International Journal of Computer Engineering and Technology (IJCET)
Volume 16, Issue 1, Jan-Feb 2025, pp. 3747-3768, Article ID: IJCET_16_01_258
Available online at https://iaeme.com/Home/issue/IJCET?Volume=16&Issue=1
ISSN Print: 0976-6367; ISSN Online: 0976-6375; Journal ID: 5751-5249
Impact Factor (2025): 18.59 (Based on Google Scholar Citation)
DOI: https://doi.org/10.34218/IJCET_16_01_258

PERFORMANCE ANALYSIS OF STATE-OF-THE-ART MODELS FOR POSE-GUIDED PERSON IMAGE GENERATION

Biponjot Kaur[1], Sarbjeet Singh[2]
[1]University Institute of Engineering and Technology, Panjab University, Chandigarh, India.
[2]University Institute of Engineering and Technology, Panjab University, Chandigarh, India

ABSTRACT

Pose-guided person image generation is now a central field of study in computer vision, where sophisticated deep-learning methods are used to generate realistic images of people in a given pose. This work compares the performance of current state-of-the-art models on two benchmark datasets: DeepFashion and Market-1501. These datasets provide dense pose, clothing, and background variations and therefore are appropriate for quantifying model robustness. Evaluation is focused on key metrics such as Structural Similarity Index (SSIM), Fréchet Inception Distance (FID), and Inception Score (IS) to estimate the quality, realism, and diversity of the generated images. Our results identify the strengths and weaknesses of each model, providing important insights for future development in pose-guided image synthesis. We also bring into focus the challenges presented by human deformation and structural alignment, which are still the areas of utmost need for improvement.

Keywords: DeepFashion, Evaluation metrics, Market-1501, Performance analysis, Pose-guided person image generation.

Cite this Article: Biponjot Kaur, Sarbjeet Singh. Performance Analysis of State-of-The-Art Models for Pose-Guided Person Image Generation. *International Journal of Computer Engineering and Technology (IJCET)*, 16(1), 2025, 3747-3768.

https://iaeme.com/MasterAdmin/Journal_uploads/IJCET/VOLUME_16_ISSUE_1/IJCET_16_01_258.pdf

1. Introduction

Pose-guided human image generation is a complex process that leverages advanced deep-learning techniques to create photorealistic images of individuals in specific poses. At its core, this method relies on human pose estimation, which detects and identifies key body parts, providing crucial information for generating high-quality images. The process typically involves a generator that produces the images and a discriminator that evaluates their authenticity. This adversarial interplay, characteristic of Generative Adversarial Networks (GANs), is central to producing realistic outputs. Techniques like image-to-image translation and pose transfer have facilitated applications in fields such as fashion, entertainment, and virtual try-on systems. Furthermore, pose transfer techniques enhance the flexibility of image generation by allowing the adaptation of an individual's pose from one image to another. The Fig. 1 describes the basic workflow of pose-guided person image generation.

Pose-guided human image generation has been widely applied in virtual try-on, person reidentification, and character animation. The first pose-guided person image generation model, introduced in 2017, employed GANs to generate images. PG2 [1], an early model, adopted a two-stage framework to first produce a coarse image in the target pose and subsequently refine

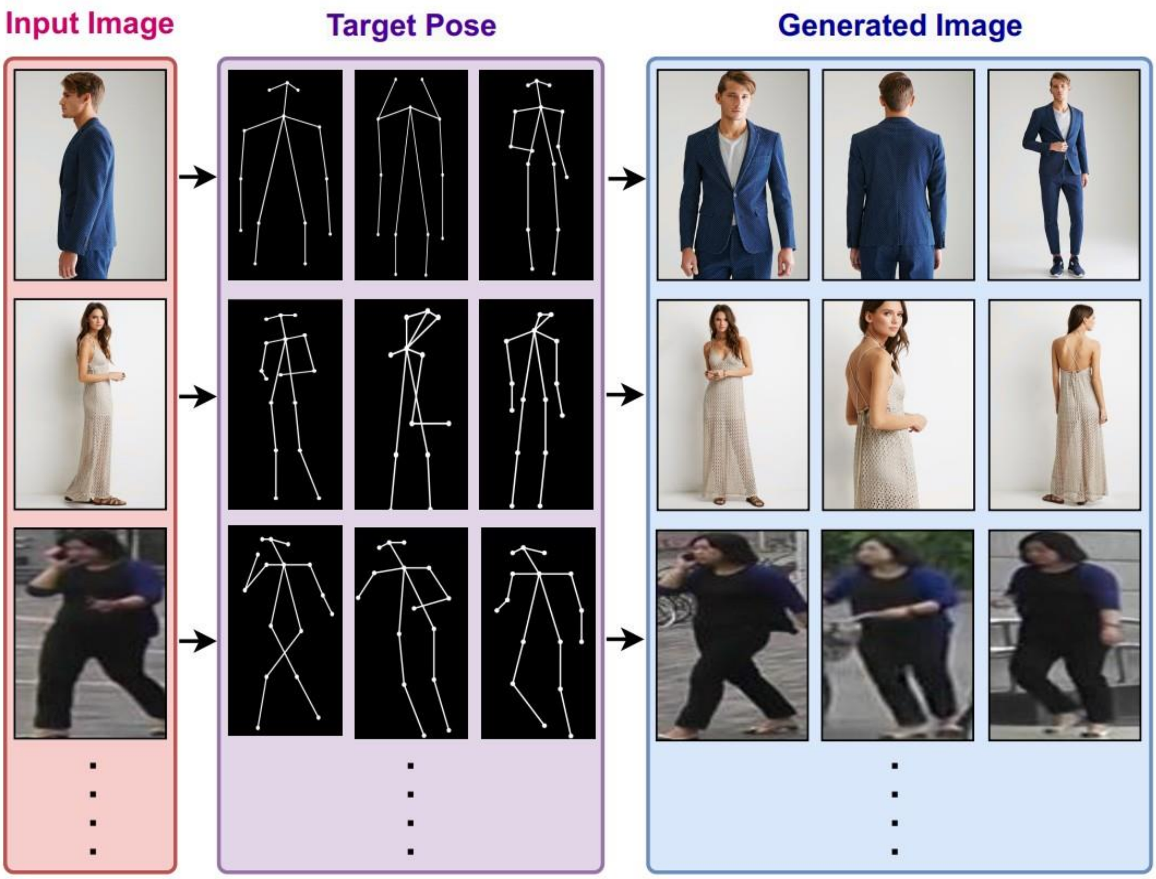


Figure 1: basic workflow of pose-guided person image generation

it with textures and details, paving the way for future advancements. Later models, such as Liquid Warping GAN [2], introduced superior warping flows to ensure consistency in clothing and texture between source and target poses. Challenges like human deformation and structural alignment were addressed by models such as deformable GANs (Def-GAN) [3] and Structure-Preserving Generative Networks (SPG-Net), enabling more precise transformations of body parts. Methods, like Pose Attention Transfer Network (PATN), improved realism through attention mechanisms, while variational approaches like VUNet [4] enabled pose and appearance disentanglement. Domain-specific models such as ClothFlow and Adaptive Content Generating Networks (ACGPN) [5] further advanced virtual try-on applications by effectively deforming clothing to accommodate novel poses.

Performance evaluation of pose-guided person image generation models is essential to assess their ability to produce realistic and pose-accurate outputs. This paper compares the performance of several models on two datasets: DeepFashion and Market-1501. Metrics such as Inception Score (IS), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Fréchet Inception Distance (FID), and are used to evaluate similarity, fidelity, realism, and diversity. These quantitative evaluations, combined with qualitative assessments, form a comprehensive framework for evaluating the effectiveness of pose-guided person image generation models. This paper makes the following contributions:

- We provide a performance comparison of three major models in pose-guided person image generation: CrossingGAN (XingGAN), Dual-task Pose Transformer Network (DPTN), and Progressive Attention Transfer Network (PATN).
- These models are evaluated on two datasets: DeepFashion and Market-1501. Performance is measured using metrics such as SSIM, IS, and FID.
- We present observations regarding the strengths and weaknesses of these models and conclude that the PATN model outperforms the others on these two datasets.
- We also highlight significant challenges in the field, including pose variations, identity retention, and maintaining texture consistency.

2. Related Work

Pose-based person image synthesis involves morphing a person's appearance from one source image into a desired pose in another image. Applications for this technology include motion transfer, posture correction, and virtual try-on experiences. Unsupervised learning approaches, including autoregressive models, diffusion models, transformer models, Variational Autoencoders (VAEs), and GANs, dominate this field. GANs, in particular, excel at creating realistic images using two networks: a generator that produces realistic-looking images and a discriminator that distinguishes real images from fake ones [6].

One of the early methods [1] for pose-guided person image generation involved a coarse-to-fine approach. This method starts by generating an initial image with the desired pose and then refines it using adversarial techniques. However, it faced challenges with feature misalignment. Ma et al. [7] proposed PG2, a two-stage pipeline that learns a disentangled representation of various image factors, such as foreground, background, and pose. While PG2 improved results, it still struggled with retaining fine details in appearance. Enhancements to PG2 involved data augmentation [8] and more informative inputs [9][10].

Grigorev et al. [11] introduced a fully convolutional network with deformable skip connections to address pose-guided image resynthesis, estimating body surface texture from a single photograph. Neverova et al. [12] combined the SMPL model [13] with neural synthesis techniques for more accurate pose transfer. This method utilized DensePose [14] to map image pixels to a common surface-based coordinate system. Si et al. [15] used recurrent neural networks with multistage adversarial losses to create realistic human images with clear foreground and background details. Pumarola et al. [16] developed an unsupervised approach using new loss functions and a bidirectional generator. Siarohin et al. [17] introduced deformable skip connections in GANs, employing nearest-neighbor loss for better detail matching.

Balakrishnan et al. [18] proposed separating the scene into body parts and background layers for improved synthesis. Dong et al. [19] introduced Warping-GAN, which handled large geometric transformations with soft-gated warping and attention layers. Dong et al. [20] later developed PP-GAN, focusing on controllable person image generation using part-preserving generators and multiscale discriminators. Li et al. [21] presented a method that integrated 3D geometry from 2D representations for accurate pose transfer, using a variant of the U-Net. Liang et al. [22] introduced PCGAN, which partitioned the human body into sub-parts and applied affine transformations. Lakhal et al. [23] proposed VDG, a two-stage encoder-decoder framework that processed input and target images in separate branches.

Sun et al. [24] combined convolutional LSTM with U-Net for realistic image generation conditioned on poses. Song et al. [25] focused on semantic parsing between poses for unsupervised image generation. PATB [26] introduced progressive pose transfer using intermediate representations. Han et al. [27] developed ClothFlow, which used dense flow fields to capture clothing deformation. Xu et al. [28] tackled unconventional perspectives with pose-guided multi-branch encoders. Shen et al. [29] generated person images with accurate pose transformations using a two-stream feature fusion module.

Zhao et al. [30] incorporated geometric constraints into pose generation using 3D convolution in GAN generators. Ma et al. [31] proposed a multi-level statistics transfer model for disentangling and transferring appearance features. Karmakar et al. [8] enhanced GAN architecture with residual learning and data augmentation techniques. Li et al. [32] introduced PoNA blocks for cross-modal feature transfer. XingGAN [33] used appearance and shape-guided discriminators for accurate pose transformations. Ren et al. [34] combined flow-based operations and attention mechanisms for localized feature sampling.

Hu et al. [35] proposed the p-Norm regression for versatile pose and appearance feature modeling. Yang et al. [36] developed FHPT, focusing on preserving fine-grained details. Li et al. [37] used semantic maps with attention mechanisms for pose-guided generation. PSG-GAN [38] emphasized incremental image generation with region-focal transfer blocks. Albahar et al. [39] used a pose-conditioned StyleGAN to preserve fine details. Khatun et al. [40] extracted and recombined appearance, local details, and pose components. SPAN [41] inferred regions of interest based on human pose using interconnected pathways and semantic parsing attention blocks.

Tang et al. [42] employed structurally aware flow-based methods for high-quality person image generation. Zhang et al. [43] introduced the Dual-task Pose Transformer Network (DPTN) with auxiliary source-to-source tasks. Chen et al. [44] developed TFJR-Net, which improves pose-guided generation with separate pathways for clothing data and external textures. Wang et al. [45] introduced SCM-Net, utilizing semantic-aware style features and correlation mining. Wu et al. [46] proposed a method of disassembling the human body into distinct parts for realistic synthesis. Liu et al. [47] introduced PCE-GAN, using global and local correspondence transformation branches.

Nakada et al. [48] employed an attention mechanism to produce images depicting individuals in any pose. Yan et al. [49] developed a semantic-driven dual-attention network for accurate and detailed image generation. Chen et al. [50] introduced a multi-level feature fusion strategy for bidirectional guidance between pose and image. Zhang et al. [51] addressed texture correlation with a texture correlation network (TCN). Jain et al. [52] introduced VGFlow for reposing human images with visibility-aware detail extraction. Lu et al. [53] tackled multi-source image generation using a flow-based strategy and hierarchical feature confidence prediction.

Wei et al. [54] focused on highlighting content details and maintaining spatial context for identity and clothing features. Ma et al. [55] proposed Multi-scale Cross-domain Alignment (MCA) with a Global Context Aggregation Transformer (GCAT). Zhang et al. [56] introduced DPTN-TA, enhancing visual quality with auxiliary tasks and texture affinity loss. Huang et al. [57] developed CPD-GAN with dynamic transfer fusion blocks and deformable convolution. Liu et al. [58] proposed CoGAN for simultaneous training of conditional and unconditional GANs. Wang et al. [59] introduced DSAT-GAN with multiscale semantic mapping and adaptive semantic attention mechanisms. Roy et al. [60] proposed an improved attention-guided progressive generation approach, achieving significant improvements over existing methods.

3. Methodology

The term ‘pose-guided person generation’ was first introduced in [1] in 2017. Numerous models have been proposed for this task, but some prominent models serve as benchmarks for further research. We implemented these models on various datasets and compared their performance based on predefined criteria. All these models are generative models and use attention networks, flow-based networks, transformers, diffusion models, and GANs.

The Progressive Attention Transfer Network (PATN) [26], a GAN model, transfers a given image from one pose to a conditioned pose. Unlike previous work, where pose transformation was performed in one step, PATN uses intermediate representations of pose. A sequence of Pose Attentional Transfer Blocks (PATBs) is used for these transformations. Experiments were conducted on two common datasets: Market-1501 and the In-Shop Cloth Retrieval benchmark of DeepFashion, with resolutions of 128×64 and 256×256, respectively. Along with standard metrics, the model was evaluated using the PCKh score [61].

CrossingGAN [33], or XingGAN, transfers the source pose into the desired pose using two pathways that contain information about shape and appearance. The model consists of three branches: Appearance-guided Shape-based generation (AS), Shape-guided Appearance-based generation (SA), which models a person’s shape and appearance, respectively, and a co-attention Fusion (CAF) block to concatenate AS and SA blocks. The model adopts the same evaluation metrics as used in the AD-GAN network. Most current methods fail to achieve accurate texture mapping. To overcome this limitation, the Dual-task Pose Transformer Network (DPTN) [43] was introduced. The approach incorporates an auxiliary task, and the source-to-source task, and leverages the correlation between the dual tasks to enhance performance. The DPTN features a Siamese structure with two branches: one for source-to-source self-reconstruction and another for source-to-target generation. By sharing some weights between these branches, the knowledge gained from the source-to-source task supports the source-to-target learning process. Additionally, the two branches are connected using a Pose Transformer Module (PTM), which adaptively explores the correlation between features from both tasks.

Every model included in the analysis uses two losses: adversarial loss and perceptual loss. Other common losses, such as style loss and $L1$ loss, are used differently in each model. These losses are used to update the discriminator and train the model efficiently so that it can differentiate well between real and fake images. Table 1 provides a brief summary of the methodologies used in this work.

4. Experimental Results

4.1 Datasets

Several datasets are used in the field of pose-guided person generation, but only a few are widely adopted. The most common datasets used in the literature are DeepFashion [62] and Market-1501 [63], which are also used in our work to evaluate their performance. DeepFashion is a large-scale dataset introduced to address the lack of annotations in previous fashion datasets. It contains over 800,000 diverse fashion images, each annotated with 50 categories, 1,000 attributes, bounding boxes, and clothing landmarks.

Table 1: Comparison of PATN, DPTN, and XingGAN model.

Feature	PATN	DPTN	XingGAN
Key Idea	Progressive pose transfer using attentional blocks	Dual-task learning with self-reconstruction	Two-way interaction between shape and appearance features
Architecture	Pose-Attentional Transfer Blocks (PATBs)	Siamese network with self-reconstruction	Two-branch network with shape-appearance interactions
Pose Representation	Keypoint-based pose heatmaps	Keypoint-based pose heatmaps	Keypoint-based pose heatmaps
Discriminators	Appearance and Shape Discriminators	Pose Transformer Module (PTM)	Appearance-Guided and Shape-Guided Discriminators
Attention Mechanism	Local attention per PATB	Transformer-based PTM for feature correlation	Cross-attention between shape and appearance
Feature Processing	Progressive pose transformation	Dual-task learning with CABs and TTBs	SA and AS blocks for bidirectional feature learning
Texture Consistency	Learned progressively with local pose attention	Enhanced through dual-task correlation	Enhanced by crossing shape and appearance features
Performance on Large Pose Variations	Moderate	Better due to feature refinement	Best due to bidirectional feature fusion
Complexity	Moderate	Low (9.79M parameters)	High due to dual-branch interactions
Strengths	Smooth pose transformation	More accurate texture mapping	Best appearance and pose consistency
Weaknesses	Limited global pose relationships	Requires dual-task training	High computational cost

Originally, the dataset had various benchmarks to evaluate different methods: Category and Attribute Prediction, In-Shop Clothes Retrieval, and Consumer-to-Shop Clothes Retrieval. Later, another benchmark, 'Fashion Landmark Detection', was added to predict the position of fashion landmarks. Previous research on pose-guided person generation utilized the In-Shop Clothes Retrieval benchmark, which includes over 300,000 cross-pose image pairs. This benchmark has 54,642 images, each annotated with bounding boxes labeled as cloth type and pose type. The pose type label consists of side, front, and back views. The DeepFashion dataset is superior in terms of scale and annotations compared to previous datasets.

Another dataset used in this research is Market-1501, a large-scale dataset primarily used for person re-identification. It consists of 1,501 identities with 32,668 annotated bounding boxes. The annotations used in the dataset are bounding boxes containing pedestrians in the given images. The images are low-resolution (128 × 64) and exhibit diversity in poses, backgrounds, and viewpoints due to multiple cameras. It is suitable for pose-guided person generation because of its diversity in poses, although these poses cannot be classified into specific categories. It is also challenging due to its low-resolution images. Details about the datasets used in this research are listed in Table 2.

4.2 Evaluation Metrics

Evaluating generative models is challenging, as performance depends on the specific model being evaluated. Common metrics used to assess image quality include Structural Similarity Index Metric (SSIM) [64], Inception Score (IS) [65], and Fréchet Inception Distance (FID) [66]. Evaluation metrics are classified into quantitative and qualitative methods.

Table 2: Comparison Between DeepFashion and Market-1501 Datasets

Feature	**DeepFashion**	**Market-1501**
Purpose	Clothing recognition and retrieval	Person re-identification
Number of Images	800,000+	32,668 + 500,000 distractors
Annotations	Categories, attributes, landmarks, cross-domain image pairs	Bounding boxes (BBoxes), identity labels, camera IDs, distractors
Categories	50 categories, 1,000 attributes	Not applicable
Landmarks	4–8 landmarks per image	None
Dataset Source	Shopping websites, Google Images	Surveillance cameras

Type of Images	Fashion images (store, street, consumer photos)	Surveillance images in front of a campus supermarket
Evaluation Protocol	Category classification, attribute prediction, retrieval accuracy	Mean Average Precision (mAP), Cumulative Matching Characteristics (CMC)
Key Benchmark Tasks	Attribute prediction, in-shop clothes retrieval, cross-domain retrieval	Person re-identification across multiple cameras
Public Availability	Yes	Yes
Detection Method	Manual annotation	Deformable Part Model (DPM) detector
Additional Features	Rich attribute metadata, consumer-to-shop image pairs	Large distractor set, multiple queries per identity

Quantitative methods calculate numerical scores based on metrics such as IS, FID, etc., while qualitative methods evaluate generated images visually, which can be inspected by humans. In this research, we used SSIM, IS, FID, Peak Signal-to-Noise Ratio (PSNR), and Learned Perceptual Image Patch Similarity (LPIPS) [67] to evaluate the models.

SSIM: The SSIM evaluates image quality based on three factors: luminance Li,j, contrast Ci,j, and structure Si,j, where i and j are two image signals. It calculates the score by comparing the intensity values of the image and their neighborhood pixels [68]. Variants of SSIM include Multi-scale SSIM (MS-SSIM) and Multi-component SSIM (3-SSIM & 4-SSIM). Luminance is defined as the average of all intensity values of the image, and contrast is calculated using the standard deviation of the image. Mathematically, equation (1) is used to calculate SSIM:

$$SSIMi,j = Li,j^{\alpha}.Ci,j^{\beta}.Si,j^{\gamma} \qquad (1)$$

IS: The training goal function in CatGAN [69] and the Inception Score, first introduced in [65], are quite similar. The Inception Score measures the diversity and quality of generated images. IS can be calculated using equation (2):

$$exp\mathbb{E}_x KLpy|x||py \qquad (2)$$

Here, KL is the Kullback-Leibler divergence between the distributions of $py|x$ and py for synthesized image samples x and their associated labels y [70]. This metric is widely used but has drawbacks, such as high sensitivity to model weights, meaning small changes in weights can significantly impact the score [71].

FID: An improved evaluation metric over the inception score, which determines how similar the original and synthesized images are, is the Fréchet Inception Distance (FID). While the FID provides the distance between the photos, the inception provides the score. The difference in resemblance between two Gaussian distributions, *Gn,S* obtained from the synthetic sample distribution *P* and Gn_x,S_x obtained from the actual sample distribution P_x, is measured by the Frechet Inception Distance. The Frechet Distance is represented by the equation (3):

$$d^2 n, S, n_x, S_x = ||n - n_x||_2^2 \ trS \ \ S_x - 2SS_x{}^{12} \tag{3}$$

Compared to the Inception Score, there is more consistency with noise levels in the Fréchet distance.

PSNR: PSNR is a ratio that assesses the quality of the generated image *Y* relative to its original image *X*, as specified in [68]. This ratio offers information on the image’s noise and distortion. Mean Squared Error (MSE), as given in equation (4), is used in the computation. The calculation of PSNR is given by:

$$\begin{gathered} PSNRX, Y = 20 \log_{10} \left(\frac{MAX_X}{\sqrt{MSE}} \right) \\ MSE_{X,Y} = \frac{1}{ij} \sum_{m=0}^{i-1} \sum_{m=0}^{j-1} Xi, j - Yi, j^2 \end{gathered} \tag{4}$$

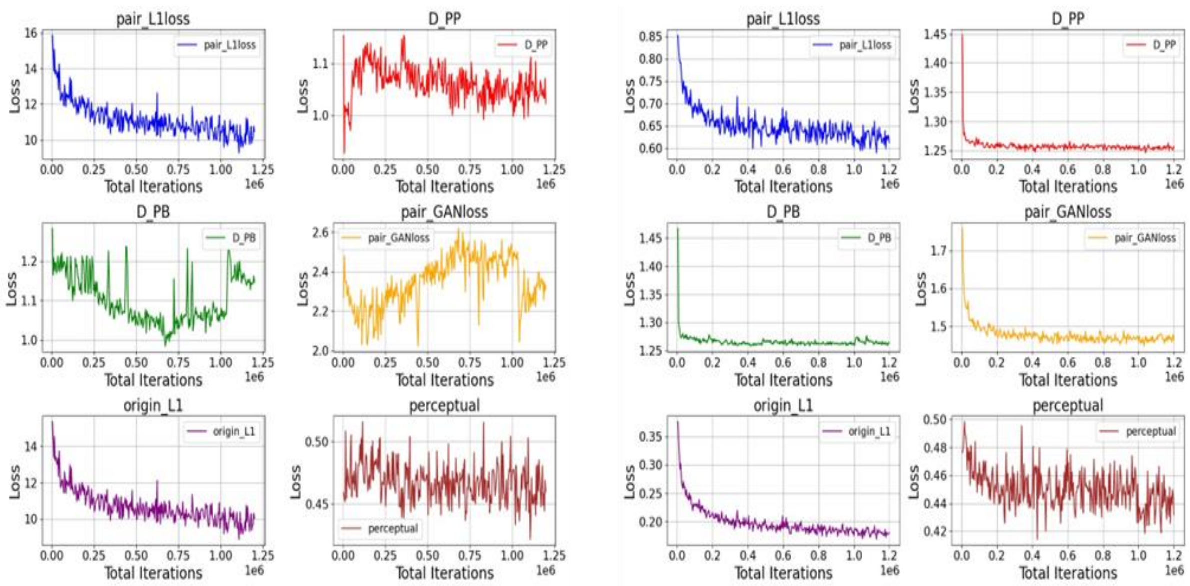
Here, MSE_X stands for the highest possible pixel value in an image, where *i* and *j* are the image’s height and width, respectively. Better image quality is correlated with a higher PSNR value.

LPIPS: Image perceptual similarity is quantified using LPIPS, which was first introduced in [67]. A Deep Convolutional Neural Network is used to extract feature vectors from the photos. The average distance *l*2 between these features is then computed to determine the perceptual similarity. It determines how similar two picture patches, each with the shape *N*,3,*H*,*W*, are in terms of activation functions. Equation (5) can be used to calculate LPIPS:

$$dx, x_0 = {}_{l}\frac{1}{H_l W_l}{}_{h,w} ||w_l \odot \hat{y}^l_{hw} - \hat{y}^l_{0hw}||_2^2 \tag{5}$$

4.3 Results

The models were trained and tested on an Intel Xeon(R) Silver 4214R CPU with an NVIDIA Corporation TU104GL [Quadro RTX 4000] GPU having 32GB of memory. The PyTorch library was used for implementation to ensure consistency across models. The Market-1501 dataset consists of 263,362 training pairs and 12,000 testing pairs. For the DeepFashion dataset, 101,966 pairs were used for training and 8,570 pairs for testing. Both datasets used the Human Pose Estimator (HPE) [72] to estimate poses. From an implementation perspective, the batch size for both datasets was 4, and each model was trained for 300 epochs, including 4000 iterations. The models used loss functions essential for directing model training to generate realistic, high-quality images. The loss functions for XingGAN and PATN are shown in Fig. 2. The results of the evaluation metrics for both datasets are shown in Table 3.



(a) **(b)**

Figure 2: loss functions of (a) XingGAN and (b) PATN models for the DeepFashion dataset.

Table 3: Performance Metrics for Market and DeepFashion Datasets

	Market-1501			DeepFashion		
	DPTN	**PATN**	**XING**	**DPTN**	**PATN**	**XING**
Inception Score (↑)	-	3.37498	3.20917	-	3.32950	3.12115
SSIM Score (↑)	0.0689	0.30135	0.13606	0.348	0.74082	0.72301
L1 Score (↓)	0.2213	0.29722	0.38293	0.1295	-	-
Masked Inception (↑)	-	3.69566	3.02500	-	-	-
Masked SSIM (↑)	-	0.80606	0.71311	-	-	-
PCKh (↑)	-	0.13568	0.05869	-	0.17437	0.08722
LPIPS (↓)	0.4583	-	-	0.4941	-	-
PSNR (↑)	11.2751	-	-	12.8363	-	-
FID (↓)	48.3663	-	-	50.6464	-	-

*↓ = Lower value is considered better.

*↑ = Higher value is considered better.

The metrics table clearly shows that PATN outperforms the other models in SSIM, masked SSIM, and Inception Score values for both datasets. High SSIM values provide evidence of strong structural similarity between the generated images and the ground truth. Therefore, PATN represents a model that achieves strong visual realism and diversity compared to the other two models. However, despite its superior performance, PATN has limitations. In some instances, it fails to produce highly realistic images, with outputs exhibiting texture inconsistencies and occlusions. These issues could be improved in future work. Some examples illustrating these limitations are shown in Fig. 3.

The generated image shown in Fig. 3 lacks detail and suffers from significant occlusion, reducing its overall clarity. Similar problems occurred with other models. Person re-identification (ReID) is another metric used in quantitative analysis. The objective is to determine whether the generated images are sufficiently realistic and discriminative to be used successfully in subsequent ReID tasks. Qualitative measures consist of visual quality and realism of the images. The models produce satisfactory results but lack realism in the generated images, a limitation that could be addressed by incorporating a different network architecture into the model.

Figure 3: some failure cases of the PATN model.

The twin-task learning design of the DPTN model, which includes a supporting source-to-source task with enhanced texture mapping and training stabilizing effects at a minimal parameter cost, makes it distinctive. Nevertheless, this method adds complexity and is highly dependent on precise pose prediction. Conversely, PATN employs a progressive pose attention transfer mechanism, which is dependent on pose estimator performance and suffers from overfitting but produces better look and shape consistency with a more streamlined architecture. Incurring higher complexity and processing cost, XingGAN employs a double-branch architecture coupled with crossing attention processes to generate truly realistic images with high visual quality. These models together establish new benchmarks for producing person images but are challenged by issues such as architectural complexity, risk of overfitting, and access to reliable pose data.

5. Conclusion

This paper analyzes pose-guided image generation techniques from a performance perspective, evaluating the degree to which they produce realistic and pose-accurate outputs. The analysis includes models such as Progressive Attention Transfer Network (PATN), Dual-task Pose Transformer Network (DPTN), and CrossingGAN (XingGAN), tested across two widely used datasets: DeepFashion and Market-1501. Metrics such as the structural similarity index, the inception score, the Fréchet inception distance, and the peak signal-to-noise ratio were used to assess the quality and diversity of the generated images.

The results show that PATN outperformed the other models in the SSIM, masked SSIM, and Inception Score values, indicating that it generates high-quality realistic images that are structurally similar to the ground truth. However, it faces challenges with texture consistency and occlusion. DPTN excels in factual texture mapping due to dual-task learning, while XingGAN is robust in bidirectional feature fusion of shape and appearance. Despite breakthroughs, challenges such as pose variations, identity preservation, and texture consistency remain vital to address. Future work should focus on developing more robust techniques for semantic parsing and multiscale reasoning to further enhance the realism and flexibility of pose-guided person image synthesis.

References

[1] Liqian Ma et al. "Pose guided person image generation". In: *Advances in Neural Information Processing Systems*. 2017, pp. 405–415.

[2] Wen Liu et al. "Liquid Warping GAN: A Unified Framework for Human Motion Imitation, Appearance Transfer and Novel View Synthesis". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, Nov. 2019, pp. 5903–5912. doi: 10.1109/ICCV.2019.00600. url: https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00600.

[3] Aliaksandr Siarohin et al. "Deformable GANs for Pose-Based Human Image Generation". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, Dec. 2018, pp. 3408–3416. isbn: 9781538664209. doi: 10.1109/CVPR. 2018.00359.

[4] Patrick Esser and Ekaterina Sutter. "A Variational U-Net for Conditional Appearance and Shape Generation". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, June 2018, pp. 8857–8866. doi: 10.1109/CVPR.2018.00923. url: https://doi. ieeecomputersociety.org/10.1109/CVPR.2018.00923.

[5] Han Yang et al. “Towards Photo-Realistic Virtual Try-On by Adaptively Generating ↔ Preserving Image Content”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, June 2020, pp. 7847–7856. doi: 10.1109/CVPR42600.2020.00787. url: https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00787.

[6] Kang Yuan and Sheng Li. “2.5D pose guided human image generation”. In: *Proceedings of the 2021 International Conference on Multimedia Retrieval*. Association for Computing Machinery, Inc, Aug. 2021, pp. 501–505. isbn: 9781450384636. doi: 10.1145/3460426. 3463580.

[7] Liqian Ma et al. “Disentangled Person Image Generation”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Dec. 2018, pp. 99–108. isbn: 9781538664209. doi: 10.1109/CVPR.2018.00018.

[8] Karmakar. “A Robust Pose Transformational GAN for Pose Guided Person Image Synthesis”. In: *Computer Vision, Pattern Recognition, Image Processing, and Graphics*. Ed. by Babu. Singapore: Springer Singapore, 2020, pp. 89–99. isbn: 978-981-15-8697-2.

[9] Meichen Liu et al. “Person image generation with semantic attention network for person re-identification”. In: *arXiv preprint arXiv:2008.07884* (2020).

[10] Stéphane Lathuilière et al. “Attention-based Fusion for Multi-source Human Image Generation”. In: *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2020, pp. 428–437. doi: 10.1109/WACV45572.2020.9093602.

[11] Artur Grigorev et al. “Coordinate-based Texture Inpainting for Pose-Guided Image Generation”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Nov. 2018). url: http://arxiv.org/abs/1811.11459.

[12] Natalia Neverova, Rıza Alp Güler, and Iasonas Kokkinos. “Dense Pose Transfer”. In: *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part III*. Munich, Germany: Springer-Verlag, 2018, pp. 128– 143. isbn: 978-3-030-01218-2. doi: $10.1007/978-3-030-01219-9_8$. url: https: //doi.org/10.1007/978-3-030-01219-9_8.

[13] Matthew Loper et al. “SMPL”. In: *ACM Transactions on Graphics* 34.6 (Oct. 2015), pp. 1–16. doi: 10.1145/2816795.2818013. url: https://doi.org/10.1145%2F2816795.2818013.

[14] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. “Densepose: Dense human pose estimation in the wild”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7297–7306.

[15] Chenyang Si et al. “Multistage Adversarial Losses for Pose-Based Human Image Synthesis”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, Dec. 2018, pp. 118–126. isbn: 9781538664209. doi: 10.1109/ CVPR.2018.00020.

[16] Albert Pumarola et al. “Unsupervised Person Image Synthesis in Arbitrary Poses”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Dec. 2018, pp. 8620–8628. isbn: 9781538664209. doi: 10.1109/ CVPR.2018.00899.

[17] Aliaksandr Siarohin et al. “Deformable GANs for Pose-Based Human Image Generation”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, June 2018. doi: 10.1109/cvpr.2018.00359. url: https://doi.org/10.1109%2Fcvpr. 2018.00359.

[18] Guha Balakrishnan et al. “Synthesizing Images of Humans in Unseen Poses”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Dec. 2018, pp. 8340–8348. isbn: 9781538664209. doi: 10.1109/CVPR.2018.00870.

[19] Haoye Dong et al. “Soft-gated warping-GAN for pose-guided person image synthesis”. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS’18. Montréal, Canada: Curran Associates Inc., 2018, pp. 472–482.

[20] Haoye Dong et al. “Part-preserving pose manipulation for person image synthesis”. In: *2019 IEEE International Conference on Multimedia and Expo (ICME)*. Vol. 2019-July. IEEE Computer Society, July 2019, pp. 1234–1239. isbn: 9781538695524. doi: 10.1109/ ICME.2019.00215.

[21] Yining Li, Chen Huang, and Chen Change Loy. “Dense intrinsic appearance flow for human pose transfer”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2019-June. IEEE Computer Society, June 2019, pp. 3688–3697. isbn: 9781728132938. doi: 10.1109/CVPR.2019.00381.

[22] Dong Liang et al. “PCGAN: partition-controlled human image generation”. In: *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI’19/IAAI’19/EAAI’19. Honolulu, Hawaii, USA: AAAI Press, 2019. isbn: 978-1-57735-809-1.doi: 10.1609/aaai.v33i01.33018698.url: https://doi.org/10.1609/aaai.v33i01.33018698.

[23] Mohamed Ilyes Lakhal, Oswald Lanz, and Andrea Cavallaro. “Pose Guided Human Image Synthesis by View Disentanglement and Enhanced Weighting Loss”. In: *Computer Vision – ECCV 2018 Workshops*. Ed. by Laura Leal-Taixé and Stefan Roth. Cham: Springer International Publishing, 2019, pp. 380–394. isbn: 978-3-030-11012-3.

[24] Wei Sun et al. “Pose Guided Fashion Image Synthesis Using Deep Generative Model”. In: *ArXiv* abs/1906.07251 (2019). url: https://api.semanticscholar.org/CorpusID: 189998692.

[25] Sijie Song et al. “Unsupervised Person Image Generation with Semantic Parsing Transformation”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, June 2019, pp.

2352– 2361. doi: 10.1109/CVPR.2019.00246. url: https://doi.ieeecomputersociety.org/ 10.1109/CVPR.2019.00246.

[26] Zhen Zhu et al. “Progressive pose attention transfer for person image generation”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2019June. IEEE Computer Society, June 2019, pp. 2342–2351. isbn: 9781728132938. doi: 10.1109/CVPR.2019.00245.

[27] Xintong Han et al. “ClothFlow: A flow-based model for clothed person generation”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Vol. 2019October. Institute of Electrical and Electronics Engineers Inc., Oct. 2019, pp. 10470– 10479. isbn: 9781728148038. doi: 10.1109/ICCV.2019.01057.

[28] Chengming Xu et al. “Pose-Guided Person Image Synthesis in the Non-Iconic Views”. In: *IEEE Transactions on Image Processing* 29 (2020), pp. 9060–9072. doi: 10.1109/TIP. 2020.3023853.

[29] Chengkang Shen, Peiyan Wang, and Wei Tang. “Two-Stream Appearance Transfer Network for Person Image Generation”. In: *ArXiv* abs/2011.04181 (2020). url: https:// api.semanticscholar.org/CorpusID:226282116.

[30] Wenbin Zhao et al. “Pose Guided Person Image Generation Based on Pose Skeleton Sequence and 3D Convolution”. In: *2020 IEEE International Conference on Image Processing (ICIP)*. 2020, pp. 1561–1565. doi: 10.1109/ICIP40778.2020.9190773.

[31] Tianxiang Ma et al. “MUST-GAN: Multi-level Statistics Transfer for Self-driven Person Image Generation”. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, June 2021, pp. 13617–13626. doi: 10.1109/CVPR46437.2021.01341. url: https://doi.ieeecomputersociety org/10.1109/CVPR46437.2021.01341.

[32] Kun Li et al. “PoNA: Pose-Guided Non-Local Attention for Human Pose Transfer”. In: *IEEE Transactions on Image Processing* 29 (Oct. 2020), pp. 9584–9599. issn: 1057-7149. doi: 10.1109/tip.2020.3029455.

[33] Hao Tang et al. “XingGAN for Person Image Generation”. In: *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV*. Glasgow, United Kingdom: Springer-Verlag, 2020, pp. 717–734. isbn: 978-3-03058594-5. doi: $10.1007/978-3-030-58595-2_43$. url: https://doi.org/10.1007/ 978-3-030-58595-2_43.

[34] Yurui Ren et al. “Deep Image Spatial Transformation for Person Image Generation”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 7687–7696. url: https://api.semanticscholar.org/CorpusID:211677243.

[35] Ting-yao Hu and Alexander G. Hauptmann. “Pose Guided Person Image Generation with Hidden P-Norm Regression”. In: *2021 IEEE International Conference on Image Processing (ICIP)* (2021), pp. 2423–2427. url: https://api.semanticscholar.org/ CorpusID:231979349.

[36] Lingbo Yang et al. “Towards Fine-Grained Human Pose Transfer with Detail Replenishing Network”. In: *IEEE Transactions on Image Processing* 30 (2021), pp. 2422–2435. issn: 19410042. doi: 10.1109/TIP.2021.3052364.

[37] Yuelong Li, Tongshun Zhang, and Jianming Wang. “SPMPG: ROBUST PERSON IMAGE GENERATION WITH SEMANTIC PARSING MAP”. In: *2021 IEEE International Conference on Image Processing (ICIP)*. Vol. 2021-September. IEEE Computer Society, 2021, pp. 1364–1368. isbn: 9781665441155. doi: 10.1109/ICIP42928.2021.9506397.

[38] Zhou Xiaomao, Wang Wei, and Du Bing. “PSG-GAN: Progressive Person Image Generation with Self-Guided Local Focuses”. In: *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*. Vol. 2021-November. IEEE Computer Society, 2021, pp. 763–769. isbn: 9781665408981. doi: 10.1109/ICTAI52525.2021.00121.

[39] Badour Albahar et al. “Pose with style”. In: *ACM Transactions on Graphics* 40 (6 Dec. 2021). issn: 15577368. doi: 10.1145/3478513.3480559.

[40] Amena Khatun et al. “Pose-driven Attention-guided Image Generation for Person ReIdentification”. In: *Pattern Recogn.* (Apr. 2021). url: http://arxiv.org/abs/2104. 13773.

[41] Meichen Liu et al. “Pose transfer generation with semantic parsing attention network for person re-identification”. In: *Knowledge-Based Systems* 223 (July 2021). issn: 09507051. doi: 10.1016/j.knosys.2021.107024.

[42] Jilin Tang et al. “Structure-aware person image generation with pose decomposition and semantic correlation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 2021, pp. 2656–2664.

[43] P. Zhang et al. “Exploring Dual-task Correlation for Pose Guided Person Image Generation”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, June 2022, pp. 7703–7712. doi: 10.1109/CVPR52688.2022.00756. url: https://doi.ieeecomputersociety. org/10.1109/CVPR52688.2022.00756.

[44] Jiaxiang Chen et al. “Exploring Kernel-based Texture Transfer for Pose-guided Person Image Generation”. In: *IEEE Transactions on Multimedia* (2022). issn: 19410077. doi: 10.1109/TMM.2022.3221351.

[45] Zijian Wang et al. “Self-supervised correlation mining network for person image generation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 7703–7712.

[46] A. Kumar Bhunia et al. “Person Image Synthesis via Denoising Diffusion Model”. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, June 2023, pp. 5968–5976. doi: 10.1109/ CVPR52729.2023.00578. url: https://doi.ieeecomputersociety.org/10.1109/ CVPR52729.2023.00578.

[47] Ji Liu and Yuesheng Zhu. “Precise Correspondence Enhanced GAN for Person Image Generation”. In: *Neural Processing Letters* 54 (6 Dec. 2022), pp. 5125–5142. issn: 1573773X. doi: 10.1007/s11063-022-10853-2.

[48] Hidemoto Nakada and Hideki Asoh. “A Method to Generate Posed Person Image with few Context Images”. In: *2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM)*. Institute of Electrical and Electronics Engineers Inc., 2022. isbn: 9781665426787. doi: 10.1109/IMCOM53663.2022.9721635.

[49] Zhengbin Yan et al. “SDAN: Semantic-Driven Dual Attentional Network for Image Generation”. In: *2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*. Institute of Electrical and Electronics Engineers Inc., 2022, pp. 521– 525. isbn: 9781665499163. doi: 10.1109/PRAI55851.2022.9904248.

[50] Baoyu Chen et al. “PMAN: Progressive Multi-Attention Network for Human Pose Transfer”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 32 (1 Jan. 2022), pp. 302–314. issn: 15582205. doi: 10.1109/TCSVT.2021.3059706.

[51] Pengze Zhang et al. “Lightweight Texture Correlation Network for Pose Guided Person Image Generation”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 32 (7 July 2022), pp. 4584–4598. issn: 15582205. doi: 10.1109/TCSVT.2021.3131738.

[52] Rishabh Jain et al. “VGFlow: Visibility guided Flow Network for Human Reposing”. In: 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, June 2023, pp. 21088–21097. doi: 10.1109/CVPR52729.2023.02020. url: https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.02020.

[53] Jiawei Lu et al. “Pose guided image generation from misaligned sources via residual flow based correction”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 2022, pp. 1863–1871.

[54] Wei Wei et al. “Style-content-aware Adaptive Normalization Based Pose Guided for Person Image Synthesis”. In: *IEEE Access* (June 2023), pp. 1–1. issn: 21693536. doi: 10.1109/access.2023.3290102.

[55] Liyuan Ma et al. “Multi-scale cross-domain alignment for person image generation”. In: *CAAI Transactions on Intelligence Technology* (2023). issn: 24682322. doi: 10.1049/ cit2.12224.

[56] Pengze Zhang et al. “Pose Guided Person Image Generation via Dual-task Correlation and Affinity Learning”. In: *IEEE Transactions on* Visualization *and Computer Graphics* (2023). issn: 19410506. doi: 10.1109/TVCG.2023.3286394.

[57] Yuan Huang et al. “CPD-GAN: Cascaded Pyramid Deformation GAN for Pose Transfer”. In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10096856.

[58] Yang Liu et al. "CoGAN: Cooperatively trained conditional and unconditional GAN for person image generation". In: *IET Image Processing* 17.10 (June 2023), pp. 2949–2957. doi: 10.1049/ipr2.12843. url: https://doi.org/10.1049%2Fipr2.12843.

[59] Meng Wang, Jiaxing Chen, and Haipeng Liu. "A novel Multi-scale architecture driven by decoupled semantic attention transfer for person image generation". In: *Computers and Graphics (*Pergamon*)* 111 (Apr. 2023), pp. 24–36. issn: 00978493. doi: 10.1016/j.cag. 2023.01.003.

[60] Prasun Roy et al. "Multi-scale attention guided pose transfer". In: *Pattern Recognition* 137 (May 2023). issn: 00313203. doi: 10.1016/j.patcog.2023.109315.

[61] Mykhaylo Andriluka et al. "2D Human Pose Estimation: New Benchmark and State of the Art Analysis". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 3686–3693. doi: 10.1109/CVPR.2014.471.

[62] Ziwei Liu et al. "DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations". In: 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 1096–1104. doi: 10.1109/CVPR.2016.124.

[63] Liang Zheng et al. "Scalable Person Re-identification: A Benchmark". In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 1116–1124. doi: 10.1109/ICCV.2015.133.

[64] Zhou Wang et al. "Image quality assessment: from error visibility to structural similarity". In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612. doi: 10.1109/TIP. 2003.819861.

[65] Tim Salimans et al. "Improved techniques for training gans". In: Advances *in neural information processing systems* 29 (2016).

[66] Martin Heusel et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium". In: *Advances in neural information processing systems* 30 (2017).

[67] Richard Zhang et al. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018). doi: 10.1109/cvpr.2018.00068.

[68] Ali Borji. "Pros and cons of GAN evaluation measures". In: *Computer vision and image understanding* 179 (2019), pp. 41–65.

[69] Jost Tobias Springenberg. "Unsupervised and semi-supervised learning with categorical generative adversarial networks". In: *arXiv preprint arXiv:1511.06390* (2015).

[70] Zhen Jia et al. "Human image generation: A comprehensive survey". In: *ACM Computing Surveys* 56.11 (2024), pp. 1–39.

[71] Shane Barratt and Rishi Sharma. "A note on the inception score". In: *arXiv preprint arXiv:1801.01973* (2018).

[72] Zhe Cao et al. "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields". In: *IEEE Transactions on Pattern Analysis & Machine Intelligence*

43.01 (Jan. 2021), pp. 172–186. issn: 1939-3539. doi: 10.1109/TPAMI.2019.2929257. url: https://doi.ieeecomputersociety.org/10.1109/TPAMI.2019.2929257.

Citation: Biponjot Kaur, Sarbjeet Singh. Performance Analysis of State-of-The-Art Models for Pose-Guided Person Image Generation. International Journal of Computer Engineering and Technology (IJCET), 16(1), 2025, 3747-3768.

Abstract Link: https://iaeme.com/Home/article_id/IJCET_16_01_258

Article Link:
https://iaeme.com/MasterAdmin/Journal_uploads/IJCET/VOLUME_16_ISSUE_1/IJCET_16_01_258.pdf