# Deep Spatial Transformation For Pose-Guided Person Image Generation and Animation

03CS6901 Seminar I

08/MCS/2020 CHN20CSIP03 Praveena K M
M. Tech. Computer Science & Engineering
(Image Processing)

Department of Computer Engineering
College of Engineering Chengannur
Alappuzha 689121
Phone: +91.479.2165706
http://www.ceconline.edu
hodcs@ceconline.edu

# College of Engineering Chengannur
# Dept. of Computer Engineering

## CERTIFICATE

This is to certify that, this report titled ***Deep spatial Transformation For Pose-Guided Person Image Generation and Animation*** is a bonafide record of the **03CS6901 Seminar I** presented on March 16, 2021 by

**08/MCS/2020 CHN20CSIP03 Praveena K M**

First Semester M. Tech. Computer Science & Engineering (Image Processing )

scholar, under our guidance and supervision, in partial fulfillment of the requirements for the award of the degree, **M. Tech. Computer Science & Engineering (Image Processing)** of **APJ Abdul Kalam Technological University**.

Guide

Radhu Krishna
Asst. Professor
in Computer Engineering

Coordinator

Ahammed Siraj K K
Associate Professor
in Computer Engineering

Head of the Department

March 16, 2021

Dr. Smithadharan
Professor
in Computer Engineering

## Acknowledgments

First of all, I am indebted to the GOD ALMIGHTY for giving me an opportunity to excel in my efforts to complete this seminar on time.

I express my sincere thanks to **Dr. Jacob Thomas V.**, Principal, College of Engineering Chengannur for extending all the facilities required for doing my seminar. My heartfelt words of gratitude to **Dr. Smitha Dharan**, Professor and Head of Department of Computer Engineering, for providing constant support.

Now I express my gratitude to my seminar co-ordinator **Mr. Ahammed Siraj K K**, Associate Professor in Computer Engineering and my seminar guide **Ms. Radhu krishna**, Assistant Professor in Computer Engineering who played a great role for valuable suggestions and expert guidance.

Praveena K M

# Abstract

Human pose transfer (HPT) is an emerging research topic, aiming at synthesizing person images under new target poses with respect to the appearance of a given source , with huge potential in fashion design, media production, online advertising and virtual reality. Pose guided person image generation and animation aims to transferring the pose of a given person to a target poses. However, existing human pose transfer methods often suffer from detail deficiency, content ambiguity and style inconsistency,which severely degrade the visual quality and realism of generated images.These are the condition generation tasks, which requires spatial manipulation of source data.However, Convolutional Neural Networks are limited by the lack of ability to spatially transform the inputs. Here,the proposed model,a differentiable global-flow local-attention framework, able to reassemble the inputs at the feature level. This framework first estimates global flow fields between sources and targets. Then, corresponding local source feature patches are sampled with content-aware local attention coefficients. It shows that this framework can spatially transform the inputs in an efficient manner. Meanwhile, it can be further model the temporal consistency for the person image animation task to generate coherent videos. The experiment results of both image generation and animation tasks demonstrate the superiority of this model. Besides, additional results of novel view synthesis and face image animation show that this model is applicable to other tasks requiring spatial transformation.

# Contents

# Chapter 1

# Introduction

Pose-guided person image generation and animation aim to transform a source person image to target poses. These tasks require spatial manipulation of source data. Aiming at synthesizing person images under new target poses with respect to the appearance of a given source image, HPT contains huge potential in empowering numerous creative applications, such as automatic fashion design, creative media production, online advertising and virtual reality.Pose transferring is a condition generation task,where the target images are the spatial deformation versions of the source images. Such deformation can be caused by object motions or viewpoint changes. This task is the core of many image/video generation problems.these tasks can be tackled by reasonably reassembling the input data in the spatial domain.

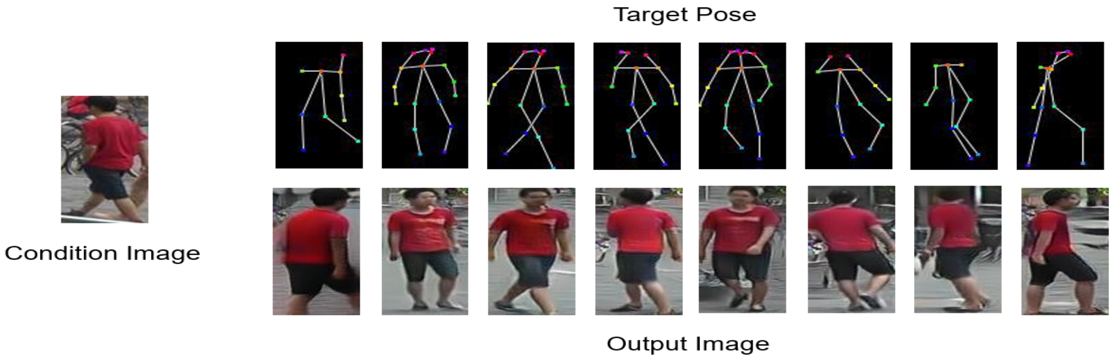

Figure 1: Illustration of pose transfer

However, Convolutional Neural Networks (CNNs) lack the ability to spatially transform the input features in a parameter efficient manner. One important property of CNNs is the equivariance to transformation , which means that if the input spatially shifts, then the output shifts in the same way. However, it limits the networks by the lack of ability to deal with the deformable-object generation task which requires spatially rearranging the input data. In order to enable spatial transformation capabilities of CNNs, Spatial Transformer Networks

(STN) [6] introduces a Spatial Transformer module to standard neural networks. This module regresses transformation parameters and warps the input features using a global affine transformation. However, the global affine transformation is not sufficient in representing the complex deformations of non-rigid objects.

The attention mechanism is able to transform information beyond local regions. It gives networks the ability to build long-term dependencies by allowing networks to use non-local features. However, for spatial transformation tasks in which target images are the deformation results of source images, each output position has a clear one-to-one relationship with the source position. Therefore, each output feature is only related to a local region of the source features. i.e., the attention coefficient matrix between the source and target should be a sparse matrix instead of a dense matrix.Flow-based operation forces the attention coefficient matrix to be a sparse matrix by sampling a very local source patch for each output position. It predicts 2D coordinate offsets for the target features specifying the sampling source positions. However, networks struggle to find reasonable sampling locations when warping the inputs at the feature level.

The proposed differentiable Global-Flow Local-Attention (GFLA) framework, enable CNNs to reasonably sample and reassemble source features without using any labeled flow fields.This network can be divided into two parts: Global Flow Field Estimator and Local Neural Texture Renderer. The Global Flow Field Estimator is responsible for extracting the long-term dependencies between sources and targets. It estimates flow fields that assign a local source feature patch for each target position. The Local Neural Texture Renderer uses the extracted flow fields to sample the vivid source neural textures.The image-based pose transformation can be further extended for the pose animation task by coherently rendering an input skeleton video.The temporal consistency is further modeled for the person image animation task. To extract clean skeleton sequences from the corresponding noise data,a Motion Extraction Network is proposed. Meanwhile, an improved GFLA model to generate video clips in a recurrent manner. It allows the model to explicitly extract the correlations between adjacent frames.

# Chapter 2

# Literature Survey

1. **Pose guided person image generation,** [2] in Advances in Neural Information Processing Systems, 2017, pp. 406–416

   Here,presents the novel Pose Guided Person Generation Network (PG2 ) that allows to synthesize person images in arbitrary poses, based on an image of that person and a novel pose. Pose Guided Person Generation Network (PG2 ) that consist of two stages, pose integration and image refinement.In the first stage,pose integration stage in which the condition image and the target pose are fed into a U-Net-like network to generate an initial but coarse image of the person with the target pose. The second stage which is the image refinement stage ,it then refines the initial and blurry result by training a U-Net-like generator in an adversarial way.

   At stage-I, a variant of U-Net is employed to integrate the target pose with the person image. It outputs a coarse generation result that captures the global structure of the human body in the target image. A masked L1 loss is proposed to suppress the influence of background change between condition image and target image. However, it would generate blurry result due to the use of L1. At stage-II, a variant of Deep Convolutional GAN (DCGAN) model is used to further refine the initial generation result. The model learns to fill in more appearance details via adversarial training and generates sharper images. Different from the common use of GANs which directly learns to generate an image from scratch, in this work w, the GAN is trained to generate a difference map between the initial generation result and the target person image. The training converges faster since it is an easier task. Besides, a masked L1 loss is added to regularize the training of the generator such that it will not generate an image with many artifacts.

   At stage-I, the encoder of generator G1 consists of N residual blocks and one fully-connected layer , where N depends on the size of input. Each residual block consists of two convolution layers with stride=1 followed by one sub-sampling convolution layer with stride=2 except the last block. At stage-II, the encoder of generator G2 has a fully convolutional architecture including N-2 convolution blocks. Each block consists of two convolution layers with stride=1 and one sub-sampling convolution layer with stride=2. Decoders in both G1 and G2 are symmetric to corresponding encoders. Besides, there are shortcut connections between decoders and encoders, . In G1 and G2, no batch normalization or dropout are applied. All convolution layers consist of 3×3 filters and the number of filters are increased linearly with each block. they apply rectified linear unit (ReLU) to each layer except the fully connected layer and the output convolution layer. For the discriminator, we adopt the same network architecture as DCGAN except the size of the input convolution layer due to different image resolution.Experiments on two dataset, a low-resolution person re-identification dataset and a high-resolution fashion photo dataset, demonstrate the effectiveness of the proposed method.This model ignores the spatial distribution of the original appearance, which limits the network to generate complex textures.

2. **Progressive pose attention transfer for person image generation**,[4] in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2347–2356

   In contrast to the one-step transfer scheme ,here adopted a scheme to transfer a condition pose by transferring through a sequence of intermediate pose representations before reaching the target. The transfer is carried out by a sequence of Pose Attentional Transfer Blocks (PATBs) .The generator of the network comprises a sequence of Pose-Attentional Transfer Blocks that each transfers certain regions it attends to, generating the person image progressively.On the input side, the condition image is encoded by N down-sampling convolutional layers (N = 2 in this case). The condition pose S and target pose heat maps are stacked along their depth axes before being encoded, also by N downsampling convolutional layers.Inside the block, there is an attention mechanism that infers the regions of interest based on the human pose. The block outputs the updated image and pose representations, so that such blocks can be cascaded in sequence to form a PAT network (PATN). The proposed network exhibits superior performance both qualitatively and quantitatively on challenging benchmarks, and substantially augments person dataset for person re-identification application.However, information may be lost during multiple transfers, which may result in blurry details.

3. **Synthesizing Images Of Humans In Unseen Poses** [6] in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2018

   Here adopted a modular generative neural network that synthesizes unseen poses using training pairs of images and poses taken from human action videos.The key idea is to decompose this complex problem into simpler, modular subtasks, trained jointly as one generative neural network.The network, split into four modular subtasks.These subtasks, implemented with separate modules, are trained jointly using only a single target image as a supervised label.First ,the source segmentation module separates the person's body parts from the background,Second is the spatial transformation module which spatially moves the body parts to target locations Next is the foreground synthesis module which then fuses body parts into a coherent foreground .Fourth subtask is background hole-filling where the parts of the image disoccluded by the body are filled in with realistic texture Finally, the foreground and background are composited to produce an output image . Designs the network such that these modules are learned jointly and trained using only the target image as a label This network separates a scene into different body part and background layers, moves body parts to new locations and refines their appearances, and composites the new foreground with a hole-filled background .. The layering approach decouples the foreground and background synthesis tasks, helps to synthesize better backgrounds and by segmenting the foreground into body parts, generate complex movements. The strategy of decomposing a network into modular subtasks has proven useful in recent learning models for visual reasoning.Here uses an adversarial discriminator to force the network to synthesize realistic details conditioned on pose.

4. **Generating long sequences with sparse transformers** ,[9] Corpus ID: 129945531, arXiv:1904.10509, 2019, [online] Available: http://arxiv.org/abs/1904.10509

Here developed the Sparse Transformer, a deep neural network which sets new records at predicting what comes next in a sequence—whether text, images, or sound. It uses an algorithmic improvement of the attention mechanism to extract patterns from sequences 30x longer than possible previous existing methods. Introduced sparse transformers which can separate the full attention operation across several steps of attention. For each step, only a subset of input positions is attended for calculation.Sparse transformers attain better performance than dense attention with significantly fewer operations . The Sparse Transformer method utilizes an improved algorithm based on the attention mechanism, which can predict a length 30 times longer than the previous maximum. Even computing a single attention matrix, however, can become impractical for very large inputs. It use sparse attention patterns, where each output position only computes weightings from a subset of input positions. Sparse Transformer reduces the computational complexity of the traditional attention mechanism model and can be applied directly to different data types. It can be used to model sequences with more than tens of thousands of elements. While many layers displayed sparse structure, some layers clearly display dynamic attention that stretch over the entirety of the image. In order to preserve the ability of this network to learn such patterns, it implemented a two-dimensional factorization of the attention matrix, where the network can attend to all positions through two steps of sparse attention.i.e. strided attention and fixed attention.The strided attention, is roughly equivalent to each position attending to its row and its column. The column attention can be equivalently formulated as attending to the row of the transposed matrix .The fixed attention, attends to a fixed column and the elements after the latest column element, a pattern that found useful for when the data didn't fit into a two dimensional structure (like text). Implementing the sparse attention would involve slicing query and key matrices in blocks, so to ease experimentation they implemented a set of block-sparse kernels which efficiently perform these operations on the the GPU.Sparse Transformer achieves lower errors and faster training speeds than Transformer with full attention. This new method can be qualitatively evaluated in image completion tasks. It can also be used to generate raw audio by simply changing position embeddings.

5. **3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training** [7] ,in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7753-7762

Introduced a method, focusing on 3D human pose estimaation in video based on the dilated temporal convolutions applied on 2D keypoints . 2D keypoints can be obtained using any person keypoint detector, but here,Mask RCNN with ResNet-101 backbone, is used as keypoint detector.The poses are presented as 2D keypoint coordinates in contrast to using heatmaps (i.e. Gaussian operation applied at the keypoint 2D location). Thus, 1D convolutions over the time series are applied, instead of 2D convolutions over heatmaps. The model is a fully convolutional architecture with residual connections that takes a sequence of 2D pose (concatenated (x,y) coordinates of the joints in each frame ) as input and transforms them through temporal convolutions.There is a Slice layer in the residual connection which performs padding (or slicing) the sequence with replicas of boundary frames (to both left and right) to match the dimensions with the main block as zero-padding is not used in the convolution operations. The 3D pose estimation is a difficult task particularly due to the limited data available online. Therefore, here adopted a semi-supervised approach of training the 2D for 3D pose estimation by exploiting unlabeled video.Specifically, 2D keypoints are detected in the unlabeled video with any keypoint detector, then 3D keypoints are predicted from them and these 3D points are reprojected back to 2D (camera intrinsic parameters are required).In the semi-supervised part training penalizes when the reprojected 2D keypoints are far from the original input. Weighted mean per-joint position error (WMPJPE) loss is weighted by the inverse of the depth to the object (since far objects should contribute less to the training than close ones) is used as the optimization goal. Basically, the semi-supervised approach becomes more effective when less labeled data is available.

# Chapter 3

# GFLA for Pose-Guided Person Image Generation and Animation

The human pose transferred image generation is a conditional generation task where the target images are the spatial deformation versions of the source images. Such deformation can be caused by object motions or viewpoint changes. This task is the core of many image/video generation problems.In order to overcome the limitations of the previous works,a global flow local attention framework is inttroduced to efficiently warp and reassemble source neural textures at feature level in a global predictive manner. The global flow local attention( *GFLA*) framework for person image generation by pose transfer,can be divided into two parts:Global Flow Field Estimator and Local Neural Texture Renderer. Image based pose transformation can be further extended for the pose animation task by coherently rendering an input skeleton video. A Motion Extraction Network is used to extract clean skeleton sequences from the corresponding noise data. Meanwhile, the improved GFLA model ,i.e, Sequential-GFLA is able to generate video clips in a recurrent manner.

## 3.1 Global Flow Local Attention for Person Image Generation

For the pose-guided person image generation task, target images are the deformation versions of source images. Therefore, target images can be generated by spatially transforming the source images. Here describes a GFLA model to efficiently warp and reassemble source neural textures. The architecture of this model can be found in Figure . GFLA can be divided into two modules: Global Flow Field Estimator F and Local Neural Texture Renderer G. The Global Flow Field Estimator is responsible for estimating the global motions between sources and targets. Flow fields w and occlusion masks m are estimated by this module. The Local Neural Texture Renderer renders the target images with vivid source features using the local attention blocks. Here describes the network as a single local attention block. As shown in Figure 3.1.1 , and this model can be extended to use multiple attention blocks at different scales.

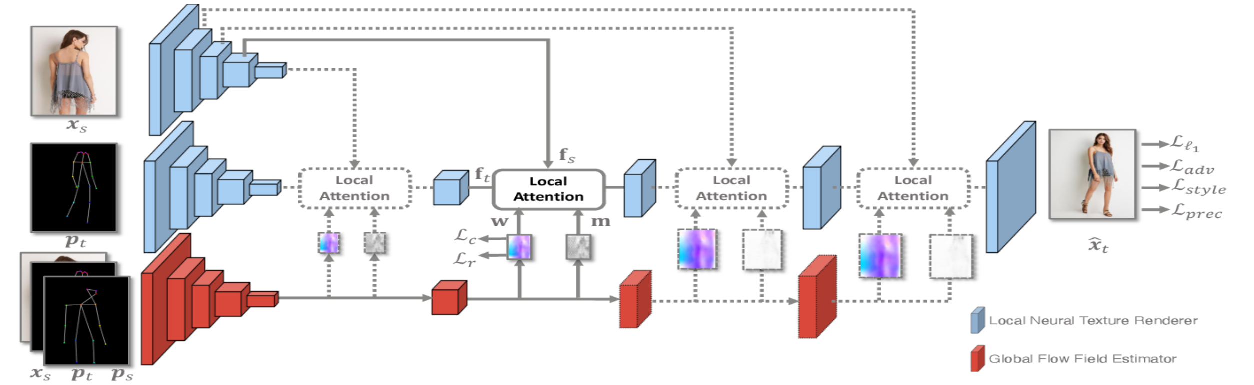

Figure 3.1.1: overview of proposed GFLA framework

### 3.1.1 Global Flowfield Estimator

The 18-channel heat map that encodes the locations of 18 joints of a human body is used as the structure guidance. Following the previous works , the human body joints are detected by the Human Pose Estimator [3].The global flow field estimator is used for extracting the long-term dependencies between sources and targets and also predicts flow fields and occlusion mask. The flowfield assign each target position with source local patch. It contains motion between source and target.The occlusion mask indicate whether the information of a target position exist in the source.
Let $p_s$ and pt denote the structure guidance of the source image $x_s$ and the target image $x_t$ respectively. The Global Flow Field Estimator F takes $x_s$, $p_s$, and $p_t$ as inputs and generates the flow fields w and occlusion masks m.

$$\text{w, m} = \text{F}(\text{x}_s, p_s, p_t) \tag{1}$$

where the flow fields w assign a source patch for each target location. The occlusion masks m with continuous values between 0 and 1 indicate whether the flowed source patches can be used to generate targets. We design F as a fully convolutional network. w and m share all weights of F other than their output layers

Warping sources at the feature level can help the model to to generate new content. Meanwhile, it relaxes the requirements of the flow field estimation since the resolutions of the generated flow fields are reduced. However,the networks may struggle to find reasonable sampling positions. An important reason is that the gradient propagation of the input features and flow fields are mutually constrained during the warping operation. The input features cannot obtain correct gradients without reasonable flow fields and vice versa. Therefore, additional losses are used to help with the training. They proposed a sampling correctness loss to constrain w in a self-supervised manner. The sampling correctness loss calculates the similarity between the warped source feature and the ground truth target feature at the VGG feature level. Let $v_s$ and $v_t$ denote the features generated by a specific layer of VGG19. $v_{s,w} = \text{w}(v_s)$ is the warped results of the source feature $v_s$ using w. The sampling correctness loss calculates the relative cosine similarity between $v_{s,w}$ and $v_t$

$$L_c = \frac{1}{N} \sum_{l \epsilon \Omega} \exp = \left( \frac{\mu\left(v_{s,w}^l, v_t^l\right)}{\mu_{max}^{'}} \right) \tag{2}$$

where $\mu(*)$ denotes the cosine similarity. Coordinate set $\Omega$ contains all N positions in the feature maps. $v_{s,w}^l$ and $v_t^l$ denote the features of $v_{s,w}$ and $v_t$ located at the coordinate l = (x, y). The normalization term $\mu_{max}^l$ is used to avoid the bias brought by occlusion. It represents the similarity between $v_t^l$ and its most similar feature in the source feature map $v_s$

### 3.1.2 Local Neural Texture Render

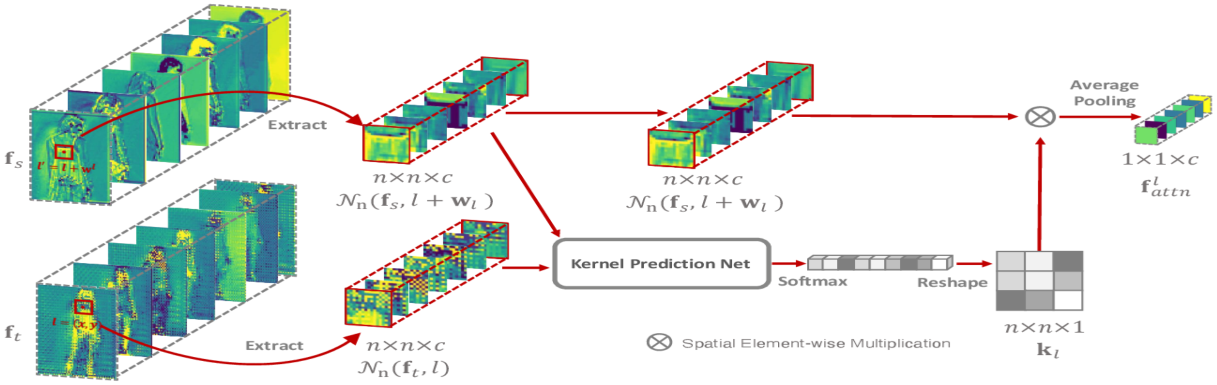

Figure 3.1.2: local attention module in local nueral texture renderer

The Local Neural Texture Renderer G ,is responsible for generating the result images by rendering target poses with the source neural textures. It takes$x_s, p_t$, w, and m as inputs and generates the result image $\hat{X}_t$.

$$\hat{X}_t = G(x_s, p_t, w, m). \tag{3}$$

To avoid the poor gradient propagation of the Bilinear sampling, here they proposed a local attention operation to sample the source features with a content-aware manner. This local attention works as a neural renderer where the source neural textures are sampled to render the target bones. The processing details are given in Figure 3.1.3. Let $f_t$ and $f_s$ represent the extracted features of target bones pt and source images $x_s$ respectively. For each location l, local patches $N_n(f_s, l + w_l)$ are first extracted from $f_t$ and $f_s$ . Then, it predict the local n × n kernel $k_l$ as the attention coefficients from the extracted local feature patch pair using a kernel prediction network M.

$$k_l = M(N_n(f_s, l + w_l), N_n(f_t, l)) \tag{4}$$

Here design M as a fully connected network. The local patch $N_n(f_s, l + w_l)$ pair are directly concatenated as the network inputs. We use the softmax function as the nonlinear activation function of the output layer of model M. This operation forces the sum of $k_l$ to 1, which enables the stability of gradient backward. Finally, the attention result localed at coordinate l = (x, y) is calculated as,

$$f^l_{attn} = P(k_l \bigotimes N_n(f_s, l + w^l)) \qquad (5)$$

where  denotes the element-wise multiplication over the spatial domain and P represents the global average pooling operation. The final warped feature$f_attn$ is obtained by repeating the previous steps for each location l. Furthermore, in order to enable the network to generate occluded contents, a mask m is used , with continuous values between 0 and 1 to select features between the warped result $f_attn$ and the target feature $f_t$. The final output feature map $f_out$ is calculated as,

$$f_{out} = (1 - m) * f_t + m * f_{attn} \qquad (6)$$

This model train the network using a joint loss consisting of a reconstruction loss, adversarial loss, perceptual loss, and style loss.
The reconstruction $l_1$ loss is written as,

$$L_{l_1} = \|x_t - \hat{x}_t\| \qquad (7)$$

The generative adversarial loss is used to mimic the distributions of the ground-truth and is given by,

$$L_{adv} = [\log(1 - D(G(x_s, p_t, w, m)))] + [(x_t)] \qquad (8)$$

where D is the discriminator of the Local Neural Texture Renderer G. The perceptual loss and style loss introduced by are used to reduce the reconstruction errors. The perceptual loss calculates $l_1$ distance between activation maps of a pre-trained network. It can be written as

$$L_{prec} =_i \|\phi_i(x_t) - \phi_i(\hat{x}_t)\|_t \qquad (9)$$

where $\phi_i$ is the activation map of the i-th layer of a pre-trained network. The style loss calculates the statistic error between the activation maps as,

$$L_{style} = \textstyle\sum_j \|G^\phi_j(x_t) - G^\phi_j(\hat{x}_t)\| \qquad (10)$$

where $G^\phi_j$ is the Gram matrix constructed from activation maps $\phi_j$ .
Thus the GFLA model is trained using the overall loss as

$$L_G = \lambda_c L_c + \lambda_r L_r + \lambda_{l_1} L_{l_1} + \lambda_a L_{adv} + \lambda_p L_{perc} + \lambda_s L_{sytle} \qquad (11)$$

## 3.2 Modelling The Temporal Consistancy For Person Image Animation

The pose-guided person image animation task refers to generating videos by rendering continuous skeletons using the neural textures of source images. Different from the generation task, it requires not only generating realistic textures for each frame but also modeling the temporal consistency between adjacent frames. Therefore, they further improved the model to generate coherent results. First, a Motion Extraction Network is proposed to extract accurate movements from the noisy input skeletons. Then there is an improved GFLA model to generate sequences in a recurrent manner

6

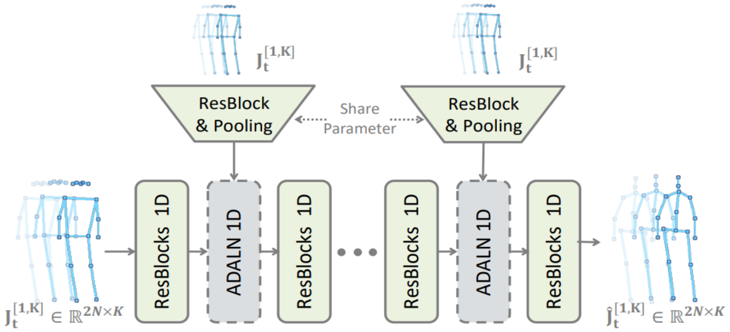

Figure 3.2.1: Architecture of Motion extraction network

### 3.2.1 Motion Extraction Network

One of the major problems is that the input skeleton sequences extracted by the popular algorithms [4], [15] are not temporally consistent.The predicted locations vibrate around the ground-truth values. This Motion Extraction Network works as a denoise model extracting accurate movements from noisy skeleton sequences. The architecture of the Motion Extraction Network is shown in Figure 3.2.1. Inspired by the method proposed in the paper [43], the network is designed by using 1D convolutional layers. The input layer of this network takes the concatenated (x, y) coordinates of the N joints for each skeleton frame instead of the 2D heat maps. Let $J_t^{[1,k]} \epsilon R^{2\times K}$ denotes the joints of K input skeletons. The output joints $\hat{J}_t^{[1,k]}$ contains the coordinates of skeletons with accurate movements. Adaptive layer normalization (ADALN) is used in this network. It has a similar architecture to that of ANAIN but using layer normalization as the normalization function. Layer normalization calculates the statistics for each single training case and normalizes the activities in a batch-wise manner. The effect of this normalization operation can be explained as to removing the unrelated factors such as global locations and scales, thereby making the network focus on motion extraction. The task is to reconstruct the coherent skeletons, it is neccassary to recover the statistics of the original sequences after reasoning about the motions. Therefore, it enables the network to recover the original statistics by calculating the affine parameters of the normalization layers from the input skeletons. The network is trained with ground-truth joints $J_{gt}^{[1,k]}$.The commonly used mean per-joint position error (MPJPE) is employed as the loss function.

$$L_{mpjpe} = \| \hat{J}_t^{[1,k]}, J_{gt}^{[1,k]} \| \tag{12}$$

Since most person animation datasets do not provide the required ground-truth skeleton labels, so this network is trained separately using the Human3.6M dataset [13]. This dataset contains accurate 3D human skeleton sequences acquired by recording the performance of 11 subjects under 4 different viewpoints. The noise skeleton inputs from the videos of the Human3.6M dataset [13] are extracted by using the pose extractor [3]. The ground-truth labels $J_{gt}^{[1,k]}$ are obtained by projecting the 3D skeletons to the corresponding viewpoints. After training the Motion Extraction Network, the clean skeletons $\hat{J}_t^{[1,k]}$ are obtained by performing inference on the animation datasets.

### 3.2.2 Sequential Gobal-Flow Local Attention Model

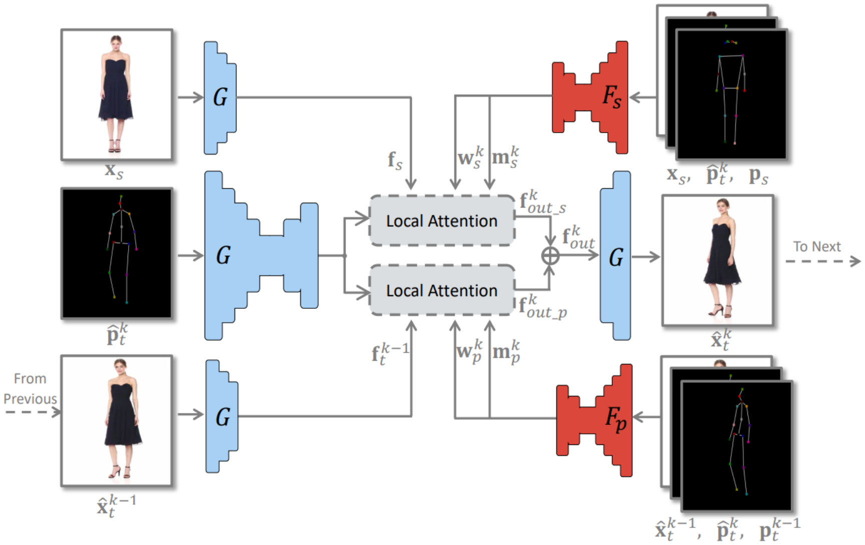

Figure 3.2.2: Generation process of a video frame

The sequential GFLA model is designed to generate result videos from the extracted accurate movements. Let $\hat{p}_t^{[1,K]} = \{\hat{p}_t^1, \hat{p}_t^2, \hat{p}_t^3, ......\hat{p}_t^K\}$ denotes the 2D heat map sequences obtained from the extracted joints $\hat{J}[1,K]_t$ . The model generates video clips $\hat{x}_t^{[1,K]} = \{\hat{x}_t^1, \hat{x}_t^2, \hat{x}_t^3, ......\hat{x}_t^K\}$ by rendering skeletons $\hat{p}[1,K]_t$ using the appearance of the source image $x_s$. This explicitly build the correlations between adjacent frames. Video clips are generated in a recurrent manner: the previously generated frames are used as the inputs of the current generation step. Specifically, Figure 3.1.2 shows the generation process of frame $\hat{x}[k]_t$ . It can be seen that there is an additional spatial transformation module responsible for transforming the information of the previously generated frame $\hat{x}[k-1]_t$ to the sequential GFLA model. The sequential GFLA model first extracts flow fields $w_s^k$ and $w_p^k$ using the Global Flow Field Estimators $F_s and F_p$ respectively.

$$w_s^k, m_s^k = F_s(x_s, p_s, \hat{p}_t^k)$$

$$w_p^k, m_p^k = F_p(x_t, \hat{p}_t^{k-1}, \hat{p}_t^k) \tag{14}$$

where the $m_s^k, m_p^k$ are occlusion mask.The Local Neural Texture Renderer G is then used to generate the result image by spatially transforming the information of $x_s$ and $\hat{x}[k-1]_t$.

$$\hat{x}_t^k = G(x_s, p_s, w_s^k, m_s^k, \hat{x}_t^{k-1}, \hat{p}_t^{k-1}, w_p^k, m_p^k, \hat{p}_t^k) \tag{15}$$

Two local attention modules are used to warp the features of the source image $x_s$ and previously generated image $\hat{x}[k-1]_t$.The processing operation is the same as that described in Section 3.1.2. The output features $f_{out_s}^k$ and $f_{out_p}^k$ are generated by these local attention modules. The final output feature $f_{out}^k$ is calculated by fusing the outputs of the two branches,

$$f_{out}^k = f_{out_s}^k + f_{out_p}^k \tag{16}$$

The animation model is trained by using both spatial and temporal losses. The spatial losses can constrain the model to generate realistic frames. The joint loss which is same (Equation 15) as that of our image generation model are used for each resultant frame. The temporal loss is used to model the correlations between different frames. A temporal discriminator $D_v$ is used to calculate this loss. The temporal discriminator $D_v$ takes image sequences as inputs and estimates the probabilities that the inputs are sampled from real video clips.

$$L_{adv_v} = E[\log(1 - D_v(\hat{x}_t^{[1,K]}))] + E[\log D_v(x_t^{[1,K]})] \tag{17}$$

Therefore, the overall loss function of the animation model can be written as,

$$L_A = \frac{1}{K}\sum_{k=1}^{K} L_G^k + \lambda_v L_{adv_v} \tag{18}$$

where $L_G^k$ represents the spatial loss of frame $\hat{x}_t^k$.

# Chapter 4

# Conclusions

## 4.1 Results and Observations

### 4.1.1 Implementation Details

Auto-encoder structures are employed to design the networks. The residual block is used as the basic component of the model. Unless otherwise specified, the model is trained using 256 × 256 images. Local attention modules used for feature maps with resolutions of 32 × 32 and 64 × 64. The extracted local patch sizes are 3 and 5 respectively. For the person image generation task, GFLA model is trained in stages. The Flow Field Estimator is first trained to generate flow fields. Then train the whole model in an end-to-end manner. For the image animation task, first train the Motion Extraction Network using the Human3.6M dataset [13] .Then train the sequential GFLA model using the predicted clean skeletons.The ADAM optimizer is adopted with the learning rate as $10^{-4}$

### 4.1.2 Metrices

Both the image-based metrics and video based metrics are employed to evaluate the results. Learned Perceptual Image Patch Similarity (LPIPS) is used to calculate the reconstruction errors of generated images. This metric computes perceptual distances between input image pairs.Meanwhile, the Frechet Inception Distance (FID)is used to to measure the realism of the generated images. It calculates the Wasserstein-2 distance between distributions of the generated data and real data. For video results, in order to model the temporal consistency errors, the I3D model is used to extract the video features. Average Euclidean Distance (AED) is used as the perceptual reconstruction error indicator. It calculates the Euclidean distance between features of generated videos and ground-truth videos. FID Video takes the extracted video features as inputs and evaluates the realism of generated videos. Besides,performs a Just Noticeable Difference (JND) test to evaluate the subjective quality.

### 4.1.3 Datasets

For the person image generation task, here uses the public datasets: DeepFashion Inshop Clothes Retrieval Benchmark. The **DeepFashion dataset** contains 52712 high-quality model images with clean backgrounds. The personal identities of the training and testing sets do not overlap. The **FashionVideo dataset** used for animation task contains 500 training and 100 test videos, each containing roughly 350 frames. Videos have static viewpoints and clean backgrounds. The iPER dataset contains 206 high-resolution videos. Human subjects in this dataset have different conditions of shape, height, and gender.

### 4.1.4 Results of the proposed system

The GFLA method is compared with several state-of the-art models on both generation and animation tasks. For the person image generation task, popular methods Def-GAN [9], VU-Net [**?**], Pose-Attn[6] and Intr-Flow [10] are selected as the competitors. The quantitative evaluation results are shown in Table 4.1 .

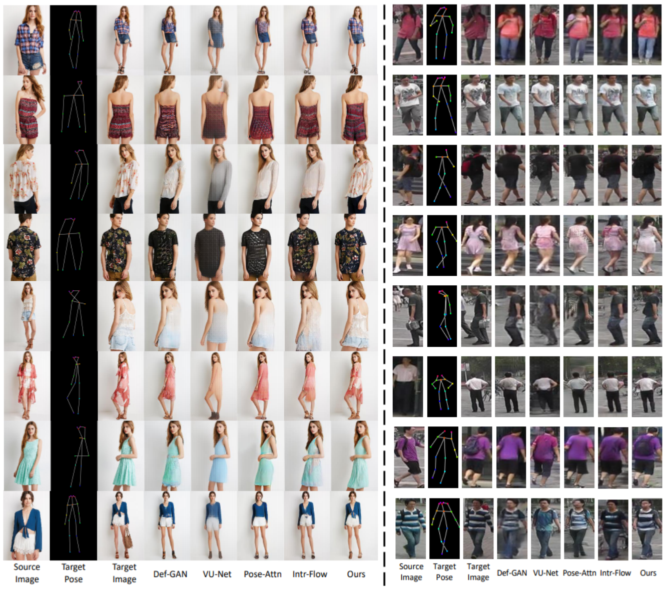

Figure 4.1.1: Qualitative comparisons with several state-of-the-art person image generation models including Def-GAN [], VU-Net [8].

For this comparisons ,the model is trained with the Market-1501 dataset using their original 128×64 images. To alleviate the influence of the backgrounds on the reconstruction errors, here follow the previous work [1] to provide the mask-LPIPS. It can be seen that the model achieves competitive evaluation results, which means that this model can generate realistic results with fewer perceptual reconstruction errors. Since subjective metrics have their own limitations, their results may mismatch with the actual subjective perceptions . Therefore, a human objective evaluation test is performed. A JND test is implemented

Table 4.1: Comparison with state-of-the-art person image generation methods and GFLA model

| Networks | FID | LPIPS | JND | No of parameters |
|---|---|---|---|---|
| Def-GAN | 18.457 | 0.2330 | 9.12% | 82.08M |
| VU-Net | 23.667 | 0.2637 | 2.96% | 139.6M |
| Pose-Attn | 20.739 | 0.2533 | 6.11% | 41.36M |
| Intr-Flow | 16.314 | 0.2131 | 12.61% | 49.58M |
| GFLA | 10.573 | 0.2341 | 24.80% | 14.04M |

on Amazon Mechanical Turk (MTurk).The test is performed over 800 image pairs for each model and dataset. To avoid individual bias, each image pair is compared 5 times by different volunteers. The results can be found in Table. It can be seen that the model achieves the best result in the challenging Fashion dataset and competitive results in the Market-1501 dataset. Besides, it provides a numbers of model parameters to evaluate the computation complexity. Because of the efficient spatial transformation blocks, the model does not require a large number of convolution layers. It helps to achieve high performance with less than half of the parameters of the competitors.

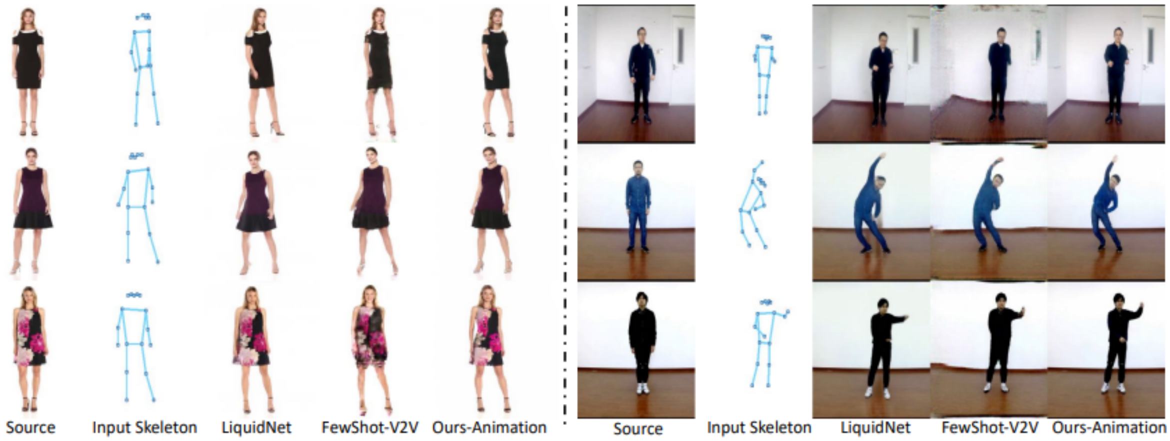

Figure 4.1.2: Qualitative comparisons with several state-of-the-art person image animation models including LiquidNet[12] and FewShot-V2V[11] .

For the pose-guided animation task,the sequential GFLA model is compared with FewShot-V2V [11] and LiquidNet[12]. The comparison results are shown in Table 4.2. Different from the competitors which employ either face refine models or background inpainting models to improve their results, they do not use any post-processing methods. It can be seen that the Seq-GFLA model achieves the best results on Fashion Video dataset. LiquidNet achieves good FID and LPIPS scores, which means that it can generate realistic video frames.

Table 4.2: Comparison with state-of-the-art person image animation methods and proposed model over Fashion Video Dataset

| Networks | FID | LPIPS | FID Videos | AED | No of parameters |
|---|---|---|---|---|---|
| LiquidNet | 17.681 | 0.0897 | 5.174 | 0.0184 | 97.45M |
| FewShot-V2V | 27.803 | 0.0816 | 5.096 | 0.0188 | 97.96M |
| Seq-GFLA | 14.95 | .0651 | 3.246 | 0.0126 | 23.51M |

## 4.2 Conclusions

Human pose transfer (HPT) is an emerging research topic with huge application potential in creative media applications. Yet current HPT methods typically introduce detail deficiency, content ambiguity or style inconsistency in synthesized person images due to the suboptimal integration between low level feature transfer and high-level semantic-guided content synthesis. Here, the person image generation and animation tasks are implemented using deep spatial transformation of inputs at feature level using local attention in a global predictive manner. The possible reasons causing poor gradient propagation are analyzed when warping sources at the feature level.The model can generate coherent results with realistic frames. Meanwhile, significantly it can use fewer model weights than competitors.

Targeted solution GFLA framework , first estimates the flow fields between sources and targets and then sample the source features in a content-aware manner.Empirically demonstrated that the GFLA model can provide improved gradients, leading to accurate spatial transformations. Meanwhile, further proposed a sequential GFLA model to extract the correlations between adjacent frames for the animation task. Experiments show that the model can efficiently build temporal dynamics and generate coherent videos. Finally, demonstration on other tasks shows that the model is versatile which requiring spatial transformation such as face image animation and novel view synthesis.

## 4.3 Future Scope and Suggestions

Although the model generates impressive results, it also observes some failure cases . These typical failure cases are due to the severe occlusions of source images, which misleads the model to sample incorrect neural textures. One solution is to add additional constraints to flow fields. For example, loss functions can be designed to penalize sampling between different semantic regions. Another solution is to perform multi-step warping operations to gradually warp source images to targets by using additional video datasets.

# References

[1] Yurui Ren;Ge Li;Shan Liu;Thomas H. Li ,Deep Spatial Transformation for Pose-Guided Person Image Generation and Animation , IEEE Transactions on Image Processing Year: 2020, Volume: 29 ,Journal Article , Publisher: IEEE,pp-8622 - 8635 , DOI: 10.1109/TIP.2020.3018224

[2] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in Advances in Neural Information Processing Systems, 2017, pp. 406–416

[3] OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields Zhe Cao;Gines Hidalgo;Tomas Simon;Shih-En Wei;Yaser Sheikh IEEE Transactions on Pattern Analysis and Machine Intelligence Year: 2021 ,Volume: 43, Issue: 1 , Journal Article , Publisher: IEEE

[4] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai, "Progressive pose attention transfer for person image generation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2347–2356.

[5] M. Jaderberg, K. Simonyan, A. Zisserman et al., "Spatial transformer networks," in Advances in neural information processing systems, 2015, pp. 2017–2025

[6] Synthesizing Images Of Humans In Unseen Poses in Proceedings of the IEEEConference on Computer Vision and Pattern Recognition,2018

[7] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7753–7762

[8] P. Esser, E. Sutter, and B. Ommer, "A variational u-net for conditional appearance and shape generation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp

[9] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe, "Deformable ' gans for pose-based human image generation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3408–3416

[10] Y. Li, C. Huang, and C. C. Loy, “Dense intrinsic appearance flow for human pose transfer,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3693–3702.

[11] T.-C. Wang, M.-Y. Liu, A. Tao, G. Liu, J. Kautz, and B. Catanzaro, “Few-shot video-to-video synthesis,” arXiv preprint arXiv:1910.12713, 2019

[12] W. Liu, Z. Piao, J. Min, W. Luo, L. Ma, and S. Gao, “Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 5904–5913.

[13] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, “Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments,” IEEE transactions on pattern analysis and machine intelligence, vol. 36, no. 7, pp. 1325–1339, 2013

[14] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, “Deepfashion: Powering robust clothes recognition and retrieval with rich annotations,” in14 Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1096–1104

[15] Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, “Crowdpose: Efficient crowded scenes pose estimation and a new benchmark,” arXiv preprint arXiv:1812.00324, 2018.