.v34i07.6775>



Emna Ben Baoues received a Research Master's degree in Computer Science from the Higher Institute of Computer Science of Kef (ISI Kef), Tunisia. She is currently a PhD student in Electrical Engineering at the National Engineering School of Sousse (ENISO), Tunisia, affiliated with the LATIS Laboratory. Her research interests include person re-identification, generative models, deep learning, and intelligent surveillance systems.

E-mail: emna.benbaoues@eniso.u-sousse.tn

ORCID iD: <https://orcid.org/0009-0004-3961-8139>



Taher Slimi is a Professor in Medical Imaging at the University of Jendouba, Tunisia, with extensive academic and industry experience. Formerly a technical leader at Samsung Healthcare and Toshiba Medical, his work spans ultrasound and medical IT systems. Trained in Grenoble-France, he now focuses on advancing diagnostic imaging and training professionals in innovative technologies.

E-mail: taher.slimi@intek.u-jendouba.tn

ORCID iD: <https://orcid.org/0000-0002-4474-9414>



Imen Jegham is an Assistant Professor of Computer Science, specialising in Artificial Intelligence, currently affiliated with the University of Sousse, Tunisia. Her research focuses on AI applications across diverse domains, including intelligent transportation systems, industrial anomaly detection, and more. Her work has been published in well-known specialised international journals and conferences. She holds a PhD in Computer Science from the Higher Institute of Computer Science and Communication Techniques of Sousse, as well as an Engineering degree in Computer Science and a European Master's degree in Highway and Traffic Engineering from the National School of Engineers of Sousse.

E-mail: Imen.Jegham@isitc.u-sousse.tn

ORCID iD: <https://orcid.org/0000-0003-1531-438X>



Anouar Ben Khalifa received the engineering degree from the National Engineering School of Monastir, an Msc degree and a PhD degree in Electrical Engineering from the National Engineering School of Tunis (Tunisia). In 2021, he received a university accreditation to direct research (HdR) from the University of Sousse. He was a Professor in Electrical and Computer Engineering at the National Engineering School of Sousse, Tunisia (from 2007 to 2022). He is currently the Director of the National Institute of Technologies and Sciences of Kef and the President of the Ranking Committee of the University of Jendouba.

E-mail: anouar.benkhalfifa@eniso.u-sousse.tn

ORCID iD: <https://orcid.org/0000-0002-9946-0829>