UNIVERSITY OF BIRMINGHAM

MULTI-HYPOTHESIS 3D HAND MESH SEQUENCE ESTIMATION FROM BLURRY IMAGE

YUMING CHEN

supervised by *Dr. Hyung Jin Chang*

A dissertation submitted to the University of Birmingham for the degree of

Master of Science in Artificial Intelligence and Machine Learning

School of School of Computer Science

College of

University of Birmingham

September 2024

Abstract

Hand images are frequently suffer from blurriness. Such blurriness causes ambiguity, which leads to multiple feasible hand mesh sequences corresponding to a given blurry image. We propose a Generation Model to make multiple predictions to one image. Further, we propose a Selection Model to select candidates with high quality. Selection Model evaluates all hypotheses and discriminates they are the top-n or not. Experimental result shows that out Method is comparable to the state-of-the-arts.

Acknowledgements

I would first like to thank my project supervisor, Dr. Hyung Jin Chang, for his support throughout this project. In addition to my supervisor I would also like to thank Zhongqun Zhang, Yihua Li and Rongyu Chen for their insight in this topic, and the help in writing and experiments.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation of Research

Recovering 3D human mesh is an active and challenging problem. It is essential in Augmented Reality[3], Motion Analysis[6] and other applications involving human-interaction. Recent advancements have been made in producing 3D hand mesh from sharp image[21, 7], which is no robust against blurriness. However, hand in image is frequently blurry due to the dexterous movement and self-occlusion. Previous work on blurry image[20] makes one single estimation. Nevertheless, there are multiple feasible poses correspond to a blurry image, which is called ambiguity[29].

Making multiple feasible predictions is expected to alleviate such challenges. Interest in multi-hypothesis estimation is increasing[4, 13, 10, 16]. Some of them aggregate all hypotheses to get the final prediction[13, 16], or aggregate a subset of hypotheses[27, 29] but the diversity of hypotheses is not promoting. It is not appropriate for our scenario where blurriness causes distinct feasible poses.

To address this challenges, we proposed a two-stage network consisting of Generation Model and Selection Model. The Generation Model generates multiple plausible hypotheses given a single blurry image. The Selection Model makes selections conditioned on the image and feature of Generation Model. To this end, the Generation Model combines an Unfolder[20] and

a Multi-Hypothesis Transformer (MHT) encoder. The encoder can efficiently generate multiple hypotheses with a lightweight embedding set based on the temporal information the Unfolder extracts. The Selection Model takes a mixture of hypothesis and the feature of the Generation Model, which makes it to better capture the subtle difference among hypotheses. After the training of the Generation Model, we train the Selection Model with the recently proposed Joint Ranking and Calibration (JRC) loss[24] . The JRC loss simplifies the task to learn in related works[27, 29] without losing the ranking information of hypotheses. The experiment in Section 4.3 shows that our model is comparable to the state-of-the-art methods.

1.2 Contributions

In conclusion, our contribution are two-folded:

- We propose a multi-hypothesis Generation Model, an efficient approach to generate multiple hypotheses with minimal additional parameters compared with BlurHandNet, the state-of-the-art in the task.
- We propose the Selection Model, which makes use of the learnt feature of the Generation Model and Joint Ranking Calibration loss.

1.3 Dissertation Structure

The dissertation is structured as follow. Chapter 2 introduces the related work in task and technique. Chapter 3 describes the network and training process of our model. Chapter 4 shows the result and analysis of experiments. Chapter 5 includes the limitation and future works.

Chapter 2

Related Work

2.1 3D Hand Mesh Estimation

Works of 3D Hand Mesh Estimation inspire from Human Mesh Estimation[11, 14]. MANO[23], a parameterized hand model, is introduced to provide human prior knowledge. Following work regresses the mesh based on the estimated joint and shape parameters[8], or directly estimate the vertex parameters[5, 19]. HandOccNet[21] introduces a transformer[26]-based structure to explicitly deal with the self-occlusion problem. Deformer[7] utilizes the temporal information of video with a spatial-temporal transformer. HaMeR[22] scales up ViT and the dataset to improve the performance and generalization. However, these works are orthogonal to the ambiguity caused by blurriness. Oh et al. [20] proposes a dateset BlurHand and a network BlurHandNet to overcome the ambiguity problem. It is different from our method as it does not involve Multi-Hypothesis Estimation.

2.2 Multi-Hypothesis Estimation

The above works output only a single estimation mesh for a given image. There are a couple of works[13, 4, 12, 1] make multiple predictions to explore the plausible estimation aligned well with the image. MHFormer[13] generates and aggregates three hypotheses once to produce a refined final output. MION[16] refines the hypotheses with PNCC positional embedding and

Mesh Refine Transformer. These works aggregate all hypotheses or select the best hypothesis with the minimal distance to the ground truth (GT). GenPose[29] makes multi-hypothesis prediction given a partially observed cloud point and select the better ones with a EnergyNet. ScoreHypo[27] train a ScoreNet to select better human pose from a RGB image. GenPose and ScoreHypo show that selection can improve the final output. However, these works do not focus on the blurriness task.

Chapter 3

Method

3.1 Problem Description

In this task, our target is to estimate 3D hand mesh sequence from a single blurry RGB image. To address the challenges of pose-ambiguity caused by blurriness and self-occlusions, we introduce multi-hypothesis method. We use MANO[23] to simplify the estimation problem. MANO is a hand model that parameterizes 3D hand mesh $\mathbf{V} \in \mathbb{R}^{778\times3}$ into pose $\theta \in \mathbb{R}^{48}$ and shape $\beta \in \mathbb{R}^{10}$. Given a blurry image $\mathbf{I} \in \mathbb{R}^{256\times256\times3}$, the generation model is expected to make multiple plausible hand mesh estimations $\mathbf{H}^k, k = 1, \ldots, K$, where $\mathbf{H}^k = \theta_0 \oplus \cdots \oplus \theta_T$ is the hypothesis that contains a pose sequence. K, T is the number of hypotheses and the length of sequence, respectively. Further we train a reward model to predict the reward $\hat{r}^k \in \mathbb{R}$ of each hypothesis conditional on the image $\mathbf{I}$. During inference, the generation model generate K hypotheses, and we select the top-n of them according to the reward $\hat{r}^k$.

3.2 Multi-Hypothesis Generation

Figure 3.1 shows the overall pipeline of (our work). It contains a Generation Model and a Selection Model. The generation model consists of an Unfolder, a Multi-Hypothesis Transformer and a Regressor.

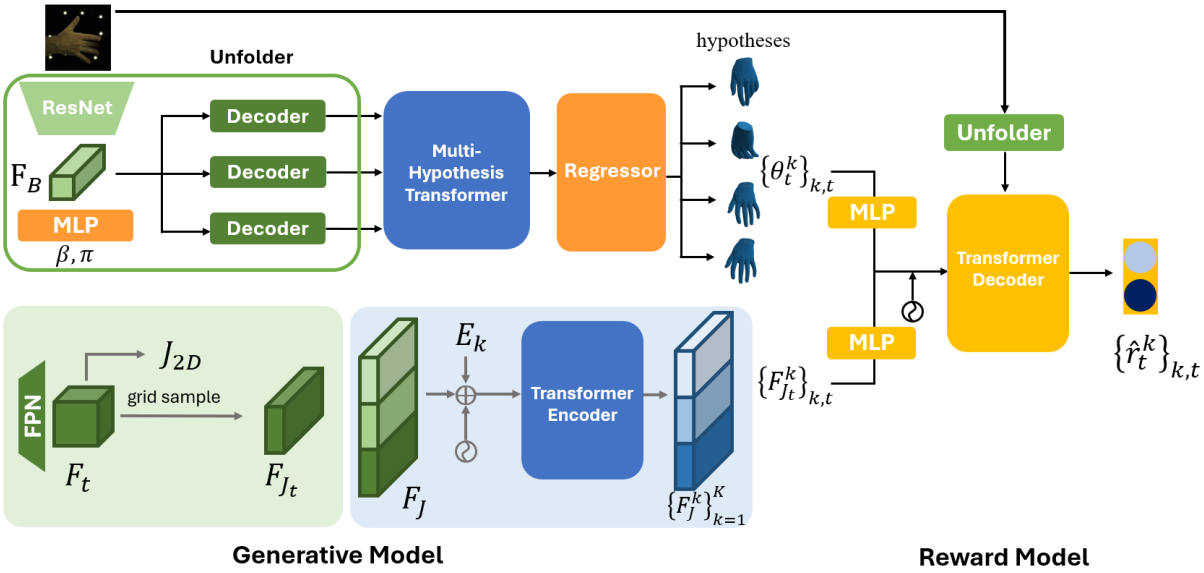

Figure 3.1: **Overall pipeline of our method.** The Generation Model generates K hypotheses given a blurry image. The Selection Model selects better hypotheses based on the outputs and features of the Generation Model.

3.2.1 Unfolder

Given a blurry image $\mathbf{I}$, the unfolder outputs a joint-wise feature F_{J_t} for each timestep. The image is first fed into a ResNet[9] backbone to get a basic feature $F_B \in \mathbb{R}^{2048}$. Then we use T separated decoders to predict the temporal feature $F_t \in \mathbb{R}^{H\times W\times E}$ where $H \times W = 32 \times 32$, $E = 512$ denote the resolution and the dimension of feature, respectively. We obtain a heatmap of each joint with a 1×1 convolution layer and regress the 2D position of joints $\mathbf{J}_t^{2D} \in \mathbb{R}^{J\times 2}$ through the differentiable soft-max operator[25], where $J = 21$ is the number of hand joints. Finally we apply grid sampling on F_t and obtain $F_{J_t} \in \mathbb{R}^{J\times E}$, the joint-wise feature of timestep t. The temporal joint-wise feature $F_J = F_{J_0} \oplus \cdots \oplus F_{J_T} \in \mathbb{R}^{T\times J\times E}$ is obtained by concatenating all joint-wise features, where $\oplus$ is the operation of concatenation.

3.2.2 Multi-Hypothesis Transformer

Multi-Hypothesis Transformer (MHT) is a transformer[26]-encoder module to generate multiple refined temporal joint-wise features $F_J^k \in \mathbb{R}^{T\times J\times E}$. Unlike works[13, 16, 1] generate each hypothesis with a weight-independent network, we find that a applying a low dimension

perturbation in the feature space is able to promote the diversity of hypotheses. Thus, the architecture of MHT is based on Kinematic Temporal Transformer[20]. Besides the temporal positional embedding $pe_T \in \mathbb{R}^{T\times E}$ and the kinematic positional embedding $pe_J \in \mathbb{R}^{J\times E}$, we apply hypothetical embedding $he_K \in \mathbb{R}^{K\times E}$ on an additional dimension. The summation $(F_J + he_K) \in \mathbb{R}^{K\times T\times J\times E}$ is then inputted into the transformer as a sequence with length of $(T\times J)$. Notice that each hypothesis is dealt independently. As a result, the length of sequence will not increase with the number of hypothesis. The output is K refined features $\{F_J^k\}_{k=1}^K$.

3.2.3 Regressor

The regressor produces MANO pose and shape parameters and camera parameter. Specifically, the pose parameter $\{\theta_t^k\}_{t=1}^T$ is estimated for each hypothesis separately based on F_J^k. The shape parameter $\{\beta_t\}_{t=1}^T$ and the camera parameter $\{\pi_t\}_{t=1}^T, \pi_t \in \mathbb{R}^3$ is estimated for all hypotheses based on F_B. The 3D hand mesh $\mathbf{V}_t^k$ and 3D joint $\mathbf{J}_t^k$ is calculated by forwarding θ_t^k, β_t into the MANO layer.

3.3 Reward Model

The trained Generation Model produces multiple plausible mesh sequences, *i.e.*, the hypotheses $\mathbf{H}^k$. Aggregating all hypotheses to obtain a single final estimation via an aggregation module[16, 13] can be computational expensive as the number of hypotheses goes large. Aggregation itself may also worsen the final estimation as feasible hypotheses are distinct in blurry image. To this end, we propose a reward model equipped with a similar architecture to the generation model. The goal of the reward model is selecting the hypotheses with high quality from all hypotheses given $\mathbf{H}^k = \{\theta_t^k\}_{t=1}^T$ and $\mathbf{I}$. Concretely, the reward model predict a reward signal $\hat{r}^k$ of hypothesis $\mathbf{H}^k$ conditional on the input image $\mathbf{I}$.

The reward model uses an unfolder (3.2.1), a transformer[26]-decoder module and a reward layer to produce reward. The unfolder is independent from the generation model and extracts the joint-wise feature $F_R \in \mathbb{R}^{T\times J\times E}$, F_R is used in the Cross-Attention (CA) block as the key

and value. To accelerate the learning process and better capture the subtle difference among hypotheses, we introduce a **Model-Specific Feature** (MSF) to produce to query feature F_Q. Concretely, $F_Q \in \mathbb{R}^{K\times T\times J\times E}$ is the concatenation of hypothesis $\mathbf{H}^k$ and corresponding refined feature F_J^k from the Generation Model. The hypothesis and the refined feature are projected by MLPs respectively. That is,

$$F_Q = \left(h(\mathbf{H}^0) \oplus g(F_J^k)\right) \oplus \cdots \oplus \left(h(\mathbf{H}^K) \oplus g(F_J^K)\right)$$

where h and g are MLP layers. The reward $\hat{r}^k$ is produced from the Reward Layer which is implemented as an MLP.

3.4 Training

The training process is split into two parts. We train the Generation Model by supervising the estimation with GT in the first part and train the Reward Model by learning to rank[2] in the second part. In contrast with BlurHandNet[20], we simply supervise the estimated sequence along temporal order instead of employing temporal order-invariant loss[20] as a diverse set of hypotheses is able to generate hypotheses with different temporal order.

3.4.1 Generation Model

The overall training loss of generation model is

$$\mathcal{L}_{gen} = \lambda_{3D}\mathcal{L}_{3D} + \lambda_{proj}\mathcal{L}_{proj} + \lambda_{MANO}\mathcal{L}_{MANO} + \lambda_{aux}\mathcal{L}_{aux} + \lambda_{div}\mathcal{L}_{div} \tag{3.1}$$

where

$$\mathcal{L}_{joint} = \sum_{t=1}^{T} \min_{k} \left\| \mathbf{J}_t^k - \mathbf{J}_t^* \right\|_1 \tag{3.2}$$

$$\mathcal{L}_{proj} = \sum_{t=1}^{T} \min_{k} \left\| \Pi \mathbf{J}_t^k - \Pi \mathbf{J}_t^* \right\|_1 \tag{3.3}$$

$$\mathcal{L}_{MANO} = \sum_{t=1}^{T} \left(\min_{k} \left\| \theta_t^k - \theta_t^* \right\|_1 + \left\| \beta_t - \beta_t^* \right\|_1 \right) \tag{3.4}$$

where $\mathbf{J}_t^*$ is GT 3D joints, Π is the operation that obtain 2D reprojective joints. θ_t^* and β_t^* is the GT MANO parameters.

Notice that the loss only propagates the gradient of the best hypothesis as we encourage the hypothesis set to be diverse. Different hypotheses are able to capture different patterns of ambiguity in this way.

We also employ an auxiliary task

$$\mathcal{L}_{aux} = \sum_{t=1}^{T} \left\| \mathbf{J}_t^{2D} - \Pi \mathbf{J}_t^* \right\| \tag{3.5}$$

to stabilize training, where $\mathbf{J}_t^{2D}$ is the 2D joints regressed by the unfolder. To explicitly enhance the diversity, we use a modified diversity promoting loss[18, 28]

$$\mathcal{L}_{div} = \frac{1}{K(K-1)} \sum_{k=1}^{K} \sum_{m=k+1}^{K} \sum_{t=1}^{T} e^{\frac{-\left\| F_{J_t}^k - F_{J_t}^m \right\|_1}{\alpha}} \tag{3.6}$$

where α is a normalizing factor. Compared with the original version[18, 28], we promote the distance in feature space instead of the output space, which less influence the quality of outputs. λ_{3D}, λ_{proj}, λ_{MANO}, λ_{aux} and λ_{div} are constant coefficients.

3.4.2 Reward Model

The reward model is expected to assign a reward $\hat{r}^k$ to the hypothesis $\mathbf{H}^k$ corresponding to its quality. We measure the quality with Mean Per Joint Position Error (MPJPE). A perfect reward

should satisfy

$$\{r^k > r^j\} \Leftrightarrow \{\mathbf{H}^k \succ \mathbf{H}^j\} \tag{3.7}$$

where $\succ$ means $\mathbf{H}^k$ has a better quality than $\mathbf{H}^j$. However, learning the desired reward function with a pairwise ranking loss[2] exceeds our demand, which only aims at selecting the top-n hypotheses regardless of the rank. A pointwise loss[30] widely used in Click-Through-Rate (CTR) tasks that discriminates whether hypotheses are in the top-n does not utilize the ranking relationship in the hypotheses set in our task. Without losing the information of rank hypotheses endowed, we train the reward model with Joint Ranking Calibration (JRC) loss[24]

$$\mathcal{L}_{jrc} = \alpha \mathcal{L}_{calib} + (1 - \alpha)\mathcal{L}_{rank} \tag{3.8}$$

where $\alpha \in [0, 1]$ is the hyper-parameter for balance.

To apply JRC loss, the reward model produces a two-dimensional vector $f^k := f(\mathbf{H}^k|\mathbf{I}) \in \mathbb{R}^2$. Follow the notations in Sheng et al. [24], let $y = \{0, 1\}$ indicates the hypothesis is top-n or not. $f^k[y]$ indicates the y-th index of f^k. $f^k[y]$ is $f^k[1]$ if hypothesis $\mathbf{H}^k$ is the top-n one and $f^k[0]$ if not. In this way, the reward is the sigmoid value of $f^k[1] - f^k[0]$, which is the predicted click probability in CTR prediction task[24].

$$\hat{r}^k = \frac{1}{1 - \exp\left(-(f^k[1] - f^k[0])\right)} \tag{3.9}$$

The pointwise calibration loss is written to be

$$\begin{aligned} \mathcal{L}_{calib} &= -\log \hat{r}^k \\ &= -\log \frac{\exp(f^k[y^k])}{\exp(f^k[0]) + \exp(f^k[1])} \end{aligned} \tag{3.10}$$

The listwise ranking loss is written to be

$$\mathcal{L}_{rank} = -\log \frac{\exp(f^k[y])}{\sum_{m=1}^{K} \exp(f^m[y])} \tag{3.11}$$

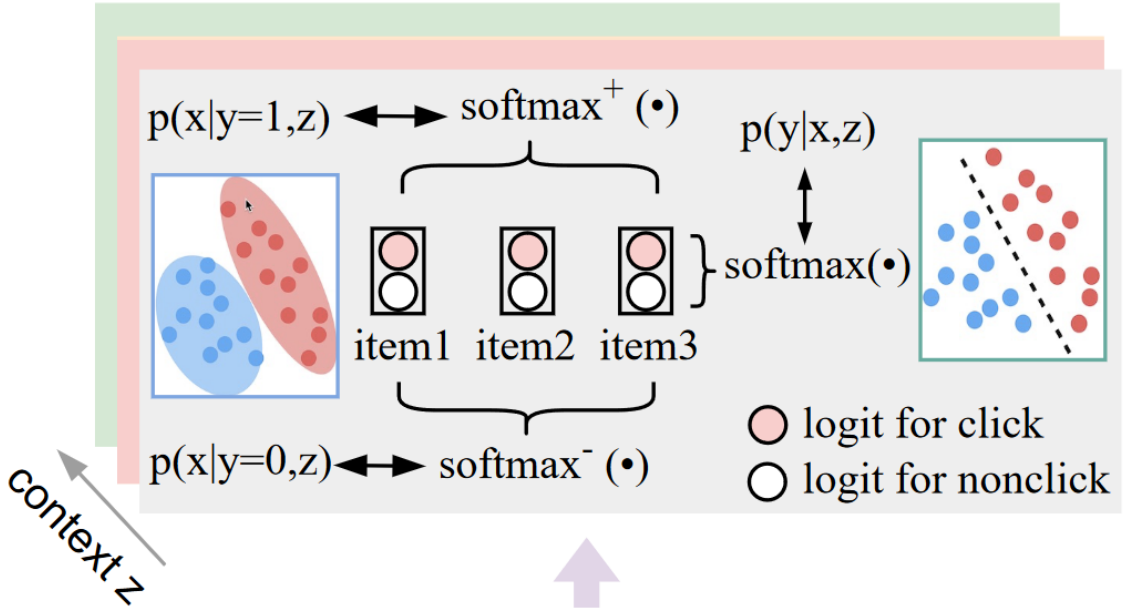

Figure 3.2: The architecture of JRC head. The graphics is from Sheng et al. [24].

Figure 3.2 illustrates the process to compute JRC loss.

The overall training loss of reward model is

$$\mathcal{L}_{rew} = \lambda_{jrc}\mathcal{L}_{jrc} + \lambda_{aux}\mathcal{L}_{aux} \tag{3.12}$$

where auxiliary loss $\mathcal{L}_{aux}$ is the same as Eq.3.5 in generation model training process. λ_{jrc} and λ_{aux} are constant coefficients.

Chapter 4

Experiment

4.1 Datasets and Metrics

We introduce the dataset and metrics we use in this section.

BlurHand. BlurHand (BH) is a 3D hand pose dataset proposed by Oh et al. [20]. It contains blurry images and corresponding 3D hand mesh sequence. We train and test our model following the train-test split.

Metrics. Consistent with previous work[20], we use Mean Per Joint Position Error (MPJPE) and Mean Per Vertex Position Error (MPVPE). The metrics measure the L^2 distance between estimation and the ground truth as Eq.4.1 shows.

$$\text{MPJPE} = \sum_{j=1}^{J} \sqrt{\left(\mathbf{J}_j - \mathbf{J}_j^*\right)^2}, \quad \text{MPVPE} = \sum_{v=1}^{V} \sqrt{\left(\mathbf{V}_v - \mathbf{V}_v^*\right)^2} \tag{4.1}$$

4.2 Implement Details

We use ResNet[9] as the backbone of unfolder, and a bottom-up Feature Pyramid Network (FPN)[15] decoder with decovolution layer. The size of embedding is $E = 512$. The length of sequence is $T = 3$ following Oh et al. [20]. The number of hand joints is $J = 21$. The number of hypotheses is $K = 16$. Both transformer encoder and decoder have $B = 4$ blocks.

We train the generation model for 13 epochs with AdamW optimizer[17] and the batch size

Method	MPJPE (mm) ↓	MPVPE (mm) ↓
BlurHandNet[20]	<u>16.80</u>	<u>15.30</u>
I2L-MeshNet†	24.32	23.08
METRO†	20.54	27.03
Pose2Pose†	18.80	17.42
Ours (final)	**16.72**	**15.22**
Ours (best)	**14.12**	**12.59**

Table 4.1: Comparison to the state-of-the-arts on BlurHand dataset[20]. Metrics (MPJPE and MPVPE) are calculated on the hand in the middle frame. The best result is **bold** and the second best one is <u>underlined</u>. Results of daggered methods † is from Oh et al. [20]

.

K	MPJPE ↓
1	16.78
4	15.27
16	14.12
32	13.78

Table 4.2: MPJPE of Ours (best) with different value of K, the number of generated hypotheses.

of 48. The learning rate is initially set to be 0.0001, and decayed by 0.1 at epochs 10 and 13. We train the reward model for 30 epochs with a cosine annealing learning rate. We set the top-n number to be $n = 4$.

4.3 Comparison to State-of-the-arts

Table 4.1 shows the comparison results with state-of-the-art mothods on BlurHand dataset. "Ours(final)" refers to the aggregated output, "Ours(best)" refers to the output with the minimum distance to ground truth.

As shown in Table 4.1. Our method is comparable to the state-of-the-art algorithms. And the best hypothesis significantly outperforms them.

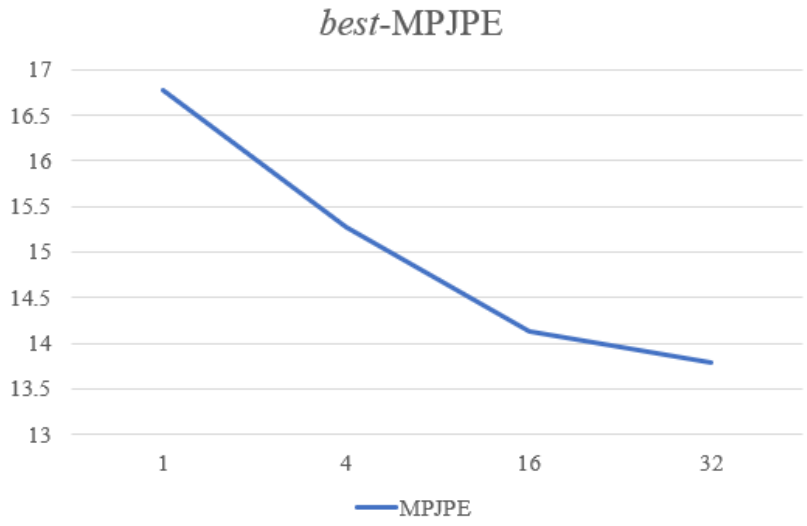

Figure 4.1: MPJPE of Ours (best) along the value of K.

4.4 Ablation Study

4.4.1 Number of the hypothesis

We are interested in the effectiveness of multi-hypothesis method. Table 4.2 shows that the minimal distance to the GT falls as the number of hypotheses K grows. When $K = 1$, our method, which is not a Multi-Hypothese method, is still comparable to the state-of-the-arts. Thus, the Multi-Hypothesis architecture does not worsen the performance of the model itself.

4.4.2 Effect of Model-Specific Feature

It is much more challenge to select good hypotheses conditioned on blurry images compared with sharp image[27] and cloud point[29]. Hence we use MSF to improve the learning process of Selection Model. Table 4.3 shows the learning process with and without MSF. Model with MSF learns faster and enjoys a better result.

4.4.3 Diversity of Hypotheses

The propose we promote diversity of Hypotheses is to obtain more predictions with meaningful difference, while these hypotheses are not too distant to the GT given an less ambiguous image.

Figure 4.2: Qualitative results on different level of blurriness.

Method	Epoch		
	10	20	30
w/o MSF	17.72	17.18	16.98
w/ MSF	17.58	17.05	16.72

Table 4.3: The MPJPE of the final prediction during the training process. Method *w/o* MSF take the original hypothesis $\mathbf{H}^k$ as input instead of the corresponding feature from the Generation Model F_J^k.

Figure 4.2 shows a qualitative result of diversity. Given nearly sharp image, our method makes similar predictions. When the blurriness goes stronger, the hypotheses become more diverse. In extreme cases, the model is able to generate highly diverse results. It is notable that the image at the bottom of Figure 4.2 is even hard for human to distinguish the trajectory of hand. Hence unsuccessful prediction on it is expectable.

Chapter 5

Conclusion

5.1 Conclusion

We present an framework which generates multiple diverse predictions efficiently to recover 3D hand mesh sequence from blurry image. Our Generation Model generates multiple hypotheses with a simple but effective embedding vector. We additionally propose a Selection Model, a module to select the top-n among hypotheses. The experiments show that out method is comparable to the state-of-the-arts, while the Selection Model is able to select the hypotheses with high quality.

5.2 Limitations and Future Work

Our method does not significantly outperform BlurHandNet[20], which can be improved with a better trained Selection Model. And the number of hypotheses is fixed. In the future, we plan to employ probabilistic generation method including diffusion model[27, 29] and Normalizing Flow[12] to flexibly generate an arbitrary number of hypotheses.

References

[1] Benjamin Biggs et al. “3D Multi-bodies: Fitting Sets of Plausible 3D Human Models to Ambiguous Image Data”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 20496–20507. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/ebf99bb5df6533b6dd9180a59034698d-Paper.pdf.

[2] Chris Burges et al. “Learning to rank using gradient descent”. In: *Proceedings of the 22nd International Conference on Machine Learning*. ICML ’05. Bonn, Germany: Association for Computing Machinery, 2005, pp. 89–96. ISBN: 1595931805. DOI: 10.1145/1102351.1102363. URL: https://doi.org/10.1145/1102351.1102363.

[3] Julie Carmigniani and Borko Furht. “Augmented Reality: An Overview”. In: *Handbook of Augmented Reality*. Ed. by Borko Furht. New York, NY: Springer New York, 2011, pp. 3–46. ISBN: 978-1-4614-0064-6. DOI: 10.1007/978-1-4614-0064-6_1. URL: https://doi.org/10.1007/978-1-4614-0064-6_1.

[4] Rongyu Chen, Linlin Yang, and Angela Yao. “MHEntropy: Entropy Meets Multiple Hypotheses for Pose and Shape Recovery”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2023, pp. 14840–14849.

[5] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. “Pose2Mesh: Graph Convolutional Network for 3D Human Pose and Mesh Recovery from a 2D Human Pose”. In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi et al. Cham: Springer International Publishing, 2020, pp. 769–787. ISBN: 978-3-030-58571-6.

[6] Steffi Colyer et al. “A Review of the Evolution of Vision-Based Motion Analysis and the Integration of Advanced Computer Vision Methods Towards Developing a Markerless System”. In: *Sports Medicine - Open* 4 (June 2018), p. 24. DOI: 10.1186/s40798-018-0139-y.

[7] Qichen Fu et al. “Deformer: Dynamic Fusion Transformer for Robust Hand Pose Estimation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2023, pp. 23600–23611.

[8] Liuhao Ge et al. “3D Hand Shape and Pose Estimation From a Single RGB Image”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[9] Kaiming He et al. “Identity Mappings in Deep Residual Networks”. In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Cham: Springer International Publishing, 2016, pp. 630–645. ISBN: 978-3-319-46493-0.

[10] Karl Holmquist and Bastian Wandt. “DiffPose: Multi-hypothesis Human Pose Estimation using Diffusion Models”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2023, pp. 15977–15987.

[11] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. “VIBE: Video Inference for Human Body Pose and Shape Estimation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020.

[12] Nikos Kolotouros et al. “Probabilistic Modeling for Human Mesh Recovery”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 11605–11614.

[13] Wenhao Li et al. “MHFormer: Multi-Hypothesis Transformer for 3D Human Pose Estimation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 13147–13156.

[14] Kevin Lin, Lijuan Wang, and Zicheng Liu. "End-to-End Human Pose and Mesh Reconstruction with Transformers". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 1954–1963.

[15] Tsung-Yi Lin et al. "Feature Pyramid Networks for Object Detection". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017.

[16] Zhiwei Liu et al. "Multi-initialization Optimization Network for Accurate 3D Human Pose and Shape Estimation". In: *Proceedings of the 29th ACM International Conference on Multimedia*. MM '21. Virtual Event, China: Association for Computing Machinery, 2021, pp. 1976–1984. ISBN: 9781450386517. DOI: 10.1145/3474085.3475355. URL: https://doi.org/10.1145/3474085.3475355.

[17] Ilya Loshchilov and Frank Hutter. "Fixing Weight Decay Regularization in Adam". In: *CoRR* abs/1711.05101 (2017). arXiv: 1711.05101. URL: http://arxiv.org/abs/1711.05101.

[18] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. "Generating Smooth Pose Sequences for Diverse Human Motion Prediction". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 13309–13318.

[19] Gyeongsik Moon and Kyoung Mu Lee. "I2L-MeshNet: Image-to-Lixel Prediction Network for Accurate 3D Human Pose and Mesh Estimation from a Single RGB Image". In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi et al. Cham: Springer International Publishing, 2020, pp. 752–768. ISBN: 978-3-030-58571-6.

[20] Yeonguk Oh et al. "Recovering 3D Hand Mesh Sequence From a Single Blurry Image: A New Dataset and Temporal Unfolding". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023, pp. 554–563.

[21] JoonKyu Park et al. "HandOccNet: Occlusion-Robust 3D Hand Mesh Estimation Network". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 1496–1505.

[22] Georgios Pavlakos et al. "Reconstructing Hands in 3D with Transformers". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 9826–9836.

[23] Javier Romero, Dimitrios Tzionas, and Michael J. Black. "Embodied Hands: Modeling and Capturing Hands and Bodies Together". In: *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*. 245:1–245:17 36.6 (Nov. 2017).

[24] Xiang-Rong Sheng et al. "Joint Optimization of Ranking and Calibration with Contextualized Hybrid Model". In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD '23. Long Beach, CA, USA: Association for Computing Machinery, 2023, pp. 4813–4822. ISBN: 9798400701030. DOI: 10.1145/3580305.3599851. URL: https://doi.org/10.1145/3580305.3599851.

[25] Xiao Sun et al. "Integral Human Pose Regression". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Sept. 2018.

[26] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[27] Yuan Xu et al. "ScoreHypo: Probabilistic Human Mesh Estimation with Hypothesis Scoring". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 979–989.

[28] Ye Yuan and Kris Kitani. "DLow: Diversifying Latent Flows for Diverse Human Motion Prediction". In: *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX*. Glasgow, United Kingdom: Springer-Verlag, 2020, pp. 346–364. ISBN: 978-3-030-58544-0. DOI: 10.1007/978-3-030-58545-7_20. URL: https://doi.org/10.1007/978-3-030-58545-7_20.

[29] Jiyao Zhang, Mingdong Wu, and Hao Dong. "GenPose: generative category-level object pose estimation via diffusion models". In: *Proceedings of the 37th International Con-*

ference on Neural Information Processing Systems. NIPS '23. New Orleans, LA, USA: Curran Associates Inc., 2024.

[30] Weinan Zhang et al. "Deep Learning for Click-Through Rate Estimation". In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. Ed. by Zhi-Hua Zhou. Survey Track. International Joint Conferences on Artificial Intelligence Organization, Aug. 2021, pp. 4695–4703. DOI: 10.24963/ijcai.2021/636. URL: https://doi.org/10.24963/ijcai.2021/636.

Appendix A

Sourcecode

Our training codes are available at https://git.cs.bham.ac.uk/projects-2023-24/yxc487.

It should be noted that our codes are based on BlurHandNet, ScoreHypo, and MHFormer. There may be the same code as in these repositories.

To run the code, the BlurHand dataset should be downloaded following the guidance in BlurHandNet. Then the libraries can be installed with *requirements.txt*. The version of python is 3.8.10. The training script is *train_script.sh* and *rm_train_script_full.sh*.