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 severely 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.