Volume 10, Issue 3, March – 2025 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar2007

Pose-Based Human Image Generation Using PATN

Dr. Suresh Babu Chandolu[1]; Tejaswi Nandeti[2]; Dimpul Deepthi Pippalla[3]; Gundareddy Gali Reddy[4]; Yallamalli Sasi Vadana Rao[5]; Mohammad Raheem[6]

[1][ORCID- ID: 0009-0006-5871-3822]

[1]CSE (Artificial Intelligence & Machine Learning)
Dhanekula Institute of Engineering and Technology (JNTUK) Vijayawada, India
[2]CSE (Artificial Intelligence & Machine Learning)
Dhanekula Institute of Engineering and Technology (JNTUK) Vijayawada, India
[3]CSE (Artificial Intelligence & Machine Learning)
Dhanekula Institute of Engineering and Technology (JNTUK) Vijayawada, India
[4]CSE (Artificial Intelligence & Machine Learning)
Dhanekula Institute of Engineering and Technology (JNTUK) Vijayawada, India
[5]CSE (Artificial Intelligence & Machine Learning)
Dhanekula Institute of Engineering and Technology (JNTUK) Vijayawada, India
[6]CSE (Artificial Intelligence & Machine Learning)
Dhanekula Institute of Engineering and Technology (JNTUK) Vijayawada, India

Publication Date: 2025/04/16

Abstract: This work presents a novel framework for personal image generation by transferring poses from a target image to a source image. Using a Pose Attention Transfer (PAT) network, our approach synthesizes realistic images of a person in the target pose while preserving identity and appearance details from the source image. The PAT network leverages attention mechanisms to focus on key regions, ensuring accurate pose transfer and high-quality texture preservation. Experimental results demonstrate that our method generates visually coherent and realistic images, outperforming existing state-of-the-art techniques. This framework has significant potential for virtual try-on, animation, and video synthesis applications.

Keywords: *Pose Image Generation, Pose Attention Transfer Network (PATN), Generative Adversarial Network (GAN).*

How to Cite: Dr. Suresh Babu Chandolu; Tejaswi Nandeti; Dimpul Deepthi Pippalla; Gundareddy Gali Reddy; Yallamalli Sasi Vadana Rao; Mohammad Raheem (2025). Pose-Based Human Image Generation Using PATN. *International Journal of Innovative Science and Research Technology*, 10(3), 3104-3112.
https://doi.org/10.38124/ijisrt/25mar2007

I. INTRODUCTION

Generating realistic pix of non-rigid objects, consisting of human beings, is a challenging mission due to the wide variety of deformations and articulations. One particularly valuable problem inside this area is the pose switch, wherein the intention is to generate a photograph of someone in a new goal pose primarily based on a supply photograph. This trouble has important programs in regions like video synthesis, where a series of poses may be used to create dynamic animations, and in data augmentation for improving man or woman-identity systems. A most important assignment in pose transfer is coping with partial observations of the person. When generating a target pose, the version must infer and reconstruct occluded or unobserved body components, which calls for information on human anatomy and spatial relationships. Additionally, the advent of a person can range notably across special poses and viewpoints, making it tough for models to hold steady texture, lighting fixtures, and identity. To address those demanding situations, we advise a singular technique based totally at the idea of manifold mastering. We treat the set of all feasible poses and perspectives of a person as mendacity on an excessive-dimensional manifold. Pose transfer may be visible as navigating this manifold, transitioning from a source pose to a goal pose. While the worldwide structure of the manifold is complicated, its nearby structure is easier, making it less complicated to model incremental changes in pose. Building

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar2007

in this perception, we introduce a modern pose-transfer framework that decomposes the pose-transfer procedure into a series of smaller, manageable steps. In contrast to standard one-step transfer strategies, our method makes use of a *sequence of Pose-Attentional Transfer Blocks (PATBs) to refine the pose iteratively, making sure clean and accurate.*

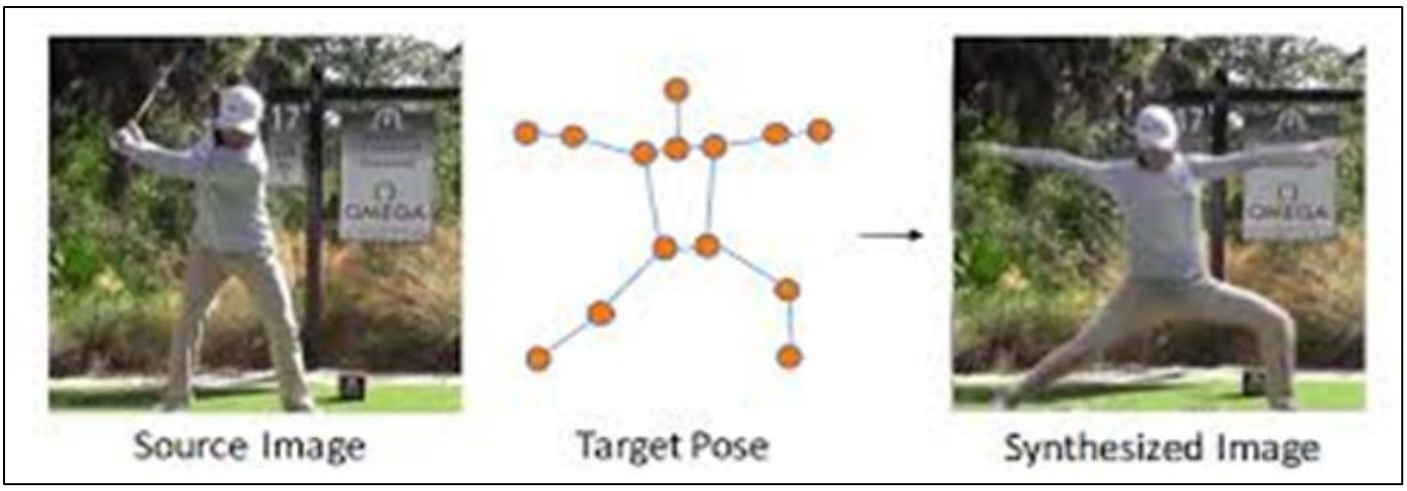

Fig 1 Pose Transfer Process

➢ *Obstacles in Pose-Based Synthesis Using GANs*

Aside from their exceptional accomplishments, GANs continue to face difficulties regarding instability in training, mode collapse, and the generation of high-resolution outputs that also maintain structural integrity. To counter these concerns, several developments ranging from Deep Convolutional GANs (DCGANs), Wasserstein GANs (WGANs), Progressive Growing of GANs (ProGANs) to StyleGANs have been introduced, each increasing the stability, diversity, and fidelity of the generated samples. In addition, domain-specific changes such as CycleGAN for image's translation and StackGAN for generating images from text have augmented the use of GANs in various domains.

➢ *Progressive Pose Attention Transfer Network (PATN)*

First, allow us to discover the sector of GANs with the Progressive Pose Attention Transfer Network (PATN), devoted to pose switch responsibilities. PATNs practice attention strategies to iteratively enhance the alignment of the supply pose and target appearance info. Rather than the use of traditional techniques of pose transfer in one single step, PATNs use a series of Pose-Attentional Transfer Blocks "PATBs" that permit gross to excellent incremental pose adjustments. In doing so, adjustments to pictures for a selected man or woman emerge as extra sensible and identification steady. Human picture synthesis, digital try-ons, and statistics augmentation for man or woman re-identity are example packages in which PATNs have demonstrated useful. PATNs apply a cascaded approach in which each Pose-Attentional Transfer Block (PATB) makes a speciality of, and local-poses capabilities in the pix the usage of an interest mechanism. This cages the results of massive pose alternate imparting greater sensible pix. Enforcing pose alignment poses extensively more final results in pose-based photograph synthesis for PATNs as compared to older GAN-based totally techniques.

➢ *Improvement in the Training of GANs on a Larger Scale*

The improvement of BigGANs has progressed photo first-rate and range, permitting high-decision picture synthesis through techniques like orthogonal regularization and truncation

Additionally, StyleGAN enhances controlled pose-based photograph synthesis with higher interpolation and spatial manipulate over stochastic and high-degree attributes.

➢ *Scope of Project*

This work aims to improve the PATN-enabled pose transfer model through the Progressive Pose Attention Transfer Network. By using attention-driven transfer blocks together with progressive refinement, we aim to improve the realism, identity, and structure fidelity of the generated human images. These improvements will be achieved by applying sophisticated loss functions, deformable skip connections, and largescale GAN training, which will make PATN-based models suitable for practical applications like virtual try-on, animation, and person re-identification.

II. LITERATURE SURVEY

➢ *The Generative Adversarial Networks were first articulated by Ian Goodfellow and others in 2014 (Goodfellow et al, 2014).*

Goodfellow and others published a paper in 2014 on what they called Generative Adversarial Networks (GANs) which is used for training generative models using an innovative method of deep learning known as adversarial learning. In a nutshell, the architecture of GAN does have two components, a generator (G) and a discriminator (D). D classifies incoming data as real or fake while G generates realistic samples of data. Both these networks set out to accomplish our goal in what can be termed in gaming as a minimax game. Here, G's intent is to outsmart D, and D constantly progresses in his ability to tell real from fake data.

Both networks incorporate multilayer perceptron's (MLPs) and use backpropagation combined with dropout for optimization. Unlike traditional generative models, GANs does not depend on Markov chains or any explicit chance

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar2007

calculating, which is certainly more resource efficient. The model was subjected to evaluation from multisource datasets of MNIST and CIFAR-10 as well as the Toronto Face Database (TFD) and was able to outperform the older capturing models such as Deep Belief Networks (DBNs) and Restricted Boltzmann Machines (RBMs).

- *Conditional Generative Adversarial Networks (cGANs) – Mehdi Mirza & Simon Osindero (2014)*

Mirza and Osindero (2014) modified the original Generative Adversarial Networks (GANs) framework by adding cGANs which allow for greater data generation flexibility with the use of additional input conditions. Unlike regular GANs, where the generator output is devoid of any context control, cGANs allow for context-aware generation through the intake of auxiliary information (y) such as class labels and modality data into both the generator and discriminator.

- *New Methods for Training GANs: Pushing the Limits – Tim Salamans, et al, 2016.*

In 2016, Salamans et al expanded the capacity of Generative Adversarial Networks by introducing a new approach involving configuring a few parameters to improve their stability, convergence, and sample quality. Training instability, mode collapse, and Nash equilibrium attainment are common issues GANs face. This paper aims to provide several methods to alleviate these concerns and boost semi-supervised learning capabilities with the use of GANs.The most important features of the work are:One of the most important contributions to learning semi-supervised feature matching is shallow network salient feature extraction, a technique used to train weakly supervised generative discriminative models. The discriminator's output is not directly targeted, but rather, as in shallow networks, its output intermediate feature activations must be captured.

- *StyleGAN: A Style-Based Generator Architecture for Generative Adversarial Networks – Tero Karras et al. (2019)*

The novel term GAN arose from Tero Karras and his colleagues' work (2019) on a new image synthesis and manipulation architecture, which incorporates visual style-transferring processes known as StyleGAN. A conventional GAN model utilizes only one single inputted latent code, but StyleGAN, in its architecture, takes advantage of an intermediate latent space in addition to an Adaptive Instance Normalization (AdaIN). With such a structure, StyleGAN achieves an unprecedented level of disentangling requisite high-level elements like pose or identity from the stochastic details of the image, such as freckles or hair texture.

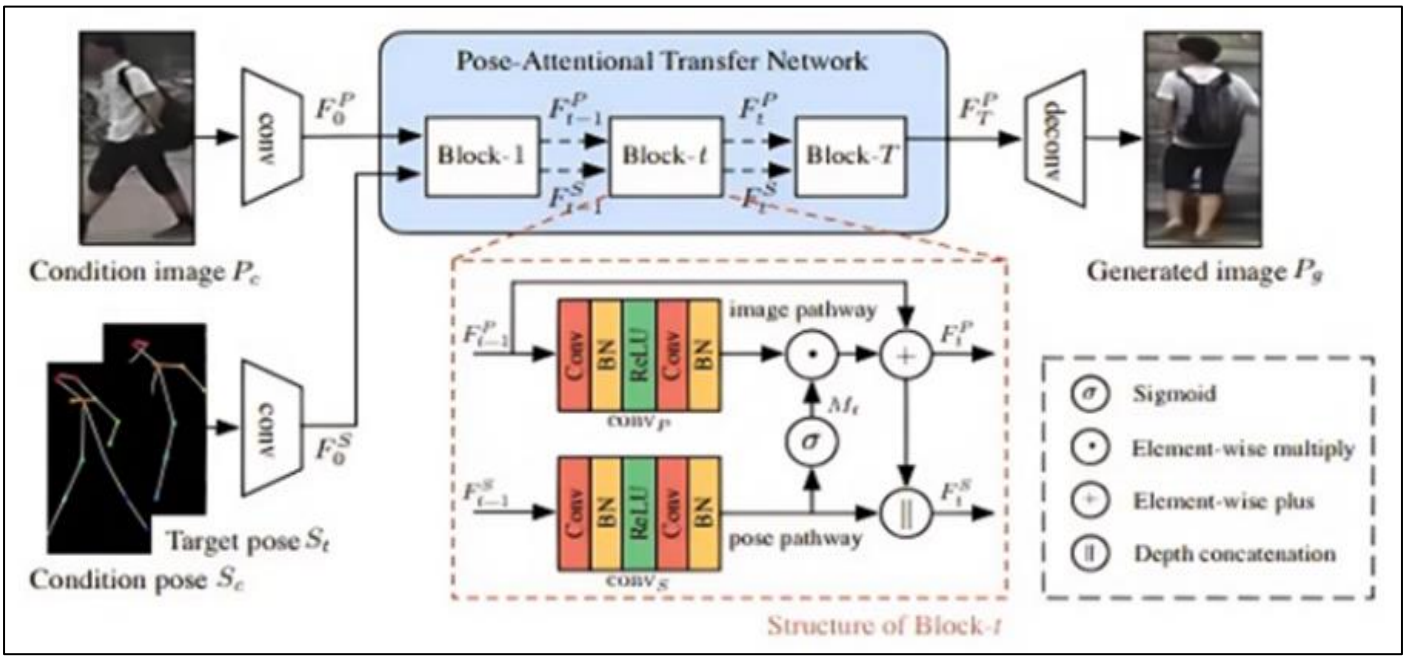

Fig 2 Architecture of Pose Attention Transfer Network (PATN)

III. PROPOSED SYSTEM

The proposed machine is based on a Pose-Attentional Transfer Network (PATN), a specialized Generative Adversarial Network (GAN) designed for pose-transfer responsibilities. The version goals to generate realistic pictures of someone in a goal pose while preserving the identification and appearance of the source photo. The system takes inputs: a source picture of someone and a goal pose represented with the aid of key points. The version then synthesizes a new photograph of the person within the goal pose. This is accomplished through interest mechanisms, perceptual loss, and adverse education, making sure of top-notch and realistic outputs. Generative Adversarial Networks (GANs) were extensively used for photograph synthesis tasks due to their ability to generate fantastically realistic photographs. A GAN includes components: a generator and a discriminator. The generator synthesizes photographs, even as the discriminator distinguishes between actual and generated photos. Through antagonistic schooling, the generator learns to supply pix that are indistinguishable from actual ones. In the context of the pose switch, the Pose-Attentional Transfer Network (PATN)

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar2007

extends the traditional GAN framework by incorporating attention mechanisms to awareness of relevant areas of the picture and pose. This permits the version to control complicated spatial modifications and hold minute information within the generated pics. The PATN structure is in particular designed to fuse statistics from the source picture and goal pose, allowing correct and realistic pose switches.

IV. MODEL ARCHITECTURE

We start by describing the dataset. The dataset consists of a collection of personal images, where each person has multiple images associated with them. The total number of persons in the dataset is denoted as **M**, and each person has a specific number of images.

Each image is represented using a keypoint-based method, which consists of an 18-channel heat map. This heatmap encodes the positions of 18 joints of the human body, capturing the overall pose and structure of the person. To estimate these joints, we use the **Human Pose Estimator.**

- **(HPE)**, which is consistent with methods used in related works.
- During training, the model takes two images as input:
- **Condition Image (Pc)**: The original image with the initial pose.
- **Target Image (Pt)**: The image representing the desired target pose.
- Along with these images, the model also requires their corresponding pose heat maps:
- **Condition Pose Heat Map (Sc)**: The pose map of the condition image.
- **Target Pose Heat Map (St)**: The pose map of the target image.

The generator uses these inputs to produce a new person image with the target pose. To ensure that the generated image looks realistic, discriminators are used to evaluate its authenticity, helping the model learn to create more convincing and natural-looking results.

A. Generator

➢ *Encoders*

The generator is designed to transfer the pose of a person from a given condition image to a target pose, producing a realistic-looking output image. The main inputs to the generator are:

- **Condition Image (Pc)**: The image of the person whose pose needs to be changed.
- **Condition Pose (Sc)**: The initial pose of the person in the condition image.
- **Target Pose (St)**: The desired pose to which the person's position should be transformed.

The generator first encodes the condition image using two down-sampling convolutional layers. Simultaneously, it encodes both the condition pose and the target pose by stacking them along their depth axes before feeding them through another set of two down-sampling convolutional layers. This encoding process effectively combines and preserves information from both poses, reducing computational complexity while maintaining dependencies.

Instead of encoding the two poses separately and concatenating the resulting vectors at the end, this integrated encoding approach works more efficiently and requires less computation.

Table 1 Comparison with Previous Work: Our model Achieves State-of-the-Art Results, Surpassing Existing Methods

Model	Market1501						DeepFashion			
	SSIM	*IS*	*mask-SSIM*	*mask-IS*	*DS*	*PCkh*	SSIM	IS	DS	PCkh
Ma etal.	0.253	3.460	0.792	3.435	-	-	0.762	3.090	-	-
Ma etal.	0.099	3.483	0.614	3.149	-	-	0.614	3.228	-	-
Siarohin et al	0.290	3.185	0.805	3.502	0.720	-	0.756	3.439	0.960	-
Ours	0.311	3.323	0.811	3.773	0.740	0.94	0.773	3.209	0.970	0.96

➢ *Pose-Attentional Transfer Network*

At the core of the generator is the **Pose-Attentional Transfer Network (PATN)**, which is composed of multiple **Pose-Attentional Transfer Blocks (PATBs)** arranged sequentially. The PATN progressively updates the encoded image and pose information throughout these blocks.

The network starts with initial image and pose codes and updates them progressively through the PATBs. At the final stage, the network outputs the updated image code to decode the generated image, while the final pose code is discarded.

➢ *Structure of Pose-Attentional Transfer Block (PATB)*

Each PATB has an identical structure and is designed to update both the image and pose codes. The block has two pathways:

- **Image Pathway:** Updates the image code.
- **Pose Pathway:** Updates the pose code.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar2007

These pathways interact and exchange information to ensure that both the image and pose are synchronized during each update step.

➢ *Pose Attention Masks*

The core idea of pose transfer is to move image patches from the positions indicated by the condition pose to those indicated by the target pose. To achieve this, the network uses **attention masks**, which are essentially maps that indicate how important each element of the image code is for the transformation.

These attention masks are generated from the combined pose information using convolutional layers followed by normalization and activation functions. The mask values range between 0 and 1, signifying the importance of each element in the image code.

➢ *Image Code Update*

Once the attention masks are generated, they are applied to the image code to either preserve or suppress certain regions. This helps maintain critical information from the original image while allowing flexible transformation. The network also uses residual connections to retain the original image information, which aids in achieving stable training and effective transformation, especially when using multiple PATBs.

➢ *Pose Code Update*

As the image code gets progressively updated through the network, the pose code must also be updated to remain in sync. The pose code update involves combining the transformed pose code with the updated image code to ensure consistent guidance for patch movement.

By continuously updating both the image and pose codes through the PATBs, the generator can produce realistic-looking transformed images that maintain visual coherence and natural appearance.

➢ *Decoder*

After the updates in the Pose-Attentional Transfer Network (PATN), the final output of the generator is the **final image code**, while the final pose code is discarded. The decoder then generates the output image using the final image code through several **deconvolutional layers**. This process allows the network to reconstruct a high-quality person image from the transformed feature representation.

B. Discriminators

To ensure the generated image looks realistic and matches the desired pose, we use two discriminators:

➢ *Appearance Discriminator (DA):*

Evaluates whether the generated image retains the same identity and appearance as the original condition image.

➢ *Shape Discriminator (DS):*

Assesses whether the generated image matches the target pose correctly.

Both discriminators have a similar structure. The generated image is concatenated with either the original image (for appearance checking) or the target pose (for shape checking) along the depth axis. This combined input is fed into a **Convolutional Neural Network (CNN)**, Which outputs **consistency scores**:

- **RA**: Appearance consistency score
- **RS**: Shape consistency score

The final consistency score (**R**) is calculated as the product of the two individual scores, ensuring both appearance and shape consistency are evaluated together.

To enhance their effectiveness, the discriminators are constructed with **three residual blocks** after two down-sampling convolutional layers. This structure improves the ability of the discriminators to distinguish between real and generated images, especially as the model training progresses.

➢ *Training*

The training process involves minimizing a **full loss function**, which combines **adversarial loss** and **L1 loss**. The adversarial loss encourages the generator to produce realistic images, while the L1 loss ensures that the generated image closely matches the target image at the pixel level.

The **adversarial loss** is calculated using both the appearance and shape discriminators, while the **combined L1 loss** includes:

- **Pixel-wise L1 loss (LL1)**: Measures the difference between the generated and target images.
- **Perceptual L1 loss (LperL1)**: Improves visual quality by considering high-level features from a pre-trained VGG-19 model, which captures texture and style differences effectively.

The **training process** alternates between training the generator and the two discriminators. The generator takes the **Condition image**, **condition pose**, and **target pose** as input, producing a transformed image. The discriminators then evaluate the consistency of the generated image with both the original identity and the target pose.

➢ *Implementation Details*

The model is implemented using the **PyTorch framework** and trained using the **Adam optimizer** for approximately **90,000 iterations**. The learning rate starts at **0.0002** and linearly decays to zero after **60,000 iterations**.

We use **9 Pose-Attentional Transfer Blocks (PATBs)** in the generator.

- *For Normalization:*

✓ **Instance normalization** is used for the **DeepFashion dataset**. **Batch normalization** is used for the **Market-1501 dataset**.
✓ **Dropout** is applied only within the PATBs, with a rate of **0.5** to prevent overfitting.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar2007

✓ **Leaky ReLU** activation (with a negative slope of **0.2**) follows each convolution or normalization layer within the discriminators to maintain stability.

By following this structured training and implementation process, the model achieves high-quality, natural-looking pose transfer with enhanced visual consistency and smoothness.

V. PERFORMANCE EVALUATION AND BENCHMARKING

In this section, we carry out comprehensive experiments to assess the effectiveness and efficiency of our proposed network. Our experiments demonstrate the superiority of our approach through both quantitative metrics and visual quality comparisons.

➢ *Datasets*

We primarily conduct our experiments on two challenging person re-identification datasets: **Market-1501** and **DeepFashion**.

- **Market-1501**: This dataset presents a tough challenge due to its low-resolution images (128 × 64) and significant variations in pose, viewpoint, background, and lighting conditions.
- **DeepFashion**: In contrast, DeepFashion images are high-resolution (256 × 256) and have clean backgrounds, making them more visually consistent.

For pose detection, we use the **Human Pose Estimator (HPE)** and filter out any images where no human body is detected. This results in:

- **Market-1501**: 263,632 training pairs and 12,000 testing pairs.
- **DeepFashion**: 101,966 training pairs and 8,570 testing pairs.

To ensure a fair evaluation of the model's generalization ability, the training and testing sets do not contain overlapping person identities.

➢ *Evaluation Metrics*

Evaluating the appearance and shape consistency of generated images is an ongoing challenge. Previous works have used several metrics, but they have limitations:

- **SSIM (Structure Similarity)**: Measures global structural similarity but does not effectively quantify shape consistency.
- **IS (Inception Score)** and **DS (Detection Score)**: Use classifiers and detectors to assess image quality, but they don't directly evaluate shape consistency.

To address this gap, we introduce a new metric to explicitly assess shape consistency. Our metric is based on **pose joints alignment**, evaluated using the **PCKh measure**. This score calculates the percentage of correctly aligned keypoints, considering the head segment as a reference.

➢ *Comparison with Previous Work*

We compare our method against existing approaches, both quantitatively and qualitatively. The results, summarized in **Table 1**, demonstrate that our method consistently outperforms previous works across most metrics.

Despite the possibility of overlap between our testing set and some training images from previous methods (due to the lack of public data splits), our method shows steady improvements. Notably, our approach achieves the highest **PCKh score** for shape consistency, outperforming previous methods by a significant margin, especially on the DeepFashion dataset, where we observe a 2% improvement.

Effectively evaluating the appearance and shape consistency of generated images remains an open problem. Previous approaches have used metrics such as Structural Similarity (SSIM) and Inception Score (IS) to assess image quality. To reduce the influence of background elements, masked versions of these metrics—mask-SSIM and mask-IS—were introduced. Additionally, Detection Score (DS) was proposed to measure whether a person in the generated image can be correctly detected by a detector. However, these metrics have limitations in explicitly quantifying shape consistency. For instance, SSIM relies on global covariance and means of images, making it inadequate for evaluating shape consistency. Similarly, IS and DS depend on image classifiers and object detectors, which are unrelated to shape consistency.

To address these limitations, we introduce a new metric to explicitly assess shape consistency. The proposed metric represents person's shape using 18 pose joints obtained from a Human Pose Estimator (HPE). Shape consistency is then approximated by evaluating the alignment of these pose joints using the PCKh measure. According to the PCKh protocol, the score is calculated as the percentage of key point pairs whose offsets are below half the size of the head segment. The head segment is estimated using a bounding box that tightly covers key points related to the head. This approach provides a more direct and accurate measure of shape consistency in generated images.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar2007

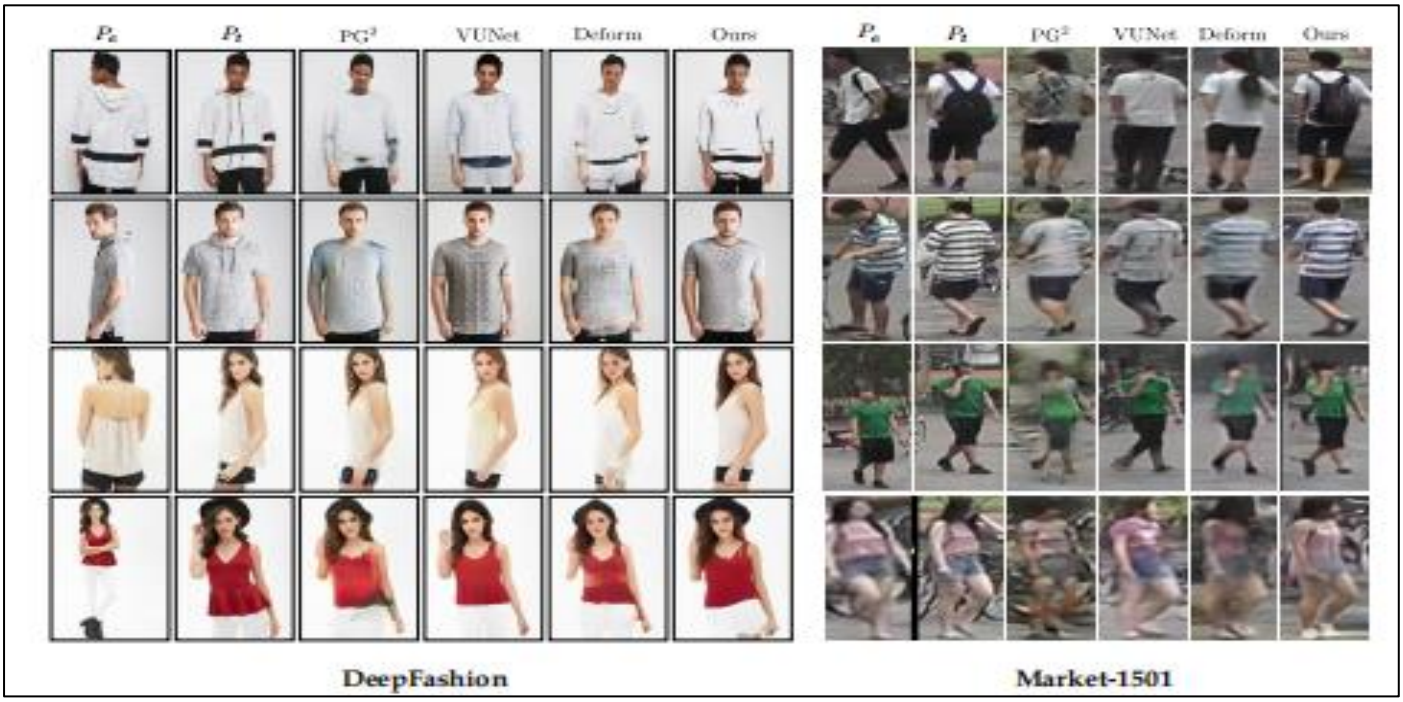

Fig 3 Qualitative Comparisons on Deep Fashion (Left) and Market- 1501 (Right) Dataset.

Fig 3 (Left) showcases some typical qualitative examples from the high-resolution **DeepFashion** dataset [19], highlighting scenarios with significant pose changes and scale variations. Our proposed method consistently preserves the person's integrity, particularly noticeable around challenging areas like the wrists, while maintaining the most natural posture among all compared methods.

- *Moreover, our approach captures fine details exceptionally well. For instance:*

- The skin tone in the first row looks realistic and consistent.
- The whiskers in the second row are clear and sharp.
- The hair details in the third row are well-defined.
- The hat in the last row appears naturally integrated.

Additionally, our method produces more refined and visually appealing facial features compared to other approaches.

We also tested our model on the **Market-1501** dataset, which is known for its poor image quality. As shown in Figure 3 (Right), our method generates the sharpest and most accurate person images, while other methods tend to produce somewhat blurred outputs. Notably, our model accurately replicates complex leg layouts that align with the target poses, even when the legs are crossed (as seen in the second and third rows). It also handles blurred input images more effectively (as seen in the last row).

One particularly impressive aspect of our model is its ability to maintain **appearance consistency**. For example, the bag visible in the first row of our results is entirely lost in other methods, demonstrating our model's superior ability to preserve essential details.

VI. RESULT ANALYSIS

Our network's generator, PATN, has two key design features: the Pose-Attentional Transfer Block (PATB) and cascaded building blocks. The PATB is carefully designed to optimize both appearance and pose simultaneously using an attention mechanism. The cascaded building blocks progressively guide the deformable transfer process, improving image quality. To evaluate the effectiveness of these design choices, we conducted two comparison experiments. In the first experiment, we replaced the PATB with a standard residual block , creating a generator named the ResNet generator. In the second experiment, we tested varying numbers of PATBs to examine the impact of the progressive design.

The qualitative comparison results are illustrated in Figure 4. When comparing images generated by the ResNet generator and those produced by our PATN generator with the same number of building blocks, it becomes evident that the PATN generator consistently produces images with shapes and appearances that closely match the target. On the other hand, the ResNet generator tends to overlook subtle but crucial appearance details, especially when they occupy only a small portion of the image. It also struggles to accurately generate foreground shapes when the target pose is uncommon. For instance, in the first row, the ResNet generator misses the red ring at the bottom of the sweater. In the third row, it fails to reproduce the white cap correctly. In the fourth row, the T-shirt color is mistakenly generated as black due to the presence of a black backpack. Additionally, in the second row, the shapes of the sitting girls appear incomplete since the sitting pose is relatively rare in the DeepFashion dataset.

In contrast, our PATN generator, with just 5 PATBs, outperforms the ResNet generator even when the latter uses 13 residual blocks. We attribute this improvement to the pose

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar2007

attention mechanism, which significantly enhances the model's ability to capture and leverage essential features. Moreover, using 9 PATBs results in even more refined and visually pleasing images. Although increasing the number of PATBs to 13 provides a slight performance boost, the improvement is marginal, so we opted for 9 PATBs as the default to balance efficiency and quality.

Quantitative results, demonstrate that our PATN generator with only 5 PATBs consistently surpasses the ResNet generator across various configurations and evaluation measures. These outcomes strongly validate the advantages of our PATN design. To further investigate the effect of each component within the PATB, we performed additional experiments by removing the addition operation (w/o add), concatenation operation (w/o cat), and residual blocks from the discriminators (w/o resD)

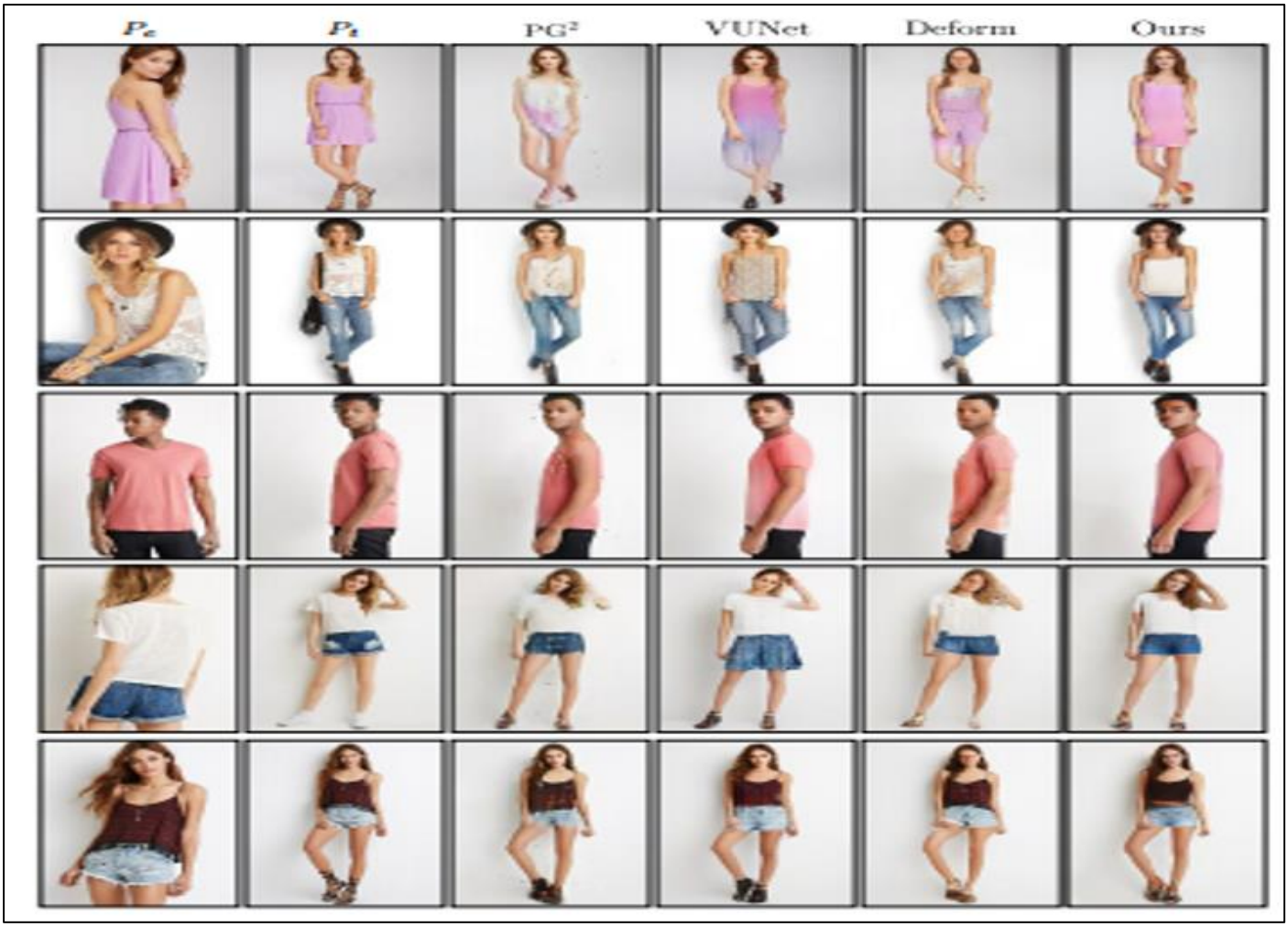

Fig 4 Quantitative results of the Deep Fashion

Both qualitative and quantitative results, presented in Figure 4, reveal that eliminating any component of the PATB leads to a noticeable performance decline. This reduction in quality is apparent through visual inconsistencies such as colour distortion and unnatural details. Moreover, omitting residual blocks from the discriminators adversely affects local detail quality and the integrity of the person's body, underscoring the importance of our architectural decisions.

VII. CONCLUSION

We put forward in this paper a progressive pose attention transfer network to address the difficult pose trans²fer. The network cascades multiple Pose Attentional Transfer Blocks (PATBs), each of which can optimize appearance and pose simultaneously through the attention mechanism, hence directs the deformable transfer process progressively.

Our network outperforms existing work in both subjective visual realness and objective quantitative scores simultaneously and meanwhile enhances computational efficiency and also decreases the complexity of the model considerably. In addition, our suggested network may be employed to solve the lacking training data problem for person re-identification extensively. Furthermore, our progressive pose-attentional transfer process could be easily seen by its attention masks, thereby making our network more interpretable. Additionally, our network's design has been experimentally verified by the ablation studies.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar2007

Our progressive pose-attentional transfer network not only is exclusive to generating person images but can also be potentially adapted to create other non-rigid objects. Additionally, we hope the concept of our progressive attention transfer method could be useful for other GAN-based image generation methods as well.

REFERENCES

[1]. Mihai Zanfir, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. Human appearance transfer. In *Proc. CVPR*, pages 5391–5399, 2018.

[2]. Bo Zhao, Xiao Wu, Zhi-Qi Cheng, Hao Liu, and Jiashi Feng. Multi-view image generation from a single-view. *CoRR*, abs/1704.04886, 2017.

[3]. Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jing dong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proc. ICCV*, pages 1116–1124, 2015.

[4]. Liang Zheng, Yi Yang, and Alexander G. Hauptmann. Person re-identification: Past, present and future. *CoRR*, abs/1610.02984, 2016.

[5]. Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *Proc. ICCV*, pages 3774–3782, 2017.

[6]. Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle consistent adversarial networks. In *Proc. ICCV*, pages 2242–2251, 2017.

[7]. Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. ICML*, pages 448–456, 2015.

[8]. Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. *CoRR*, abs/1712.02621, 2017.

[9]. Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Recti-fier nonlinearities improve neural network acoustic models. In *Proc. ICML*, volume 30, page 3, 2013.

[10]. Liang Mei, Jingen Liu, Alfred O. Hero III, and Silvio Savarese. Robust object pose estimation via statistical mani fold modeling. In *Proc. ICCV*, pages 967–974, 2011.

[11]. Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

[12]. Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proc. ICML*, pages 807–814, 2010.

[13]. Natalia Neverova, Riza Alp Guler, and Iasonas Kokkinos.Dense pose transfer. *arXiv preprint arXiv:1809.01995*, 2018.

[14]. Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proc. ICML*, pages 2642–2651, 2017.

[15]. Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Unsupervised person image synthesis in arbitrary poses. In *Proc. CVPR*, pages 8620–8628, 2018.

[16]. Alec Radford, Luke Metz, and Soumith Chintala. Un supervised representation learning with deep convolu tional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[17]. Amit Raj, Patsorn Sangkloy, Huiwen Chang, James Hays, Duygu Ceylan, and Jingwan Lu. Swapnet: Image based garment transfer. In *Proc. ECCV*, pages 679–695, 2018.

[18]. Aliaksandr Siarohin, Enver Sangineto, Stephane Lathuili ´ ere, ` and Nicu Sebe. Deformable gans for pose-based human image generation. *CoRR*, abs/1801.00055, 2018.

[19]. Zhen Zhu , Tengteng Huang, Baoguang Shi , Miao Yu , Bofei Wang and Xiang Bai . Progressive Pose Attention Transfer for Person Image Generation.*arXiv:1904.03349v3,2019.*