Machine Perception Report Marineprotection

Agisilaos Politis Leonardo Barberi Guglielmo Bonifazi Giovanni Acampa

ABSTRACT

In this project we are trying to predict 3D SMPL pose and shape parameters of humans from RGB images. We use 2 datasets, 3DPW that contains 3D annotations and MPII that contains 2D annotations. To solve the problem we took inspiration from the PARE paper, modifying the backbone, tuning hyperparameters and adapting their architecture to our training datasets [4]. After intensive training, our results passed the baseline.

1 INTRODUCTION

In the realm of computer vision and graphics, accurately estimating the 3D pose and shape of human bodies is a fundamental challenge. It holds immense significance in applications ranging from virtual reality and gaming to motion capture and human-computer interaction. To tackle this complex task, researchers have developed the SMPL model [5]. The 3D SMPL pose and shape estimation approach leverages a data-driven model to estimate the 3D pose and shape of a human body given a 2D image or video input. Despite substantial advancements, cutting-edge 3D human posture and form estimate techniques are still prone to partial occlusion and can make alarmingly inaccurate predictions, even when the majority of the body is visible. The PARE model was developed to solve the partial occlusion problem in 3D human pose and shape estimation, and we decided to adapt its architecture for this project [4]. Compared to other state-of-the-art 3D body model creation models, not only was PARE more robust to occlusions, but we also found that it was more interesting and intuitive to understand given what we had learned in the Machine Perception course.

Our model uses a pre-trained backbone, that is a variation of ResNet50 and fine-tunes the weights with the data from the 3DPW and MPII datasets that were provided [6] [1]. After the backbone, we split the features into a 2D branch and a 3D branch, then we applied a part-attention mechanism, and combined the features again to finally predict the body pose and shape using fully connected layers. Originally, we had implemented one MLP to predict all outputs for all joints. To improve our score, we attempted to create one linear layer per joint, and while the scores were encouraging, we only found the time to fully train this new architecture on the whole dataset for a total of 70 epochs.

2 METHOD

Regarding the backbone, we used a ResNet-50 neural network, available in the PyTorch library, and tested different pre-trained weights [3]. One approach, was to use the weights of the DINO version of ResNet-50 [2]. DINO, whose stands for Self-**di**stillation with **no** labels, creates a teacher and a student network, both with the same architecture. It is very flexible and can be used with a Vision Transformers (ViT) or a ConvNet, such as the popular ResNet-50, and it is nowadays a popular choice as backbone. It can be found at https://dl.fbaipublicfiles.com/dino/dino_resnet50_pretrain/dino_resnet50_pretrain.pth. We trained our model for a limited number of epochs trying both the DINO and the ImageNet weights. While theoretically promising, we didn't end up using DINO weights for the training of our final model since, the validation score was significantly lower compared to the ImageNet case.

The other backbone weights that we tried was ResNet-50 with the standard weights of ImageNet. In both cases, we extracted volumetric features cutting before the global average pooling layer. After this step, we divided the features in two branches. For both the 2D and 3D branches, we used three 2× upsampling followed by 3×3 convolutional layers, applied with batch-norm and ReLU. To obtain part attention maps, we applied $J + 1$ 1×1 convolutional kernels to 2D part features to reduce the channel dimension, where J is the number of joints (that is 24). Each channel in the 2D output then represents one joint, and, upon applying a softmax channel-wise to these 2D features, we could consider each of these channels as an attention matrix, giving weights to pixels that are most important for predicting specific joints. We then multiplied the 2D features matrix with the 3D features matrix. This acts as a part-attention on the image features that are extracted from the 3D branch.

Finally, using 24 MLP for the poses, 1 for the shape and 1 for the camera parameters, we predicted the SMPL human poses and shapes. We used Adam optimizer with learning rate of 3×10^{-4} and batch size of 64. We tried to train the model for 120 epochs, but as time was running low we settled on the checkpoint after 70 epochs.

3 EVALUATION

We started to work studying some papers that already have good results on this task. Our first naive try was to predict the results only using ResNet-50 for feature extraction and some MLPs for pose, shape, and camera parameter regression. Of course the results were not good, since the model's architecture was very naive to tackle the complexity of the problem. We decided to follow the PARE model, adding the 2D, the 3D branches and the part attention, with a shallower network respect to PARE, with only two Convolution blocks. We obtained better results, but they were still far from the baseline. Then, we tried to change the weights in the backbone, trying for example the ImageNet weights, and the DINO weights. Following the tips for training from the course, we also tried to implement stochastic weight averaging as a optimizer. However, due to the dependencies with other libraries, we saw fit to only update the pytorch-lightning library to version 1.2.1, where stochastic weight averaging was implementable as a callback on the trainer class. Unfortunately, we received errors which we weren't able to debug (about ctypes objects containing pointers that cannot be pickled), so we had to settle on only using Adam as optimizer. Finally, we made the network deeper, we increased the learning rate and we trained the model for longer, also relying on tensorboard to verify whether the training was rolled out smoothly.

4 DISCUSSION

All the results are related to the metric MPJPE and to the validation set that is classified as 'public'. Remember that in MPJPE

metric, lower is better. The first result that we obtained during the experiments is a score of 112.74 for the ResNet + MLPs. After implementing the 2D and 3D branches and training only for 4 epochs, we got a 79.19, meaning that this kind of features improve the model very much. After modifying the parameters and the backbone, we obtained a result of 68.45, very close to the baseline. We finally passed the baseline with a score of 66.32, making the model deeper and training for a higher number of epochs.

5 CONCLUSION

In conclusion, we tried to solve the task starting from the already existing work PARE on it, making modifications and adapting it to our datasets. We also tried different ways and not everything worked, but we finally obtained good results.

REFERENCES

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2014. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9650–9660.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[4] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. 2021. PARE: Part attention regressor for 3D human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11127–11137.

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

[6] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. 2018. Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera. In *European Conference on Computer Vision (ECCV)*.