Improving Texture Integrity through Second-Order Constraints on Warping Maps

Mohsen Tabejamaat[a] (mohsen.tabejamaat@inria.fr), Farhood Negin[a] (farhood.negin@inria.fr), François Bremond[a] (francois.bremond@inria.fr)

[a] INRIA, France

Improving Texture Integrity through Second-Order Constraints on Warping Maps

Mohsen Tabejamaat[a], Farhood Negin[a], François Bremond[a]

[a]*INRIA, France*

Abstract

Pose Transfer has recently gained significant attention, particularly for its user-friendly applications in the animation industry. The primary objective is to transform a given RGB image into a new target pose. This process involves two consecutive tasks: initially, warping the image to approximately align with the target pose, and subsequently using this rough estimation to generate a photorealistic image of the input in the desired pose.

The primary challenge lies in the first task, where the image undergoes a rough transformation to its new location in the target pose. Current deep learning approaches rely on first-order warping, employing an affine transformation to move all image pixels. Despite yielding promising results, this approach has significant challenges when dealing with complex deformations, mainly due to the simplistic nature of its linear function. In contrast, we suggest transferring patches using a set of correlation layers. In each layer, the warping for each pixel of the image is individually estimated. We additionally introduce a constraint aimed at minimizing the energy of second derivatives across the entire warping map of the pixels. This allows for keeping the integration of local textures following the warping process, a feature already ensured in the affine-based transformation by restricting the transition to a linear function for all the image pixels. Our approach not only preserves the integrity of local textures, akin to the affine transformation, but achieves this by individually estimating the warping for each image

*Corresponding author.

Email addresses: mohsen.tabejamaat@inria.fr (Mohsen Tabejamaat), farhood.negin@inria.fr (Farhood Negin), francois.bremond@inria.fr (François Bremond)

pixel, thereby enabling finer adjustments of the input sample to the target pose. We illustrate the superior performance of this technique compared to affine-based strategies on the renowned DeepFashion database.

Keywords: Pose transfer, Image to image translation, Deep spatial transformation

1. Introduction

Deep learning has become a prominent topic in both science and technology, demonstrating widespread applicability across various fields He et al. (2023); Liu & Zhang (2024); Dhar et al. (2023); Al Ka'bi (2023), including the animation industry Mourot et al. (2022); Pham et al. (2017); Wan & Ren (2021). Notably, Pose Transfer Zhang et al. (2021); Ren et al. (2020); Zhang et al. (2022); Zhou et al. (2022); Ren et al. (2022) has emerged as a focal point within animation, drawing attention for its user-friendly application and potential transformative effects on animation processes.

Animation is a method for introducing movement into a static drawing, wherein a series of consecutive drawings are produced and captured as a complete video. The generation of movement typically involves maintaining the constant shape of the primary drawing and implementing alterations only at key positions within the scene. Computer animation, an emerging form of mixed media art, empowers animating objects in real-world images and not just drawings, with diverse applications in the realms of entertainment, education, and training. The process of crafting a single frame of animated video from a static image is known as pose transfer, where the new pose is determined by a set of spatial landmarks, such as skeletal keypoints (Figure 1). It has been shown that other specialized guides, including depth, edge, and segmentation maps, can also function as a pose guide. However, the generation of these alternative maps necessitates substantial manual efforts, impeding their practical implementation in real-time systems.

Pose transfer typically involves two distinct tasks: (i) relocating those regions of the input image that are visible in both the source and target poses. To do so, we simply displace thses regions to their respective positions in the target pose, and (ii)

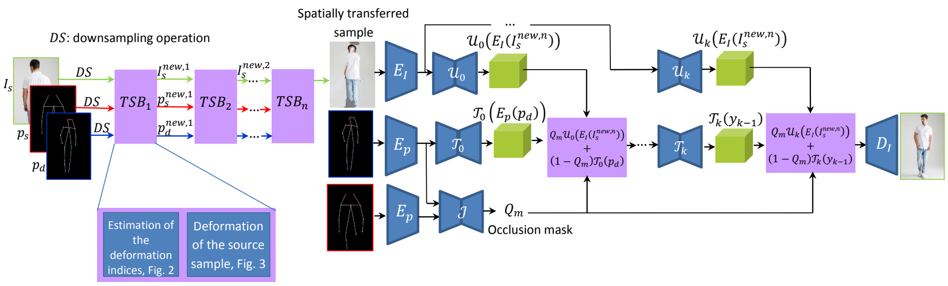

DS: downsampling operation
I_s
p_s
p_d
DS
DS
DS
TSB_1
$I_s^{new,1}$
$p_s^{new,1}$
$p_d^{new,1}$
TSB_2
$I_s^{new,2}$
TSB_n
Spatially transferred sample
E_I
E_p
E_p
$\mathcal{U}_0$
$\mathcal{T}_0$
$\mathcal{J}$
$\mathcal{U}_0(E_I(I_s^{new,n}))$
$\mathcal{T}_0(E_p(p_d))$
$Q_m\mathcal{U}_0(E_I(I_s^{new,n})) + (1-Q_m)\mathcal{T}_0(p_d)$
$\mathcal{U}_k$
$\mathcal{T}_k$
$\mathcal{U}_k(E_I(I_s^{new,n}))$
$\mathcal{T}_k(y_{k-1})$
$Q_m\mathcal{U}_k(E_I(I_s^{new,n})) + (1-Q_m)\mathcal{T}_k(y_{k-1})$
D_I
Q_m
Occlusion mask
Estimation of the deformation indices, Fig. 2
Deformation of the source sample, Fig. 3

Figure 1: The overarching structure of our approach involves taking an image, its corresponding skeletal pose, and a driving pose as inputs. Following a series of Texture Selection Blocks, the image undergoes deformation to align with the pose depicted in the driving map. Subsequently, the deformed image, along with the source pose and the driving pose, is processed through a generator. This generator encompasses the encoders E_I and E_p, in addition to a collection of combination blocks ($\mathcal{U}i$ and $\mathcal{T}i$).

inpaiting those parts that are visible in the target pose but not in the source pose. The primary challenge in pose transfer networks lies in accurately displacing each visible region to its corresponding part, a task referred to as spatial transformation. Current methodologies for implementing spatial transformations in deep neural networks include affine-based image warping and Transformers. However, affine-based warping faces significant challenges in generating complex poses, primarily due to the simplicity of its affine transformation. Moreover, retrieving image patches becomes impractical after passing through a transformer[1]. Consequently, this results in a sub-optimal transformation with no discernible similarity between the patches of the source sample and their estimations in the target pose.

In this paper, we propose a novel pose transfer network that leverages a set of non-affine deformation blocks as opposed to the commonly employed affine transformation. Our warping scheme capitalizes on correlation layers, wherein each layer computes a cost volume of matches between the source and target pose of a given sample. This

[1] In transformers, each patch of the output sample is a combination of all the pixels in the source image.

approach ensures coverage of all potential displacements for each image patch. Next, we propose to minimize the energy of the second derivative for the estimated warping map across all pixels which ensures to keep the locality of textures after the warping. Subsequently, like the affine-based strategies we employ an image matching task as a measure for estimating the parameters of our pixel displacement blocks.

Additionally, our approach leverages a gradual estimation of warping maps instead of relying on a single-shot estimation. This choice is made due to the inherent limitation of skeletal poses, which lack sufficient information about the body volume in a target pose.

The overall structure of our method is illustrated in Figure 1. It incorporates a series of Texture Selection Blocks (TSBs), each responsible for a localized transformation on the coordinates of the source pixels. The sequential application of these local blocks enables a more versatile transformation, enhancing the model's ability to learn a broad spectrum of displacements without confusion between small and large displacements. Each TSB takes multiple inputs, including the source image, source pose, and target pose. Subsequently, a local transformation is derived from this input set and applied to the source sample, resulting in a transferred image that serves as an estimate of the source sample in the target pose. This estimation includes the body volume in the target pose (p_d), providing a novel representation of the source sample that incorporates the body volume rather than relying solely on the skeletal representation of p_s. In summary, the output of each TSB consists of a novel estimation of the source image, along with updated source and target poses. Following the application of n TSBs, the source image is entirely displaced to the target pose. This displaced image is then fed into a generator to incorporate regions that are not visible in the source sample but are introduced in the target pose.

The effectiveness of our method is verified through a set of extensive experiments across two different applications, including a pose transfer and a novel view synthesis task. This way, we demonstrate that our network is not an application-oriented framework specifically designed for a single task. This evaluates the ability of our method both to add extra details on the same view of a scene and also to generate the scene

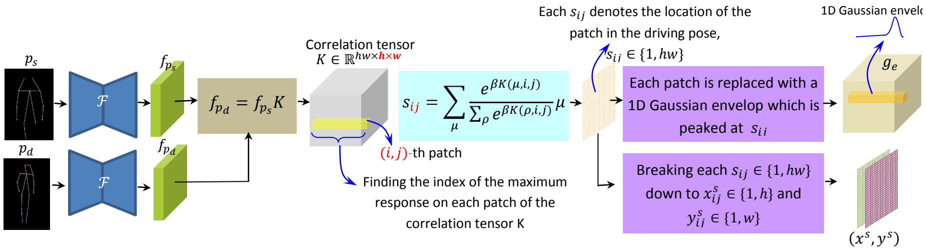

p_s
p_d
$\mathcal{F}$
$\mathcal{F}$
f_{p_s}
f_{p_d}
$f_{p_d} = f_{p_s}K$
Correlation tensor
$K \in \mathbb{R}^{hw \times h \times w}$
(i,j)-th patch
Finding the index of the maximum response on each patch of the correlation tensor K
$s_{ij} = \sum_{\mu} \frac{e^{\beta K(\mu,i,j)}}{\sum_{\rho} e^{\beta K(\rho,i,j)}} \mu$
Each s_{ij} denotes the location of the patch in the driving pose,
$s_{ii} \in \{1, hw\}$
Each patch is replaced with a 1D Gaussian envelop which is peaked at s_{ii}
Breaking each $s_{ij} \in \{1, hw\}$ down to $x^s_{ij} \in \{1, h\}$ and $y^s_{ii} \in \{1, w\}$
1D Gaussian envel
g_e
(x^s, y^s)

Figure 2: Estimation of deformation indices using a correlation layer. The correlation tensor K is headed by a softargmax operation which ensures a one-to-one correspondence between the patches of the target and the patches of the source samples. Two sets of deformation indices (g_e and (x_s, ys)) are predicted by our method.

from a novel viewpoint. However, other competing pose transfer strategies can not be applied to a novel view synthesis task.

Our contributions can be outlined as follows:

- We introduce a novel approach that spatially displaces image pixels to align with their target positions. In contrast to affine-based transformations, our method avoids applying a uniform spatial operation to all the image pixels.
- We introduce a softargmax-based operation for aligning the corresponding parts of two skeletal images. In contrast to widely used softmax-based operations, our approach relies solely on the positional information of the matching points, without considering their specific values.
- We suggest a hierarchical warping map estimation, enabling the incorporation of the target body volume in the warping map estimation process.

2. Related work

Pose transfer: Deep pose transfer was initially introduced by Ma et al. Ma et al. (2017). In their approach, a primary network offers a comprehensive estimation of the image in the target pose, while a secondary network supplements this estimate with image details. Despite yielding promising results, the absence of mutual influence between these networks leads to suboptimal texture generation. Several methods, such as Li et al. (2019); Ren et al. (2020); Tabejamaat et al. (2021), incorporate a flow estimator predicting clothing textures at the target pose location, followed by a network that adds the human body details onto these textures. Alternatively, approaches like Tang et al. (2020b); Esser et al. (2018) propose separate streams within a unified network to estimate clothing texture and body appearance. This design allows for the inclusion of texture-appearance correlation in the image generation task. For example, Tang et al. Tang et al. (2020b) proposed a dual-stream network that generates the pose and appearance of samples using a Siamese network. To increase the diversity of the generated samples, Esser et al. Esser et al. (2018) proposed to combine these appearance and pose features in the latent space of a variational autoencoder Kingma & Welling (2013). Pose Attention Transfer Network (PATN) Zhu et al. (2019), Patch Transfer (PT) Tabejamaat et al. (2021), and CoCosNet Zhou et al. (2021) employ an attention mechanism Bahdanau et al. (2014) for the fusion task, leading to enhanced performances. Ren et al. Ren et al. (2022) introduce a double attention approach to extract semantic entities from the source image, using each as a dictionary element for target sample generation. Additionally, Roy et al. Roy et al. (2023) incorporate multi-scale attention to enhance the image quality in a pose transfer scenario.

Recent research predominantly leverages additional human parsing to enhance the quality of semantic image generation. In this vein, PISE Zhang et al. (2021), SPGnet Cheng et al. (2019), CASD Zhou et al. (2022), and ADGAN Men et al. (2020) propose incorporating semantic segmentation of samples as an additional input for pose transfer models. This approach simplifies the matching operation between corresponding areas of the source and target samples. However, it necessitates the estimation of semantic segmentation for the target sample based on a sparse set of skeletal keypoints, presenting an under-constrained problem. Furthermore, pixel annotation is required for supervision, posing challenges, particularly for more intricate garments.

Additionally, Albahar et al. AlBahar et al. (2021), Neverova et al. Neverova et al. (2018), and Chang et al. Chen et al. (2022) advocate for the estimation of a dense UV map for individual input samples. This approach facilitates a precise transformation of clothing items to their target positions. However, the challenge lies in estimating a comprehensive dense UV map that encompasses all body parts and clothing items, making it notably more intricate than estimating a semantic segmentation map in the target pose. Consequently, this requires the utilization of large networks that consume substantial amounts of memory.

Recently, diffusion models Sohl-Dickstein et al. (2015) have achieved significant interest due to their capacity to synthesize high-fidelity images Rombach et al. (2022). In contrast to Generative Adversarial Networks (GANs) Goodfellow et al. (2014), these models excel at generating uncommon textures, a valuable trait in tasks like pose transfer, where clothing textures may be infrequent in the training data. Bhunia et al. Bhunia et al. (2023) introduced a pose-guided diffusion model specifically designed for the pose transfer task.

Novel view synthesis: Novel View Synthesis involves estimating the appearance of a scene from a novel viewpoint. The process can be supervised by a variety of guiding points including 3D points Huang et al. (2023); Zhang et al. (2023) and camera angle Wiles et al. (2020). InfoGAN Chen et al. (2016) suggests manipulating the latent space of an image's view. Kulkarni et al. Kulkarni et al. (2015) introduce a unique approach that learns a transformation by averaging the latent space values when applying a specific transformation to various training samples. Park et al. Park et al. (2017) present a disocclusion-aware appearance flow network, which learns a visibility map and a rotation matrix to shift visible parts of an image to corresponding positions in another view of that image. Sun et al. Sun et al. (2018) adopt a similar strategy but utilize a set of viewpoints instead of a single view from a source sample, enabling a more accurate estimation of the flow map. They also propose replacing the rotation matrix with an unsupervised flow estimation network, closely resembling the approach in Park et al. (2017). In Tatarchenko et al. (2016), a set of depth maps is predicted from a given image, each corresponding to a specific viewpoint. These maps are then utilized to estimate a point cloud of the object, which is subsequently rendered from the target viewpoint of the sample. Similar to Kulkarni et al. (2015), Worrall et al. Worrall et al. (2017) proposed applying a 3D transformation to the latent space of an autoencoder network. Sitzmann et al. Sitzmann et al. (2019) suggested mapping arbitrary points in world coordinates to a feature representation and then employing a ray-marching LSTM and a convolutional pixel generator to render this feature representation from a novel viewpoint. Notably, these approaches all rely on synthetic datasets containing single-object images captured at various view angles.

Deep warping estimation: In recent years, the application of deep neural networks for estimating the flow function between two signals or images has gained attention. For instance, Kazlauskait et al. Kazlauskaite et al. (2019) introduced a method based on a probabilistic model constructed with nonparametric priors, offering general estimates for the matching of two signals. Another approach, Deep Canonical Time Warping (DCTW) Trigeorgis et al. (2017), focuses on estimating the temporal alignment of time series within a common subspace. Oh et al. Oh et al. (2018) utilized sequence transformers to learn functions for stretching, compressing, rotating, and/or transforming signals to fit clinical time series data. For the pose transfer task, the flow function is computed between two sets of skeletal joints and then applied to the source sample to shift it to the position of the target pose. Early work on this was introduced by Siarohin et al Siarohin et al. (2018). The authors suggested estimating an affine transformation for each subset of skeletal joints and predicting the overall flow function through softmax voting among these local transformations. In Siarohin et al. (2019), the authors presented a first-order warping strategy closely linked to Siarohin et al. (2018). Meanwhile, in Zhao & Zhang (2022), the authors propose using a thin plate spline for flow estimation between two RGB images, which has demonstrated greater effectiveness than the first-order modeling of motions. However, it involves matrix inversion, posing challenges with certain complex motions.

Similarities and Deviations: Our approach is a pose transfer network, distinguishing itself from existing strategies by introducing a constrained warping maps rather than the widely used affine transformations. This feature ensures accurate deformation even for complex deformations between the source and target poses of the samples. Similar to previous approaches, we conduct our evaluations using the established Deepfashion database.

3. Method

Given a source image I_s along with a source pose p_s and a driving pose p_d, we aim to transfer the source image (Figure 4(a)) into the novel pose of p_d (Figure 4(b)). Each pose is a volumetric stack of 2D heatmaps, where each heatmap is a 2D Gaussian envelope centered at the location of a skeletal keypoint.

Our method consists of two different modules; (a) spatial transformation and (b) generator. This is illustrated in Figure 1. TSBs indicate the spatial transformation, outputting a Spatially transferred image. This image along with the source and the target pose is given to the generator to add the photorealistic details on top of the spatially transferred image. The network in Figure 2 is the lower stream of the network in Figure 1 (overall network) and Figure 3 depicts the strategy by which we use the two outputs of Figure 2 to spatially sample the pixels of the input image. The Spatial Transform learns to displace the patches that are visible in both the source pose and the driving pose of the sample (Figure 4(c)). Given these transformed patches, our generator learns to generate the remaining patches that are invisible in the source sample but newly introduced in the driving pose (Figure 4(d)).

3.1. Spatial transformation

This module includes a cascade of Texture Selection Blocks (TSBs). Each block takes a collection of the source images, the source pose, and the target pose as input and makes an update in their representation. The update has two different aspects: (a) displacing the patches of the source sample according to the driving pose, and (b) incorporating the body volume of the sample in the source pose and also in the driving pose.

We first compute a cost volume of matches between the patches of the source pose p_s and the patches in the driving pose p_d. This volume is then used to assign a positional code to each patch of the driving pose. To benefit from a more expressive function, we first project the pose maps to an intermediate feature space.

$$\begin{aligned} p_s \in R^{m\times h\times w} \xrightarrow{\mathcal{F}} f_{p_s} \in R^{c\times h\times w} \\ p_d \in R^{m\times h\times w} \xrightarrow{\mathcal{F}} f_{p_d} \in R^{c\times h\times w} \end{aligned} \quad (1)$$

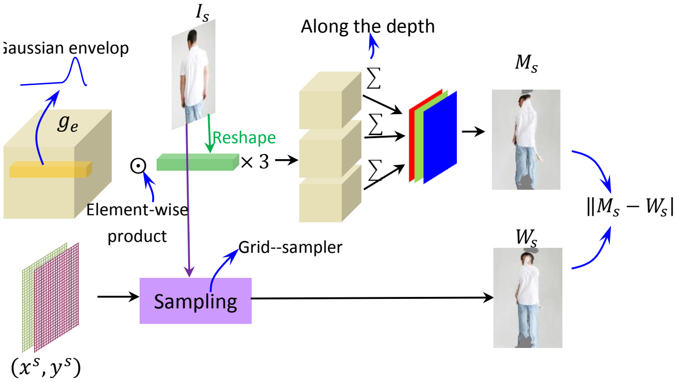


Figure 3: We consider two spatial deformations of the input sample and try to minimize their difference using an optimization process. The sampling block at the bottom represents a differentiable sampling operator proposed by Jaderberg et al. (2015). In the top row, we consider a direct sampling operation based on the multiplication of the spike vector, denoted by the softtargmax operation, in the pixels of the image.

$\mathcal{F}$ is the same for both the maps p_s and p_d. To capture a one-to-many relation, we assume each patch of the driving pose f_{p_d} is a linear combination of the entire patches in the source pose f_{p_s}. This is formulated as: $f_{p_d}^{c \times hw} = f_{p_s}^{c \times hw} K^{hw \times hw}$. The superscripts denote the dimension of the matrices. Accordingly, $K \in R^{hw \times h \times w}$ is a correlation tensor whose channels represent the similarity of the driving patches to the patches of the source pose. Then, we estimate the location of the maximum response in

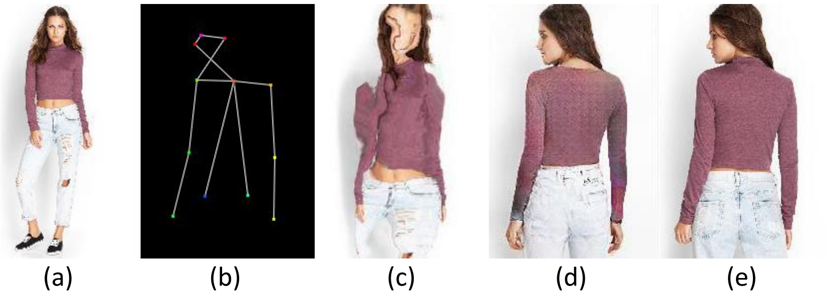


Figure 4: The deformation and the final result of our strategy, (a) source sample, (b) target pose, (c) deformation result, (d) final result, (e) ground-truth.

each spatial location of $K \in R^{hw \times i \times j}$, which is computed by the softargmax operation (Figure 2).

$$s_{i,j} = \sum_{\mu} \frac{e^{\beta K(\mu,i,j)}}{\sum_{\rho} e^{\beta K(\rho,i,j)}} \mu \tag{2}$$

where $s(i, j)$ is the spatial index of the maximum response, indicating the index of the source patch which is assigned to (i, j)-th patch in the driving pose. In practice, we consider $s_{i,j}$ as the flow value and add it to the coordinate index (x_s, y_s) to obtain the novel location of the pixel (s_x, s_y) after the spatial transformation.

$$\begin{aligned} s_x^{'} &= s_i + x_s[i] \\ s_y^{'} &= s_j + y_s[j] \end{aligned} \tag{3}$$

Then, we directly sample the source image according to the sampling coordinates $(s_x^{'}, s_y^{'})$, $\mathcal{S}$ is considered as the sampling operation (Figure 3):

$$W_s(i, j) = \mathcal{S}((s_x^{'}, s_y^{'}), I_s) \tag{4}$$

We also consider a second form of sampling operation which is directly applied to the source images. This process is completely inspired by the attention mechanism in Transformers, considered to stabilize the optimization of the warping operations. For this purpose, we first create a Gaussian envelope for each of the patches in the target pose, (i, j). The length of this one-dimensional envelope is equal to hw. The Gaussian envelope denotes a Gaussian vector peaked at the location of $s^{'}(i, j)$. This vector already denotes the location of the maximum similarity between the ith patch of the source pose and the jth patch of the target pose.

$$g_e(i, j) = e^{\zeta(\mu - s(i,j))^2} \tag{5}$$

where $g_e(i, j) \in R^{hw}$. Having this envelope created for the entire locations $i, j \in \{(1, h), (1, w)\}$, we have a tensor of dimension $hw \times h \times w \times 1$. Then for sampling, we simply rescale the source image to the dimension of $hw \times 1 \times 1 \times 3$ and multiply it by this Gaussian envelope. This operation is individually performed on each spatial location (i, j), where it samples the corresponding location of the source image. This sampling operation is formulated as follows:

$$M_s(i, j) = \sum_{\mu} g_e(\mu, i, j) I_s(\mu) \tag{6}$$

where $I_s \in R^{hw}$. $M_s \in R^{3\times h\times w}$ is an updated version of the source sample which is transformed into the target pose. In Section 3.3, we encourage M_s and W_s to be exactly the same. This allows for reducing the instability issue of the differentiable sampling introduced by Jaderberg et al. (2015), because it does not allow the sampling values to go beyond the spatial size of images.

Next, we propose to update the pose maps, p_s and p_d in a way that $p_{d]}$ includes some information about the body silhouette in the target pose. This aims to include two different characteristics: (1) similar to p_s and p_d, the new pose maps need to distinguish between different parts of the human body, (2) the maps need to differentiate between the background and the silhouette of the samples (as an estimation of the body volume). The first condition is already satisfied by f_{p_s} and f_{p_d}, as their characteristics are directly extracted from the skeletal keypoints. For the second condition, we have M_s as an estimation of the source sample in the target pose, because it is already the source image spatially warped to the target pose. Therefore, we simply have $I_s^{new} = M$. Given that, it is clear that a function[2] of this body estimate can represent the silhouette of the body volume in the target pose. Accordingly, applying the same function on the source sample provides an estimation of the body silhouette in the source pose.

$$z_s = \mathcal{P}(I_s), \quad z_d = \mathcal{P}(M_s) \tag{7}$$

where z_s and z_d are the estimation of the silhouette respectively in the source and also in the target pose. $\mathcal{P}$ is a convolutional function which is headed by a sigmoid operation.

Finally, multiplying these silhouettes (z_s and z_d) by f_{p_s} and f_{p_d} provides us with the novel estimation of the source and the driving pose maps:

$$p_s^{new} = z_s * f_{p_s}, \qquad p_d^{new} = z_d * f_{p_d} \tag{8}$$

After n TSBs, the source patches are spatially transferred to their corresponding locations in the target pose.

[2] like a convolutional network

3.2. Image generation

Given the transferred image $I_s^{new,n}$, the source pose I_s and the driving pose I_d, this module learns to generate a photorelistic image of the source sample in the driving pose. To accomplish this, we first project $I_s^{new,n}$ and I_d to a feature space, then combine these two features using the following equation:

$$y = Q_m * \mathcal{U}(E_I(I_s^{new,n})) + (1 - Q_m) * \mathcal{T}(E_p(I_d)) \tag{9}$$

where $\mathcal{U}$ and $\mathcal{T}$ are two convolutional networks which transfer the $I_s^{new,n}$ and I_d to a feature space. $I_s^{new,n}$ is the deformed sample of the n-th TSB. Q_m is an occlusion matrix indicating which patches are visible in both the source and the target pose of the sample and which patches are newly introduced in the target pose. In practice, Q_m is roughly estimated from the feature maps of p_s and p_d. The estimation is performed using the function J on a concatenation of these feature maps (Figure 1).

To benefit from a progressive function, we propose the combination of $E_I(I_s^{new,n})$ and $E_p(I_d)$ to be performed using a cascade of functions $\mathcal{U}_i$ and $\mathcal{T}_i,\ \ i = 1, ..., k$:

$$\begin{aligned} y_0 &= Q_m * \mathcal{U}_0(E_I(I_s^{new,n})) + (1 - Q_m) * \mathcal{T}_0(E_p(I_d)) \\ y_i &= Q_m * \mathcal{U}_i(E_I(I_s^{new,n})) + (1 - Q_m) * \mathcal{T}_i(y_{i-1}) \end{aligned} \tag{10}$$

y_k is then passed to a decoder to generate the output of our model, which is a photorealistic image of the source sample in the target pose.

3.3. Learning

Our model is trained in a supervised manner in which two images with the same identity and garments are utilized as the source and the target samples. Moreover, the skeletal keypoints of the target sample is utilized as the driving pose. For training, we benefit from three loss functions, a perceptual loss, an adversarial loss, and a reconstruction loss. For the perceptual loss, we use a pretrained VGG19 model and minimize the distance between y_k and the target sample I_t in its intermediate layers.

$$L_{vgg} = \mathbb{E}\big[\sum_{r}^{R} \|\mathcal{V}_r(y_N) - \mathcal{V}_r(I_t)\|_1\big] \tag{11}$$

where, $\mathcal{V}_r$ stands for the r-th layer of VGG19. Moreover, we consider a reconstruction loss which ensures a holistic similarity between the generated image and the ground-truth data:

$$L_{l1} = \mathbb{E}\big[\|y_k - I_t\|_1\big] \tag{12}$$

Further, we consider an adversarial loss which ensures the photo-realism of the generated samples:

$$L_{GAN} = \mathbb{E}\big[log(1 - \mathcal{D}(y_k))\big] + \mathbb{E}\big[\mathcal{D}(I_t)\big] \tag{13}$$

where, $\mathcal{D}$ is a fully convolutional discriminator. Moreover, we also benefit from the l_1 and the perceptual functions to minimize the distance between the deformed W_s and the target sample. Additionally, we minimize the l_1 loss between the two estimations of the deformed sample (W_s and M_s).

$$\begin{aligned} L_{def} =& \mathbb{E}\big[\sum_{r}^{R} \|\mathcal{V}_r(W_s) - \mathcal{V}_r(I_t)\|_1\big] + \lambda_0 \mathbb{E}\big[\|W_s - I_t\|_1\big] \\ &+ \lambda_1 \mathbb{E}\big[\|W_s - M_s\|_1\big] \end{aligned} \tag{14}$$

It is noteworthy that the deformation part is quite disentangled from the generator of our model. This way, there are no mutual dependencies between L_{def} and L_{Gen} of our model which ensures a faster convergence for each of these modules.

To ensure a smooth transition, we minimize the second derivatives of the angles between s_i, s_j and x, y-axes.

$$L_{smooth} = \mathbb{E}\big[\frac{\partial^2 \theta(s_i, s_j)}{\partial x^2}\big] + \mathbb{E}\big[\frac{\partial^2 \phi(s_i, s_j)}{\partial y^2}\big] \tag{15}$$

where $\theta(s_i, s_j)$ is the angle between (s_i, s_j) and the x-axes, $\phi(s_i, s_j)$ is the angle between (s_i, s_j) and the y-axes, and $\frac{\partial^2 f}{\partial x^2}$ is the partial second order derivative with respect to x.

3.4. Inference

For inference, we require a source image I_s along with its skeletal keypoints p_s and also a set of novel keypints as the driving pose p_d. The skeletal keypoints of the source sample can be easily extracted from the source image using a pose estimation network. However, the driving pose can be manually provided by a human observer, making the strategy a user-friendly framework in applications like image animation.

4. Experiments

The experiments encompass two distinct tasks: pose transfer and view synthesis, conducted on the Deepfashion and Phototourism datasets, respectively. The evaluations are comprehensive, relying on three widely recognized metrics: FID (Fréchet Inception Distance) Heusel et al. (2017), LPIPS (Learned Perceptual Image Patch Similarity) Zhang et al. (2018), and SSIM (Structural Similarity Index) Wang et al. (2004).

Both LPIPS and SSIM serve as pair-wise metrics, gauging the fidelity of the generated samples to their corresponding targets. They provide insights into the perceptual and structural quality of the generated images in relation to the ground truth. On the other hand, FID offers a comprehensive perspective by comparing the overall distributions of the generated and target samples. This metric provides a measure of photorealism, offering valuable insights into the global similarity between the generated samples and the ground truth targets. The combined use of these metrics ensures a thorough and nuanced evaluation of the model's performance across various dimensions.

4.1. Pose transfer

Deepfashion is a fashion show dataset Liu et al. (2016) with $101,966$ image pairs as training samples and 8570 pairs as test ones. The same split is used by Tang et al. (2020a); Ren et al. (2020); Zhu et al. (2019); Tang et al. (2019); Men et al. (2020); Zhang et al. (2021), allowing a fair comparison between our method and these competing algorithms. All the images are of the size 256×176, captured in an indoor scene against a white background. One challenging issue of this database is the small number of training samples per identity, making it difficult for the network to generalize over the variety of poses for a single texture. The skeletal joints are extracted by the HPE algorithm Cao et al. (2017). The joints are further represented as a volumetric heatmap. Each map is generated by a Gaussian envelope with $\sigma = 0.1$. The qualitative results of the competitors are generated by the pretrained models or directly from the images provided by the authors. For quantitative performance, the values are either reported from the papers or through evaluations of their generated samples.

Input image	Groundtruth	ADGAN	PISE	GFLA	DPTN	CASD	NTED	PIDM	Ours
Single iteration	Single iteration	Single iteration	Single iteration	Single iteration	Single iteration	Single iteration	Single iteration	50 iterations	Single iteration

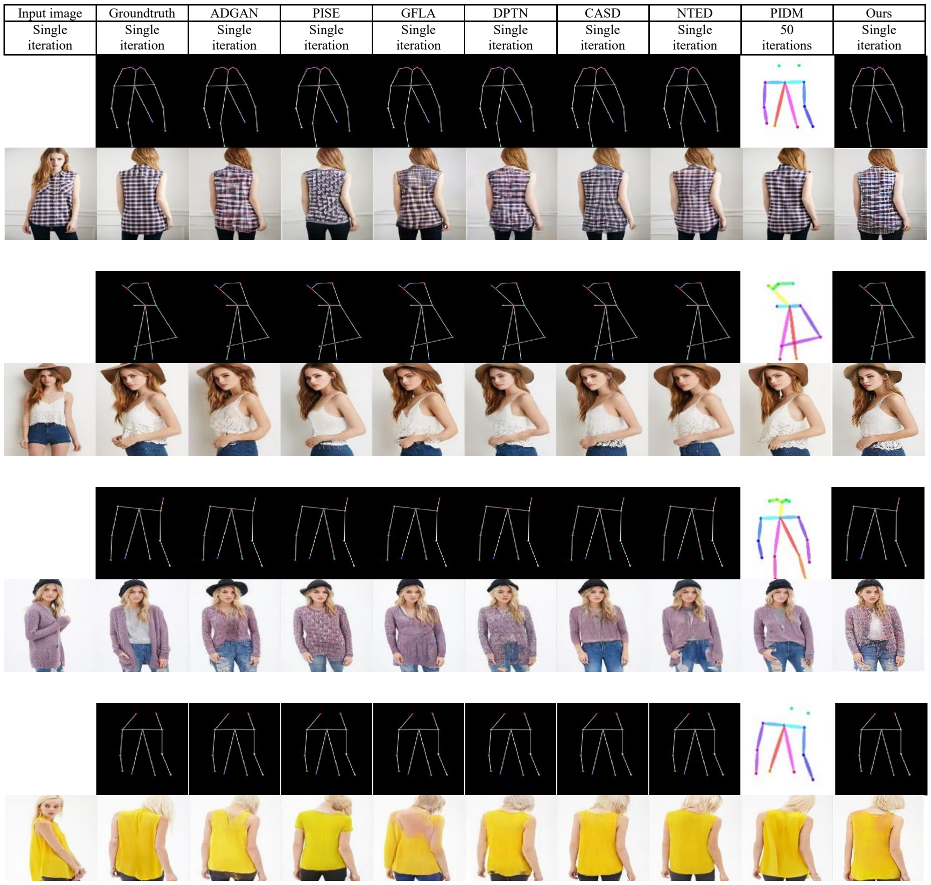


Figure 5: Illustrations of reconstructed samples from an alternate perspective are presented. The second row showcases the iteration number required to generate the image at test time. Each subsequent even row displays the respective guide map employed by each method. Notably, PIDM leverages the skeletal visualization of the target samples as its guide pose. However, a drawback emerges during inference, as it necessitates more than one iterations, limiting its suitability for real-time image animation. The competing images are provided by Bhunia et al. (2023)

4.1.1. Quantitative performance

In this section, we conduct a quantitative comparison between our method and several state-of-the-art algorithms, including ADGAN Men et al. (2020), PISEZhang et al. (2021), GFLARen et al. (2020), DPTNZhang et al. (2022), CASDZhou et al. (2022), NTEDRen et al. (2022), and PIDM Bhunia et al. (2023). The comparison results are detailed in Table 1. Notably, our method demonstrates competitive or superior performance when measured using both LPIPS and SSIM metrics, underscoring its effectiveness in generating semantically similar content to the ground-truth data. Benefiting from non-affine warping map estimation, our method effectively deals with complex transformations between different poses. Consequently, this ensures that the generative component receives more accurate warped images, ultimately leading to the generation of samples with textures that faithfully represent the original data.

It is crucial to emphasize that PIDM excels in generating high-quality images. However, it is noteworthy that the inference time of this model is considerably longer than other competing algorithms as it is a multi-iteration method based on the diffusion model. Even without any external segmentation map, our approach exhibits a remarkable 26.7% superiority over the average FID of single-shot strategies. Additionally, we observe a 15% and 42.9% improvement in performance on the LPIPS and SSIM metrics, respectively.

It's crucial to note that the decreased FID, compared to the diffusion model PIDM, primarily stems from challenges faced by our deformation module. This is particularly evident when handling infrequent poses absent in the training data. The deformation module's difficulty with these uncommon poses can lead to misguidance of the generator, presenting inaccurate information about the input sample's position and occasionally causing model failures.

4.1.2. Qualitative performance

In this section, we conduct a visual analysis of samples produced by our proposed method and several competing approaches, aiming to provide insights into visual accuracy, stylistic attributes, and overall performance. The results, depicted in Fig. 5, reveal that images generated by PISE, ADGAN, and GFLA lack consistent shapes in hats and garments. While GFLA benefits from warping, enabling high fidelity in texture related to the source image, it faces challenges generating satisfactory results for regions newly introduced in the target pose but invisible in the source image. NTED and CASD, leveraging cross-attention mechanisms for long pixel displacement, exhibit good performance but struggle to preserve precise details of source textures. PIDM generates high-fidelity samples but requires a substantial execution time due to multiple iterations during inference. In contrast, our method not only accurately preserves garment and hat shapes from the source image but also generates images with high texture fidelity. The visual results robustly support our method's capability in producing high-fidelity samples for intricate differences between the source and target pose, confirming its effectiveness in preserving exact color tones in the generated images.

	Strategy	Single iteration	FID↓	LPIPS↓	SSIM↑
Our method	-	✓	8.5	0.17	0.729
ADGANMen et al. (2020)	Segmentation map	✓	14.4	0.22	0.672
PISEZhang et al. (2021)	Segmentation map	✓	13.6	0.20	0.662
GFLARen et al. (2020)	Affine-based warping	✓	10.5	0.23	0.707
DPTNZhang et al. (2022)	Cross-attention	✓	11.3	0.19	0.711
CASDZhou et al. (2022)	Cross-attention	✓	11.3	0.19	0.724
NTEDRen et al. (2022)	Cross-attention	✓	8.6	0.17	0.718
PIDM Bhunia et al. (2023)	Diffusion model	✗	6.3	0.16	0.731

Table 1: An evaluation of various pose transfer methods on the Deepfashion database.

4.1.3. Shift invariant pose transfer

Shift invariance is one of the critical advantages of a pose transfer network, especially in practical applications like image animation. In this case, projecting different views of a source sample to the same target pose should generate the same result. This causes integrity over the whole space of the image generation. To ensure this property, we propose to augment the source samples using a random affine transformation and then encourage the network to generate the same results as before the transformation. We experiment to evaluate the effectiveness of this strategy on the shift-invariance of the generated samples. For evaluation, we first make a random shift in all the driving

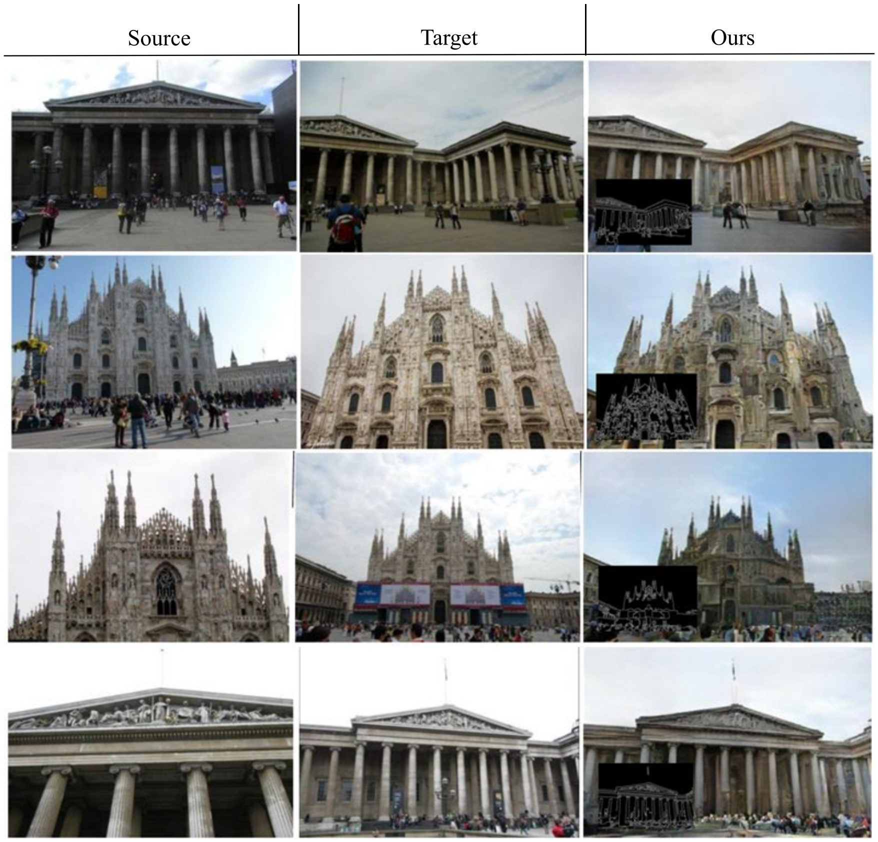


Figure 6: Examples of the reconstructed samples of monuments given a target pose.

keypoints of the target samples. All the points are shifted by the same value but it randomly changes for different samples.

The results are shown in Table 2. As can be seen, our method outperforms NTED in handling the shift of the samples. It is evident that the augmentation is quite effective in generating the same result from different views of a sample. Compared to NTED, our method benefits from a spatial deformation module, whose output is further used as the input of our generator. This makes it necessary for our method to use a huge number of training samples to learn about the correct deformation of samples. Therefore, compared to NTED, the augmentation technique has a greater influence on our strategy, leading to a better performance than the original technique.

	Our method	NTED
Vanilla model	0.729	0.718
Shift(up to 20 pixel)	0.714	0.702
Aug+shift (up to 20 pixel)	0.719	0.709

Table 2: Evaluation of the shift invariance based on the SSIM score

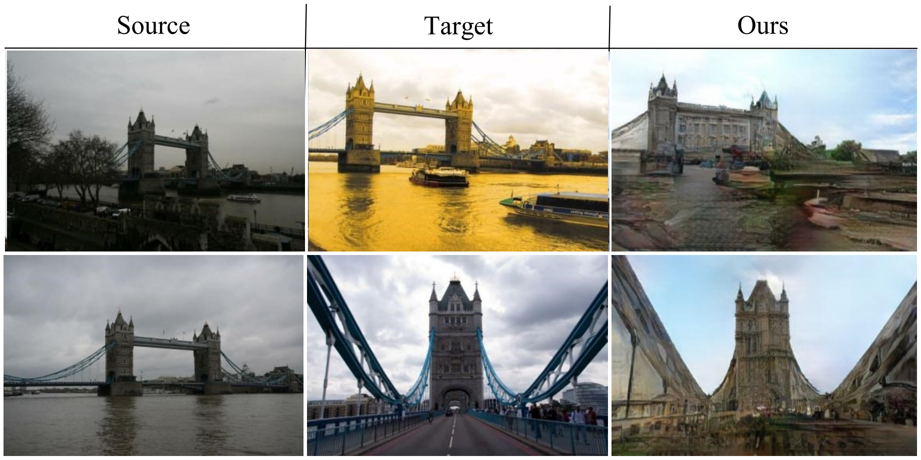


Figure 7: Examples of the reconstructed samples where the network fails to generate proper output as the network uses an edge drawing map in the absence of guiding keypoints map.

4.2. Ablation Study

In this section, we evaluate the effectiveness of our proposed strategy in displacing the image patches to their correct locations in the target pose. To do so, we first replace each of the TSB blocks with a Vision Transformer block implemented by a kernel size of 16×16. The performance of this architecture is evaluated on the same split of the Deepfashion database that we introduced in Section 4.1. The results are shown in Table 3. As can be seen, our proposed method is quite superior to the Transformer-based model. This results from the linear combination of transformer blocks, whereby each patch of an output sample is a combination of all the patches in the input image. This makes it almost impossible to retrieve the original patches from these combined textures.

Moreover, we conducted another experiment to evaluate the effectiveness of each term in our collections of the loss functions. Accordingly, we consider a baseline model comprising of l_1, L_{smooth}, and perceptual loss functions for the deformation module and l_1 loss for the image generation module. Then, we progressively add the remaining terms to this baseline model. For the sake of notational simplicity, we consider the following abbreviations: A: baseline model, B: GAN+baseline model, C:GAN loss+perceptual+baseline. The results are listed in Table 4. By comparing A and B, it is clear that the GAN loss has a significant impact on the FID score, more precisely 1.9 lower than the FID score of the baseline model. In contrast, by comparing A and C, we can observe the effectiveness of the perceptual loss on the SSIM score. L_{smooth} is the only condition to guarantee the uniformity of our spatial transformation and without it, the network fails to generate a meaningful deformation image for our image generator. For this reason, we included L_{smooth} in all of the ablation studies.

	FID↓	SSIM↑
Our method w TSBs	8.5	0.729
Our method w/o TSBs+VITs	9.1	0.715

Table 3: An ablation study for evaluating the effectiveness of TSBs in our proposed method

	FID↓	SSIM↑
A: baseline model	11.4	0.692
B: GAN+baseline model	9.2	0.703
C: GAN loss+perceptual+baseline	8.5	0.729

Table 4: An ablation study for evaluating each term of loss functions in our method

4.3. Novel view Synthesis

Unlike the pose generation, novel view synthesis aims to regenerate a novel view of a stationary scene. The regeneration is guided by an edge map of samples in the target view. In practice, these maps are extracted from the target samples by using the

Canny edge detection algorithm Canny (1986) (lower left side of the generated images in Figure 6). The thresholds of the Canny operator are set to 10 and 200. This approximates a simple drawing of the pose by a human user which is applicable in an image animation task. The experiments are conducted on the Phototourism dataset which contains more than $25,000$ training samples from 15 monuments and more than 4000 test samples from 9 different monuments. For evaluation, we consider $100,000$ training pairs randomly selected from the training samples. This allows for a reasonable training time. Additionally, 2014 samples are paired as our test samples. Figure 6 shows some examples reconstructed by our method. As can be seen, our method can well reconstruct the texture of the samples. The invisible parts of the source samples are correctly reconstructed in their positions and correspond to the texture of the visible parts of the sample. However, sometimes the color tone of the generated image is not exactly matched with that of the source sample. This is more related to the diversity of the color tones between the images that we paired in our database. This encourages the network to learn a different color tone for the target pose of an image. However, collecting outdoor images with the same color tone is not an easy task. This requires further research on the color correction of the images. Figure 7 illustrates some examples where our method fails to keep the fidelity of the generated images to the source samples in the absence of guiding keypoint map. In practice, guided view synthesis is a very challenging task as there is no consistency between the number of guiding points in different samples. As can be seen, our method has also some difficulties with generating some parts of the images including the bridges and spaces with different textures to their neighboring areas.

4.4. User study

Similar to Zhu et al. (2019), we asked 30 volunteers to provide us with their opinions about the quality of the generated samples. The experiment is conducted over 116 images randomly selected from the test pairs of the Deepfashion database. Each target sample is generated by 7 different algorithms. We did a warm-up session with 17 additional images so that the participant became familiar with the experimental protocol. Table 5 lists the results of our questionnaire which contains two different questions: (1)

which algorithm best preserves the fine details of the source sample? (2) how well does the generated pose match the ground-truth data? As can be seen, our method provides a competitive performance with PISE Zhang et al. (2021) and GFLA Ren et al. (2020). PISE takes advantage of the segmentation maps which allows for a more accurate estimation of the target pose. However, the need for a segmentation map hinders the application of this strategy in some real-world scenarios like image animation, where it is more convenient for a human to provide the network with the landmarks of the samples[3] rather than a dense segmentation map.

	Q1: which algorithm best ...	Q2: how well does ...
Ours	26.1	89.9
NTED	25.3	90.1
PISE	22.1	92.7
ADGAN	7.3	90.1
GFLA	19.2	72.9

Table 5: User study on two different questions conducted on Deepfashion dataset

4.5. Implementation details

In this section, we provide a comprehensive overview of the implementation details employed in our pose transfer methodology. The outlined details encompass key aspects of our approach, including network architecture, training procedure, and specific techniques utilized to enhance the performance of the model.

4.5.1. Network Architecture

The architecture of our pose transfer network is crafted to handle the complex task of generating realistic and precise pose-transferred images. As a result, it incorporates a diverse array of architectures and layers. We leverage a combination of convolutional layers, attention mechanisms, and spatialy transformation layers carefully tailored to capture spatial dependencies and the semantic information crucial for successful pose transfer. E_I and E_a both utilize three convolutional layers with a kernel size of 7×7

[3] the source and target pose of the samples.

(illustrated in Figure 8(a) and 8(b)). The same E_p is used for both the source pose and the target pose. The architecture of each $\mathcal{U}_i$ and $\mathcal{T}_i$ is illustrated in Figure 8(c) and 8(d). For $\mathcal{F}$, we benefit from an autoencoder with 2 convolutional layers, each including one Resnet headed by a self-attention layer. For the self-attention, motivated by Bhunia et al. (2023) it is included to enable the model to attend to critical spatial and semantic cues during the pose transfer process. For $\mathcal{J}$ we use an 8-layer autoencoder, each layer is a convolutional operator with the kernel size 3×3 which is followed by a Batch normalization and ReLU operation. There is no skip connection between the encoder and decoder of the autoencoder. All the residual connections of our method are followed by a Batch normalization and a ReLU activation function. Our discriminator is a 6-layer convolutional neural network with the same architecture as the encoder of $\mathcal{J}$

4.5.2. Training Procedure

The training process assumes a pivotal role in ensuring the effectiveness of our pose transfer model. Utilizing the Adam optimizer Kingma & Ba (2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.99$, we conduct training for all Texture Selection Blocks, the generator, and the discriminator. Initially, a pretraining phase is executed on The Deepfashion database. During this phase, affine transformations are applied to each image, serving as inputs to the generator. The model attempts to regenerate a displaced version of the input sample, with shifts of up to 20 pixels both vertically and horizontally, randomly applied. A learning rate of 0.0001 is employed for the generator, and 0.00001 for the discriminator during this pretraining phase, spanning 250 epochs.

Subsequently, we fix the parameters of the generator and proceed to train the Texture Selection Blocks for 50 epochs, employing paired images and a learning rate of 1e-6. Subsequently, upon locking the parameters of the Texture Selection Blocks, we engage in a fine-tuning process for both the generator and discriminator parameters, once again utilizing paired images. During this fine-tuning phase, the learning rate for the generator is established at 1e-4, while for the discriminator, it is set to 1e-5, spanning 250 epochs. This multi-phase training approach is designed to ensure that the network attains resilient features and effectively generalizes across a diverse range of poses.

4.5.3. Correlation Layer

An important aspect of our implementation involves integrating a correlation layer, a feature that bolsters the model's capacity to accurately deform images, particularly in situations marked by substantial displacements between the source and target poses. We delve into the application and influence of the correlation layer in our methodology. The correlation is computed based on the intermediate feature vector, with a spatial size of 64×64. This computation involves evaluating cosine similarities between every pair of locations within the input samples. Subsequently, two convolutional layers, each with a kernel size of 7, are applied, each followed by an instance normalization layer.

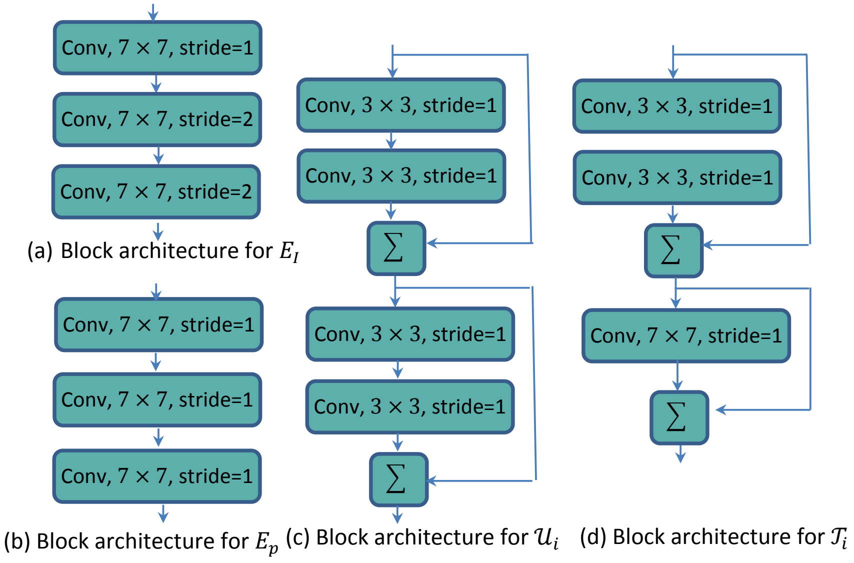
4.5.4. Computational Resources

We performed our experiments utilizing eight V100 GPUs with the PyTorch framework. These computational resources were chosen to facilitate effective training and evaluation of our pose transfer model. The complete training process spans a duration of 9 days within this framework.

5. Conclusion

In conclusion, our study introduces an innovative approach for estimating deformation maps in images through the integration of a correlation layer and minimization of the second derivatives of the warping maps. This method presents a notable departure from conventional warping-based strategies by overcoming limitations associated with kernel size while eliminating reliance on affine transformations. A key enhancement is the hierarchical computation of deformation indices, enabling a more expressive function for handling complex transformations in image samples. Our methodology undergoes rigorous evaluation across two distinct tasks: pose transfer and novel view synthesis. We conduct extensive experiments on well-established databases, including Deepfashion and Phototourism.

The outcomes consistently showcase the superiority of our proposed method in accurately generating pose variations based on deformation maps. The approach's effi-



(a) Block architecture for E_I

(b) Block architecture for E_p (c) Block architecture for $\mathcal{U}_i$ (d) Block architecture for $\mathcal{T}_i$

Figure 8: Block architecture of our method

cacy is notably demonstrated through its robust performance in both pose transfer and novel view synthesis tasks. These findings affirm the effectiveness of our methodology and its potential to contribute to advancements in the field of image generation and deformation mapping

Acknowledgements

The authors thank Amin Rahmaninezhad for his help with the datasets. We also would like to thank the National Research Agency (ANR) for the partial funding through the Investments in the Future project with the reference number ANR-19-P3IA-0002.

References

Al Ka'bi, A. (2023). Proposed artificial intelligence algorithm and deep learning techniques for development of higher education. *International Journal of Intelligent Networks*, *4*, 68–73.

AlBahar, B., Lu, J., Yang, J., Shu, Z., Shechtman, E., & Huang, J.-B. (2021). Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan. *ACM Transactions on Graphics (TOG)*, *40*, 1–11.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, .

Bhunia, A. K., Khan, S., Cholakkal, H., Anwer, R. M., Laaksonen, J., Shah, M., & Khan, F. S. (2023). Person image synthesis via denoising diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5968–5976).

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (pp. 679–698).

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7291–7299).

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, *29*.

Chen, Z., Yin, K., & Fidler, S. (2022). Auv-net: Learning aligned uv maps for texture transfer and synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1465–1474).

Cheng, B., Chen, L.-C., Wei, Y., Zhu, Y., Huang, Z., Xiong, J., Huang, T. S., Hwu, W.-M., & Shi, H. (2019). Spgnet: Semantic prediction guidance for scene parsing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 5218–5228).

Dhar, T., Dey, N., Borra, S., & Sherratt, R. S. (2023). Challenges of deep learning in medical image analysis—improving explainability and trust. *IEEE Transactions on Technology and Society*, *4*, 68–75.

Esser, P., Sutter, E., & Ommer, B. (2018). A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8857–8866).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

He, Z., Chen, Y., Zhang, H., & Zhang, D. (2023). Wkn-oc: a new deep learning method for anomaly detection in intelligent vehicles. *IEEE Transactions on Intelligent Vehicles*, *8*, 2162–2172.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, *30*.

Huang, S., Gojcic, Z., Wang, Z., Williams, F., Kasten, Y., Fidler, S., Schindler, K., & Litany, O. (2023). Neural lidar fields for novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 18236–18246).

Jaderberg, M., Simonyan, K., Zisserman, A. et al. (2015). Spatial transformer networks. *Advances in neural information processing systems*, *28*.

Kazlauskaite, I., Ek, C. H., & Campbell, N. (2019). Gaussian process latent variable alignment learning. In *The 22nd International Conference on Artificial Intelligence and Statistics* (pp. 748–757). PMLR.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, .

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

Kulkarni, T. D., Whitney, W. F., Kohli, P., & Tenenbaum, J. (2015). Deep convolutional inverse graphics network. *Advances in neural information processing systems*, *28*.

Li, Y., Huang, C., & Loy, C. C. (2019). Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3693–3702).

Liu, Y., & Zhang, J. (2024). Service function chain embedding meets machine learning: Deep reinforcement learning approach. *IEEE Transactions on Network and Service Management*, .

Liu, Z., Luo, P., Qiu, S., Wang, X., & Tang, X. (2016). Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1096–1104).

Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., & Van Gool, L. (2017). Pose guided person image generation. *Advances in neural information processing systems*, *30*.

Men, Y., Mao, Y., Jiang, Y., Ma, W.-Y., & Lian, Z. (2020). Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5084–5093).

Mourot, L., Hoyet, L., Le Clerc, F., Schnitzler, F., & Hellier, P. (2022). A survey on deep learning for skeleton-based human animation. In *Computer Graphics Forum* (pp. 122–157). Wiley Online Library volume 41.

Neverova, N., Guler, R. A., & Kokkinos, I. (2018). Dense pose transfer. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 123–138).

Oh, J., Wang, J., & Wiens, J. (2018). Learning to exploit invariances in clinical time-series data using sequence transformer networks. In *Machine learning for healthcare conference* (pp. 332–347). PMLR.

Park, E., Yang, J., Yumer, E., Ceylan, D., & Berg, A. C. (2017). Transformation-grounded image generation network for novel 3d view synthesis. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3500–3509).

Pham, H. X., Cheung, S., & Pavlovic, V. (2017). Speech-driven 3d facial animation with implicit emotional awareness: A deep learning approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 80–88).

Ren, Y., Fan, X., Li, G., Liu, S., & Li, T. H. (2022). Neural texture extraction and distribution for controllable person image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13535–13544).

Ren, Y., Yu, X., Chen, J., Li, T. H., & Li, G. (2020). Deep image spatial transformation for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7690–7699).

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10684–10695).

Roy, P., Bhattacharya, S., Ghosh, S., & Pal, U. (2023). Multi-scale attention guided pose transfer. *Pattern Recognition*, *137*, 109315.

Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., & Sebe, N. (2019). First order motion model for image animation. *Advances in Neural Information Processing Systems*, *32*, 7137–7147.

Siarohin, A., Sangineto, E., Lathuiliere, S., & Sebe, N. (2018). Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3408–3416).

Sitzmann, V., Zollhöfer, M., & Wetzstein, G. (2019). Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, *32*.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning* (pp. 2256–2265). PMLR.

Sun, S.-H., Huh, M., Liao, Y.-H., Zhang, N., & Lim, J. J. (2018). Multi-view to novel view: Synthesizing novel views with self-learned confidence. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 155–171).

Tabejamaat, M., Negin, F., & Bremond, F. F. (2021). Guided flow field estimation by generating independent patches. In *BMVC 2021-32nd British Machine Vision Conference*.

Tang, H., Bai, S., Torr, P. H., & Sebe, N. (2020a). Bipartite graph reasoning gans for person image generation. *arXiv preprint arXiv:2008.04381*, .

Tang, H., Bai, S., Zhang, L., Torr, P. H., & Sebe, N. (2020b). Xinggan for person image generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16* (pp. 717–734). Springer.

Tang, H., Xu, D., Sebe, N., Wang, Y., Corso, J. J., & Yan, Y. (2019). Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2417–2426).

Tatarchenko, M., Dosovitskiy, A., & Brox, T. (2016). Multi-view 3d models from single images with a convolutional network. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14* (pp. 322–337). Springer.

Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2017). Deep canonical time warping for simultaneous alignment and representation learning of sequences. *IEEE transactions on pattern analysis and machine intelligence*, *40*, 1128–1138.

Wan, Y., & Ren, M. (2021). New visual expression of anime film based on artificial intelligence and machine learning technology. *Journal of Sensors*, *2021*, 1–10.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, *13*, 600–612.

Wiles, O., Gkioxari, G., Szeliski, R., & Johnson, J. (2020). Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7467–7477).

Worrall, D. E., Garbin, S. J., Turmukhambetov, D., & Brostow, G. J. (2017). Interpretable transformations with encoder-decoder networks. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 5726–5735).

Zhang, J., Huang, S., Liu, J., Zhu, X., & Xu, F. (2023). Pyrf-pcr: A robust three-stage 3d point cloud registration for outdoor scene. *IEEE Transactions on Intelligent Vehicles*, .

Zhang, J., Li, K., Lai, Y.-K., & Yang, J. (2021). Pise: Person image synthesis and editing with decoupled gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7982–7990).

Zhang, P., Yang, L., Lai, J.-H., & Xie, X. (2022). Exploring dual-task correlation for pose guided person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7713–7722).

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 586–595).

Zhao, J., & Zhang, H. (2022). Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3657–3666).

Zhou, X., Yin, M., Chen, X., Sun, L., Gao, C., & Li, Q. (2022). Cross attention based style distribution for controllable person image synthesis. In *European Conference on Computer Vision* (pp. 161–178). Springer.

Zhou, X., Zhang, B., Zhang, T., Zhang, P., Bao, J., Chen, D., Zhang, Z., & Wen, F. (2021). Cocosnet v2: Full-resolution correspondence learning for image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11465–11475).

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