Machine Perception Report AMAI

Mason Minot Aashish Singh Jeremy Marbach Jonas Binder

ABSTRACT

3D hand surface reconstruction from monocular images has a wide range of potential use cases including AR and robotics. Previous efforts have focused on inputs with a single hand. In the wild, however, both hands frequently interact with one another and perception techniques can be confounded due to the occurrence of occlusion or highly interacting hand poses. To this end, leveraging a portion of the InterHand2.6M dataset and components and concepts from the MANO, HMR, and an Interacting Attention Graph model, we attempt to predict 3D hand surfaces from 2D images.

1 INTRODUCTION

Reconstructing 3D hand surfaces from 2D images is a challenging problem. Recent approaches to this task have focused on a single hand. However, an additional layer of complexity occurs when two hands are considered. This results as both hands are often interacting with one another and can form complex poses in which, for example, they are intertwined or causing a large degree of occlusion. Previously, the publication of SMPL, MANO, and HMR, have spurred rapid progress in creating 3D surfaces from 2D images. Additionally, the publication of InterHand2.6M [2], a growing number of groups are tackling this problem. Our work attempts to build on these findings by developing a MANO-Attention-HMR-MANO cascade, which we term MAHM. This cascade is built with a U-net style design with large dimensions in the initial layers that progressively decrease before progressively increasing in size again. Resnet is used to extract basic features from the image. The MANO and HMR layers are used to compute the mesh and camera parameters. The attention layer is included to allow our model to consider both hands simultaneously.

2 METHODS

To develop our model we first used the skeleton code to identify hyperparameters of importance and implemented image cropping and loss weighting to be beneficial. Other areas of exploration that did not make it into the final model included adding additional loss functions, testing different optimizers, learning rates and learning rate schedules, and techniques used in other papers including joint heatmaps and alternative attention-based schemas. We also found memory limits on Euler to be a limiting factor in our experiments and constrained our model size as a result. In general, we found that training for 20 epochs or greater was beneficial, but resulted in training times ranging from 10 to greater than 24 hours using 1 GPU.

2.1 MAHM Layer

Our MAHM layer takes in the MANO parameters and image features from the previous layer as input. An MLP is used to split the image features (from resnet) into left and right hand vectors of 256 dimension each. These reduced image features are then concatenated with the MANO camera, joints, beta, and pose parameters. Depending on the stage of the U-net like MAHM cascade, this concatenated feature vector may be compressed to reduce the size using an MLP. Finally, this feature vector is fed to the interacting hands attention layer from Li et al. [1].

2.2 MAHM Cascade

The best model's feature vectors are cascaded from 1024, 512, 256, 128, 256, 512, 1024, these values were found empirically where increasing the depth correlated with a better validation score.

2.3 Loss Weighting

It has been found beneficial to prioritise certain losses such that they are all within the same order of magnitude. Thus, losses for the 3D and 2D keypoints were upscaled, as well as the loss for the mano pose. The loss responsible for the translation measurement only experienced a slight change, whereas the loss for the beta values did not require any scaling as it already was within the desired order of magnitude.

2.4 Cropping

The skeleton code provided a cropping feature which has been enabled and adjusted slightly, yielding somewhat better results.

3 EVALUATION

We evalutaed the MAHM layers against predecessors models, indicating how the current cascade has been found. Self attention plays an important role as well, but is subject to a memory issue, its potential is highlighted nonetheless.

3.1 Evaluation of MAHM

As seen in section 2, the MAHM layers are an integral part of our model and the specific Cascade was found via empirical evaluations, which can be seen in Figure: 1. It is noticable that in this case, increasing the depth of the model yielded a better validation score. Our best model which uses the cascading format mentioned in 2.2, achieves a public score of 48.7.

3.2 Self Attention

While our best model makes use of MAHM layers, its most competitive counterpart is a model which uses self-attention in addition. It occured that the model with self attention reached similar validation scores already after half of the training time of our best model. However, due to memory limits, the model aborted training, thus it is likely that it could not reveal its true potential Figure: 2.

4 DISCUSSION

As mentioned in section 2, various approaches were tried. The most prominent alternative solutions was the integration of heatmaps. However, in our case their effective performance was poor. We assume that the way we chose to implement heatmaps did not enrich the features handed over to the MANO layers. The other prominent approach is the creation of the MAHM layer which yielded good results and was adjusted empirically, as seen in Figure 1. Another branch emerged from that approach, namely applying self attention to the already existing model, however, as seen in Figure 2, despite depicting promising results, it suffered from a memory issue due to the already large model size it had before applying self attention. Ultimately we arrived at a model which scores 48.7 on the leaderboard which makes heavy use of the presented MAHM layers, loss weighting and image cropping.

5 CONCLUSION

The report presents insights into the different major ideas developed during this project and highlights the best workding one, namely the application of MAHM layers. Furthermore, it indicates the potential of self attention in combination with these layers.

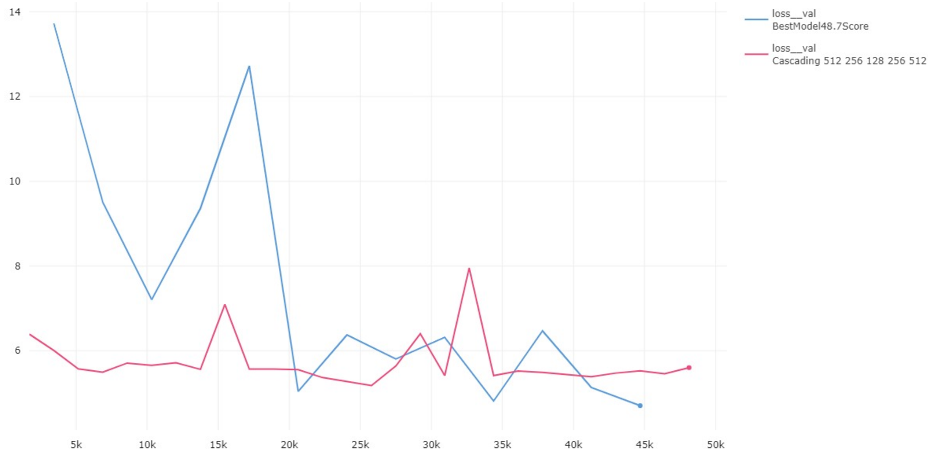

Figure 1: Evaluation of different cascading, in our case, deeper models yielded better results. (Blue) is our best model and (Red) is a predecessor model with shallower cascading. Y-Axis: Score, X-Axis: steps.

REFERENCES

[1] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. 2022. Interacting Attention Graph for Single Image Two-Hand Reconstruction. https://doi.org/10.48550/ARXIV.2203.09364

[2] Gyeongsik Moon, Shoou-i Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. 2020. InterHand2.6M: A Dataset and Baseline for 3D Interacting Hand Pose Estimation from a Single RGB Image. https://doi.org/10.48550/ARXIV.2008.09309

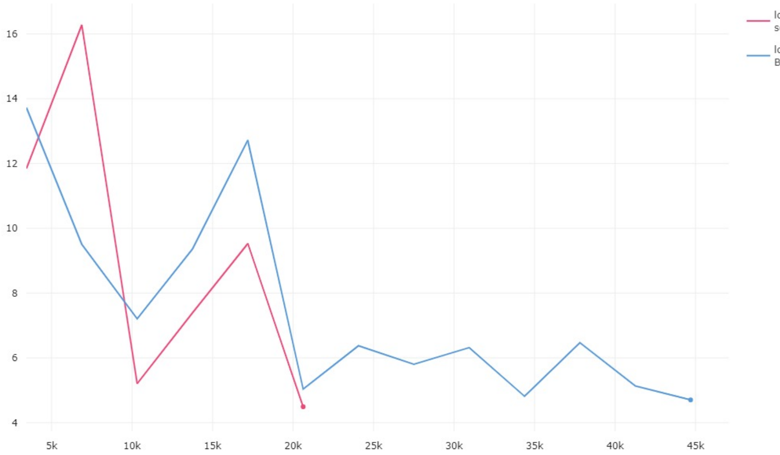

Figure 2: Self attention added to the best model (Red), way faster convergence and potentially better validation score than our best model (Blue), but memory issue. Y-Axis: Score, X-Axis: steps.