View Reviews

Paper ID
99

Paper Title
Collaborative Neural Rendering using Anime Character Sheets

Reviewer #1

Questions

2. Summarize the paper's claimed primary contributions: In 5-7 sentences, describe the key ideas, results, findings, and significance as claimed by the paper's authors.
The paper proposes a DNN model for appearance transfer of anime images, namely generating a novel anime image of a specific character, given a desired pose (a pose reference) and a small collection of images showing the character at different viewpoints (the character sheet). The model task is designed to assist anime video creation by human artist. The model is based on UNet-type architecture with parallel decoder streams, which encode-decode different images from the support collection. The pose representation is integrated into these streams. In addition, the paper presents a dataset of character sheets containing 700K hand-drawn and synthetic images of diverse poses to further research.

3. What do you see as the main strengths of this work? Consider, among others, the significance of critical ideas, validation, writing quality, and data contribution. Explain clearly why these aspects of the paper are valuable. ACs are instructed to ignore unsupported responses.
Appearance transfer on anime images may have several useful applications. The collected dataset is useful to further explore the differences between appearance transfer in real and anime images.

4. What do you see as the main weaknesses of this work? Clearly explain why these are weak aspects of the paper, e.g., why a specific prior work has already demonstrated the key contributions, why the experiments are insufficient to validate the claims, etc. ACs are instructed to ignore unsupported responses.
The model is not quantitatively evaluated against other models, which might be tuned to perform the same task.

There is no other evaluation metric except the training loss. It is difficult then to compare performance between different models.

The technical novelty of the proposed model seems to be limited. It contains a UNet-based encoder and several decoders, in parallel streams but connected to each other. A pose representation as color map is fed to the first decoder on each stream in addition to the encoder output. The structure of parallel connected encode-decoder streams appears in various forms in previous work (e.g., Wang et al., 2020, HRNet).

6. [Rate the paper as it stands now (pre-rebuttal). Borderline will not be an option for your final post-rebuttal recommendation, and so it should only be used rarely now.]
Reject

7. Justify your rating. Be specific: What are the most critical factors in your rating? What points should the authors cover in their rebuttal? Your reply should clearly explain to the authors what you need to see in order to increase your rating.
The model is not quantitatively evaluated against other models. See more comments in Sec. 4

11. Justify your post-rebuttal assessment. Acknowledge any rebuttal and be specific about the final factors for and against acceptance that matter to you. (Will be visible to authors after author notification)
Given the feedback and other reviews, I still think that the contributions of this paper are limited.

12. Give your final rating for this paper. Don't worry about poster vs oral. Consider the input from all reviewers, the authors' feedback, and any discussion. (Will be visible to authors after author notification)

Weak Reject. I tend to vote for rejecting this submission, but accepting it would not be that bad.

Reviewer #2

Questions

2. Summarize the paper's claimed primary contributions: In 5-7 sentences, describe the key ideas, results, findings, and significance as claimed by the paper's authors.
In this paper, a novel CoNR method is proposed with creating new images from a few arbitrarily posed reference images available in character sheets. Firstly, the watershed algorithm is used to separate the anime character from images to build the dataset, and the dataset is annotated by manual and UDP methods. Next, the renderer is employed to generate the character images of the desired pose. Finally, the existing physics engine is exploited to produce the pose of the anime character, where the CINN is used as the decoder. The experiment results demonstrate the superiority of the proposed method.

3. What do you see as the main strengths of this work? Consider, among others, the significance of critical ideas, validation, writing quality, and data contribution. Explain clearly why these aspects of the paper are valuable. ACs are instructed to ignore unsupported responses.
1) A UDP representation is designed in this paper to build a large character sheet dataset containing diverse poses, which is an interesting idea.
2) A collaborative inference method of feed-forward neural network is exploited to model character sheets as a dynamically-sized set of images.
3) Comprehensive evaluation on a variety of anime video creation demonstrates the effectiveness of the proposed CoNR.

4. What do you see as the main weaknesses of this work? Clearly explain why these are weak aspects of the paper, e.g., why a specific prior work has already demonstrated the key contributions, why the experiments are insufficient to validate the claims, etc. ACs are instructed to ignore unsupported responses.
1) How to generate more than one desired pose of the anime character by using the proposed method?
2) The authors should increase more comparative experiments and set more evaluation indicators to verify the effectiveness and superiority of the method proposed in this paper.
3) Some details are not very clear.
a. The authors claim that "the high diversity of body shapes of anime characters defies the employment of universal body models for real-world humans, like SMPL". What is SMPL? How is it different from that of anime characters? What influence will it have?
b. The authors mention the image translation in Section 2.1, but the reviewer could not find the any relevant description.
c. Can the target pose ${{P}_{tar}}$ be taken as one of the ${{I}_{t}}$? In addition, is there noise in the data? How is it generated? How to deal with it?
d. What is the purpose of the weights output by the CNN? Is it to highlight the importance of the reference images? Why not use the attention mechanism?
e. The authors should state the principles of the proposed model CoNR in more details (such as equations, flowchart, pseudocode) .

5. Reproducibility: Could the work be reproduced by a talented graduate student from the information in the paper?
Agreement accepted

6. [Rate the paper as it stands now (pre-rebuttal). Borderline will not be an option for your final post-rebuttal recommendation, and so it should only be used rarely now.]
Weak Accept

7. Justify your rating. Be specific: What are the most critical factors in your rating? What points should the authors cover in their rebuttal? Your reply should clearly explain to the authors what you need to see in order to increase your rating.

Well written. Good empirical analysis. But in some places the paper lacks the justifications and proper reasoning. Experiment/results section need some better explanations.

11. Justify your post-rebuttal assessment. Acknowledge any rebuttal and be specific about the final factors for and against acceptance that matter to you. (Will be visible to authors after author notification)
I have checked the rebuttal and the authors clearly address all my concerns.

12. Give your final rating for this paper. Don't worry about poster vs oral. Consider the input from all reviewers, the authors' feedback, and any discussion. (Will be visible to authors after author notification)
Weak Accept. I tend to vote for accepting this submission, but rejecting it would not be that bad.

Reviewer #3

Questions

2. Summarize the paper's claimed primary contributions: In 5-7 sentences, describe the key ideas, results, findings, and significance as claimed by the paper's authors.
This paper has two primary contributions: a) a new dataset that contains anime character images in which parts of the synthetic dataset include UDP label, a two UV map represents the semantics; b) a neural network called CoNR, that is able to take a set of reference images of an anime characters and a new UDP to "render" a new pose of the same character.

3. What do you see as the main strengths of this work? Consider, among others, the significance of critical ideas, validation, writing quality, and data contribution. Explain clearly why these aspects of the paper are valuable. ACs are instructed to ignore unsupported responses.
1. The writing is clear and easy to understand;
2. The task is novel;
3. The contributed datasets and tools can be useful for both the research community and the animation industry.

4. What do you see as the main weaknesses of this work? Clearly explain why these are weak aspects of the paper, e.g., why a specific prior work has already demonstrated the key contributions, why the experiments are insufficient to validate the claims, etc. ACs are instructed to ignore unsupported responses.
1. The UDP representation is not very usable for artists. It might take more time for artists to draw a UDP than to draw the character directly. I guess one has to resort to rigging the character in 3D in order to "render". However, if that is the case, one can directly use the traditional animation-style render engine. In this sense, the provided method and dataset are not very useful in practice.
2. When you want to transfer the pose from one character to another character (as demoed in Fig 4), the proposed UDP requires that the body shape and dressing are similar. For example, if the template character has a significantly different hair style, then the whole method built on UDP will not work well.

5. Reproducibility: Could the work be reproduced by a talented graduate student from the information in the paper?
Agreement accepted

6. [Rate the paper as it stands now (pre-rebuttal). Borderline will not be an option for your final post-rebuttal recommendation, and so it should only be used rarely now.]
Borderline

7. Justify your rating. Be specific: What are the most critical factors in your rating? What points should the authors cover in their rebuttal? Your reply should clearly explain to the authors what you need to see in order to increase your rating.
I fully understand that the authors put a lot of efforts to make this paper, the method, and the dataset. However, given the intrinsic limitation (I believe) of the UDP, I'd like to hear more from the authors before make my final recommendation.

Reviewer #4

Questions

2. Summarize the paper's claimed primary contributions: In 5-7 sentences, describe the key ideas, results, findings, and significance as claimed by the paper's authors.
Given a set of hand-drawn or synthesized images of anime characters, the paper presents an approach for synthesizing images of the character in novel poses. This is achieved via a CNN architecture that takes the source images of the characters and a target pose (represented using ultra-dense-pose) as input and generates the images of the person in the target pose. The CNN architecture is designed to be permutation invariant via a message passing mechanism. The experiments are performed on a new dataset consisting of 70k images of hand-drawn and synthesized anime characters. The experiments demonstrate that the proposed approach outperforms existing baselines i.e., Liquid Warping GAN.

3. What do you see as the main strengths of this work? Consider, among others, the significance of critical ideas, validation, writing quality, and data contribution. Explain clearly why these aspects of the paper are valuable. ACs are instructed to ignore unsupported responses.
- The paper addresses the challenging problem of novel pose synthesis of anime characters (or any other articulated object in general). The problem has a wide variety of applications and is of great importance to the community.

- The qualitative video is very good and entertaining.

4. What do you see as the main weaknesses of this work? Clearly explain why these are weak aspects of the paper, e.g., why a specific prior work has already demonstrated the key contributions, why the experiments are insufficient to validate the claims, etc. ACs are instructed to ignore unsupported responses.
-- Writing
- The writing of the paper is not fully polished. There are many typos and incorrect sentence structures.

-- Novelty
The proposed approach is not novel and the idea of using dense pose has been explored in many existing works for novel pose synthesis.

-- Missing Related Work
1. A. Raj et al., ANR: Articulated Neural Rendering for Virtual Avatars, CVPR'21
2. N Neverova, Continuous surface embeddings, NeurIPS'19
[1] also proposes an approach for novel pose synthesis and has the advantage that it doesn't require any reference image during inference.
[2] also proposes a continous and high-res representations of DensePose.

-- Experiments
Experiments on Anime Characters are not sufficient for the comparison with state-of-the-art. The authors should also perform some experiments on humans or other articulated objects since the underlying framework will remain unchanged. The paper [1] can be used to obtain Ultra-Dense-Pose for humans or other animals. Also, [2] shows that highly detailed DensePose is not required if the features are processed using a UNet since the network can learn to hallucinate the required details condition on the pose.

6. [Rate the paper as it stands now (pre-rebuttal). Borderline will not be an option for your final post-rebuttal recommendation, and so it should only be used rarely now.]
Weak Reject

7. Justify your rating. Be specific: What are the most critical factors in your rating? What points should the authors cover in their rebuttal? Your reply should clearly explain to the authors what you need to see in order to increase your rating.
The paper in its current form is not ready to be published at ECCV. The paper requires a major revision in terms

of writing and experiments. Also, I cannot find any significant novel contribution in the paper. Important related works are also missing.

11. Justify your post-rebuttal assessment. Acknowledge any rebuttal and be specific about the final factors for and against acceptance that matter to you. (Will be visible to authors after author notification)
I thank the authors for their rebuttal.

While the authors have addressed most of my comments, I think the paper requires a significant revision to incorporate them in the paper.

Regarding author's arguments for "Relation to ANR" and "Why not let UNet hallucinate", I agree with authors but we need to prove this empirically. Simple baselines often tend to perform better than we expect.

Hence, I am keeping my original rating of Weak Reject and encourage the authors to address reviewer's comments and submit to a future venue.

Here is another related work that might be useful:

https://youngjoongunc.github.io/nhp/

12. Give your final rating for this paper. Don't worry about poster vs oral. Consider the input from all reviewers, the authors' feedback, and any discussion. (Will be visible to authors after author notification)
Weak Reject. I tend to vote for rejecting this submission, but accepting it would not be that bad.

Thanks for the questions of all reviewers. This research is about an unexplored direction of combining AI and art. The creative nature of our rendering makes it very different from seemingly-related tasks like digital human or style transfer, which begin by detecting existing images. Our large collection containing thousands of raw 3D anime models may expand research in AI-assisted anime creation and related fields to an unprecedented level.

To R1, R2: Quantitative comparisons.

A: Our focus is on solving a new task. We quantitatively compare multiple baselines (UNet, CINN, etc.) (Line 559-566) to find out how well they could fit in this task. At present, there is no neural digital system designed for anime characters. We include typical but unaligned methods from the field of digital human [30] and style-transfer [35] and compare the visual effects in our paper and supplementary. We will add quantitative results of these results.

To R1: Q1: Evaluation metric is only training loss?

A: As explained in Line 404, all losses are measured on the validation set rather than the training set. We will make this clear in tables and report additional PSNR metrics.

Q2: Relation to HRNet and previous model designs?

A:We use a straightforward baseline to deal with a dynamic set of input images in a commutation-invariant manner and align their features, which is required for the proposed task. HRNet explores the design of high and low resolution features, which are completely orthogonal to our design.

To R2: Q1: How to generate more than one pose?

A: We repeatedly dump UDP P_{tar} from the physics engine and feed it into CoNR to get each image. We keep the 4 images in $\mathbf{S}_{ref}$ the same in our demo.

Q2: Questions about details.

A: We will add the description of image translation and SMPL to the revised paper. SMPL is a realistic 3D mesh of the naked human body that is based on skinning and blendshapes. Anime characters are different from SMPL meshes and from each other in shape and topology.

All code will be public. We will include some pseudo-code and chart as documentation. The generation of our dataset is described in the appendix. In short, we drive existing 3D models into different poses in rendering software, and record the annotations. The noise of dataset may not be a major issue.

The weights are to fuse information from different views and may explain the contribution of each view. Other implicit attention mechanisms are also interesting but less straightforward in our early exploration.

To R3: Q1: Practicability and limitation?

A: Our statements of limitations (Line 606-616) assume using CoNR alone. In practice, most computer games have tons of rigged meshes with proper physics and motion configured. CoNR is a drop-in replacement for a traditional deferred renderer. It renders directly with user-uploaded free-form hand drawings instead of the special UV texture drawn by 3D artists. People could also chain MonsterMash, RigNet, and CoNR into an automated pipeline. Anthor practical applications is to select one of our collected 3d models and make a character designed by hand-drawings get moving (shown in our uploaded demo).

To R4: Q1: Relation to ANR, digital human, etc.

A: By learning a cross-modality matching between geometry and hand-drawn images using a large dataset, CoNR implicitly models the mapping between a set $\mathbf{S}_{ref}$ and a UDP. The concurrent work, ANR (will cite) deals with a very different and degenerated problem where $\mathbf{S}_{ref} = \varnothing$, requires a large number of accurately-aligned poses during training, and disallows replaceable $\mathbf{S}_{ref}$ in run-time which is critical to fast iterative try-and-error required by artists. Also, ANR-like methods require UV mapping from SMPL, which assumes topology similarity. Complicated anime character design requires the model to model the geometry rather than reciting the vertex ID on a template. Different characters may not correspond to the same topological meshes. Our UV-free modeling handles this well. We will enrich related work descriptions.

Q2: Why UDP is novel comparing with CSE?

A: CSE-like methods (will cite) address a vastly different task of detecting a continuous surface coding from an image. Similar "define-by-training" methods are inapplicable in anime creation, since one has to first manually draw the desired output and then infer its CSE (starting a chicken-and-egg recursion). In short, CSE can only be used for reproduction currently, not for generation task. UDP can be obtained directly in all existing 3D editors, game engines, and many other up-streams.

Q3: Why not let UNet hallucinate?

A: "Hallucination" means to model both the physics (inertia, elasticity, etc) and rendering (geometry to texture) with insufficient input variables. It is possible to build a CoNR variant for other data modalities or other fields (Fig 7). The high diversity and high degree-of-freedom nature of UDP imposes a significant challenge for the neural networks, but such disentanglement will connect methods in different fields and give artists desired artistic control over garments and details.

Q4: Novelty of CoNR? Why not use NeRF?

A: CoNR decouples the geometric and textural information from NeRF and makes it a generalized method by moving the computation from training to inference time. Few-shot character images, arbitrary poses, complicated clothing geometry/texture, run-time changeable input, and limited inference time, make ACS rendering an extremely challenging task with no successful previous attempts. This paper may inspire further explorations that would result in exciting insights and profound impact from both academic and industrial perspectives.