Collaborative Neural Rendering using Anime Character Sheets

Anonymous CVPR 2022 submission

Paper ID 6818

Figure 1: **Collaborative Neural Rendering** (CoNR) takes multiple arbitrarily ordered and posed reference images as the **Character Sheet** input, and an **Ultra-Dense Pose** (UDP) representation as the pose input. The UDP is generated from another image providing a desired pose. Then, CoNR outputs an image of the given character with the desired pose.

Abstract

Drawing images of characters at desired poses is an essential but laborious task in anime production. In this paper, we present the Collaborative Neural Rendering (CoNR) method to create new images from a few arbitrarily posed reference images available in character sheets. In general, the high diversity of body shapes of anime characters defies the employment of universal 3D body models for real-world humans, like SMPL. To overcome this difficulty, CoNR uses a novel and compact form of landmark encoding to avoid requiring a unified UV mapping in the pipeline. In addition, CoNR's performance can be significantly increased when having multiple reference images by using feature space cross-view dense correspondence and warping in a specially designed neural network construct. Moreover, we collect a character sheet dataset containing over 700,000 hand-drawn and synthesized images of diverse poses to facilitate research in this area. The code and data will be released.

1. Introduction

Artists commonly use character sheets to show their design of a character. Character sheets are the image collections of a specific character with multiple postures observed from different viewpoints. Thus they cover all the appearance details and are widely used to assist image creation of animations or their derived media. Moreover, these sheets allow other artists to draw together while maintaining the consistency of the design of this character.

However, creating images of the characters with new poses is still an uphill task during most animation, comic, and game production. This is especially true for anime, a prominent art form that traditionally requires human imagination and expertise to draw every character image manually. Drawing a sequence of anime frames with desired poses is extremely time-consuming, and therefore does not generalize easily for interactive applications like games or live streaming with virtual avatars. Due to the semantic gap between the character sheets and the desired pose, it is very challenging for computers to automatically draw character images using character sheets like human artists. Non-photorealistic rendering, a recently emerged computer graphics technique, enables fast and on-demand anime frame generation. However, it requires significantly more expertise and effort to design special 3D models with textures and complicated shading programs to resemble the drawing style of different characters. Thus it is less accessible to the anime industry and the vast majority of the fan-art communities who are more familiar with traditional hand-drawing processes.

We formulate the task of rendering an image of a particular character in the desired pose from character sheets.

Instead of modeling character sheets as sequences, which suffer from the ordering issues in various aspects, our formulation allows them to be dynamically-sized sets, better matching the established convention in the anime industries.

Based on this formulation, we develop a **Collaborative Neural Rendering (CoNR)** model. CoNR fully exploits the information available in a provided set of reference images by using feature space cross-view dense correspondence and warping.

In addition, CoNR uses the Ultra-Dense Pose, a novel and compact form of landmark encoding designed specifically for anime characters to avoid requiring a unified UV mapping in the pipeline. It can represent the fine details of characters, such as hairstyles or clothing, thus allowing better artistic control and adjustment over the desired pose for anime production purposes. It can also be easily generated with existing computer graphics pipelines, allowing a wide range of interactive applications like anime-based games or virtual assistants.

The contributions in this work are three-fold:

- We present a baseline method, CoNR, for a novel task of rendering anime character images with desired pose using character sheets.
- We explore the collaborative inference method of feed-forward neural networks to model character sheets as a dynamically-sized set of images.
- We introduce an Ultra-Dense Pose representation designed specifically for anime characters and build a character sheet dataset containing images of diverse poses.

2. Related Works

2.1. Image Generation & Translation for Anime

Recent years have seen significant advancement in applying deep learning techniques [17, 43, 49] to assist the creation of anime. The generative modeling of anime faces has achieved exciting results [23, 18, 40, 25, 21]. There are also attempts to produce vectorized anime images [37], similar to the step-by-step manual drawing process. A modified StyleGAN2 model [4] is proven to be able to generate full-body anime images. However, it still suffers from artifacts, including weirdly connected body parts, and it is not efficient or straightforward to control the generated poses of the character.

2.2. Human Pose & Appearance Transfer

In the real world, however, there has been a great success in the human pose or appearance transfer tasks [11, 5]. Most of these works create vivid body motions or talking heads from only one single image [44, 15, 34, 41, 45, 35, 26].

The learned prior of the human body [28], head [8], or real-world clothing shape and textures [2] enable the model to solve ill-posed problems like imagining and inpainting the back view even if only the frontal view of the human is available. Unfortunately, anime has long been featuring a flexible character design leading to high diversity in clothing, body shapes, hairstyles, and other appearance features, making it challenging to adopt real-world human priors in the domain of anime characters.

There are also some attempts [27] to extend the pose transfer task by utilizing SMPL [28], a parametric 3D human model, to combine appearance from different views. Using multiple reference images would, in principle, allow the model to follow the original appearance of the given person instead of finding similar posterior estimates, and better suit the needs of anime production.

Some recent works utilize NeRF [30], a category of neural rendering models which are trained using photometric reconstruction losses and optional camera poses over multiple images of a 3D object. Due to their ray-marching nature and capability to in-paint in 3D, they are promising methods in modeling real-world 3D data [32, 31], which are not depending on or being influenced by any prior knowledge other than the object to be modeled. However, they have not yet made much progress in modeling hand-drawn data like anime character sheets, which are less following strict geometric and physical constraints.

2.3. Representation of Human Pose

We analyze the representation of the human body posture in the real world in multiple data modes.

In the real world, stick-figure of skeletons [11], SMPL vectors [28], and heat maps of joints [10, 36] are well-defined and widely-used representations that can be obtained from objective data sources, including motion capturing equipment. However, noisy manual annotations, occlusion caused by diverse styles of garment or other body decorations, and ambiguity caused by diverse body shapes impose significant challenges when migrating these sparse representations from human to anime [1].

Anime characters usually require flexible artistic control over fine details or pose exaggerating on additional moving parts like floating hair or flowing skirt, which are not directly driven by human joints. The aforementioned human pose representations, however, are very abstract and fuzzy in the inverse task of pose-guided anime rendering.

Human parsers or clothing segmenters [46, 45, 15, 12] are robust to the uncertainty of joint positions. However, the provided semantic masks are not informative enough for representing pose, or even the orientation of a person.

DensePose [19], UV texture mapping [44, 45, 16], greatly enhance the detail of pose representation on the human body or face by imposing a universal definition that essentially unwarps the 3D human body surface into a 2D coordinate system. However, three problems may emerge

when anime characters start to annotate themselves accordingly. Anime girls have trouble finding the precise location where they should cut their skirts and flatten this cone-like object in the same way as others. Anime boys get stuck as they are unsure how to consistently handle jeans, kilts, and other non-homeomorphic body shapes. Last but not least, they have no idea of the number of key points they should use. Due to the diverse body shapes of anime characters, every region of the body can require more key points than others. Therefore, existing dense representations that are not designed for anime-related tasks may still not serve as an off-the-shelf solution.

3. Method

3.1. Task Formulation

We explore this task by first observing real artists drawing anime images. While drawing different body parts on the canvas, artists will usually refer to multiple images in the character sheets because appearance details required at the desired pose are usually distributed across different reference images.

Given a sequence of reference images $\mathbf{I}_1$ ... $\mathbf{I}_n$ in the character sheet, human artists drawing anime images can be seen a sequence of operations on the canvas after seeing each $\mathbf{I}_t$, similar to the widely adopted formulation of tasks about drawing or painting art [48, 22, 37]. As $\mathbf{I}_t$ differ from each other only by the pose of the character, the order of sequence $\mathbf{I}_t$ can also be seen as the order of poses. However, the underlying mathematics of existing pose representations prevents an easy definition of their order or a satisfying way to discretize them into finite canonical categories. Even if an algorithm or a human successfully finds a place in the sequence for a 45-degree left side view, it may still struggle at a standing character with his head looking back, which is a frequently seen pose in the domain of anime.

Therefore, the sequence formulation is not favorable for character sheets. For freeing the users from putting $\mathbf{I}_t$ into a sequence during inference, arbitrary ordering should be allowed in the character sheet.

Here we present a novel task formulation by considering one character sheet $\mathbf{S}_{ref}$ in a whole as an input sample and ignoring the order of reference images $\mathbf{I}_n \in \mathbf{S}_{ref}$. To give rendering directives to the model, a target pose $\mathbf{P}_{tar}$ representation is also required in the input. The task can be formulated as mapping input sample $\mathbf{S}_{ref}$ to target image y that follows the desired target pose $\mathbf{P}_{tar}$.

$$y = f(\mathbf{P}_{tar}, \mathbf{S}_{ref}) \tag{1}$$

We also notice that complicated poses, motions, or characters may require a larger collection of references in $\mathbf{S}_{ref}$ than others. A dynamically sized $\mathbf{S}_{ref}$ should therefore also be allowed.

3.2. Modeling of Character Sheets

To properly model the task in 3.1, we propose a Collaborative Inference method for convolutional Neural Networks (CINN).

Usually, multiple images can be fed in arbitrary order into multiple copies of the same classical convolutional neural network to obtain corresponding inference results. In CINN, however, multiple images in a set are defined in a whole as one single input sample. Adding feature-averaging on outputs of all corresponding blocks in multiple copies of an existing convolutional neural network, we obtain a network of a dynamical number of sub-networks that share the same weight and are inter-connected by message passing mechanisms. Due to the commutative nature of addition, changing the order of the sub-networks won't affect the inference result of the entire pipeline.

When performing a collaborative inference on such a network, n reference images (or views) in a set are fed into n weight-shared sub-networks, respectively. The sub-networks form a fully connected graph, as illustrated in Figure 3, so that each block of a sub-network would share part of its outputs as messages to corresponding successive blocks in all other sub-networks, in addition to forwarding other outputs to its following blocks like in a classical neural network. To further modulate the message sent at each block of each view, we use half of the messaging outputs of each block as edge weights to perform a weighted averaging.

3.3. Ultra-Dense Pose

Here we present Ultra-Dense Pose (UDP), a compact landmark representation designed specifically for anime characters. It allows better compatibility across a broader range of anime body shapes and enables better artistic control over body details like garment motions.

3D meshes are widely used data representations for anime characters in their game adaptations. Vertex in a mesh usually comprises corresponding texture coordinate (u, v) or a vertex color (r, g, b). Interpolation over the barycentric coordinates allows triangles to form faces filled by color values or pixels looked up from textures coordinates.

Taking a bunch of anime body meshes standing at the center of the world, we ask them to perform the same T-pose to align the joints, as shown in the Figure 2(a). To construct UDP, we remove the original texture and overwrite the color (r, g, b) of each vertex with a landmark, which is the current world coordinate(x, y, z) of this vertex, as shown in Figure 2(a). When the anime body changes its pose, the vertex on the mesh may move to a new position in the world coordinate system, but the landmark at the corresponding body part will remain the same, shown as the same color in Figure 2(b).

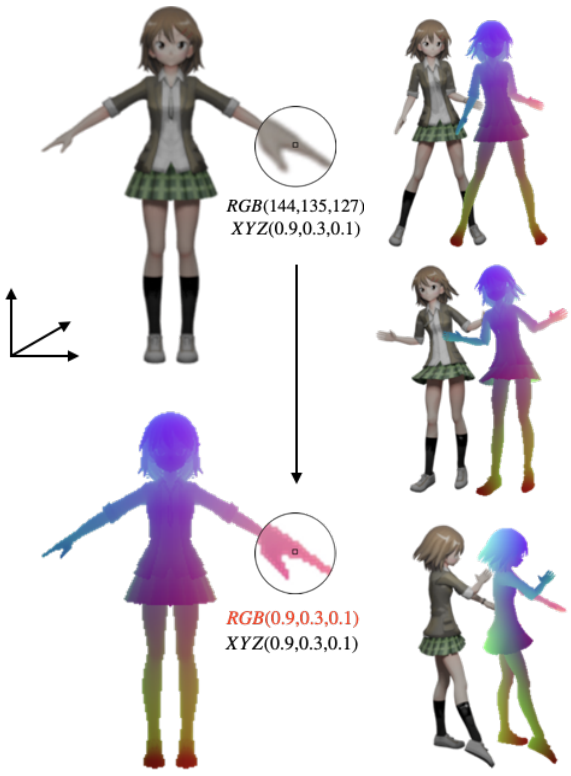

Figure 2: **UDP of Kurei Kei, an anime character.** (a) UDP uses 3D coordinates as landmarks to avoid the difficulties of unwarping 3D surfaces into a 2D UV map. (b) When the anime body changes its pose, the landmark at the corresponding body part will remain the same.

To avoid the difficulty of down-sampling and processing meshes, we convert the modified meshes into 2D images, which are more friendly to neural networks. This is done by introducing a camera, performing culling on occluded faces and projecting only the faces visible from the camera into an image. The resulting UDP representation is a $u \times v \times 4$ shaped image recorded in floating-point numbers ranging from 0 to 1. The four channels include three body landmark encodings and one occupancy for whether the pixel is body or empty.

We find three properties of this representation that could alleviate the difficulties mentioned in 2.3.

1) UDP is a detailed 3D pose representation since every tiny piece of surface on the anime body, no matter if it is from the hair or the garment, could be automatically assigned with a unique encoding without complicated handcrafted annotations.

2) UDP is also a widely compatible pose representation with acceptable exchangeability since the anime characters with similar body shapes will also get out-fits that are consistently pseudo-colorized.

3) the encoding defined in UDP also serves a dual role of describing the local 3D shape of the human body, which could provide additional geometric information to downstream tasks, although it may not be necessary for the task of this work.

3.4. Data Preparation

As character sheets used in the anime-related industries are not yet available to the computer vision community, we build a dataset containing more than $20,000$ hand-drawn anime characters by selecting human-like characters from public datasets [3, 24]. We manually perform matting to remove the background from the character with the help of the watershed algorithm. With a similar method, we also obtain a background dataset for use in augmentations.

Manually annotating hand-drawn anime images with UDP involves intolerable levels of hardship. To alleviate the problem of label scarcity, we constructed a synthesized dataset from anime-styled 3D meshes in the way described in 3.3.

Finally, we randomly split from a synthesized dataset, which has high-quality UDP labeling, and a hand-drawn dataset, which has higher diversity in styles and characters, by a 16:1 ratio into the training and validation sets. The split is done on a per-character basis so that the validation set contains characters unseen during training.

3.5. Collaborative Neural Rendering

Overview CoNR consists of a renderer and an optional Ultra-Dense Pose detector. Figure 3 shows the pipeline of the proposed approach. The renderer generates character images of the desired pose taking the target pose's UDP representation $\hat{\mathbf{P}}_{tar}$ and a character sheet $\mathbf{S}_{ref}$, as the inputs.

The input UDP representation can be produced by a UDP detector from reference images or videos. For interactive applications like games, the existing physics engine can be used as a drop-in replacement for the UDP detector to compute body and cloth dynamics for the anime character directly.

Renderer The renderer is based on a simple U-Net [33]. We apply the following modifications.

Firstly, to allow efficient inference on videos, we remove the UDP input from the encoder side but instead concatenate the UDP input rescaled with the nearest-sampling method into each skip channel from the encoder to the decoder, as shown in Figure 3. This allows us to checkpoint the evaluated results from the encoder that can be reused when inference on multiple target UDPs in a video.

Secondly, inspired by [13, 14], we use two extra channels in each decoder block to generate a flow-field and perform a grid sampling over other output features of the block, for enhancing the long-range look-up ability for CNNs.

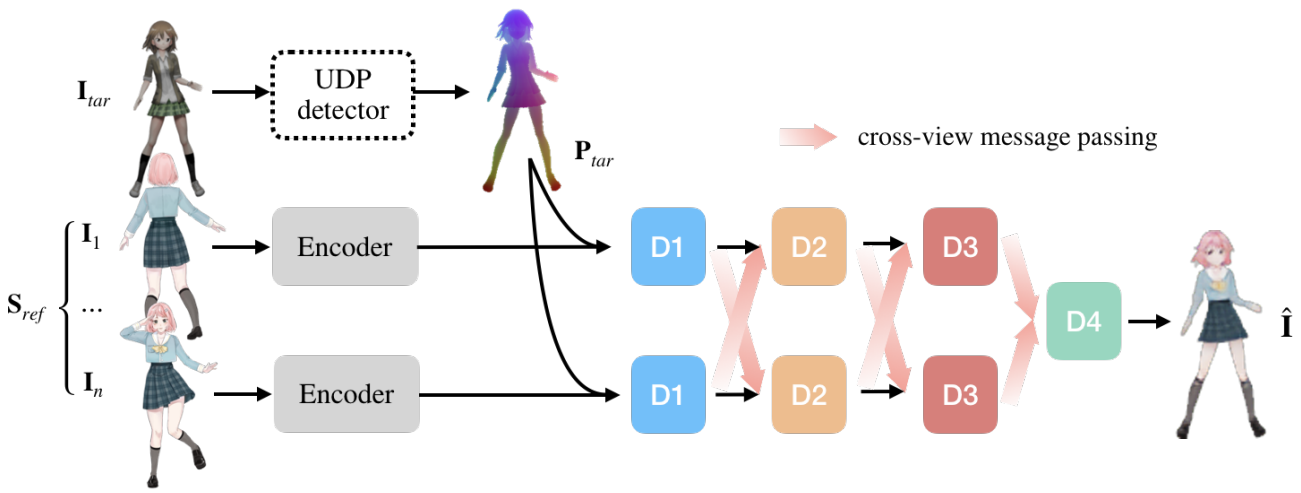

Figure 3: **Overview of CoNR.** Reference images $\mathbf{I}_1 \cdots \mathbf{I}_n \in \mathbf{S}_{ref}$ from the input character sheet are feed into a CINN using modified U-Nets as sub-networks. The same UDP representation $\mathbf{P}_{tar}$ detected from $\mathbf{I}_{tar}$ is resized and concatenated into each scale of the encoder outputs in all sub-networks. Blocks with the same color share weights. D1 to D4 refer to block 1 to block 4 of the decoder. The sub-networks form a fully connected graph using cross-view message passing. Each block will receive the averaged message from corresponding blocks in all other sub-networks.

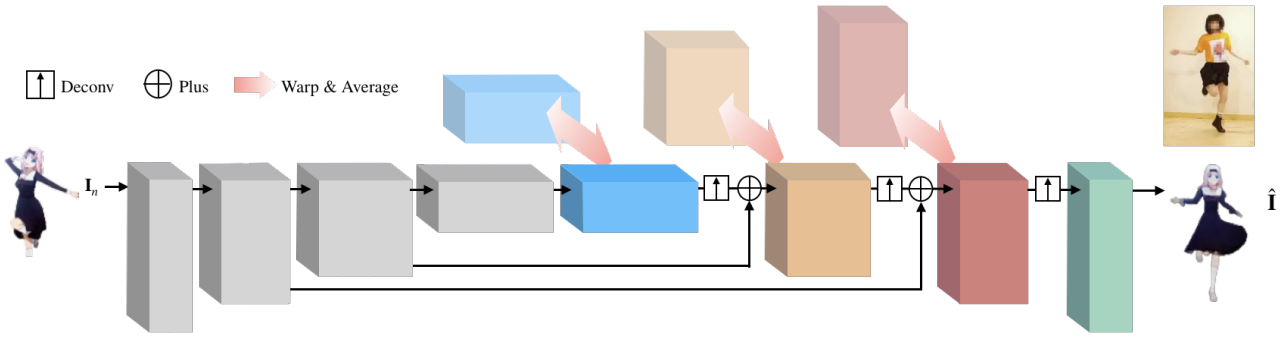

Figure 4: **Sub-network structure of CoNR**. Every input $\mathbf{I}_n \in \mathbf{S}_{ref}$ goes into a U-Net shaped sub-network. The warping and averaging operation are performed on the output of every decoder block. The semi-transparent blocks represent corresponding blocks in all other sub-networks. Blocks with the same color share weights. Each block in a sub-network would pass part of the outputs as messages to corresponding successive blocks in all other sub-networks, in addition to forwarding outputs to its following blocks in the current sub-network. The path of UDP is omitted.

Finally, we apply the CINN method to the decoders of the network. We split the original up-sampling output feature channels by half, one as the remote branch and the other for the local branch. The warping is first performed only on the remote branches. Then the remote branch of output features from all sub-networks are averaged and passed to the next block, as described in 3.2. The local branch remains unchanged. The local branch output and the remote branch output from the previous block are concatenated with the encoder output and fed into the next block, as shown in Figure 4. The last decoder block will collect averaged output features from all previous decoder block in all sub-networks and decode them as the final RGB output $\hat{\mathbf{I}}_{tar}$.

Ultra-Dense Pose Detector To prepare the UDP representation $\widehat{\mathbf{P}}_{tar}$ of target pose, we use a simple U-Net [33] consists of a ResNet-50 [20] encoder and a decoder with 5 Residual Blocks to detect it from an RGB image $\mathbf{I}_{tar}$, The detector gives out four channels of UDP values, in the same way as the UDP representation and the raining dataset.

The detector can be trained independently on the synthesized dataset or trained jointly with the renderer in an end-to-end manner. In the second case, the detector could provide the renderer with augmentations on UDPs, which gradually decrease when the detector converges. To further reduce the memory footprint of the model, we can optionally share the weight of the ResNet-50 backbone in the en-

coder of UDP detector with the encoder of the renderer. The UDP detector is only evaluated for on the target $\mathbf{I}_{tar}$.

4. Experiments

4.1. Training Strategy

We train CoNR models of m sub-networks (views) on our dataset. To create one training sample, we randomly select a character from the dataset and then randomly select $m+1$ arbitrary poses of that character. These images are split as m character sheet inputs $\mathbf{I}_1 \cdots \mathbf{I}_m \in \mathbf{S}_{ref}$ and one image of the target pose as the ground truth of CoNR's final output.

We apply random image augmentations to the target pose image, paste them onto k random backgrounds and run them through the UDP detector. We use the average of the k UDP detection results of the same target pose, $\widehat{\mathbf{P}}_{tar} = 1/k \sum_{j=1}^{k} \widehat{\mathbf{P}}_j$, as the final UDP detection results. We also obtain the corresponding UDP of the target pose from the dataset as the ground truth to train the UDP detector. We compute losses at both the output of the detector and the end of the CoNR pipeline. We use L1 loss on the 3 lanmark encodings and BCE loss on the mask to train the detector if the ground truth label is available.

$$\mathcal{L}_{udp} = ||\widehat{\mathbf{P}}_{tar} - \mathbf{P}^{gt}_{tar}||_1 + BCE(\widehat{\mathbf{P}}_{tar\_mask}, \mathbf{P}^{gt}_{tar\_mask}) \tag{2}$$

In addition, we evaluate a consistency loss by computing the standard deviation of k UDP detector outputs.

$$\mathcal{L}_{cons} = \sqrt{\frac{1}{k-1}\sum_{j=1}^{k}(\widehat{\mathbf{P}}_j - \widehat{\mathbf{P}}_{tar})^2} \tag{3}$$

At the end of the collaborated renderer Φ, we use L1 loss to ensure a correct photometric reconstruction of the character in the desired target pose,

$$\mathcal{L}_{photometric} = ||\Phi(\widehat{\mathbf{P}}_{tar}, \mathbf{S}_{ref}) - \mathbf{I}^{gt}_{tar}||_1 \tag{4}$$

Optionally, we compute the perceptual loss using LPIPS [47],

$$\mathcal{L}_{perception} = lpips(\Phi(\widehat{\mathbf{P}}_{tar}, \mathbf{S}_{ref}), \mathbf{I}^{gt}_{tar}) \tag{5}$$

The UDP detector and renderer are trained from scratch simultaneously in an end-to-end manner. The total loss function is evaluated as a weighted sum of all these losses, where $\alpha = 0.5, \beta = 0.05, \gamma = 1.0, \delta = 1.0$.

$$\mathcal{L} = \alpha\mathcal{L}_{perception} + \beta\mathcal{L}_{photometric} + \gamma\mathcal{L}_{udp} + \delta\mathcal{L}_{cons} \tag{6}$$

Our model is optimized by AdamW [29] with weight decay 10^{-4} for 2 epochs on the training set. We choose $m = 4, k = 4$ during training and $m = 4, k = 1$ during inference in all experiments unless otherwise specified. The training process uses a batch size of 6 with all input resolution set to 128×128. CoNR pipeline is implemented in PyTorch and trained on four 2080Ti GPUs for about four days. For quantitative evaluation, we measure the training losses on the validation split.

Table 1: **Comparison on the number of input reference images.** We use character sheets of m reference images to train the CoNR model, and then use character sheets of n reference images to evaluate the trained model.

Setting	epoch1		epoch2	
	$\mathcal{L}_1$	LPIPS	$\mathcal{L}_1$	LPIPS
$m=1, n=1$	0.0247	0.0832	0.0238	0.0801
$m=4, n=1$	0.0249	0.0865	0.0237	0.0827
$m=1, n=4$	0.0219	0.0798	0.0211	0.0764
$m=4, n=4$	**0.0187**	**0.0659**	**0.0179**	**0.0612**

4.2. Comparative Study

Effectiveness of the collaboration. Unlike most previous methods for similar tasks that allow only one reference image as input, CoNR uses a dynamically-sized set of reference inputs during training and inference.

The experiment results in Table 2 show that using an additional number of views $m > 1$ during training will enhance the accuracy of generated images. On the opposite, removing images from the character sheet will reduce coverage of the body surface. When CoNR is trained with $m = 1$, the chances are high that the provided character sheets do not allow a valid solution, such as providing the character's backside and asking the network to imagine the frontal side. In this case, the photo-metric reconstruction and perceptual losses may encourage the network to learn the wrong solution. Therefore even if enough information in the $n > 1$ images is provided during inference, the network may not generate the target image accurately. Similarly, keeping the training view count $m = 4$ while reducing the inference view count $n = 1$ will also harm the quality of the resulting image.

However, CoNR does not require the inference view count n to match the m during training. Adding additional views during inference will, on the contrary, enhance the quality of the generated images as shown in Figure 5. As illustrated in 3.2, CoNR models the input character sheet as a dynamically-sized set so that it can leverage information distributed across all images without favoring the first m inputs. This allows CoNR to scale seamlessly from a pose transfer task, when only a few shots of the character are given, to frame compiling or interpolating when a lot of footage of a character are available.

The example in Figure 5 shows the behavior of CoNR when it does not have enough information to draw missing parts during the inference correctly. Even if very few ref-

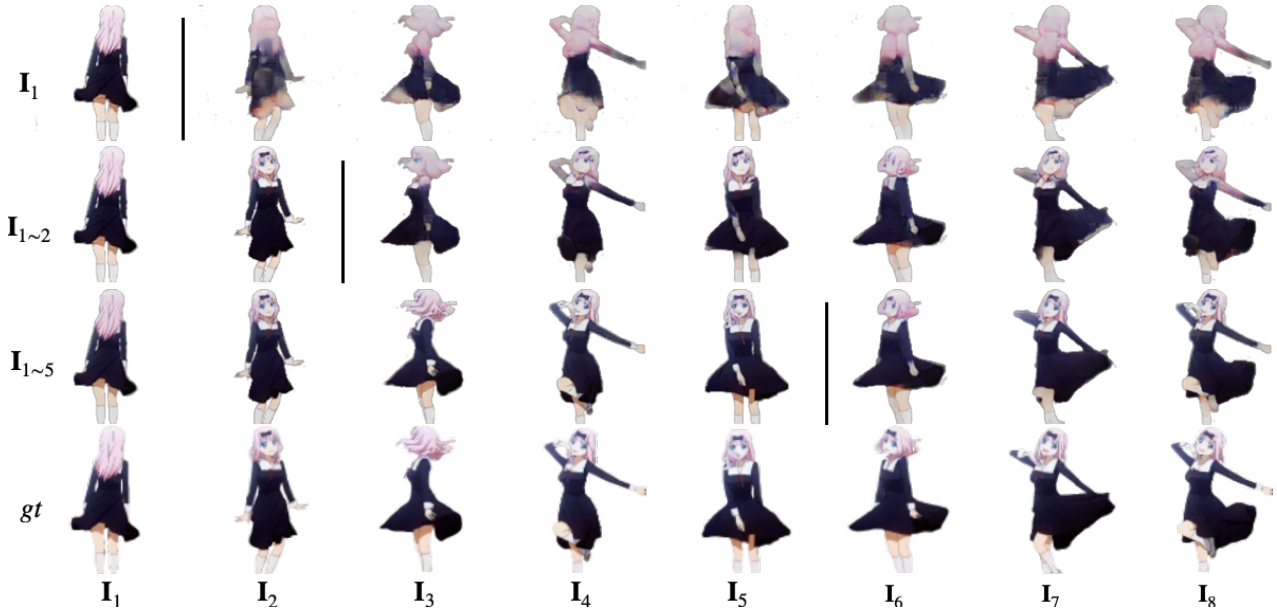

Figure 5: **Effectiveness of the collaboration.** We perform a reconstruction test with random video in the wild, with unseen pose and appearance, to understand the effectiveness of the collaboration. The last row shows 8 frames $\mathbf{I}_j^{gt} \in \mathbf{S}_{vid}^{gt}$ from an Youtube video. Starting form the first row shows the result of $\Phi(\widehat{\mathbf{P}}_j, \mathbf{S}_{ref})$. $\widehat{\mathbf{P}}_j$ are UDP detected from $\mathbf{I}_j^{gt}$. The used subsets of character sheet $\mathbf{S}_{ref} \subset \mathbf{S}_{vid}^{gt}$ are shown in the first column. Even if the provided reference images are not sufficient for drawing the target pose, CoNR will generate a conservative guess by trying to in-paint the unprovided area. Generated images for novel poses unseen in the character sheets are shown on the right of vertical lines per row.

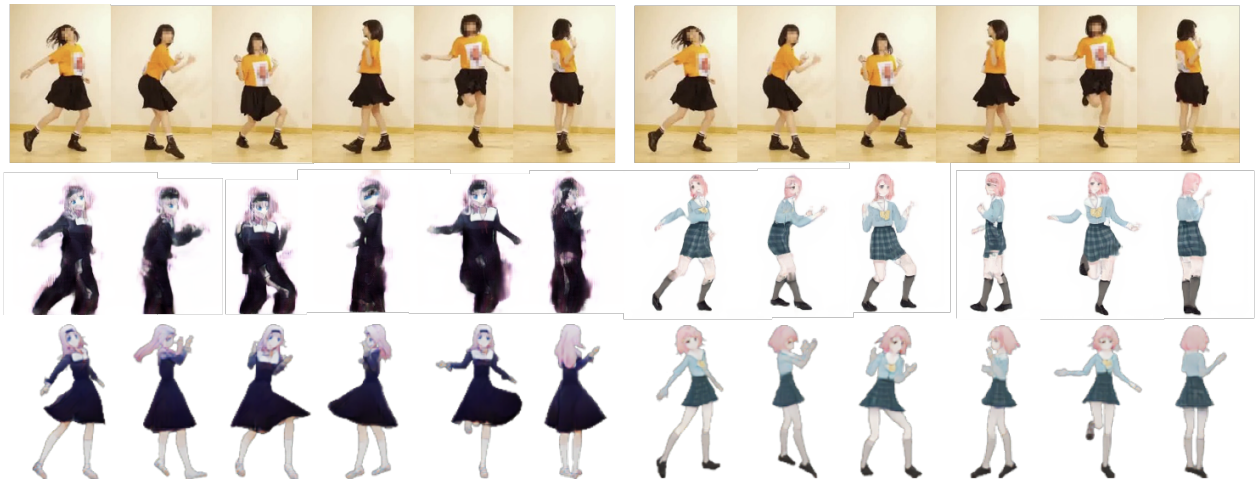

Figure 6: **Comparison with the SMPL-based method.** We compare results of our method (the 3nd row) with results of [27] (the 2rd row) when trying to resemble the target poses shown in the first row. Further diagnosis shows that parametric 3D human models like SMPL may not represent diverse clothing in anime. The long skirt of the first character is incorrectly mapped to the legs, and the short skirt of the second character seems to be tightened on the legs instead of sagging naturally.

erence image is provided, the target pose that finds a similar reference in the character sheets will be accurate, as expected.

Comparison with SMPL-based pose representations. We compare the results produced by CoNR to a real-world

Table 2: **Comparison on the number of input reference images.** We use character sheets of m reference images to train the CoNR model and then use character sheets of n reference images to evaluate the trained model.

Setting	epoch1		epoch2	
	L1	LPIPS	L1	LPIPS
$m=1, n=1$	0.0247	0.0832	0.0238	0.0801
$m=4, n=1$	0.0249	0.0865	0.0237	0.0827
$m=1, n=4$	0.0219	0.0798	0.0211	0.0764
$m=4, n=4$	**0.0187**	**0.0659**	**0.0179**	**0.0612**

digital human system [27] using the same target poses [1] as used in their demo. To perform the test, we use two characters sheets[2] with both pose and appearance unseen during training. One contains the same 4 images as in Figure 1, the other character sheet taken from a random Youtube video contains $\mathbf{I}_{1\sim5}$ as used in Figure 5. Figure 6 shows that the long skirt prevents a high accuracy estimation of the leg joints and that parametric 3D human models like SMPL may not may not handle the body shape of anime characters correctly.

The visual comparison of inference results in Figure 6 also indicates that CoNR can produce images at desired target poses with better quality.

4.3. Ablative Studies

We performed ablation studies on the UDP detector and renderer. Table 3 shows that UDP representation can be inferred from images using a U-Net [33]. A näive U-Net, which takes a concatenated tensor of 4 reference images and the target UDP as the input, does not provide acceptable results on this task, as shown in Table 4. The proposed CoNR method with both the feature warping and the CINN method significantly increases the performance, thus establishing a baseline for this task.

5. Limitations

The proposed method is unable to model the environmental lighting effects. The inputs of CoNR, neither the character sheet nor the UDP representation, could provide any information about the environmental or contextual information that could be utilized to infer any lighting effects asserted on the character. Even if the current form of CoNR may be used as a drop-in deferred shader for a portion of interactive applications, users still have to look for sketch relighting techniques [49, 17, 38] to properly deal with the lighting.

[1]The video is from https://download.impersonator.org/iper_plus_plus_latest_samples.zip

[2]A random anime character from a random Youtube video https://www.youtube.com/watch?v=m6k_t8yEyvE and another character illustrated by an amateur artist for this paper.

Table 3: **Ablation on UDP Detector.** U refers to a näive U-Net without batch-norms, b refers to a ResNet34 Encoder, $R50$ refers to a ResNet50 Encoder. The **red** setting is used in CoNR, the baseline proposed in this paper.

Setting	epoch1		epoch2	
	$\mathcal{L}_1$ †	BCE ‡	$\mathcal{L}_1$ †	BCE ‡
U	0.1247	0.0856	NaN	NaN
$U+R34$	0.1051	0.1068	0.1004	0.0747
$U+R50$	0.0969	0.0792	0.0971	0.0736

NaN: fail to converge, loss become NaN.
†: $\mathcal{L}_1$ of UDP.
‡: BCE of Mask.

Table 4: **Ablation on CoNR Renderer.** U refers to a näive U-Net without backbone or batch-norms. C refers to CINN method, G refers to the gird-sampling operation to perform output feature warping. R50 refers to ResNet50. The **red** setting is used in CoNR, the baseline proposed in this paper.

Setting	epoch1		epoch2	
	$\mathcal{L}_1$	LPIPS	$\mathcal{L}_1$	LPIPS
U	0.0313	0.1100	0.0311	0.1038
$U+G$	0.0309	0.1090	0.0308	0.1036
$U+C+R50$	0.0315	0.1097	0.0286	0.0977
$U+C+G$	0.0194	0.0673	0.0180	0.0612
$U+C+G+R50$	**0.0187**	**0.0659**	**0.0179**	**0.0612**

The proposed method is unable to model the dynamics of the character. The CoNR model requires detecting UDP from the images or videos used as the provider of the target body pose. This will limit the use of the proposed method in the animation or games as CoNR requires creating a storyboard beforehand to drive the generation of character images. To bypass the UDP detector, users have to rely on additional technologies like garment captures [9], physics simulations [6] or use learning based methods [42, 7, 39] to obtain a synthesized UDP.

The dataset may not fully reflect the distribution of anime characters in the wild. The collected dataset contains only human-like anime characters from the year 2014 to 2018. As building UDPs requires aligning character meshes according to joints, the models trained on this dataset may not be applied to animal-like characters, which follow joint hierarchies different from humans. Training with larger datasets may alleviate this limitation.

6. Conclusion

In this work, we introduce a novel task to render anime character images with desired pose using multiple images in character sheets. We developed a simple feed-forward baseline, CoNR, for this task. We anticipate the methods and the datasets presented in this paper would inspire further research.

References

[1] Pose estimation of anime/manga characters: A case for synthetic data. In *Proceedings of the 1st International Workshop on coMics ANalysis, Processing and Understanding*, 2016. 2

[2] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[3] Anonymous, Danbooru community, and Gwern Branwen. Danbooru2020: A large-scale crowdsourced and tagged anime illustration dataset. 4

[4] aydao. This anime does not exist. https://thisanimedoesnotexist.ai/. 2

[5] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[6] David Baraff and Andrew Witkin. Large steps in cloth simulation. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '98*, pages 43–54. ACM Press. 8

[7] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. Pbns: Physically based neural simulator for unsupervised garment pose space deformation. *arXiv preprint arXiv:2012.11310*, 2020. 8

[8] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '99*, pages 187–194. ACM Press. 2

[9] Derek Bradley, Tiberiu Popa, Alla Sheffer, Wolfgang Heidrich, and Tamy Boubekeur. Markerless garment capture. In *ACM SIGGRAPH 2008 Papers on - SIGGRAPH '08*, page 1. ACM Press. 8

[10] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 43(1):172–186, 2019. 2

[11] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[12] Chien-Lung Chou, Chieh-Yun Chen, Chia-Wei Hsieh, Hong-Han Shuai, Jiaying Liu, and Wen-Huang Cheng. Template-free try-on image synthesis via semantic-guided optimization. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2021. 2

[13] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 4

[14] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 4

[15] Oran Gafni, Oron Ashual, and Lior Wolf. Single-shot freestyle dance reenactment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[16] Chen Gao, Yichang Shih, Wei-Sheng Lai, Chia-Kai Liang, and Jia-Bin Huang. Portrait neural radiance fields from a single image. *arXiv preprint arXiv:2012.05903*, 2020. 2

[17] Zhengyan Gao, Taizan Yonetsuji, Tatsuya Takamura, Toru Matsuoka, and Jason Naradowsky. Automatic Illumination Effects for 2D Characters. In *NIPS Workshop on Machine. Learning for Creativity and Design.*, 2018. 2, 8

[18] Aaron Gokaslan, Vivek Ramanujan, Daniel Ritchie, Kwang In Kim, and James Tompkin. Improving shape deformation in unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[19] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5

[21] Zhenliang He, Meina Kan, and Shiguang Shan. Eigengan: Layer-wise eigen-learning for gans. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[22] Zhewei Huang, Shuchang Zhou, and Wen Heng. Learning to Paint With Model-Based Deep Reinforcement Learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[23] Yanghua Jin, Jiakai Zhang, Minjun Li, Yingtao Tian, Huachun Zhu, and Zhihao Fang. Towards the Automatic Anime Characters Creation with Generative Adversarial Networks. In *NIPS Workshop on Machine. Learning for Creativity and Design.*, 2017. 2

[24] Jerry Li. Pixiv dataset. 4

[25] Minjun Li, Yanghua Jin, and Huachun Zhu. Surrogate gradient field for latent space manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[26] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua 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 (ICCV)*, 2019. 2

[27] Wen Liu, Zhixin Piao, Zhi Tu, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan with attention: A unified framework for human image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021. 2, 7, 8

[28] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 2

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. 6

[30] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[31] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[32] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, 2015. 4, 5, 8

[34] Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural Re-Rendering of Humans from a Single Image. Lecture Notes in Computer Science. 2

[35] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Advances in Neural Information Processing Systems (NIPS)*, 2019. 2

[36] Aliaksandr Siarohin, Oliver J. Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion Representations for Articulated Animation. 2

[37] Hao Su, Jianwei Niu, Xuefeng Liu, Jiahe Cui, and Ji Wan. Vectorization of raster manga by deep reinforcement learning. *arXiv preprint arXiv:2110.04830*, 2021. 2, 3

[38] Wanchao Su, Dong Du, Xin Yang, Shizhe Zhou, and Hongbo Fu. Interactive Sketch-Based Normal Map Generation with Deep Neural Networks. 1(1):1–17. 8

[39] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-gif: Neural generalized implicit functions for animating people in clothing. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 8

[40] Hung-Yu Tseng, Matthew Fisher, Jingwan Lu, Yijun Li, Vladimir Kim, and Ming-Hsuan Yang. Modeling artistic workflows for image generation and editing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[41] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[42] Tuanfeng Y Wang, Tianjia Shao, Kai Fu, and Niloy J Mitra. Learning an intrinsic garment space for interactive authoring of garment animation. *ACM Transactions on Graphics (TOG)*, 38(6):1–12, 2019. 8

[43] Xinrui Wang and Jinze Yu. Learning to cartoonize using white-box cartoon representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[44] Pengfei Yao, Zheng Fang, Fan Wu, Yao Feng, and Jiwei Li. Densebody: Directly regressing dense 3d human pose and shape from a single color image. *arXiv preprint arXiv:1903.10153*, 2019. 2

[45] Jae Shin Yoon, Lingjie Liu, Vladislav Golyanik, Kripasindhu Sarkar, Hyun Soo Park, and Christian Theobalt. Pose-guided human animation from a single image in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[46] Mihai Zanfir, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. Human appearance transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[48] Song-Hai Zhang, Tao Chen, Yi-Fei Zhang, Shi-Min Hu, and Ralph R Martin. Vectorizing cartoon animations. *IEEE Transactions on Visualization and Computer Graphics*, 15(4):618–629, 2009. 3

[49] Qingyuan Zheng, Zhuoru Li, and Adam Bargteil. Learning to shadow hand-drawn sketches. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 8

Figure 1: **Evaluation results of Swapping Autoencoder for Deep Image Manipulation (SwapAE)** We consider alternative modeling on the proposed task using SwapAE[6]. We trained a Pytorch implementation of SwapAE [7] for more than two weeks with pairs of images with black background from random characters in our dataset. This figure, from the first to the last row, shows (a) the target pose $\mathbf{I}_{tar}$, (b) the reconstruction of $\hat{\mathbf{I}}_{tar}$, (c) the reference images $\mathbf{I}_{ref}$, (d) the reconstruction of $\hat{\mathbf{I}}_{ref}$, (e) swapped image using structure of $\mathbf{I}_{tar}$, and texture of $\mathbf{I}_{ref}$, (f) swapped image using structure of $\mathbf{I}_{ref}$, and texture of $\mathbf{I}_{tar}$. SwapAE performs a style transfer to swap the textures and structures on two characters directly and fails to construct an intermediate pose representation for the target character. We believe the parameter size of SwapAE (3x the size of CoNR already) may not be enough to learn an N-to-N swapping to handle the high diversity of textures, pose, and body structure for anime. The black background $RGB = (0, 0, 0)$ is replaced white background when making this figure.

Figure 2: **Comparision between detection results of hand-drawn anime character images in the wild using OpenPose[4], SMPLify[3] and the UDP Detector in CoNR.** The images are from the validation split of the hand-drawn dataset [2] with the corresponding ID number shown on the left. Three different representations, including joints, UDP, and SMPL, are detected on each image. The joint detection results consist of a lot of missing or wrong joints. The detected SMPL body can not fully handle the diverse body shapes of anime characters. HOWEVER, the UDP landmarks detected by the UDP detector in CoNR can cover fine details of the anime body, including hair, garments, and other accessories. This figure also indicates that the UDP detector could generalize well on unlabeled hand-drawn data.

Submission Summary

Conference Name
IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022

Track Name
CVPR2022

Paper ID
6818

Paper Title
Collaborative Neural Rendering using Anime Character Sheets

Abstract
Drawing images of characters at desired poses is an essential but laborious task in anime production. In this paper, we present the Collaborative Neural Rendering~(CoNR) method to create new images from a few arbitrarily posed reference images available in character sheets. In general, the high diversity of body shapes of anime characters defies the employment of universal 3D body models for real-world humans, like SMPL. To overcome this difficulty, CoNR uses a novel and compact form of landmark encoding to avoid requiring a unified UV mapping in the pipeline. In addition, CoNR's performance can be significantly increased when having multiple reference images by using feature space cross-view dense correspondence and warping in a specially designed neural network construct. Moreover, we collect a character sheet dataset containing over 700,000 hand-drawn and synthesized images of diverse poses to facilitate research in this area. The code and data will be released.

Created on
2021/11/6 15:13:28

Last Modified
2021/11/18 23:16:41

Authors
Zuzeng Lin (Tianjin University, Megvii Inc) < linzuzeng@hotmail.com>
Ailin Huang (Wuhan University, Megvii Inc.) < p2oileen@whu.edu.cn>
Zhewei Huang (MEGVII) < hzwer@pku.edu.cn>
Chen Hu (Megvii Technology) < ccchendada@gmail.com>
Shuchang Zhou (MEGVII Technology) < zsc@megvii.com>

Primary Subject Area
Image and video synthesis and generation

Secondary Subject Areas
Vision + graphics

Conflicts of Interest
Shuchang Zhou - zsc@megvii.com
- a co-author

Zhewei Huang - hzwer@pku.edu.cn
- a co-author

Submission Files
CoNRv10.pdf (2.6 Mb, 2021/11/18 23:14:44)

Supplementary Files
CoNRv2.zip (32 Mb, 2021/11/23 20:06:18)

Submission Questions Response
1. Dual/double submissions policy
By submitting a manuscript to CVPR 2022, the authors acknowledge that it has not been previously published or accepted for publication in substantially similar form in any peer-reviewed venue including journal, conference,

workshop, or archival forums. Furthermore, no paper substantially similar in content has been or will be submitted to another conference or workshop during the review period of CVPR 2022. The authors also attest that they did not submit a substantially similar manuscript to CVPR 2022. Violation of any of these conditions will lead to rejection.

Agreement accepted

2. Ethics Guidelines

I have read the Ethics Guidelines. My paper conforms to them.

Agreement accepted

3. Discussion of potential negative impact of your work?

In your paper, did you discuss any potential negative societal impact of your work?

Examples of negative societal impact include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), environmental impact (e.g., training huge models), fairness considerations (e.g., deployment of technologies that could further disadvantage historically disadvantaged groups), privacy considerations (e.g., a paper on model/data stealing), and security considerations (e.g., adversarial attacks).

We expect many papers to be foundational research and not tied to particular applications, let alone deployments, but being foundational does not imply that research has no societal impact. If you see a direct path to any negative applications, you should point it out, even if it is not specific to your work. In a theoretical paper on algorithmic fairness, you might caution against overreliance on mathematical metrics for quantifying fairness and examples of ways this can go wrong. If you improve the quality of generative models, you might point out that your approach can be used to generate deep-fakes for disinformation. On the other hand, if you develop a generic algorithm for optimizing neural networks, you do not need to mention that this could enable people to train models that generate deep-fakes faster.

Consider different stakeholders that could be impacted by your work. It is possible that research benefits some stakeholders while harming others. Pay special attention to vulnerable or marginalized communities.

Consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

If there is negative societal impact, you should also discuss any mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of CV models).

No potential negative impact

4. Discussion of limitations

In your paper, did you describe the limitations of your work?

You are encouraged to create a separate "Limitations" section in your paper.

Point out any strong assumptions and how robust your results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). Reflect on how these assumptions might be violated in practice and what the implications would be.

Reflect on the scope of your claims, e.g., if you only tested your approach on a few datasets or did a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.

Reflect on the factors that influence the performance of your approach. For example, a recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting.

We understand that authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection. It is worth keeping in mind that a worse outcome might be if reviewers discover limitations that

are not acknowledged in the paper. In general, we advise authors to use their best judgement and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

Yes

5. Data contributions policy

Did you claim a dataset release as one of the core scientific contributions of your paper?

If your paper submission is claiming a dataset release as one of its contributions, it is expected that the dataset will be made publicly available no later than the camera-ready deadline. To facilitate handling papers with dataset contributions, the authors need to indicate here whether the submitted paper claims a dataset as one of its core scientific contributions. Any paper that claims a dataset contribution will need to provide a URL for the dataset when submitting the camera ready. The CVPR website will provide a list of all papers with dataset contributions and include a link to the author provided URL in order to facilitate wide dissemination of new datasets to the CVPR audience.

Note that this does NOT imply that all datasets used in CVPR submissions must be public. The use of private or otherwise restricted datasets (e.g. for training or experimentation) continues to be permitted. However, private or otherwise restricted datasets cannot be claimed as core scientific contributions of the paper as they do not become available to the scientific community.

New dataset contribution claim

6. Use of existing assets

In your main paper or supplemental, if your work uses existing assets, such as datasets or code,

(a) did you cite the creators?

Cite the original paper that produced the code package or dataset. Remember to state which version of the asset you are using. If possible, include a URL.

(b) Did you mention the license of the assets?

State the name of the license (e.g., CC-BY 4.0) for each asset. If you scraped data from a particular source (e.g., website), you should state the copyright and terms of service of that source. If you are repackaging an existing dataset, you should state the original license as well as the one for the derived asset (if it has changed).
If you cannot find this information online, you are encouraged to reach out to the asset's creators.

Yes

7. Personal data/Human subjects

If you work makes use of personal data and/or human subjects,

(a) Did you discuss whether and how consent was obtained from people whose data you are using/curating (in your main paper or supplemental)?

For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

Depending on the country in which research is conducted, Institutional Review Board (IRB) approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper. For initial submissions, do not include any information that would break anonymity, such as the institution conducting the review.

Even if you used an existing dataset, you should check how data was collected and whether consent was obtained. We acknowledge this might be difficult, so please try your best; the goal is to raise awareness of possible issues that might be ingrained in our community.

(b) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content?

There are some settings where the existence of this information is not necessarily bad (e.g., swear words occur naturally in text). This question is just to encourage discussion of potentially undesirable properties.

Explain how you checked this (e.g., with a script, manually on a sample, etc.).

No personal/human subject data collected

8. Code/Data included for review

Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)?

The instructions should contain the exact command and environment needed to run to reproduce the results. Main experimental results include your new method and baselines, where appropriate. You should try to capture as many of the minor experiments in the paper as possible. If a subset of experiments are reproducible, you are encouraged to state which ones are. At submission time, to preserve anonymity, remember to release anonymized versions.

While we encourage release of code and data, we understand that this might not be possible, hence "No, code/data is proprietary" is an acceptable answer.
Please give details in your submission for any other reason you are not submitting code/data.

Yes

View Reviews

Paper ID
6818

Paper Title
Collaborative Neural Rendering using Anime Character Sheets

Track Name
CVPR2022

Reviewer #1

Questions

2. Summary. In 5-7 sentences, describe the key ideas, experimental or theoretical results, and their significance.
The paper proposes a novel method for generating new views of anime characters given a small number of reference views in certain poses, and a novel "Ultra-dense-pose" encoding of the desired target pose. Claimed technical novelty lies in the proposed deep architecture. Furthermore, authors also propose a novel dataset of anime characters suitable for training / testing the method. Ablative studies in the experimental section reveal performance improvements.

3. Strengths. Consider the significance of key ideas, experimental or theoretical validation, writing quality, data contribution. Explain clearly why these aspects of the paper are valuable. Short bullet lists do NOT suffice.
+ A novel dataset.
+ Tackles an interesting task.
+ Compelling qualitative results.

4. Weaknesses. Consider the significance of key ideas, experimental or theoretical validation, writing quality, data contribution. Clearly explain why these are weak aspects of the paper, e.g. why a specific prior work has already demonstrated the key contributions, or why the experiments are insufficient to validate the claims, etc. Short bullet lists do NOT suffice.
1. Lack of technical novelty.
1.1.: While the application is novel, technically, the proposed method is very similar to DensePose Transfer [a] and several other works building on top of [a]. The similarities include the conditioning on the input DensePose encoding, or the warping module (also used in [a]). [a] is not cited, nor considered as a baseline in the experimental section.
1.2.: The message passing component is very similar to that of PointNet [b], where input-specific encodings are averaged over all inputs and the average is concatenated as an input to the next layers. There are several additional graph-learning architectures that propose a similar trick and build on top of PointNet.
1.3.: Ultra Dense-Pose proposes to encode the dense canonical map with xyz locations of the canonical surface. In order to evaluate the contribution of this novelty, it is important to quantitatively evaluate whether xyz-based encoding gives better results than the originally proposed UV-map encoding of DensePose.

2. Insufficient experimental section.
The quality of the experimental section is insufficient overall:
2.1.: Missing comparison to existing baselines: This includes comparison to [a] or any other existing method that transfers textures using a DensePose-like encoding. Although authors propose a novel type of pose encoding, which uses the xyz coordinates of the template mesh as opposed to DensePose's uv-map, at least one of the existing DensePose-based methods (e.g. [a]) can be retrained on top of the novel type of embedding without any additional effort.
2.2.: Missing relevant ablation studies: The paper evaluates the contribution of various deep architectures in Tab. 3 and 4. However, there are several more important design choices that should be ablated, namely: The

message-passing component and all losses described in eqs 2-5.
2.3.: I could not find the exact meaning of $\mathcal{\ell}_1$, LPIPS, epoch1, epoch2 from Tab. 2-4 anywhere in the text. What was the training / test set and the exact evaluation protocol to obtain these numbers, and what do the numbers mean actually? What does epoch1/epoch2 stand for? If these are training epochs, it is not clear what is the contribution of presenting numbers of a non-converged model only after the 1st epoch?

[a] Alp Guler et al.: Dense Pose Transfer
[b] Qi et al.: PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation

5. Paper rating (pre-rebuttal).
Borderline

7. Justification of rating. What are the most important factors in your rating?
While the applicattion and quantitative results are quite interesting, the paper lacks techincal novelty and the experimental evaluation is also severely lacking (most importnantly, missing comparison to existing works). I am leaning towards rejection but I would rather confirm this with other reviewers.

8. Are there any serious ethical/privacy/transparency/fairness concerns? If yes, please also discuss below in Question 9.
No

9. Limitations and Societal Impact. Have the authors adequately addressed the limitations and potential negative societal impact of their work? Discuss any serious ethical/privacy/transparency/fairness concerns here. Also discuss if there are important limitations that are not apparent from the paper.
All addressed sufficiently.

10. Is the contribution of a new dataset a main claim for this paper? Have the authors indicated so in the submission form?
Dataset contribution claim in the paper. Indicated in the submission form

14. Final recommendation based on ALL the reviews, rebuttal, and discussion (post-rebuttal).
Reject

15. Final justification (post-rebuttal).
It is true that the dataset is a clear contribution but all other parts, including presentation, technical contribution, novelty, and experimentation are severly lacking. The 4 points below are a justification of my vote for rejection:

1) The presentation is very confusing in some cases but authors promised to fix this in the final version - OK but this will require a lot of edits.
2) There is not a single comparison to an existing baseline in the paper, despite the fact that style-transfer literature is now very abundant. Authors did not address this concern sufficiently in the rebuttal.
3) Ablation studies: The present ablation studies in Tab 3, 4 are fairly irrelevant. As suggested in my review, authors ablated the message passing component, but did not ablate all used losses (4 loss terms in total), which is a big omission.
4) Authors propose the "novel" Ultra Dense Pose encoding. I do not see a single proof in the paper that this "xyz"-based encoding would actually be better than the standard DensePose 2D UV map. In fact, the canonical xyz coordinates have imho a bigger chance of introducing conflicts. This is because points close in 3D space can be in fact quite far in the (more meaningful) geodesic 2D UV space.

Reviewer #2

Questions

2. Summary. In 5-7 sentences, describe the key ideas, experimental or theoretical results, and their significance.
This paper presents a new framework to generate new anime-styled images based on input reference pose and the Anime Character Sheet (ACS) as a style reference. The ACS can have an arbitrary number of reference

poses, and the poses can be in arbitrary order. The proposed CINN is used for modelling the Character Sheet, and UDP is a landmark representation for the 3D poses. Results show that the proposed method outperformed a SMPL-based method. A new dataset and the code will be released.

3. Strengths. Consider the significance of key ideas, experimental or theoretical validation, writing quality, data contribution. Explain clearly why these aspects of the paper are valuable. Short bullet lists do NOT suffice.
- New application and new dataset: the new application in presented is well-motivated and future research will be supported by the new dataset which contain a significant amount of data.

- The UDP representation is novel which represent the landmark effectively. The image-based representation also facilitate the computation.

- The paper is well-written in general, with the details of the proposed framework and rationale behind the design. There are also some discussions on the different between the design and other related work.

4. Weaknesses. Consider the significance of key ideas, experimental or theoretical validation, writing quality, data contribution. Clearly explain why these are weak aspects of the paper, e.g. why a specific prior work has already demonstrated the key contributions, or why the experiments are insufficient to validate the claims, etc. Short bullet lists do NOT suffice.
- the coparison is relatively limited, although it is understandable that this is a new applciiton. On the other hand, the ultimate goal of this paper ris to generate animation, and perceptual study has been widely used in motion synthesis research.

5. Paper rating (pre-rebuttal).
Weak Accept

7. Justification of rating. What are the most important factors in your rating?
This is a new application and the new dataset will benefit the community. The proposed methodology produces good quality motions as shown in the demo video. The quantitative analysis also show positive results.

8. Are there any serious ethical/privacy/transparency/fairness concerns? If yes, please also discuss below in Question 9.
No

9. Limitations and Societal Impact. Have the authors adequately addressed the limitations and potential negative societal impact of their work? Discuss any serious ethical/privacy/transparency/fairness concerns here. Also discuss if there are important limitations that are not apparent from the paper.
Some limitations are discussed in the paper. Since the data are anime drawings, I do not see any major ethical/privacy concerns on the new dataset.

10. Is the contribution of a new dataset a main claim for this paper? Have the authors indicated so in the submission form?
Dataset contribution claim in the paper. Not indicated in the submission form

11. Additional comments to author(s). Include any comments that may be useful for revision but should not be considered in the paper decision.
see above

14. Final recommendation based on ALL the reviews, rebuttal, and discussion (post-rebuttal).
Borderline Reject

15. Final justification (post-rebuttal).
After reading all reviews and the rebuttal, some weaknesses potined out in the review have not been fully addressed in the rebuttal, especially the evaluation and ablation study. This work can start a very interesting research direction and providing others with a valuable dataset.

Reviewer #3

Questions

2. Summary. In 5-7 sentences, describe the key ideas, experimental or theoretical results, and their significance.
The paper introduces a method for 2D character synthesis conditioned by a few images with different view and pose (Character Sheet). The synthesis is driven by an input motion sequence, represented by Ultra-Dense Pose feature map. The paper proposes a novel architecture for the collaborative neural rendering network and achieve plausible results considering the character sheet only contain a few number of images. Experiments also show that although the resolution of the synthesised image is limited, the overall quality is encouraging.

3. Strengths. Consider the significance of key ideas, experimental or theoretical validation, writing quality, data contribution. Explain clearly why these aspects of the paper are valuable. Short bullet lists do NOT suffice.
The paper achieves plausible results for 2D motion retargeting. The proposed CINN network architecture utilise the multi view/multi pose information from the character sheet in a smart yet efficient way. The RGB based pose representation is straightforward but powerful. The authors also mention they collected a dataset for this task. The dataset itself seems to be very helpful for many other related research projects.

4. Weaknesses. Consider the significance of key ideas, experimental or theoretical validation, writing quality, data contribution. Clearly explain why these are weak aspects of the paper, e.g. why a specific prior work has already demonstrated the key contributions, or why the experiments are insufficient to validate the claims, etc. Short bullet lists do NOT suffice.
The paper is interesting but the writing is unfortunately problematic. Many technical details are missing from the manuscript, e.g., how is the UDP detector trained? how is the feature been averaged in CINN and how is the UDP concatenated described in math equation? The evaluation is also very insufficient. What does the output of UDP detector look like? How is the generalisation of such method to different motion source? How is the gap between synthetic training set and real test set? Baseline comparison is also missing, e.g., similar 2D based approach https://arxiv.org/pdf/2111.05916.pdf, or 3D based approach https://arxiv.org/abs/2201.04127.

I am also confused that why the demo video shows some secondary effects between 0'20" to 0'30"? Since UDP is pose based, there's no way for it to understand motion. Ln 839 also discussed this limitation. However in the demo video, looks like the long skirt is deformed not only by pose but also by the motion.

5. Paper rating (pre-rebuttal).
Borderline

7. Justification of rating. What are the most important factors in your rating?
In general, I would like to support this work to be accepted. However, this work is not properly evaluated (only 2 styles and 1 source motion is shown, no evaluation on UDP detector), critical technical details are missing, no baseline comparison. This makes me difficult to tell if the method described in the manuscript is solid or not. I'd be happy to argue for acceptance if more results/evaluation can be discussed in rebuttal.

8. Are there any serious ethical/privacy/transparency/fairness concerns? If yes, please also discuss below in Question 9.
No

9. Limitations and Societal Impact. Have the authors adequately addressed the limitations and potential negative societal impact of their work? Discuss any serious ethical/privacy/transparency/fairness concerns here. Also discuss if there are important limitations that are not apparent from the paper.
The limitations are properly discussed. Societal Impact is missing however.

10. Is the contribution of a new dataset a main claim for this paper? Have the authors indicated so in the submission form?
Dataset contribution claim in the paper. Indicated in the submission form

14. Final recommendation based on ALL the reviews, rebuttal, and discussion (post-rebuttal).
Borderline Accept

15. Final justification (post-rebuttal).
I appreciate the technical details provided by the authors in the rebuttal. I now have a better understanding of

the proposed method. However, the rebuttal makes me feel the original submission is over claimed for the technical contribution (e.g., the upstream preprocessing actually significantly reduce the difficulty of training as described in the paper). So I'm still at a borderline position but would be happy to support the acceptance of this paper if other reviewers are positive.

Thanks for the constructive suggestions of all reviewers!

Common questions: Q1: Comparisons and related work

A: Thanks for referring to DT, and two concurrent works, DIW and HNeRF. We will cite them. Both DT and DIW apply GAN to a single input, while CoNR is more of synthesizing images from a set of references that better constrain appearances (L160-168 and Fig. 5). HNeRF utilizes a NeRF-like MLP to stylize SMPL which is hand-crafted for human. Since none of the three works have released codes, exact comparisons are difficult. Nevertheless, our comparisons show going from 2D representation to UDP causes an LPIPS loss reduction as much as 16% (L211-228). Kindly refer to Fig. 6 and Supplementary for more comparative studies with existing works, and Table 1,3,4 for ablative studies.

To Reviewer #1: Q1: Definition of Epoch and notations

A: In Table 1,2 and 4, $\mathcal{L}_1$ is the $\mathcal{L}_{photometric}$ in L577, In Table 3, $\mathcal{L}_1$ is $\mathcal{L}_{UDP}$ defined in L565. We will fix the confusing notations in revision. LPIPS is defined in L579-582. We use "Epoch" just as the usual definition, meaning one round of traversal of the entire training dataset with characters used as $\mathbf{I}_{tar}$. We found 2 epochs (60k iterations) is sufficient for producing visually acceptable results, while doubling the training time (taking one more week on 4 GPUs) can only improve image quality mildly.

Q2: Relationship with PointNet

A: Thanks for pointing out the links to PointNet (will add citation). Our innovation can be roughly understood as using point features with feature space flows on the vastly different task of Anime Character Sheet (ACS) rendering.

Q3: More ablations on CINN (Message-passing)

A: We perform ablations on the number of messages passing, which will be included in the revision. More ghosting can be observed when sub-networks communicate less than three times. Here $\mathcal{L}_{ph}$ stands for $\mathcal{L}_{photometric}$.

Messaging	epoch1		epoch2	
	$\mathcal{L}_{ph}$	LPIPS	$\mathcal{L}_{ph}$	LPIPS
0 time	0.028	0.107	0.026	0.099
1 time	0.019	0.066	0.018	0.063
3 times	**0.018**	**0.065**	**0.017**	**0.061**

To Reviewer #2: Thanks for appreciating our work. We will provide the code and the dataset to help setting up an anime baseline for future quantitative and perceptual evaluations.

To Reviewer #3: Q1: Details of UDP

A: A UDP specifies a character's pose by mapping 2D viewport coordinates to feature vectors, which are 3-tuple floats that continuously and consistently encode body surfaces. In this way, a UDP can be represented as a color image $\mathbf{P}_{tar} \in \mathbb{R}^{H\times W\times 3}$ with pixels corresponding to landmarks $L_{(x,y)} \in \mathbb{R}^3$. Non-person areas of the UDP image are simply masked and ignored. A UDP can be extracted from 3D meshes (Fig. 2) or detected from an image $\mathbf{I}_{tar}$. UDP datasets can also be built for humans or animals with the same body structure. The training of the UDP detector is described in L553-572. The detector generalizes well to humans and even toys, thanks to the augmentations.

Q2: Details of messaging

A: We apply weighted averaging on cross-view messages, where the weights are predicted by CNN and normalized by the number of views. We will include the formulation.

Q3: Synthesis/Real Gap & Variety in Styles

A: Our model can cope with the diverse styles of anime, including differences between synthesized images and hand-drawn images, thanks to data augmentations. Here we list some random CoNR outputs on the evaluation split.

Inference results for synthetic (rendering) data

Inference results for real (hand-drawn) data

Q4: Details of video production

A: CoNR renders images given UDPs and ACSs. The video demonstrates that CoNR generalizes well to unseen hand-drawn ACS and novel UDPs. To obtain UDPs for CoNR to animate the ACS (at top of video), we employ two different upstream preprocessing methods. In both cases, the input human pose detected with PoseNet is first used to rig two unseen 3D character models with the help of Bullet Physics Engine (PE). In 05" to 16", the UDP detector that was used in the training pipeline, detects UDPs from rendered video (on the left of 44"). To illustrate the use case of game developers, we also feed UDPs directly computed by PE (L843-846) from an untextured mesh with long skirt into our neural renderer in 16" to 29". Unfortunately, PE introduces unwanted secondary garment motions, which leak into final CoNR results via the UDPs. We apologize for the confusion introduced by upstream preprocessing steps irrelevant to CoNR and wrong impressions that UDPs is fuzzy on garment poses. We will make the inference path with PE clear in Fig. 3, and replace all human used for illustrating poses in the paper and video with UDP, which is a rigorous and detailed pose representation.