

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <https://orca.cardiff.ac.uk/id/eprint/170150/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Zhang, Jinsong, Lai, Yu-Kun , Ma, Jian and Li, Kun 2024. Multi-scale information transport generative adversarial network for human pose transfer. Displays 84 , 102786. 10.1016/j.displa.2024.102786

Publishers page: <https://doi.org/10.1016/j.displa.2024.102786>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



Multi-scale information transport generative adversarial network for human pose transfer

Jinsong Zhang^a, Yu-Kun Lai^b, Jian Ma^{a,*}, Kun Li^{a,*}

^aCollege of Intelligence and Computing, Tianjin University, Tianjin 300350, China.

^bSchool of Computer Science and Informatics, Cardiff University, Cardiff CF24 4AG, United Kingdom.

ARTICLE INFO

Keywords:

Human pose transfer

Generative adversarial network

Image synthesis

Optimal transport

ABSTRACT

Human pose transfer, a challenging image generation task, aims to transfer a source image from one pose to another. Existing methods often struggle to preserve details in visible regions or predict reasonable pixels for invisible regions due to inaccurate correspondences. In this paper, we design a novel multi-scale information transport generative adversarial network, composed of Information Transport (IT) blocks to establish and refine the correspondences progressively. Specifically, we compute a transport matrix to warp the source image features by integrating an optimal transport solver in our proposed IT block, and use IT blocks to refine the correspondences in different resolutions to preserve rich details of the source image features. The experimental results and applications demonstrate the effectiveness of our proposed method. We further present an image-specific optimization using only a single image. *The code is available for research purposes at <https://github.com/Zhangjinso/OT-POSE>.*

1. Introduction

Human pose transfer, *i.e.*, pose-guided person image synthesis, is an active research topic in computer vision and computer graphics. Human pose transfer aims to generate a new image of the same person in a new pose, which has many applications such as video generation [1, 2, 3, 4], virtual try-on [5, 6] and person re-identification [7, 8, 9, 10].

Previous methods can be categorized into three kinds of methods: attention-based methods, transformation-based methods and modulation-based methods. All these methods are based on Generative Adversarial Network (GAN) to generate realistic images. The attention-based methods [11, 7, 12,

13, 14] adopt an image branch to extract information and a pose branch to obtain pose information. Then, taking the pose information as the guidance, they adopt an attention mechanism to transfer the source image features from the source pose to the target pose. This kind of methods can generate reasonable results but fails to generate sharp images due to the soft association in the attention mechanism. Therefore, many works research on transformation-based methods and modulation-based methods. The transformation-based methods [15, 16, 17, 18, 19, 20] adopt affine matrices or optical flow to model the correspondences between the source image and the target image. These methods can generate sharper images that contain more details, but may struggle to predict reason-

*Corresponding author.

e-mail: jianma@tju.edu.cn (Jian Ma), lik@tju.edu.cn (Kun Li)

able textures for newly exposed regions that are visible in the target image but invisible in the source image. The modulation-based methods [6, 5, 21] transfer the style/texture information of the source image by modulating the weight and bias of the generated feature, which is then decoded to produce the results. These methods adopt many blocks with modulation operations to transfer the source image information, which helps to recover the style for the regions invisible in the source image. However, the style in the output image tends to lose spatial information, leading to blurry results with little details, especially for complex textures.

The aforementioned methods encounter challenges in synthesizing photo-realistic images with rich details due to the difficulty of modeling complex correspondences between the source and target images caused by complex textures and poses, especially for regions invisible in the source image. Good correspondences should have the following two properties: 1) preserving the source information as much as possible, which helps to preserve more details from the source image; and 2) ensuring reliable transformation, which means a point from the source image can be transferred to the right places of the target image for both visible and invisible regions.

In this paper, we propose a novel multi-scale information transport GAN, which deals with human pose transfer by leveraging optimal transport theory. The optimal transport problem is a fundamental problem to transform one distribution into another using the minimal cost [22]. This problem can be solved by seeking a transport matrix to distribute the mass of each point in the source distribution to many points of the target distribution *without loss of the mass*. Besides, the transport matrix is available with minimum cost [22], which means that the correspondences are *reliable*. In this sense, to address the complex correspondence modeling problem, the key idea of our model is to treat this problem as a progressive optimal transport problem, which tries to obtain the correspondences by seeking transport matrices progressively to transform the feature of the source image from the source pose to the target pose.

There are several works [23, 24, 25] introducing optimal

transport theory to their GANs, and all of them use the standard tool, *i.e.*, Sinkhorn algorithm [26], to compute the cost as the loss function at the feature level to regularize the distribution of the generated images to be close to the distribution of the real images, and do not improve the capacity of the generator. Instead of minimizing the distance of two distributions as a loss function, the motivation of our work is to establish reliable correspondences to transform the source feature from the source pose to the target pose. Therefore, we need to explore how to improve the capacity of our generator using optimal transport theory. Besides, to achieve realistic and reasonable results, we need to model the dense correspondences at both the feature level and image level, and refine the correspondences progressively, which has not been explored before.

In order to achieve our goal, we design a novel Information Transport (IT) block to improve the capacity of our generator, which computes the matrix to transport the feature from the source pose to the target pose by integrating an optimal transport solver [26]. Besides, to model the dense correspondences at both the feature level and image level, we introduce a simple loss function to constrain the computed correspondences at the image level. In order to establish more accurate correspondences and retain features from the source image, we adopt multi-scale IT blocks to leverage the source features to obtain more accurate correspondences progressively and capture more details of the source image. Our IT block allows robust many-to-many correspondences between the source and the target, while taking the manifold structure into account. Qualitative and quantitative results demonstrate the effectiveness of our proposed model. Furthermore, different from previous methods, there are no tunable parameters while computing the correspondences, which indicates that our model has better generalization capability. An image-specific optimization method is proposed to improve the performance of generalization. We also analyze our model to show the superiority of deformation and generalization. Moreover, we perform two applications of image-based animation and exemplar-based face image synthesis to validate the versa-

tility of our model. The code is available for research purposes at <https://github.com/Zhangjinso/OT-POSE>.

Our main contributions can be summarized as follows:

- We propose a novel multi-scale information transport GAN to establish and refine the correspondences progressively for human pose transfer.
- We design an Information Transport (IT) block to compute a transport matrix by integrating the optimal transport solver, which represents a many-to-many mapping to assign the features of the source to the target. With a simple loss, the IT block can model dense correspondences at both the feature level and image level.
- To leverage the characteristic that the correspondences computed by our model have better generalization capability, we present an image-specific optimization using only a single source image from other datasets, to generate promising results with good generalization ability.
- Experimental results demonstrate the superior performance of our method. The applications on image-based animation and exemplar-based face image generation show the scalability of our model.

2. Related work

2.1. Generative adversarial networks (GANs)

Generative Adversarial Networks (GANs) [27] have achieved great success in image synthesis. A GAN consists of a generator and a discriminator. The generator tries to synthesize realistic images to fool the discriminator, while the discriminator tries to differentiate real images and generated images. Benefiting from the adversarial training, GANs can generate realistic images [28, 29, 30, 31]. For the need of controlling the generated images, conditional GANs [32] take the inputs as a condition, and try to generate the images that match the condition. U-Net [33] is a common architecture of the conditional generator due to its capability of preserving multi-scale features. Many works [34, 35] adopt conditional GANs to tackle the image-to-image translation tasks (*e.g.*,

day to night). Huang *et al.* [36] propose Adaptive Instance Normalization (AdaIN) by modulating the weight and bias of the generated images to tackle the style transfer task. Karras *et al.* [28] adopt AdaIN to compose the generator to control the style of the generated images. Karras *et al.* [37] improve their work in many ways, and one of the most important improvements is that they replace AdaIN with modulated convolution, which improves the quality of generated images. To edit the generated image using semantic information, Park *et al.* [38] use a semantic segmentation mask as the guidance to modulate the weight and bias of the generated image, and achieve flexible control of the generated images. Zhu *et al.* [39] adopt group convolution to compose a conditional GAN to tackle the segmentation-to-image task.

Unlike the task of image-to-image translation, human pose transfer is an unaligned task, in which the input is unaligned with the target image. In this paper, we adopt the GAN framework with a multi-scale setting to generate realistic images, and design a novel block for the generator to tackle this unaligned task progressively.

2.2. Human pose transfer

Human pose transfer, *i.e.*, pose-guided person image synthesis, is an active topic in computer vision and computer graphics. Ma *et al.* [40] propose a two-stage framework to alleviate the difficulty of synthesizing images from unaligned inputs. They first generate a coarse result using a U-Net generator and then predict a difference map to refine the coarse result. Esser *et al.* [41] combine the variational auto-encoder [42] and the U-Net [33] to model the relation of poses and images. The U-Net can generate reliable results for the aligned tasks, but it is difficult to synthesize sharp images for the unaligned pose-guided person image synthesis. Therefore, instead of directly learning a conditional distribution, more researchers focus on how to transfer the information of the source image from the source pose to the target pose.

We summarize previous work into three kinds: attention-based methods, transformation-based methods and modulation-based methods. The attention-based methods [11, 7, 12, 13, 14]

adopt a two-branch network with a pose branch and an image branch, and explore how to use the pose information as guidance to transfer the source image. Zhu *et al.* [11] propose a novel network with a progressive pose transfer scheme to model the change in the manifold of images. However, the local attention mechanism is difficult to capture the long-range dependence. Li *et al.* [7] propose a pose-guided non-local attention block to capture the long-range dependence to tackle this problem. Tang *et al.* [13] propose bipartite graph reasoning GANs to reason the pose information. Tang *et al.* [14] propose an appearance-guided shape-based generation branch and a shape-guided appearance-based branch to transfer the source image from the source pose to the target pose. The images generated by these attention-based methods are reasonable but often blurry.

To generate sharper and more realistic images, transformation-based methods [15, 43, 16, 17, 44] adopt affine matrices or optical flows as the transformation to warp the source image to the target pose. Li *et al.* [43] and Liu *et al.* [17] use a parametric human model to estimate the precise flow. However, these methods based on the parametric human model rely on accurate estimation of the parametric human model, which is unstable and difficult to model loose clothing. Siarohin *et al.* [15] take each body part as a rectangle region, and compute the affine matrix of each region according to its corresponding region. Ren *et al.* [16] first predict flow fields according to the poses and the source image to warp the features of the source image, and then refine the warped features using a local attention mechanism. These transformation-based methods can synthesize realistic pixels for regions visible in both the source image and the target image, but can struggle to synthesize reasonable pixels for invisible regions in the source image.

Inspired by the style transfer task, some researchers focus on how to transfer the style information using the modulation operation, which can alleviate the difficulty of invisible region generation. Men *et al.* [6] introduce human parsing as inputs and adopt AdaIN to transfer the style information of each se-

mantic region to the generated images. Zhang *et al.* [5] and Lv *et al.* [21] use a two-stage framework that first synthesizes a human parsing as an intermediate result and then generates the final image by adopting Semantic Region-Adaptive Normalization (SEAN) [45] to modulate the generated feature of each semantic region. These methods generate more reasonable results for invisible regions, but fail to preserve rich style information for complex textures. Some methods [46, 47, 48, 49, 50, 51] generate realistic images using implicit or explicit representation of human, but these methods require an image sequence or multi-view images of a specific person, which fail to generalize to other identities.

In this paper, we propose a novel multi-scale information transport GAN to establish and refine the correspondences progressively. The correspondences is established by computing a transport matrix of an optimal transport problem, and refined by the current feature and the source image feature progressively. This transport matrix is a many-to-many mapping and has theoretical support. This characteristic is not only beneficial to modeling the correspondences for both invisible and visible regions, but it also provides stable gradients for the optimization of the model [52].

2.3. Optimal transport

Optimal transport theory studies the transmission problem between two distributions, which has been used in many research areas, such as graph matching [53] and domain adaptation [54]. For more comprehensive reviews, we direct readers to [55, 56].

Several works adopt optimal transport to learn the cross-domain alignment. Puy *et al.* [57] propose a novel framework based on optimal transport to predict scene flows according to two sets of point clouds. Inspired by this work, Song *et al.* [58] propose a correspondence-refinement network to warp the points of the parametric model to transfer the pose in 3D. Besides, some works use optimal transport cost as a distance to regularize their models. Avraham *et al.* [23] adopt Wasserstein distance in a low dimensional latent space to alleviate the instability issues in GANs [59]. Salimans *et al.* [24] also im-

prove GANs by introducing optimal transport as a regularization. Geng *et al.* [25] propose to regularize the model to tackle the Non-Line-of-Sight imaging problem.

In a word, previous works [23, 24, 25] all adopt the Sinkhorn algorithm, a standard tool for optimal transport problems, to construct their losses to optimize GANs for image generation. However, these methods do not focus on improving the capacity of the generator. Besides, all these methods focus on the feature level. Other works [53, 57, 58, 60, 61] explore the effectiveness of the Sinkhorn algorithm at the graph level, feature level, and point cloud level, respectively, and validate the efficacy of optimal transport theory. However, these works only use the Sinkhorn algorithm once at a specific level without any refinement, leading to unstable results. We validate the effectiveness of optimal transport at the image level by constraining the mapping at the image level. To the best of our knowledge, our method is the first work that establishes and refines correspondences using optimal transport in both image level and feature level. Furthermore, unlike previous works [23, 24, 25] that require elaborate loss functions, we can use simple loss functions with better capacity brought in by the proposed IT block.

3. Method

The goal of our work is to generate a new image of the same person in a new pose. Specifically, given a source image I_s containing a person we first extract 18 human keypoints using a popular pose estimation method [62], and encode the keypoints following [63] as its pose representation P_s . We also assume the target pose P_t is also known, and our method aims to deliver a generated image I_g with the similar appearance with I_s but in the target pose P_t . The ground truth target image with the person from the source image in the target pose is denoted as I_t , which is only accessible during training. How to model the correspondences between the source image and the target image is the fundamental problem. With precise correspondences, we can leverage the source image to synthesize the target image.

To establish accurate correspondences, we propose a novel multi-scale information transport GAN to establish and refine

the correspondences progressively. Figure 1 shows the framework of our method. We design an Information Transport (IT) block by computing an optimal transport matrix, which indicates the correspondences between the source image and the target image. The characteristic that the optimal transport matrix can model a many-to-many mapping gives our model the ability to predict reasonable results for invisible regions in the source image. With the multi-scale setting, our model can refine the correspondences and transform more details of the source image from the source pose to the target pose.

3.1. Optimal transport problem

Due to the discreteness of images, we first introduce the Monge problem [64] between discrete measures. Given two point sets $\mathcal{X}=\{x_1, x_2, \dots, x_n\}$ and $\mathcal{Y}=\{y_1, y_2, \dots, y_n\}$, discrete measures with weights $\mathbf{a}$ of $\mathcal{X}$ and $\mathbf{b}$ of $\mathcal{Y}$ read

$$\alpha = \sum_{i=1}^n \mathbf{a}_i \delta_{x_i}, \quad \beta = \sum_{j=1}^n \mathbf{b}_j \delta_{y_j} \quad (1)$$

where $\mathbf{a}, \mathbf{b} \in \mathbb{R}_+^n$ with $\sum_{i=1}^n \mathbf{a}_i = 1$, $\sum_{j=1}^n \mathbf{b}_j = 1$, and δ_x denotes the Dirac delta function at position x , which means that a unit of mass is concentrated at location x [22]. Therefore, α and β are two discrete measures defined on point sets $\mathcal{X}$ and $\mathcal{Y}$, which can be served as the masses of these two point sets.

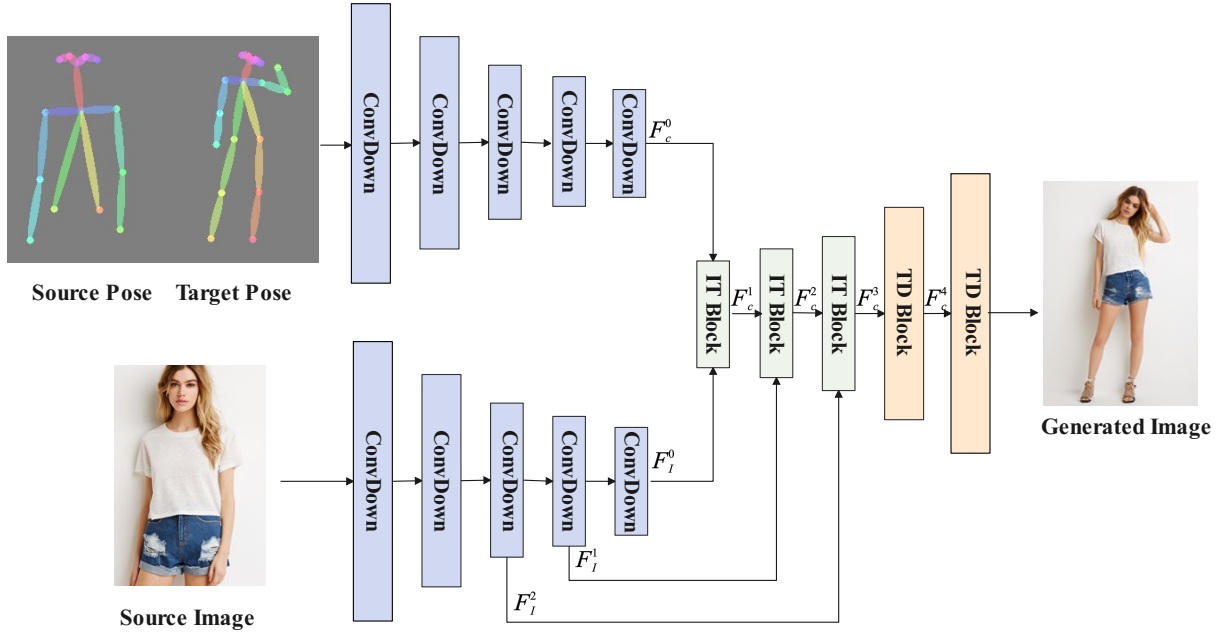
The Monge problem seeks a map that relates each point x_i to a single point y_i , and must push the mass of α to the mass of β [22]. The map $T : \{x_1, x_2, \dots, x_n\} \rightarrow \{y_1, y_2, \dots, y_n\}$, *i.e.*, the transport plan, must meet that:

$$\forall j \in \{1, 2, \dots, n\}, \mathbf{b}_j = \sum_{i: T(x_i)=y_j} \mathbf{a}_i \quad (2)$$

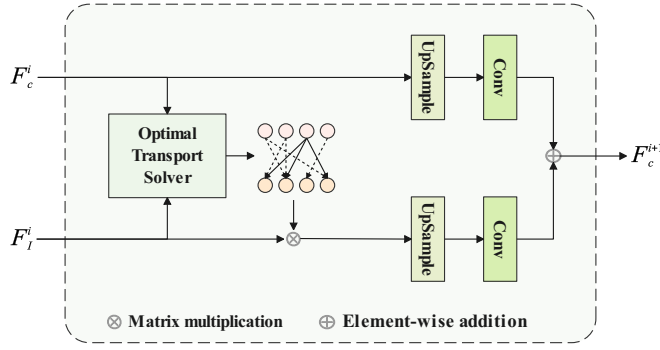
where all the elements of $\mathbf{b}$ are positive and the transport plan is surjective. Thus, we can write it in a compact form as $T_{\#}\alpha = \beta$. $T_{\#}$ is a push-forward measure, which means a measurable mapping that can transfer measure α to the measure β . Let $\mathbf{C} = \{c_{ij}\} \in \mathbb{R}^{n \times n}$ be the cost matrix, where each element C_{ij} denotes the transport cost between x_i and y_j . The transportation cost can be defined as:

$$\mathcal{L}_M(\mathcal{X}, \mathcal{Y}) = \min \sum_{i=1}^n C(x_i, T(x_i)) \quad \text{s.t. } T_{\#}\alpha = \beta \quad (3)$$

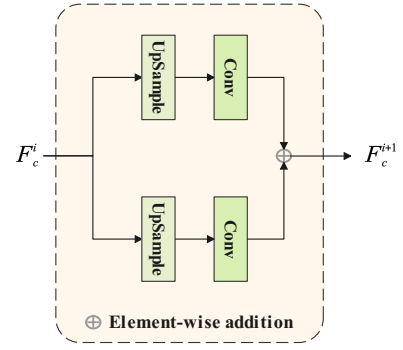
In Monge problem, the mass of x_i with $\mathbf{a}_i$ units can be assigned to a specific point y_j , and a single point y_j can receive



(a) Overview of our framework



(b) Information Transport (IT) block



(c) Two-branch Decoder (TD) block

Fig. 1. Overview of our framework: (a) overall framework of our model; (b) the architecture of our IT block; (c) the architecture of our TD block. Our IT block can model and refine the correspondences between the source image and the target image according to the source image and the poses. With the multi-scale setting, our model preserves the details after changing the pose.

the mass from many points in $\mathcal{X}$. In other words, this formulation leads to a transport plan that is a many-to-one mapping.

However, it is difficult to solve the original Monge problem. Kantorovich [65] relaxes the constraint in Monge problem that a source point can only be transported to a single point, and instead allows that a source point can be assigned to many points, which means that the transport plan is a many-to-many mapping. Specifically, the element T_{ij} of the transport plan $T \in \mathbb{R}_+^{n \times n}$ denotes the amount of the mass of x_i transported to y_j . Define the set of transport plans as $U(\mathbf{a}, \mathbf{b})$, and the formulation

of $U(\mathbf{a}, \mathbf{b})$ can be written as:

$$U(\mathbf{a}, \mathbf{b}) \stackrel{\text{def}}{=} \{T \in \mathbb{R}_+ : T\mathbb{1}_n = \mathbf{a}, T^\top \mathbb{1}_n = \mathbf{b}\} \quad (4)$$

where $\mathbb{1}_n$ is a vector of ones with n elements, and $T\mathbb{1}_n$ and $T^\top \mathbb{1}_n$ are matrix-vector multiplications. The first constraint means that the mass of each point x_i with $\mathbf{a}_i$ units will be pushed forward to many points in $\mathcal{Y}$, and the second constraint means that each point y_i receives the mass with $\mathbf{b}_i$ units from many points in $\mathcal{X}$.

Therefore, the Kantorovich's optimal transport problem can

be written as:

$$\mathcal{L}_K = \min_{T \in U(a,b)} \langle C, T \rangle \quad (5)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product operation. It can be seen that if the cost function is set to Euclidean Distance, $\mathcal{L}_K$ becomes Wasswerstein-2 distance [59]. We can obtain the optimal transport plan T by solving the problem:

$$T = \arg \min_{T \in U(a,b)} \langle C, T \rangle \quad (6)$$

3.2. Information transport GAN

3.2.1. Motivation and contribution

As described in Section 1, the fundamental problem in human pose transfer is how to model the correspondences between the source image I_s and the target image I_t for both visible and invisible regions in the source image according to the pose information. Good correspondences should have the following two properties: 1) preserving the source information as much as possible, and 2) ensuring reliable transformation. The first property guarantees the integrity of the source image, and the second property means that the transformation is appropriate, *i.e.*, a point from the source image is transferred to the right places of the target image through the correspondences. The optimal transport problem seeks a transport plan to make the cost of transferring the mass of one point set to another minimum without loss of the mass. Note that the optimal transport plan is available with minimum cost [22, 66], which means that the correspondences between the two point sets are reliable. From this perspective, we take human pose transfer as a progressive correspondence learning problem to obtain reliable correspondences. Different from previous works [23, 24, 25] that utilize optimal transport theory to minimize the distance between two distributions, our goal is to leverage the transport plan to enhance the capacity of our generator. We aim to obtain a transport plan T by solving an optimal transport problem between the image features and the pose features. This allows us to assign the features of the source image to synthesize the target image, effectively tackling this misalignment problem. Besides, we present a solution, namely the information transport block to

demonstrate how to leverage optimal transport theory for conditional image generation at both the feature level and image level. Moreover, different from previous works [60, 58, 57] that establish the correspondences only once, our framework can establish and refine the correspondence progressively thanks to the IT block. The details of our model are given in the following.

3.2.2. Encoder

As shown in Figure 1, our generator has two encoders: the image encoder and the pose encoder. These two encoders have the same structure with four down-sampling convolution layers. The pose encoder takes the source pose P_s and the target pose P_t as input, and outputs the pose feature F_p . The pose feature F_p not only contains the feature of the target pose but also the feature of the source pose, which is useful for the correspondence modeling according to the poses. The input of the image encoder is I_s and the output of the image encoder is the image feature list $F_I = \{F_I^k\}$, where F_I^k denotes the image feature of the k -th layer. The image feature F_I^k contains the feature of both the source image and the source pose, and the features with different resolutions help to preserve the details in the source image.

3.2.3. Information transport block

With the image feature list F_I and the pose feature F_p , the Information Transport (IT) block is responsible for seeking a transport plan/matrix first to deform the features of the source image. Then, the IT block fuses the original features before deformation and the warped features to generate richer features for the generated image. Figure 1 shows the architecture of the k -th IT block.

Denote the pose feature F_p as the initial current feature F_c^0 , and we give the details for the k -th IT block. Given the current feature F_c^k and the image feature F_I^k , we first define a cost function between F_I^k and F_c^k to compute a transport matrix T^k to warp F_I^k from the source pose to the target pose. Cosine similarity is a popular choice in optimal transport problems [57, 58]. We also use it as our cost function, which is defined as:

$$C^k = 1 - \frac{F_c^k F_I^{k\top}}{\|F_c^k\|_2 \|F_I^k\|_2} \quad (7)$$

In this work, we solve this problem using Sinkhorn-Knopp algorithm [26] to get an optimal transport plan/matrix. Given the cost matrix $\mathbf{C}^k$, we adopt an entropic regularization term [67] to solve the optimal transport problem. The formulation of optimal transport problem can be revised from Equation 6 as:

$$\mathbf{T}^k = \arg \min_{\mathbf{T}^k \in \mathcal{U}(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}^k, \mathbf{T}^k \rangle + \gamma \sum_{ij} \mathbf{T}_{ij}^k (\log(\mathbf{T}_{ij}^k) - 1) \quad (8)$$

where γ is the regularization parameter. In our case, taking each location of the feature grid as a point, the mass of each point can be regarded as the same with $\frac{1}{n}$ units, where n is the number of points in the point set. In other words, the elements of $\mathbf{a}$ and $\mathbf{b}$ are all the same and equal to $\frac{1}{n}$.

Then the transport matrix $\mathbf{T}$ can be solved using Sinkhorn-Knopp algorithm:

$$\mathbf{T}^k = \text{OTS}(\mathbf{C}^k, \gamma, n_{iter}) \quad (9)$$

where OTS denotes the optimal transport solver, *i.e.*, Sinkhorn-Knopp algorithm [26], γ is the weight of the regularization term, and n_{iter} is the maximum iteration number of the algorithm.

With the transport matrix $\mathbf{T}^k$ in the k -th IT block, we can warp the source image feature from the source pose to the target pose by matrix multiplication at the feature level, mathematically:

$$\mathbf{W}_c^k = \mathbf{T}^k \mathbf{F}_I^k \quad (10)$$

To aggregate the former feature and the current feature, we add these two features after convolution layers, which can be written as:

$$\mathbf{F}_c^{k+1} = \phi_k(f_k(\mathbf{W}_c^k) + g_k(\mathbf{F}_c^k)) \quad (11)$$

where ϕ_k is implemented as a convolutional layer, and f_k, g_k are implemented as convolutional layers with up-sampling layers.

3.2.4. Decoder

Due to the high computational budget of IT blocks, we only adopt them at some scales with low resolutions, and we design Two-branch Decoder (TD) blocks for the scales with high resolutions. The two-branch decoder block has the same input and output as an IT block, but does not have the process of solving

the optimal transport matrix. For the k -th TD block, the process can be formulated as:

$$\mathbf{F}_c^{k+1} = \phi_k(f_k(\mathbf{F}_c^k) + g_k(\mathbf{F}_c^k)) \quad (12)$$

this indicates that the new correspondence is solved from the current feature, which helps to make it more reliable.

The decoder is composed of IT blocks and TD blocks at different scales. We follow the architecture of StyleGAN2 [68] to design our decoder. The generated images I_g are delivered by up-sampling and summing all the RGB outputs at different scales.

3.2.5. Discriminator

In order to produce high-quality and realistic images, we utilize a discriminator D to distinguish between real and generated images. The discriminator takes as input either the real image I_t or the generated image I_g . This input passes through a convolutional layer and multiple residual blocks with down-sampling. The ultimate score is obtained from a linear layer.

3.3. Loss functions

In this work, we adopt five loss functions including an image reconstruction loss $\mathcal{L}_{image}$, a face reconstruction loss $\mathcal{L}_{face}$, a transport matrix constraint $\mathcal{L}_{con}$, a self-reconstruction loss $\mathcal{L}_{self}$ and an adversarial loss $\mathcal{L}_{adv}$ to train our model in an end-to-end manner. The total loss can be written as:

$$\mathcal{L}_{full} = \lambda_{image}\mathcal{L}_{image} + \lambda_{face}\mathcal{L}_{face} + \lambda_{con}\mathcal{L}_{con} + \lambda_{self}\mathcal{L}_{self} + \lambda_{adv}\mathcal{L}_{adv} \quad (13)$$

where $\lambda_{image}, \lambda_{face}, \lambda_{con}, \lambda_{self}, \lambda_{adv}$ are the balancing weights of different terms, respectively.

Image reconstruction loss. We adopt the perceptual loss [69] as the image reconstruction loss to encourage the similarity between the target image I_t and the generated image I_g , which is defined as:

$$\mathcal{L}_{image} = \sum_i \|\varphi_i(I_t) - \varphi_i(I_g)\|_1 \quad (14)$$

where $I_g = G(I_s, P_s, P_t)$ is the generated image and φ_i denotes the i -th layer of the pre-trained VGG-19 network [70]. In practice, we adopt relu1_1, relu2_1, relu3_1, relu4_1, and relu5_1 as feature extractors.

Face reconstruction loss. To enhance the details and the similarity of the facial region, we compute the distance between the cropped faces of the target image and the generated image using perceptual loss. It can be formulated as:

$$\mathcal{L}_{face} = \sum_i \|\varphi_i(\tau(I_t)) - \varphi_i(\tau(I_g))\|_1 \quad (15)$$

where τ is the crop operation that crops the facial region according to the target pose. For φ_i , we adopt the same layers of VGG-19 as image reconstruction loss.

Transport matrix constraint. To ensure the reliability of the transport matrix and facilitate correspondence learning at both the feature level and image level, we constrain the warped image obtained by warping the source image using the transport matrix to be similar to the target image at different scales. Specifically, with the transport matrix T^k , we down-sample the source image I_s and the target image I_t to obtain I_s^k and I_t^k with the same resolution as F_i^k . Then, we can warp I_s^k using the transport matrix T^k by

$$I_w^k = T^k I_s^k \quad (16)$$

In order to ensure reliable correspondences at the image level, we aim to make the warped image I_w^k similar to the corresponding target image I_t^k . Consequently, we compute the distance between the warped image I_w^k and the target image I_t^k . Mathematically,

$$\mathcal{L}_{con} = \sum_k \|I_t^k - I_w^k\|_1 \quad (17)$$

where T^k denotes the transport matrix obtained from the k -th IT block.

Self-reconstruction loss. Inspired by DUAL [71], we adopt the dual-task setting to train our model. Different from [71] with a multi-task setting, we simply replace the target pose with the source pose to reconstruct the source image and compute the loss using

$$\mathcal{L}_{self} = \sum_i \|\varphi_i(I_s) - \varphi_i(I'_g)\|_1 \quad (18)$$

where $I'_g = G(I_s, P_s, P_s)$ is the output in the self-reconstruction task.

Adversarial loss. To generate more realistic images, we adopt an adversarial loss by training a discriminator D to distinguish whether the image comes from the generator. The adversarial loss is defined as:

$$\mathcal{L}_{adv} = \mathbb{E}[\log(1 - D(I_g))] + \mathbb{E}[\log(D(I_t))] \quad (19)$$

3.4. Image-specific optimization

Because there are no learnable parameters in the optimal transport solver of our IT blocks, the correspondence matrix is *computed* according to the appearance feature and the pose feature, which indicates our IT block has better generalization capability. To get better results by leveraging the single source image, we present an image-specific optimization and optimize our generator with a single image, to further improve the performance for input images from new datasets. Specifically, given a source image I_s and the source pose P_s , we optimize our generator G by synthesizing the source image similar to the self-reconstruction task, mathematically:

$$I_g = G(I_s, P_s, P_s) \quad (20)$$

We adopt image reconstruction loss, face reconstruction loss, and transport matrix constraint described in Section 3.3 between I_s and I_g as our loss functions to optimize our generator. Note that we only have a single image as the input and the output, which is not suitable for adopting adversarial loss. We train the generator about 200 iterations using Adam solver [72] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and the learning rate is set to 0.0001.

3.5. Training details

We train our model with 256×176 images in an end-to-end manner. Due to the computational budget from the optimal transport solver, we only adopt IT blocks in resolutions 16×16 , 32×32 , and 64×64 for faster training speed. The transport matrix constraint loss is also computed at these resolutions. Following NTED [63], we adopt the historical average technique [73] to average the model for better performance. The hyperparameters $[\lambda_{image}, \lambda_{face}, \lambda_{con}, \lambda_{self}, \lambda_{adv}]$ are set to $[2, 1, 15, 2, 1.5]$. We adopt Adam solver [72] with $\beta_1 = 0, \beta_2 = 0.99$ to optimize our model. The learning rates of both the generator and the discriminator are the same and equal to 2×10^{-3} . We

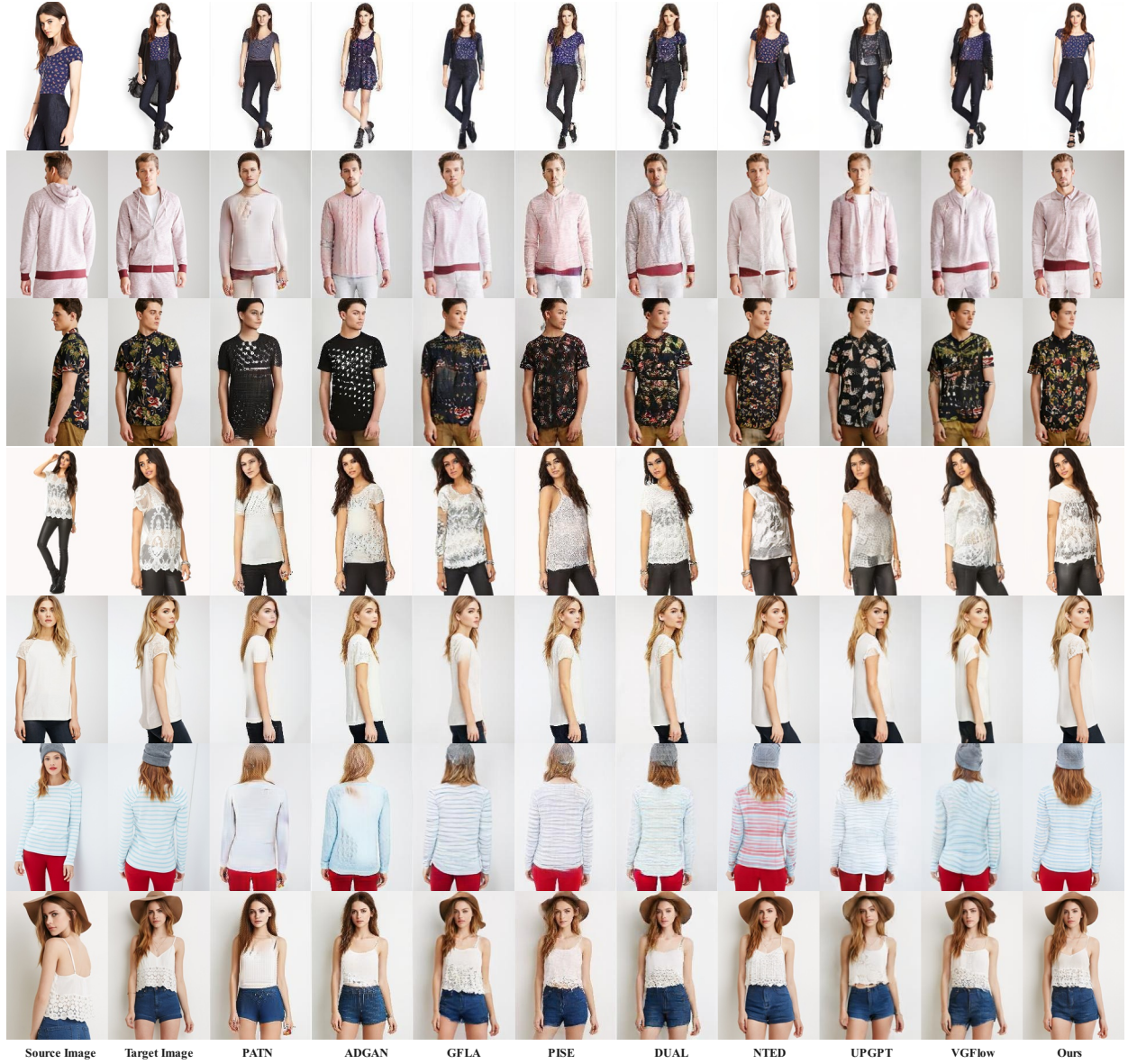


Fig. 2. Qualitative comparison on DeepFashion Dataset [74]. From left to right are the results of PATN [11], ADGAN [6], GFLA [16], PISE [5], DUAL [71], NTED [63], UPGPT [75], VGFlow [20] and ours, respectively.

first train our model without self-reconstruction loss about 120 epochs and then train our model with self-reconstruction loss about 80 epochs.

4. Experimental results

In this section, we first give the details of the dataset and metrics in Section 4.1. The comparison results with several state-of-the-art methods are presented in Section 4.2. Then, we conduct ablation studies to validate the effectiveness of our model in Section 4.3. In Section 4.4, we analyze our model by evaluating and visualizing the transport matrix in our IT

blocks and validate the performance of generalization on FashionVideo dataset. Finally, we show an application of exemplar-based face image generation in Section 4.5.

4.1. Dataset and metrics

Dataset. We conduct the evaluation of human pose transfer on In-shop Clothes Retrieval Benchmark DeepFashion [74]. This dataset contains images of the same person in the same clothes with different poses, which is suitable for human pose transfer. We use the high-resolution images, and resize images to 256×176 . This indicates that we only use the information of

Table 1. Quantitative comparison with state-of-the-art methods on DeepFashion dataset [74]

	PATN [11]	ADGAN [6]	GFLA [16]	PISE [5]	DUAL [71]	NTED [63]	UPGPT [75]	VGFlow [20]	Ours
FID ↓	22.20	14.53	9.430	10.84	10.53	8.447	9.293	<u>8.175</u>	8.092
clean-FID ↓	22.12	14.46	9.449	10.94	10.52	8.566	9.421	<u>8.325</u>	8.219
LPIPS ↓	0.2516	0.2245	0.1847	0.2070	0.1953	<u>0.1747</u>	0.1817	0.1783	0.1724
PSNR ↑	16.65	16.78	17.70	16.91	17.54	17.96	17.67	<u>17.98</u>	18.00

Table 2. Quantitative results of ablation study about the effectiveness of different components of our model.

	Baseline	IT	Full
FID ↓	8.820	8.315	8.092
clean-FID ↓	8.934	8.429	8.219
LPIPS ↓	0.1878	0.1754	0.1724
PSNR ↑	17.56	17.88	18.00

256×176 pixels. Following the dataset splits in [11], there are 101,966 training pairs and 8,570 test pairs for the training and evaluation.

Metrics. To evaluate the performance of our model, we adopt four metrics as our evaluation metrics. Peak Signal to Noise Ratio (PSNR) and Learned Perceptual Image Patch Similarity (LPIPS) [76] are used to compute the difference between the generated image and the target PSNR measures the difference in pixel level, and LPIPS calculates the distance in perceptual level. To measure the quality of the generated images, we adopt Fréchet Inception Distance (FID) [77] to compute the distance between the distributions of the generated images and the real images. To decrease the influence of image resizing and compression, we use clean-FID [78] to measure the realism of the generated images further. Besides, all images are evaluated in the JPEG format with a resolution of 256×176 pixels.

4.2. Comparison results

We compare our method with eight state-of-the-art methods to validate the performance of our method: PATN [11], ADGAN [6], GFLA [16], PISE [5], DUAL [71], NTED [63], UPGPT [75] and VGFlow [20]. Note that all the results are produced following the public repositories with the same training and test splits.

4.2.1. Quantitative comparison

The quantitative comparison results are given in Table 1. It can be seen that our model outperforms the other state-of-the-art methods in all evaluation metrics, which indicates that our method can achieve better performance at pixel level and perceptual level. Besides, the lower FID and clean-FID scores denote that our synthesized images are more realistic. The modulation-based method PISE also encounters difficulty in synthesizing realistic images. The latest attention-based method NTED shows competitive results, but it is difficult to preserve details and leads to higher scores in FID and clean-FID (*i.e.*, lower quality). The transformation-based method VGFlow also achieves promising performance, but the quality of the results is lower than ours.

4.2.2. Qualitative comparison

We give some visual results compared with several state-of-the-art methods in Figure 2. From the observation of the last two rows, the transformation-based method VGFlow can contain the details for the visible regions in the source image. However, it is difficult to predict reasonable pixels for the invisible regions in the source image. In contrast, the attention-based method NTED can generate reasonable regions for both visible and invisible regions, but it struggles to preserve the texture (in the first, fifth and seventh rows). Similarly, the modulation-based methods ADGAN and PISE are also difficult to preserve the complex details. Our model adopts the optimal transport solver to build reliable correspondences to transform the source image feature, and adopts a multi-scale setting to preserve details, which can not only synthesize complex texture (in the third and fourth rows), but also generate reasonable results for both visible and invisible regions. More results can be found in supplementary materials.

Table 3. Quantitative results of ablation study about the design of our IT block.

	Baseline	w/o add.	w concat.	Full
FID ↓	8.820	8.172	8.003	<u>8.092</u>
clean-FID ↓	8.934	8.279	8.152	<u>8.219</u>
LPIPS ↓	0.1878	0.1890	<u>0.1734</u>	0.1724
PSNR ↑	17.56	17.45	<u>17.93</u>	18.00

**Fig. 3.** Qualitative results of the ablation study about the effectiveness of different parts.

4.3. Ablation study

4.3.1. The effectiveness of different components

To validate the effectiveness of different parts in our model, we conduct experiments with several variants to compare with our full model.

Baseline model. The baseline model is trained to show the effectiveness of our idea that formulates the human pose transfer as an optimal transport problem to build reliable correspondences. We replace the optimal transport solver in our IT block with the concatenation operation. The model is trained with the image and face reconstruction losses and the adversarial loss. The baseline model is with the multi-scale setting.

IT model. The decoder of the IT model has an IT block and

**Fig. 4.** Qualitative results of the ablation study about the design of our IT block.

four TD blocks without the multi-scale setting. We train this model with the full losses. This model is used to validate the effectiveness of our multi-scale architecture.

Full model. This model is our full model presented in Section 3, and trained with the full losses.

The quantitative and qualitative results are shown in Table 2 and Figure 3, respectively. From Table 2, by comparing the results of the full model and baseline model, it can be seen that adopting the optimal transport theory to tackle the human pose transfer can boost the quality of the generated images in terms of all evaluation metrics. Besides, we can also find that the multi-scale setting can improve the quality of the generated images, which indicates that the correspondences built by our IT blocks are beneficial to different resolutions. This means that we can generate better results by adopting the IT block in higher resolution. Similar conclusions can be found in Figure 3. The images generated by the baseline model lose structure details (in the second and the third rows), and have wrong colors (in the fourth row) and artifacts (in the first and the fourth rows). The IT model can preserve the structure details (in the

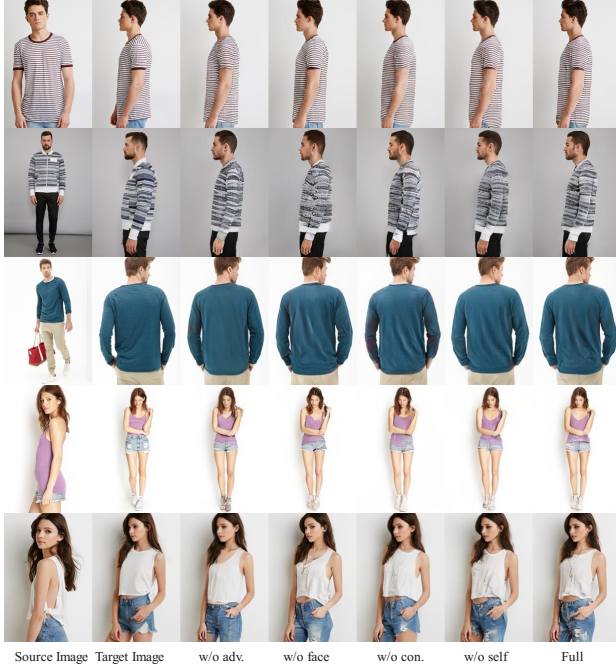


Fig. 5. Qualitative results of ablation study about the different loss functions.

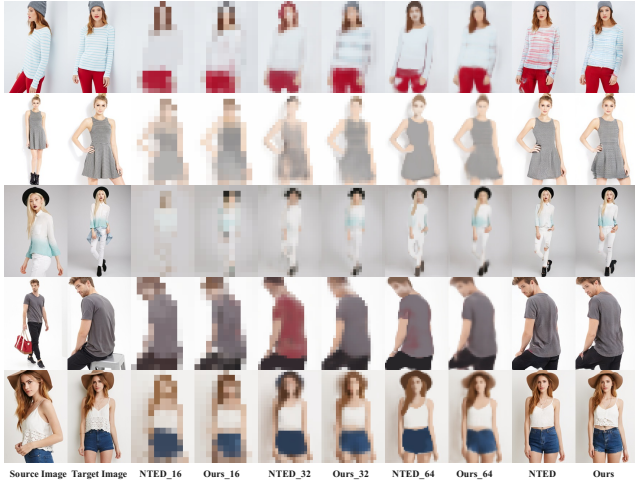


Fig. 6. Visualization of transport matrices at different resolutions by warping the source image. The first two rows show the source image and the target image. The following rows show the warped images by warping the source image with the different resolutions using the corresponding transport matrices of NTED [63] and ours. The last two rows show the final results of NTED and ours.

first and the third rows), but still fails to deal with complex texture (in the second row) and generates artifacts (in the fourth row). By comparison, the generated images by our full model can deform complex textures and preserve the structure details, Which demonstrates that our full model can synthesize more realistic images.

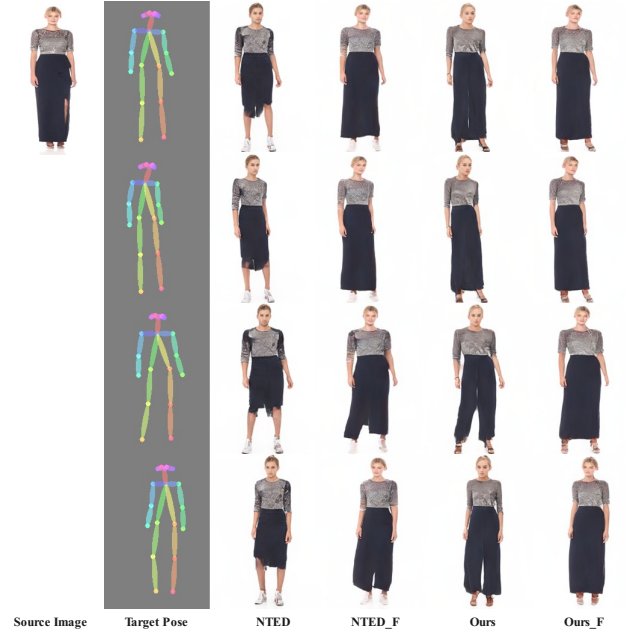


Fig. 7. Qualitative results compared with NTED [63] on FashionVideo dataset [79]. From left to right are the source image, target poses, and the results of NTED, NTED optimized using the source image, ours and ours using image-specific optimization with the source image, respectively.

4.3.2. The design of the IT block

As illustrated in Section 3, we adopt an addition operation to fuse the features before and after deformation. Therefore, we conduct several variants to prove the effectiveness of our design. Note that all the models are trained with the same training scheme.

Baseline model. This model is the same as the previous baseline model, which can prove the effectiveness of our key idea with fairer comparison by directly comparing with our full model. Due to the lack of the transport matrix, the transport matrix constraint is not used during the training of this model.

IT block without addition (w/o add.). The IT block in this model is without the addition operation, which means that the feature before deformation is not used. This model is trained with the full losses.

IT block with concatenation (w concat.). The IT block in this model replaces the addition operation with the concatenation operation in Equation 11.

The quantitative and qualitative results are shown in Table 3 and Figure 4, respectively. From Table 3, it can be seen that both addition operation and concatenation operation are useful

Table 4. Quantitative results of ablation study about the different loss functions.

	w/o adv.	w/o face	w/o con.	w/o self.	Full
FID ↓	11.09	<u>8.232</u>	8.490	8.237	8.092
clean-FID ↓	11.12	8.354	8.620	<u>8.326</u>	8.219
LPIPS ↓	<u>0.1725</u>	0.1728	0.1759	0.1732	0.1724
PSNR ↑	18.09	17.96	17.88	17.94	<u>18.00</u>

to generate better results. The results generated by IT block with concatenation operation have better FID and clean-FID scores, which shows that the concatenation operation is beneficial to synthesizing more realistic images. However, compared with the full model, the LPIPS and PSNR scores show that IT block with concatenation generates less similar images. Besides, the results generated by the model IT block without addition can obtain good FID and clean-FID scores, but worse performance on LPIPS and PSNR, which indicates that the addition operation is important to preserve the original features to synthesize more similar results. The comparison between the baseline model and the full model can demonstrate the effectiveness of the new formulation of this task further. The qualitative results also lead to similar conclusions. The results of the baseline model that uses concatenation operation to replace optimal transport solver and IT block without addition are blurry with artifacts, and fail to preserve texture or details of the source image. This shows the effectiveness of our proposed IT block. The results of IT block with concatenation are also realistic, however, some artifacts (in the first row) and dissimilar patterns (in the second row) exist in the generated images. In contrast, our full model generates more realistic images with similar textures and details of the source image.

4.3.3. The effectiveness of different loss functions

To assess the effectiveness of different loss functions, we conduct experiments using different loss functions. Table 4 presents the quantitative results, comparing our full model with the full model without the respective loss functions, including adversarial loss (adv.), face loss (face), constraint loss (con.), and self-reconstruction loss (self). Comparing the model without adversarial loss and the full model, we can see that the adversarial loss is vital to improving the realism of the generated

images. Interestingly, the model without the adversarial loss can get better performance in terms of the PSNR metric due to the blurry results. This observation aligns with previous studies [76] where blurry results tend to yield higher PSNR. The face loss, which emphasizes constraint on facial regions, exhibits improvements comparing the model without face loss and the full model. The model without constraint loss performs less effectively than the full model, highlighting the importance of learning correspondences at both the feature and image levels. Additionally, the self-reconstruction loss also improves the performance of our model.

The qualitative results are shown in Figure 5. It can be seen that the results without adversarial loss are blurry and smooth, while the full model produces more detailed and sharper images. Furthermore, the results without face loss fail to preserve the identity of the source image (in the second and the fifth rows). Moreover, the results without constraint loss may generate artifacts and lack reasonable clothing style (in the second, third, and fourth rows). The results without self-reconstruction loss also have some artifacts (in the second and the fourth rows). More qualitative results can be found in the supplementary materials.

4.4. Analysis of our model

4.4.1. Evaluation of transport matrices

We conduct a quantitative comparison to validate the effectiveness of transport matrices compared with NTED. By utilizing the transport matrices and the source image, the computation of the warped images follows the same procedure as Equation 16. Table 5 presents the comparison results in different resolutions. It can be seen that our method yields superior performance in terms of the warped image compared to NTED across various resolutions. This indicates that our method is capable of computing more accurate correspondences.

4.4.2. Visualization of transport matrices

To get an intuitive understanding of the transport matrices in our model, we visualize the transport matrices in different resolutions by warping the source image with the corresponding resolutions. As shown in Figure 6, compared with NTED

Table 5. Quantitative results of transport matrices compared with NTED in different resolutions

	16×16		32×32		64×64	
	NTED	Ours	NTED	Ours	NTED	Ours
FID ↓	38.62	35.76	34.63	32.21	27.57	26.94
clean-FID ↓	46.95	44.91	43.80	42.25	32.57	31.45
LPIPS ↓	0.5272	0.5180	0.4640	0.4568	0.3667	0.3586
PSNR ↑	15.08	15.21	16.85	17.12	18.07	18.42

Table 6. Quantitative results of generalization on FashionVideo dataset.

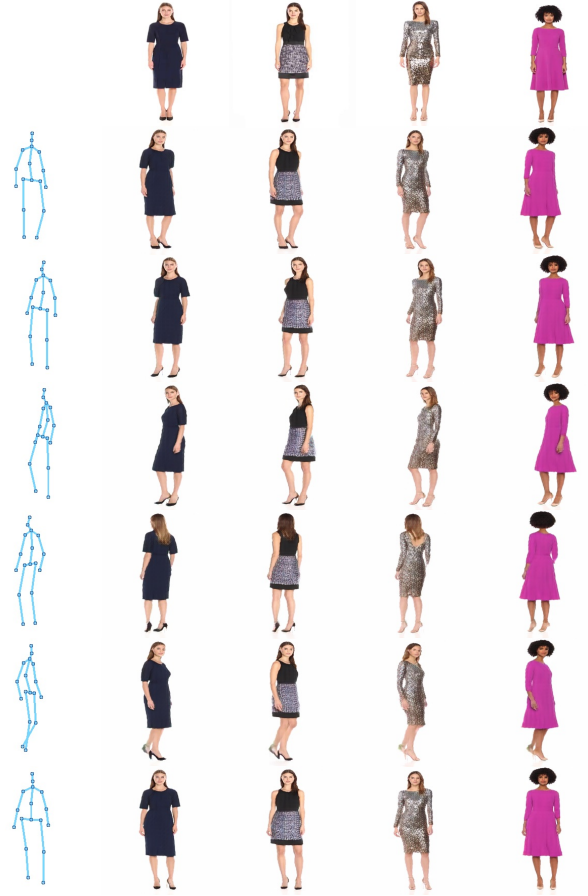
	NTED	NTED_F	Ours	Ours_F
FID ↓	31.24	21.74	28.65	19.57
clean-FID ↓	32.85	22.76	30.09	20.58
LPIPS ↓	0.1688	0.1252	0.1501	0.1182
PSNR ↑	18.49	21.13	19.49	21.42

[63], for the source image in different resolutions, the transport matrices computed by NTED may lose details (in the second and the third rows) and generate some artifacts (in the first and the fourth rows), leading to unsatisfactory results. On the contrary, the corresponding transport matrix computed in our IT block can deform that to the target pose properly and preserve the structure details, which indicates the correspondences are reliable.

4.4.3. Generalization on FashionVideo dataset

To evaluate the generalization performance on other datasets, we randomly select 20 images from each test video of FashionVideo test set [79] as our evaluation data, resulting in a total of 2000 images. Table 6 gives the quantitative results compared with the latest work, NTED [63], using the same image-specific optimization described in Section 3.4, denoted as NTED_F. Comparing NTED and Ours in Table 6, it can be seen that our model achieves superior performance without image-specific optimization, indicating its stronger generalization ability. Moreover, with the image-specific optimization, our model still has better performance compared Ours_F with NTED_F, and improves the performance of the original model without image-specific optimization.

Figure 7 shows some visual results compared with NTED. The first column shows the source image we used as the in-

**Fig. 8.** Qualitative results of image-based animation. The first row shows the source image, and the first column presents the target pose. The rest of the images are generated by our method.

put during optimization. The second column shows the target poses. Note that the poses are also from FashionVideo dataset, which are unseen poses of our model. The other columns present the results generated by different models. The third and the fifth columns show the results generated by NTED and ours without optimization. It can be seen that our model can generate more similar texture and clothes style, which indicates that our model has better performance on generalization. The fourth and sixth columns show the results generated by NTED and ours with optimization. Both the results of NTED_F and ours_F have better performance. However, the images generated by our model have fewer artifacts and more consistent texture and clothes style. More results can be found in the supplementary materials.

4.5. Applications

To demonstrate the versatility of our model, we applied our

model to image-based animation [16] and exemplar-based face image generation [80].

4.5.1. Image-based animation.

In image-based animation, the objective is to generate a video that combines the content of an input image with the movements from a driving sequence. To achieve this, we transfer the content of the input image to the driving sequence by our proposed IT block. The architecture of our model is similar to our human pose transfer model. To generate smooth sequences, we have made two improvements. Firstly, in addition to the input image, we also consider the previously generated frame as a reference, and add their features from the IT block. Secondly, we introduce an additional temporal discriminator alongside the original discriminator. By doing so, our model can generate smooth and continuous results. We train our model using the FashionVideo dataset [79], which consists of 500 training videos and 100 testing videos as described in [16]. Figure 8 shows some qualitative results. The first row displays the input images, and the first column shows the driving skeleton sequence. Our model is capable of generating plausible results, demonstrating the effectiveness of our proposed IT block for animation tasks. Dynamic results can be found in the supplementary video.

4.5.2. Exemplar-based face image generation.

In this application, an exemplar image is utilized to control the style of the generated image. Given a source image with its corresponding edge map E_s and an exemplar image I_e , the objective is to generate a target image that preserves the semantic structures of E_s while adopting the style of similar style elements present in I_e . Similar to human pose transfer, the core problem is also how to model the correspondences between the exemplar image and target image based on the edge condition, which transforms the style of the exemplar image to the edge condition. The architecture of the model employed in this task is the same as our model for human pose transfer. We train and test our model using the widely-used face dataset, CelebA-HQ [81], which comprises 30,000 facial photos. Following [80], we take 27,000 samples for training and the remaining 3,000 sam-



Fig. 9. Qualitative results of exemplar-based face image generation. The first row displays the exemplar image, while the first column presents the source image alongside its corresponding edge map. The remaining images are the results produced by our method.



Fig. 10. Failure cases generated by our method.

ples for testing. Figure 9 shows the results of exemplar-based face image generation. The first row shows the exemplar images and the first column shows the source images with their corresponding edges. It can be seen that our model consistently transfers the style of the exemplar images to the source images. Additional results can be found in the supplementary materials.

5. Conclusion and discussion

Conclusion. In this paper, we propose a novel multi-scale information transport GAN to establish and refine the correspondences between the source image and the target image. Specifically, we propose the information transport block by leveraging the optimal transport solver to establish correspondences and refine the transformed features using the following information transport blocks, which helps to obtain more accurate correspondences. We also present an image-specific optimization to improve the performance for the out-of-distribution images. The quantitative and qualitative comparison results show the effectiveness of our proposed model. Besides, we conduct ablation studies to validate the effectiveness of our proposed IT block and our proposed model. Both quantitative and qualitative results show the effectiveness of our new formulation and the other designs. Furthermore, there are no tunable parameters in the optimal transport solver, which indicates that our model has better performance for out-of-distribution images. We also give two applications of image-based animation and exemplar-based face image generation to validate the versatility of our model.

Limitations. Although our model generates promising results in most cases, there are still some failure cases that are difficult to deal with. See Figure 10 for some examples. Our model can build reliable correspondences, but for some specific patterns, our model fails to generate sharp images with specific patterns. To analyze the reasons behind these cases, we visualize the warped images using correspondences in different resolutions of the IT block and the source image. It can be seen that low resolutions pose difficulties in deforming details with specific patterns or text. As described in Section 3, we currently employ IT blocks only in resolutions of 16×16 , 32×32 , and 64×64 due to hardware limitations. In the future, we plan to explore lightweight structures for IT blocks and investigate few-shot learning methods to improve these specific cases.

6. Acknowledgments

This work was supported in part by National Key R&D Program of China (2023YFC3082100), National Natural Science Foundation of China (62171317 and 62122058), and Science Fund for Distinguished Young Scholars of Tianjin (No. 22JCJQC00040).

References

- [1] X. Li, J. Zhang, Y. Liu, Speech driven facial animation generation based on GAN, *Displays* 74 (2022) 102260.
- [2] J. Zhang, Y. Liu, X. Li, W. Li, Y. Tang, Talking face generation driven by time-frequency domain features of speech audio, *Displays* 80 (2023) 102558.
- [3] Y. Li, Z. Yu, Y. Zhu, B. Ni, G. Zhai, W. Shen, Skeleton2Humanoid: Animating simulated characters for physically-plausible motion in-betweening, in: *ACM Int. Conf. Multimedia*, 2022, pp. 1493–1502.
- [4] C. Yang, S.Y. Yao, Z.W. Zhou, B. Ji, G. Zhai, W. Shen, Poxture: Human posture imitation using neural texture, *IEEE Trans. Circuit Syst. Video Technol.* 32 (2022) 8537–8549.
- [5] J. Zhang, K. Li, Y.K. Lai, J. Yang, PISE: Person image synthesis and editing with decoupled GAN, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 7982–7990.
- [6] Y. Men, Y. Mao, Y. Jiang, W.Y. Ma, Z. Lian, Controllable person image synthesis with attribute-decomposed GAN, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 5084–5093.
- [7] K. Li, J. Zhang, Y. Liu, Y.K. Lai, Q. Dai, PoNA: Pose-guided non-local attention for human pose transfer, *IEEE Trans. Image Process.* 29 (2020) 9584–9599.
- [8] Q. Zhou, H. Fan, H. Yang, H. Su, S. Zheng, S. Wu, H. Ling, Robust and efficient graph correspondence transfer for person re-identification, *IEEE Trans. Image Process.* 30 (2019) 1623–1638.
- [9] Q. Xu, H. Yang, L. Chen, G. Zhai, Group re-identification with hybrid attention model and residual distance, in: *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 1217–1221.
- [10] E. Ning, C. Zhang, C. Wang, X. Ning, H. Chen, X. Bai, Pedestrian Re-ID based on feature consistency and contrast enhancement, *Displays* (2023) 102467.
- [11] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, X. Bai, Progressive pose attention transfer for person image generation, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 2347–2356.
- [12] J. Zhang, X. Liu, K. Li, Human pose transfer by adaptive hierarchical deformation, *Comput. Graph. Forum* 39 (2020) 325–337.
- [13] H. Tang, S. Bai, P.H. Torr, N. Sebe, Bipartite graph reasoning GANs for person image generation, in: *Proc. Brit. Mach. Vis. Conf.*, 2020, pp. 1–13.
- [14] H. Tang, S. Bai, L. Zhang, P.H. Torr, N. Sebe, XingGAN for person image generation, in: *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 717–734.
- [15] A. Siarohin, S. Lathuilière, E. Sangineto, N. Sebe, Appearance and pose-conditioned human image generation using deformable GANs, *IEEE Trans. Pattern Anal. Mach. Intell.* (2019).
- [16] Y. Ren, G. Li, S. Liu, T.H. Li, Deep spatial transformation for pose-guided person image generation and animation, *IEEE Trans. Image Process.* 29 (2020) 8622–8635.
- [17] W. Liu, Z. Piao, M. Jie, W. Luo, L. Ma, S. Gao, Liquid Warping GAN: A unified framework for human motion imitation, appearance transfer and novel view synthesis, in: *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5904–5913.
- [18] L. Ma, K. Huang, D. Wei, Z.Y. Ming, H. Shen, FDA-GAN: Flow-based dual attention GAN for human pose transfer, *IEEE Trans. Multimedia* (2021).
- [19] J. Liu, Y. Zhao, S. Chen, Y. Zhang, A 3D mesh-based lifting-and-projection network for human pose transfer, *IEEE Trans. Multimedia* (2021).
- [20] R. Jain, K.K. Singh, M. Hemani, J. Lu, M. Sarkar, D. Ceylan, B. Krishnamurthy, VGFlow: Visibility guided flow network for human reposing, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 21088–21097.

- [21] Z. Lv, X. Li, X. Li, F. Li, T. Lin, D. He, W. Zuo, Learning semantic person image generation by region-adaptive normalization, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 10806–10815.
- [22] G. Peyré, M. Cuturi, et al., Computational optimal transport, Center for Research in Economics and Statistics Working Papers (2017).
- [23] G. Avraham, Y. Zuo, T. Drummond, Parallel optimal transport GAN, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 4411–4420.
- [24] T. Salimans, H. Zhang, A. Radford, D. Metaxas, Improving GANs using optimal transport, in: *Proc. Int. Conf. Learn. Represent.*, 2018, p. 1.
- [25] R. Geng, Y. Hu, Z. Lu, C. Yu, H. Li, H. Zhang, Y. Chen, Passive non-line-of-sight imaging using optimal transport, *IEEE Trans. Image Process.* 31 (2021) 110–124.
- [26] R. Sinkhorn, Diagonal equivalence to matrices with prescribed row and column sums, *The American Mathematical Monthly* 74 (1967) 402–405.
- [27] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A.C. Courville, Y. Bengio, Generative adversarial nets, in: *Proc. Adv. Neural Inform. Process. Syst.*, volume 27, 2014, pp. 1–9.
- [28] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, *IEEE Trans. Pattern Anal. Mach. Intell.* 43 (2021) 4217–4228.
- [29] S. Wu, Y. Yan, Y. Li, Y. Cheng, W. Zhu, K. Gao, X. Li, G. Zhai, GAN-Head: Towards generative animatable neural head avatars, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 437–447.
- [30] Z. Che, A. Borji, G. Zhai, X. Min, G. Guo, P. Le Callet, GazeGAN: A generative adversarial saliency model based on invariance analysis of human gaze during scene free viewing, *arXiv preprint arXiv:1905.06803* (2019).
- [31] X. Zheng, X. Yang, Q. Zhao, H. Zhang, X. He, J. Zhang, X. Zhang, CfaGAN: Cross fusion attention and frequency loss for image style transfer, *Displays* 81 (2024) 102588.
- [32] M. Mirza, S. Osindero, Conditional generative adversarial nets, *arXiv preprint arXiv:1411.1784* (2014).
- [33] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.
- [34] P. Isola, J.Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 1125–1134.
- [35] T.C. Wang, M.Y. Liu, J.Y. Zhu, A. Tao, J. Kautz, B. Catanzaro, High-resolution image synthesis and semantic manipulation with conditional GANs, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 8798–8807.
- [36] X. Huang, S. Belongie, Arbitrary style transfer in real-time with adaptive instance normalization, in: *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1501–1510.
- [37] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 8110–8119.
- [38] T. Park, M.Y. Liu, T.C. Wang, J.Y. Zhu, Semantic image synthesis with spatially-adaptive normalization, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 2337–2346.
- [39] Z. Zhu, Z. Xu, A. You, X. Bai, Semantically multi-modal image synthesis, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 5467–5476.
- [40] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, L. Van Gool, Pose guided person image generation, in: *Proc. Adv. Neural Inform. Process. Syst.*, volume 30, 2017, p. 1.
- [41] P. Esser, E. Sutter, B. Ommer, A variational u-net for conditional appearance and shape generation, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 8857–8866.
- [42] D.P. Kingma, M. Welling, Auto-encoding variational bayes, in: *Proc. Int. Conf. Learn. Represent.*, 2014, p. 1.
- [43] Y. Li, C. Huang, C.C. Loy, Dense intrinsic appearance flow for human pose transfer, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3693–3702.
- [44] Y.T. Sun, Q.C. Fu, Y.R. Jiang, Z. Liu, Y.K. Lai, H. Fu, L. Gao, Human motion transfer with 3d constraints and detail enhancement, *IEEE Trans. Pattern Anal. Mach. Intell.* (2022).
- [45] P. Zhu, R. Abdal, Y. Qin, P. Wonka, SEAN: Image synthesis with semantic region-adaptive normalization, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 5104–5113.
- [46] Z. Zheng, H. Huang, T. Yu, H. Zhang, Y. Guo, Y. Liu, Structured local radiance fields for human avatar modeling, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 15893–15903.
- [47] F. Zhao, Y. Jiang, K. Yao, J. Zhang, L. Wang, H. Dai, Y. Zhong, Y. Zhang, M. Wu, L. Xu, et al., Human performance modeling and rendering via neural animated mesh, *ACM Trans. Graph.* 41 (2022) 1–17.
- [48] H. Zhao, J. Zhang, Y.K. Lai, Z. Zheng, Y. Xie, Y. Liu, K. Li, High-fidelity human avatars from a single rgb camera, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 15904–15913.
- [49] M. Wu, Y. Wang, Q. Hu, J. Yu, Multi-view neural human rendering, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 1682–1691.
- [50] F. Zhao, W. Yang, J. Zhang, P. Lin, Y. Zhang, J. Yu, L. Xu, Humannerf: Efficiently generated human radiance field from sparse inputs, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 7743–7753.
- [51] X. Li, J. Zhang, Y.K. Lai, J. Yang, K. Li, High-quality animatable dynamic garment reconstruction from monocular videos, *IEEE TCSVT* (2023).
- [52] Y. Ren, Y. Wu, T.H. Li, S. Liu, G. Li, Combining attention with flow for person image synthesis, in: *ACM Int. Conf. Multimedia*, 2021, pp. 3737–3745.
- [53] L. Chen, Z. Gan, Y. Cheng, L. Li, L. Carin, J. Liu, Graph optimal transport for cross-domain alignment, in: *International Conference on Machine Learning*, 2020, pp. 1542–1553.
- [54] R. Flamary, N. Courty, D. Tuia, A. Rakotomamonjy, Optimal transport for domain adaptation, *IEEE Trans. Pattern Anal. Mach. Intell.* (2016).
- [55] N. Bonneel, J. Digne, A survey of optimal transport for computer graphics and computer vision, *Comput. Graph. Forum* 42 (2023) 439–460.
- [56] B. Taşkesen, S. Shafieezadeh-Abadeh, D. Kuhn, Semi-discrete optimal transport: Hardness, regularization and numerical solution, *Mathematical Programming* 199 (2023) 1033–1106.
- [57] G. Puy, A. Boulch, R. Marlet, Flot: Scene flow on point clouds guided by optimal transport, in: *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 527–544.
- [58] C. Song, J. Wei, R. Li, F. Liu, G. Lin, 3D pose transfer with correspondence learning and mesh refinement, *Proc. Adv. Neural Inform. Process. Syst.* (2021).
- [59] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: *International Conference on Machine Learning*, 2017, pp. 214–223.
- [60] F. Zhan, Y. Yu, K. Cui, G. Zhang, S. Lu, J. Pan, C. Zhang, F. Ma, X. Xie, C. Miao, Unbalanced feature transport for exemplar-based image translation, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 15028–15038.
- [61] Y. Liu, X. Ouyang, T. Jiang, H. Ding, X. Cui, Optimal transport-based unsupervised semantic disentanglement: A novel approach for efficient image editing in GANs, *Displays* 80 (2023) 102560.
- [62] Z. Cao, T. Simon, S.E. Wei, Y. Sheikh, Realtime multi-person 2D pose estimation using part affinity fields, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 7291–7299.
- [63] Y. Ren, X. Fan, G. Li, S. Liu, T.H. Li, Neural texture extraction and distribution for controllable person image synthesis, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 13535–13544.
- [64] G. Monge, Mémoire sur la théorie des déblais et des remblais, *Mem. Math. Phys. Acad. Royale Sci.* (1781) 666–704.
- [65] L.V. Kantorovich, On the transfer of masses, *Journal of mathematical sciences* 133 (2006) 1381–1382.
- [66] G. Peyré, M. Cuturi, Computational optimal transport: With applications to data science, *Foundations and Trends in Machine Learning* 11 (2019) 355–607.
- [67] M. Cuturi, Sinkhorn distances: Lightspeed computation of optimal transport, in: *Proc. Adv. Neural Inform. Process. Syst.*, volume 26, 2013, pp. 1–9.
- [68] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 8110–8119.
- [69] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.
- [70] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.
- [71] P. Zhang, L. Yang, J.H. Lai, X. Xie, Exploring dual-task correlation for pose guided person image generation, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 7713–7722.
- [72] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in:

- Proc. Int. Conf. Learn. Represent., 2015, pp. 1–15.
- [73] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training GANs, in: Proc. Adv. Neural Inform. Process. Syst., volume 29, 2016, pp. 1–9.
 - [74] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, DeepFashion: Powering robust clothes recognition and retrieval with rich annotations, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2016, pp. 1096–1104.
 - [75] S.Y. Cheong, A. Mustafa, A. Gilbert, UPGPT: Universal diffusion model for person image generation, editing and pose transfer, arXiv preprint arXiv:2304.08870 (2023).
 - [76] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 586–595.
 - [77] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs trained by a two time-scale update rule converge to a local nash equilibrium, in: Proc. Adv. Neural Inform. Process. Syst., volume 30, 2017, pp. 1–12.
 - [78] G. Parmar, R. Zhang, J.Y. Zhu, On aliased resizing and surprising subtleties in GAN evaluation, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2022, pp. 11410–11420.
 - [79] P. Zablotskaia, A. Siarohin, B. Zhao, L. Sigal, DwNet: Dense warp-based network for pose-guided human video generation, in: Proc. Brit. Mach. Vis. Conf., 2019, pp. 1–13.
 - [80] C. Jiang, F. Gao, B. Ma, Y. Lin, N. Wang, G. Xu, Masked and adaptive transformer for exemplar based image translation, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2023, pp. 22418–22427.
 - [81] C.H. Lee, Z. Liu, L. Wu, P. Luo, MaskGAN: Towards diverse and interactive facial image manipulation, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2020, pp. 5549–5558.