Multi-Hypothesis 3D Hand Mesh Recovering from a Single Blurry Image

Anonymous ICME submission

***Abstract*—Recovery of 3D hand mesh from blurry hand images is challenging due to the ambiguity. Most existing works attempt to solve this issue by exploiting physical and temporal constraints. However, those works ignore the fact that multiple feasible solutions will exist. In this paper, we propose a two-stage Multi-Hypothesis Hand Mesh Recovery network, consisting of the Generation Model and the Selection Model. In the first stage, the Generation Model explicitly extracts the temporal information with an Unfolder. Then a Multi-Hypothesis Transformer generates multiple diverse hypotheses with a lightweight Hypothesis Embedding set. In the second stage, the Selection Model selects a subset of better-quality hypotheses. We additionally introduce combined classifying and ranking loss to better align the target of the selection model. Extensive experiments show that the proposed method produces much more accurate results on blurry images. Source codes will be available.**

***Index Terms*—multi-hypothesis, 3d reconstruction, hand mesh regression, computer vision**

I. Introduction

Recovering 3D human mesh is an active and challenging problem. It is essential in Augmented Reality [1], Motion Analysis [2] and other applications involving human-interaction. Recent advancements have been made in producing 3D hand mesh from sharp image [3], [4], However, ideal sharp images are not always obtained when deployed in the wild. For example, camera shake or fast movement of the subject during shooting can cause blurry images. This situation is prevalent in daily life for flexible hands with a lot of movement space, especially when the camera position is usually close to the hand, which will further amplify this blurriness problem. Therefore, for this important blurriness scenario, the conventional model trained on sharp images will face the problem of domain gap, resulting in poor performance in predicting on blurry images.

It is difficult to annotate blurry images that are so ambiguous that even humans cannot perceive them. Previous work [5] proposed to synthesize a similar visual effect of a blurry hand image by interpolating and averaging adjacent frames. Methodwise, the task of reconstructing the hand shape from a blurred image alone is too difficult. They also proposed a task formulation to predict a sequence of three consecutive frames from a blurred hand image instead of just a mesh, aiming to restore the blurred motion and use this temporal auxiliary to alleviate the difficulty of the task.

However, even so, blur hand pose estimation still has its inherent ambiguity [6]; it is also difficult to accurately estimate a motion sequence from a blurry image with severe feature information loss [5]. This is a one-to-many inverse problem

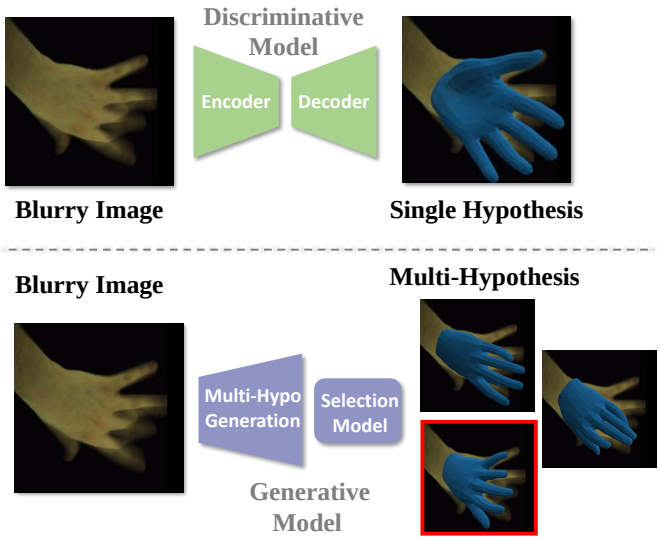

Fig. 1. Overview of results with single-hypothesis vs. multi-hypothesis. A recent state-of-the-art discriminative model [5], outputs a single solution that is inconsistent with the blurry image. Then, the proposed generative model outputs a multi-hypothesis to select the most accurate one (red box).

[7] as shown in Fig. 1 rather than a one-to-one deterministic mapping. In addition to the depth ambiguity and ambiguity caused by occlusion and self-occlusion commonly existing in monocular 3D estimation [7]–[10], the blur hand task also has its unique blur ambiguity. For example, a fast motion and its reverse sequence can be photographed to obtain a similar blur image, and it is difficult to determine the sequence from a single blur frame. In short, the blurry image corresponds to multiple reasonable and feasible motion sequences. The ambiguous nature of the task itself makes the traditional deterministic discriminant model fall into the local optimal solution or collapse the model to the mean prediction [7].

Inspired by the related field of sharp image pose estimation, we argue that making multiple feasible predictions is expected to ease such challenges, and for the first time propose a more natural and precise formulation of the blur hand task into a one-to-many multi-hypothesis generation task to model ambiguity. Unlike the conventional method of only regressing a single motion sequence, this task expects the model to simultaneously output complete, feasible, and diverse multiple motion sequences as long as they correspond to the blurred hand image, which of course includes the annotated solution in the data. Interest in multi-hypothesis estimation is increasing [7]–[10]. Some of them aggregate all hypotheses to get the

final prediction [8], [10], or aggregate a subset of hypotheses [6], [11] but the diversity of hypotheses is not promoting. These previous methods mainly focus on studying depth ambiguity and occlusion ambiguity, which are consistent with visible 2D image cues. However, this is not suitable for our scenario because blur will lead to different feasible motion sequences in time.

To address these challenges, we proposed a two-stage network consisting of the Generation Model and the Selection Model. In the first stage, the Generation Model extracts the temporal information with an encoder. Then a Multi-Hypothesis Transformer generates multiple candidate features with a lightweight Hypothesis Embedding set. The candidate features are then fed into a Regressor to produce hypotheses. To avoid hypotheses collapsing into one solution, we only backpropagate the loss gradient of the hypothesis with minimal distance to the ground truth (GT). Besides, to mitigate the impact of insufficient diversity in the data, a diversity-promoting loss is proposed to further enhance the diversity of hypotheses.

In the second stage, the Selection Model makes selections conditioned on the image with a combined classifying-ranking loss. Compared with the widely used pairwise loss [12], the combined classifying-ranking loss aligns the task better without losing the ranking information of hypotheses. Thanks to the training of CCR loss, the selection model can accurately evaluate the quality of the hypotheses generated by the first-stage generation model, and then filter out the hypotheses with higher quality that are most likely to match the input blur image based on the quality.

In conclusion, our contributions are summarized as follows:

- We observe significant ambiguity in the blur hand and for the first time propose a more intuitive way to formulate the inverse problem task into a one-to-many multi-hypothesis motion generation task to capture the unique blur ambiguity.
- A multi-hypothesis Generation Model is proposed as an efficient approach to generate multiple hypotheses with minimal additional parameters compared with BlurHandNet [5], the state-of-the-art in the task. Equipping the multi-hypothesis transformer with diversity-promoting loss, the generation model is able to propose feasible diverse motion sequences for blurry images.
- Along with the generation model, the effective Combined Classifying-Ranking (CCR) loss is designed to train a selection model to assess the score of generated hypotheses to select hypotheses with better quality conditioned on the image.
- The proposed multi-hypothesis generation and selection method achieved state-of-the-art results on the blur hand benchmark. Detailed ablation studies are conducted to better understand and reveal the advantages of our method and better model the blur ambiguity task.

II. Related Work

A. 3D Hand Mesh Estimation

Works of 3D Hand Mesh Estimation inspire from Human Mesh Estimation [13], [14]. MANO [15], a parameterized hand model, is introduced to provide human prior knowledge. The following work regresses the mesh based on the estimated joint and shape parameters [16], or directly estimates the vertex parameters [17], [18]. HandOccNet [3] introduces a transformer [19]-based structure to explicitly deal with the self-occlusion problem. Deformer [4] utilizes the temporal information of video with a spatial-temporal transformer. HaMeR [20] scales up ViT and the dataset to improve the performance and generalization. However, these works are orthogonal to the ambiguity caused by blurriness. Recently, [5] proposes a dataset BlurHand and a network BlurHandNet to overcome the ambiguity problem. It is different from our method as it does not involve Multi-Hypothesis Estimation.

B. Recovering Mesh from Blur

Most methods solve the problem of Mesh from Blur by images and video debluring [21]–[23]. Recently, the Shape from Blur [24] proposes to use an optimization method to recover 3D shapes from a single blurry image. The Human from Blur [25] extends the Shape from Blur and proposes the first method for human pose estimation from substantially blurred images. They use a 3D human model, a texture map, and a sequence of poses and solve the problem by backpropagating the pixel-wise reprojection error to recover the best human motion representation. The Deformer [26] reasons about the relationship between hand parts within the same spatial-temporal dimension from a blurry hand image. The BlurHandNet [5] unfolds a blurry input image to a 3D hand mesh sequence to utilize temporal information. However, those methods ignore that multiple feasible solutions exist in Mesh from Blur. Our method focuses on finding multiple plausible mesh hypotheses.

C. Multi-Hypothesis Estimation

The above discriminative models output only a single estimation mesh for a given image. There are a couple of works [7], [8], [27], [28] make multiple predictions generatively to explore the plausible estimation aligned well with the image. MHFormer [8] generates and aggregates three hypotheses once to produce a refined final output. MION [10] refines the hypotheses with PNCC positional embedding and Mesh Refine Transformer. These works aggregate all hypotheses with an extra module. On the other hand, some works [28], [29] simply select the best hypothesis with a minimal distance to the GT. GenPose [6] makes multi-hypothesis predictions given a partially observed cloud point and selects the better ones with an EnergyNet. ScoreHypo [11] train a ScoreNet to select a 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.

III. METHOD

A. Overview

Given a blurry image, our target is to recover a 3D hand mesh sequence that contains three key frames, the start, middle, and the end frame.

We use MANO [15] to simplify the estimation problem. MANO is a hand model that parameterizes 3D hand mesh $\mathbf{V} \in \mathbb{R}^{778\times 3}$ into pose $\theta \in \mathbb{R}^{48}$ and shape $\beta \in \mathbb{R}^{10}$. Given a blurry image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, we will produce three 3D hand meshes that represent the three key frames.

Conventional methods usually build an estimation network to directly estimate one result from blurry images. In this work, we argue that a single estimation cannot well-handle the problem of blurry images due to the ambiguity.

We instead propose a generative model pipeline where we first generate multiple hypotheses and then use a selection model to choose a reasonable subset of them.

To address the challenges of pose-ambiguity caused by blurriness and self-occlusions, we introduce the multi-hypothesis method. Given a blurry image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, the generation model is expected to make multiple plausible hand mesh estimations $\mathbf{H}^k, k = 1, \ldots, K$, where $\mathbf{H}^k = \theta_0^k \oplus \cdots \oplus \theta_T^k$ is the hypothesis that contains a pose sequence. K, and T are the number of hypotheses and the length of sequence, respectively. Further, we train a selection model to predict the quality signal $\hat{r}^k \in \mathbb{R}$ of each hypothesis conditional on the image $\mathbf{I}$. During inference, the generation model generates K hypotheses, and we select the top-n of them according to the score signal $\hat{s}^k$.

B. Multi-Hypothesis Generation

Figure 2 shows the overall pipeline of our work. It contains a Generation Model and a Selection Model. The generation model consists of an ResNet-based [30] Encoder, and a Multi-Hypothesis Transformer with a MANO parameter regressor.

Encoder. We use ResNet [30] and a bottom-up Feature Pyramid Network (FPN) [31] with deconvolution layer as the encoder. It outputs a joint-wise feature F_{J_t} for each timestep given a blurry image $\mathbf{I}$. The image is first fed into the ResNet [30] backbone to get a basic feature $F_B \in \mathbb{R}^{C\times H/32\times W/32}$. Then we use T separated deconvolution layers to predict the temporal feature $F_t \in \mathbb{R}^{H/8\times W/8\times E}$ where $H \times W$, C, E denote the resolution, the number of channels, 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 [32], 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.

Multi-Hypothesis Transformer (blue part in Fig. 2) is a transformer [19]-encoder module to generate refined temporal joint-wise features $F_J^k \in \mathbb{R}^{T\times J\times E}$ of multiple hypotheses, where k indicates different hypotheses. F_J^k feeds into the transformer as a sequence of $T \times J$ tokens. Previous works either produce each hypothesis with a weight-independent network [8], [10], [28], or samples solutions from a distribution on the parameter space of MANO [11]. We find it more efficient to generate hypotheses with a small embedding set $\{z_k\}_{k=1}^K$, whose size is the number of hypotheses. We set $z_k \in \mathbb{R}^E$. K different features $\{F_J^k\}_{k=1}^K$ are obtained by adding the hypothesis feature z_k to F_J token-wise. $\{F_J^k\}_{k=1}^K$ are forwarded to the transformer independently, i.e., they are processed as separated sequences and concatenated along the batch dimension. The output is K refined features $\{F_J^{k+}\}_{k=1}^K$.

Regressor produces MANO pose and shape parameters and camera parameters. 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 since change of pose is able to cover the main ambiguity of an image and shared shape and translation parameters prevent hypotheses deviating from each other too much, which stabilizes the learning. 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.

Multi-hypothesis losses. The overall training loss of the generation model is

$$\begin{aligned}\mathcal{L}_{gen} =& \lambda_{joint}\mathcal{L}_{joint} + \lambda_{proj}\mathcal{L}_{proj} + \lambda_{MANO}\mathcal{L}_{MANO} \\ &+ \lambda_{aux}\mathcal{L}_{aux} + \lambda_{div}\mathcal{L}_{div}\end{aligned} \tag{1}$$

where

$$\mathcal{L}_{joint} = \min_k \sum_{t=1}^{T} \left\|\mathbf{J}_t^k - \mathbf{J}_t^*\right\|_1 \tag{2}$$

$$\mathcal{L}_{proj} = \min_k \sum_{t=1}^{T} \left\|\Pi\mathbf{J}_t^k - \Pi\mathbf{J}_t^*\right\|_1 \tag{3}$$

$$\mathcal{L}_{MANO} = \min_k \sum_{t=1}^{T} \left\|\theta_t^k - \theta_t^*\right\|_1 + \sum_{t=1}^{T} \left\|\beta_t - \beta_t^*\right\|_1 \tag{4}$$

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

Notice that the loss only propagates the gradient of the best hypothesis as we do not want the hypothesis set to collapse to the mean prediction. Different hypotheses are able to capture different patterns of ambiguity in this way.

We also employ an auxiliary task

$$\mathcal{L}_{aux} = \sum_{t=1}^{T} \left\|\mathbf{J}_t^{2D} - \Pi\mathbf{J}_t^*\right\| \tag{5}$$

to improve the expressiveness of the Decoder, where $\mathbf{J}_t^{2D}$ is the 2D joints regressed by the encoder in the heatmap manner. To explicitly enhance the diversity across different generated hypotheses, we use a modified **diversity promoting loss** [33], [34]

$$\mathcal{L}_{div} = \frac{1}{K(K-1)} \sum_{k=1}^{K} \sum_{m=k+1}^{K} \sum_{t=1}^{T} e^{\frac{-\left\|F_{J_t}^k - F_{J_t}^m\right\|_1}{\alpha}} \tag{6}$$

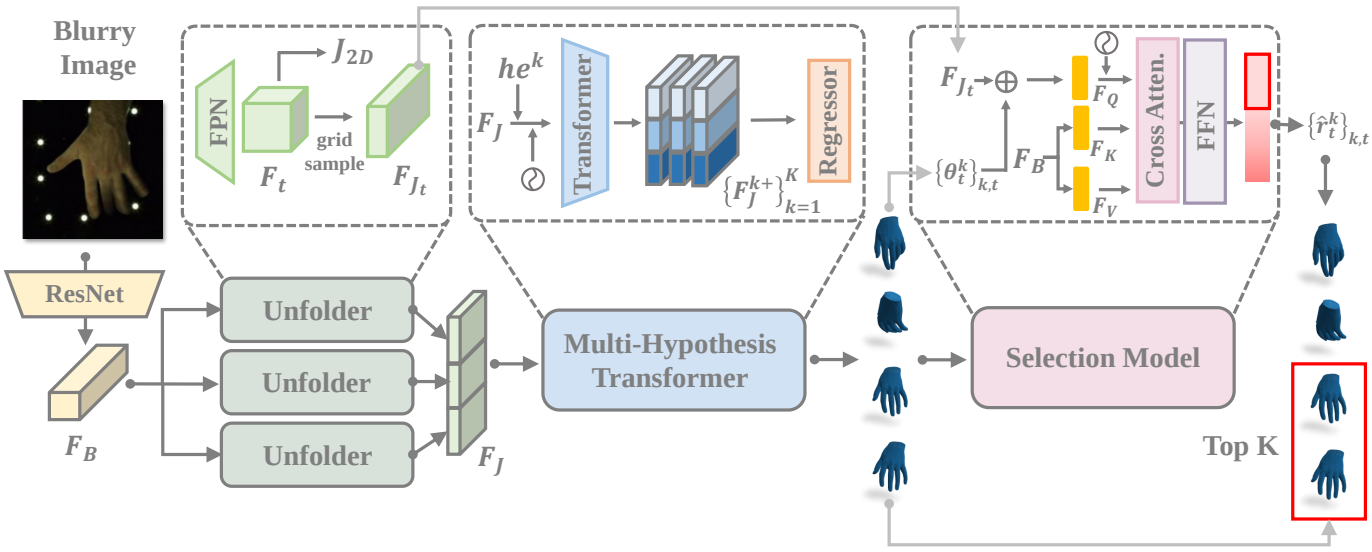

Fig. 2. **Overall pipeline of our method,** which consists of two key components, Multi-Hypothesis Generation model and Selection model. Given a blurry hand image, our method first extracts the global feature by a ResNet and then the Unfolders decodes it into three. The Generation Model generates K hypotheses given a blurry image. The Selection Model selects better hypotheses based on the outputs of the Generation Model and the blurry image.

where α is a normalizing factor. Compared with the original version [33], [34], we promote the distance in feature space instead of the output space to avoid influencing the quality of outputs. λ_{joint}, λ_{proj}, λ_{MANO}, λ_{aux} and λ_{div} are constant coefficients.

C. Selection Model

The Selection Model is to pick the better ones from the hypotheses $\{\mathbf{H}^k\}_{k=0}^K$. Aggregating all hypotheses to obtain a single final estimation via an aggregation module [8], [10] can be computationally expensive as the number of hypotheses goes large. More importantly, our method is aiming at proposing diverse plausible estimations. The aggregation manner fails to provide diversity as all hypotheses are trained to fit every sample. To this end, the Selection Model assigns a score signal to each hypothesis. During inference, the hypotheses with n-th highest score are selected as the most possible hand meshes that shape the image. We measure the possibility with Mean Per Joint Position Error (MPJPE).

Architecture. The Selection Model uses a transformer [19]-decoder module with cross attention (CA). It receives the feature F_J, F_B, and the hypotheses $\{\mathbf{H}^k\}_{k=0}^K$ from the Generation Model. Specifically, the output MLP layer of the transformer produces a two-dimensional value (μ^k, ν^k) instead of producing the score signal $\hat{s}^k$ directly. The quality signal $\hat{s}^k$ is calculated following

$$\hat{s}^k = \frac{\exp \mu^k}{\exp \mu^k + \exp \nu^k}$$

which is the softmax value of μ corresponding to softmax(μ^k, ν^k).

Combined Classifying and Ranking (CCR) Loss. Previous works [6], [11] train a Selection Model with pair-wise ranking loss [12] and take n hypotheses with the highest predicted score. However, the pair-wise loss is inconsistent with the goal of the Selection Model. The goal is to select the top-n of K without ranking the top-n. The pair-wise ranking loss exceeds the demand. On the other hand, it can be achieved by learning a binary classifier to distinguish whether the hypothesis is top-n. That is maximizing the probability of $p(y^k|\mathbf{H}^k, \mathbf{I})$, where

$$y^k = \begin{cases} 1, & \text{if } \mathbf{H}^k \text{ is top-}n \\ 0, & \text{if } \mathbf{H}^k \text{ is not top-}n \end{cases}$$

Nevertheless, such a classifying loss does not consider the ranking information hypotheses endowed.

To utilize the ranking information, combining classifying and ranking is a straightforward idea. However, $p(y^k = 1|\mathbf{H}^k, \mathbf{I})$ does not contain ranking information mathematically. $p(y^i = 1|\mathbf{H}^i, \mathbf{I}) > p(y^j = 1|\mathbf{H}^j, \mathbf{I})$ does not means $\mathbf{H}^i$ is better than $\mathbf{H}^j$. Thus, simply applying pairwise ranking loss to $p(y^k|\mathbf{H}^k, \mathbf{I})$ deviates from our expectation.

Inspired by [35], we propose Combined Classifying and Ranking loss, which simultaneously the two objects without conflict by maximizing the likelihood of joint distribution $p(\mathbf{x}, \mathbf{y})$ of hypotheses $\mathbf{x}$ and their classes $\mathbf{y}$. Further discussion refers to the supplementary. The CCR loss is formulated as a summation of two cross-entropy losses,

$$\mathcal{L}_{ccr} = \alpha \mathcal{L}_{class} + (1 - \alpha)\mathcal{L}_{rank} \tag{7}$$

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

The classifying part is the cross entropy loss to supervise the classification result given a hypothesis,

$$\mathcal{L}_{class} = -\log \frac{y^k \exp \mu^k + (1 - y^k) \exp \nu^k}{\exp \mu^k + \exp \nu^k} \tag{8}$$

Method	MPJPE (mm) ↓			MPVPE (mm) ↓
	past	middle	future	middle
I2L-MeshNet [18]	_	24.32	_	23.08
METRO [31]	_	20.54	_	27.03
Pose2Pose [36]	_	18.80	_	17.42
BlurHandNet [5]	18.08	16.80	18.21	15.30
Ours (best)	**16.95**	**15.45**	**17.07**	**13.92**
Ours (aggregate)	<u>18.02</u>	<u>16.76</u>	<u>18.20</u>	<u>15.22</u>

TABLE I
COMPARISON TO THE STATE-OF-THE-ARTS ON BLURHAND DATASET [5]. THE BEST RESULT IS **BOLD** AND THE SECOND BEST ONE IS <u>UNDERLINED</u>.

$best$-MPJPE				
K	1	4	16	32
Ours	16.78	15.27	14.12	13.78
MPJPE				
n	1	2	3	4
Ours	17.40	16.39	15.86	15.45

TABLE II
$best$-MPJPE OF THE MIDDLE FRAME WITH DIFFERENT VALUES OF K. MPJPE OF THE MIDDLE FRAME WITH DIFFERENT VALUES OF n.

Method	MPJPE
Random Selection	15.69
Projective Selection [39]	16.22
Selection Model *w/* Pair-wise loss [11], [12]	<u>15.62</u>
Selection Model *w/* CCR loss (Ours)	**15.45**

TABLE III
COMPARISON AMONG DIFFERENT SELECTION STRATEGIES.

The ranking part is the cross entropy loss to learn the quality (being top-n or not) distribution given image $\mathbf{I}$, which is

$$\mathcal{L}_{rank} = -\log \frac{y^k \exp \mu^k + (1-y^k) \exp \nu^k}{\sum_{j=1}^{K} \left[y^k \exp \mu^j + (1-y^k) \exp \nu^j \right]} \tag{9}$$

IV. EXPERIMENTS

A. Datasets and Metrics

BlurHand. BlurHand (BH) is a 3D hand pose dataset proposed by [5] based on InterHand2.6M [37]. It contains blurry images and corresponding 3D annotations. The blurry image is synthesized with five sequential sharp frames from a 30 fps video. The annotations include 3D joint coordinates, MANO pose and shape parameters of $1st$, $3rd$ and $5th$ frames. We train and test our model following the train-test split, containing 121,839 and 34,057 samples, respectively.

Metrics. Consistent with previous work [5], 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.

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

B. Implement Details

The size of embedding is $E = 512$. The length of the sequence is $T = 3$ following [5]. The number of hand joints is $J = 21$. We train the generation model for 15 epochs with AdamW optimizer [38] and the batch size 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 learning rate initialized to be 0.0002 and decayed by 0.2 at epochs 15, 20 and 25. Values of constant coefficient $\lambda_{joint}, \lambda_{proj}, \lambda_{MANO}, \lambda_{aux}, \lambda_{div}, \lambda_{ccl}$ are set to be $(10, 2, 1, 8, 0.01, 1)$, respectively. The generation model is trained to produces $K = 16$ hypotheses. The Selection Model takes the top-n of them as the feasible predictions. We set the top-n number to be $n = 4$. In the following sections, we report the minimal distance to the GT among the top-n hypotheses as "MPJPE", and the minimal distance among all K hypotheses as "$best$-MPJPE".

1) Comparisons with state-of-the-arts: Table I shows the comparison results with state-of-the-art methods on the BlurHand dataset. "Ours(best)" represents the result of the hypothesis with the minimum distance to the GT in the selected subset. "Ours(aggregate)" represents the result of aggregating selected hypotheses by simple average pooling. Our method outperforms the current best approach, BlurHandNet [5]. When a single final estimation is required, our hypothesis with a simple aggregation strategy is still competitive with BlurHandNet.

We provide a qualitative comparison of the mesh sequence of our approach and BlurHandNet at all time steps. Fig.3 shows that our method generates multiple diverse hypotheses. One of these selected hypotheses fits the groundtruth. It can be seen that for some severely ambiguous parts, like fingertips, BlurHandNet is likely to make wrong predictions. Our method is able to cover the correct estimation.

C. Ablation Study

1) Number of the hypothesis: We are interested in the effectiveness of the multi-hypothesis method. Concretely, how the number of all hypotheses K, and the number of selected hypotheses n influence the performance. Table II shows that $best$-MPJPE decreases as the number of hypotheses K grows. The $best$-MPJPE can be regarded as the lower bound of error our method can reach. When $K = 1$, our method, which no longer has the advantage of the Multi-Hypothesis method and Selection Model, degenerates back to BlurHandNet. With a fixed K, the MPJPE decreases as n grows larger, and our method outperforms BlurHandNet [5] when $n \geq 2$. Selecting a single best hypothesis is not robust due to the blurriness compared with selecting on sharp images like [11]. When the selection number n increases to 2, the model benefits from multi-hypotheses and gets a better result than BlurHandNet.

2) Hypothesis Embedding: Unlike learning a latent space for probabilistic generative decoder [11], [27], our method adds noise-like embeddings to the learnt feature F_J. However, learnable embedding is more meaningful than merely adding noise. Table IV shows the comparative results of the

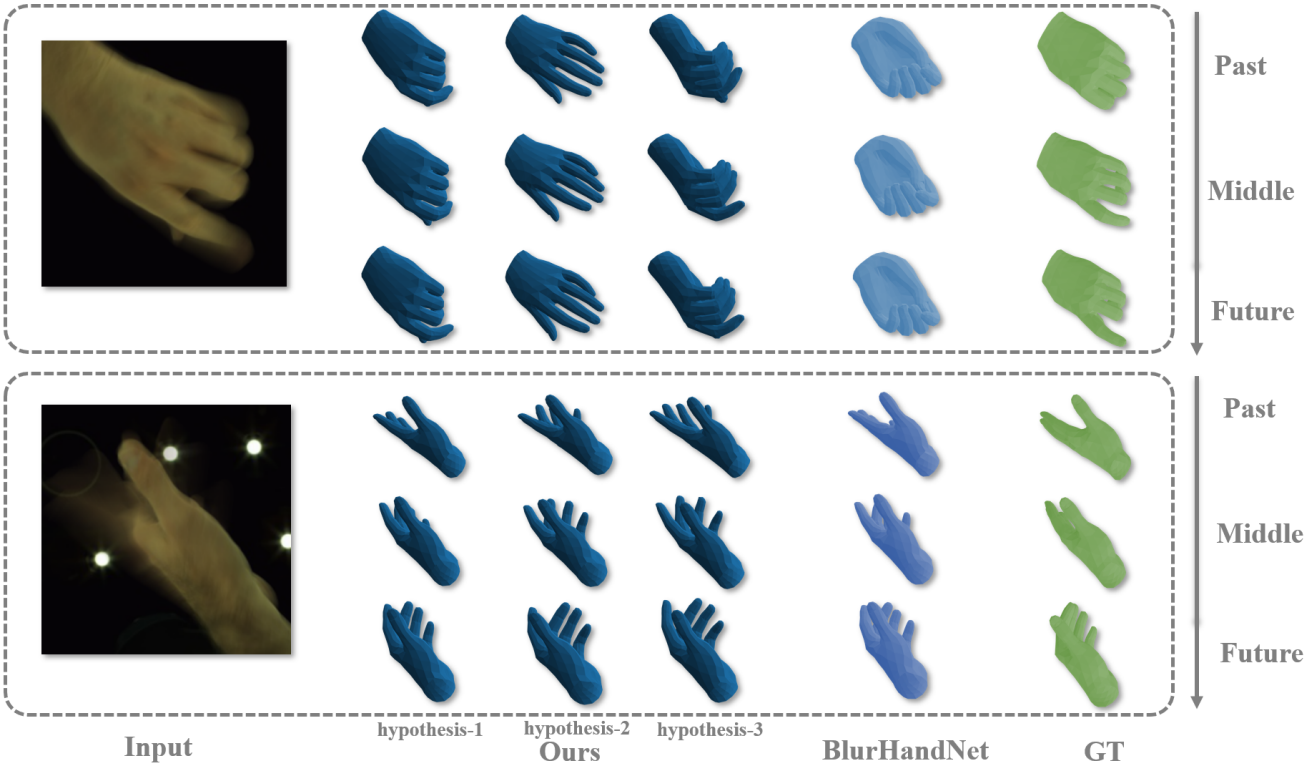

Fig. 3. Qualitative results of comparison of our method and BlurHandNet [5] on BlurHand [5] test set. Hypothesis-1, hypothesis-2 and hypothesis-3 represent the best, median, and worst hypothesis, respectively.

Method				*best*-MPJPE		
DP loss	noise	VAE	Embedding	past	middle	future
✓	✓			16.48	14.62	16.67
✓		✓		16.22	14.52	16.34
			✓	16.03	14.35	16.14
✓			✓	**15.52**	**14.12**	**15.66**

TABLE IV
ABLATION ON DIVERSITY PROMOTING LOSS.

methods to generate a hypothesis. "noise" simply adding noise sampled from a Gaussian with zero expectation and variance of $1/E$. "VAE" applies a VAE [40] which learns a Gaussian distribution $\mathcal{N}(\mu, \Sigma|F_J)$ conditional on F_J. The hypothesis embeddings F_J^k are sampled from $\mathcal{N}(\mu, \Sigma|F_J)$. We report the $best$-MPJPE to evaluate the error lower bound of each method.

3) Diversity of Hypotheses: We promote the diversity of hypotheses to cover more plausible predictions. Table IV gives a quantitative evaluation of diversity. "DP loss" indicates the model is trained with loss.6. The overall generalization ability is diminished without diversity and leads to a worse result.

4) Effect of Selection: In this part, we verify the impact of the Selection Model. Table III reports the results of different selection strategies. "Random Selection" randomly ranks hypotheses, which is seen as the baseline. "Projective Selection" indicates ranking hypotheses according to the distance to the 2D joints $\mathbf{J}^{2D}$ predicted by the Encoder. It is a basic method to select hypotheses in 2D-3D lifting tasks [39], where the GT 2D joints are known. "Selection Model" uses the Selection Model to rank hypotheses. "*w/* Pair-wise loss" trains the Selection Model with pair-wise loss as [6], [11] do. "*w/* CCR loss" trains the Selection Model with the CCR loss. All methods take $n = 4$ candidates among $K = 16$ hypotheses.

As table III shows, selecting with the Selection Model can effectively distinguish the good hypotheses. CCR loss is superior to the pair-wise loss. As we mentioned in Section III-C, the CCR loss better aligns the classification task while reserving the ranking information and utilizing it for the task. Projective Selection, as a heuristic method, gets the worst result as the predicted 2D joints are noisy, leading to incorrect selection.

V. CONCLUSION

This paper proposes a two-stage method to address the ambiguity problem of blurry images. We introduce a Generation Model that generates multiple feasible pose sequences efficiently with minimal additional parameters. To select hypotheses that better align the blurry image, we further propose a Selection Model. Our method outperforms the state-of-the-art on BlurHand Dataset [5].

Although our method provides reliable and diverse estimations for blurry images. It is not flexible enough. Change on K and n requires retraining a new model or loss of performance. Additionally, the evaluation of diversity is still indirect and vague [7]. To produce a generalized solution, we can leverage reinforcement learning [41] to directly encourage the best estimation.

REFERENCES

[1] Julie Carmigniani and Borko Furht, *Augmented Reality: An Overview*, pp. 3–46, Springer New York, New York, NY, 2011.

[2] Steffi Colyer, Murray Evans, Darren Cosker, and Aki Salo, "A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system," *Sports Medicine - Open*, vol. 4, pp. 24, 06 2018.

[3] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee, "Handoccnet: Occlusion-robust 3d hand mesh estimation network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 06 2022, pp. 1496–1505.

[4] Qichen Fu, Xingyu Liu, Ran Xu, Juan Carlos Niebles, and Kris M. Kitani, "Deformer: Dynamic fusion transformer for robust hand pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10 2023, pp. 23600–23611.

[5] Yeonguk Oh, JoonKyu Park, Jaeha Kim, Gyeongsik Moon, and Kyoung Mu Lee, "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.

[6] Jiyao Zhang, Mingdong Wu, and Hao Dong, "Genpose: generative category-level object pose estimation via diffusion models," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024, NIPS '23, Curran Associates Inc.

[7] 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)*, 10 2023, pp. 14840–14849.

[8] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool, "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.

[9] 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)*, October 2023, pp. 15977–15987.

[10] Zhiwei Liu, Xiangyu Zhu, Lu Yang, Xiang Yan, Ming Tang, Zhen Lei, Guibo Zhu, Xuetao Feng, Yan Wang, and Jinqiao Wang, "Multi-initialization optimization network for accurate 3d human pose and shape estimation," in *Proceedings of the 29th ACM International Conference on Multimedia*, New York, NY, USA, 2021, MM '21, p. 1976–1984, Association for Computing Machinery.

[11] Yuan Xu, Xiaoxuan Ma, Jiajun Su, Wentao Zhu, Yu Qiao, and Yizhou Wang, "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.

[12] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender, "Learning to rank using gradient descent," in *Proceedings of the 22nd International Conference on Machine Learning*, New York, NY, USA, 2005, ICML '05, p. 89–96, Association for Computing Machinery.

[13] 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.

[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] Javier Romero, Dimitrios Tzionas, and Michael J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, vol. 36, no. 6, Nov. 2017.

[16] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan, "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.

[17] 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*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, Eds., Cham, 2020, pp. 769–787, Springer International Publishing.

[18] 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*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, Eds., Cham, 2020, pp. 752–768, Springer International Publishing.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. 2017, vol. 30, Curran Associates, Inc.

[20] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik, "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.

[21] Grigorios G Chrysos, Paolo Favaro, and Stefanos Zafeiriou, "Motion deblurring of faces," *International journal of computer vision*, 2019.

[22] Denys Rozumnyi, Martin R Oswald, Vittorio Ferrari, Jiri Matas, and Marc Pollefeys, "Defmo: Deblurring and shape recovery of fast moving objects," in *CVPR*, 2021.

[23] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas, "Deblurgan: Blind motion deblurring using conditional adversarial networks," in *CVPR*, 2018.

[24] Denys Rozumnyi, Martin R Oswald, Vittorio Ferrari, and Marc Pollefeys, "Shape from blur: Recovering textured 3d shape and motion of fast moving objects," *NeurIPS*, 2021.

[25] Yiming Zhao, Denys Rozumnyi, Jie Song, Otmar Hilliges, Marc Pollefeys, and Martin R Oswald, "Human from blur: Human pose tracking from blurry images," in *ICCV*, 2023.

[26] Qichen Fu, Xingyu Liu, Ran Xu, Juan Carlos Niebles, and Kris M Kitani, "Deformer: Dynamic fusion transformer for robust hand pose estimation," in *ICCV*, 2023.

[27] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis, "Probabilistic modeling for human mesh recovery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 11605–11614.

[28] Benjamin Biggs, David Novotny, Sebastien Ehrhardt, Hanbyul Joo, Ben Graham, and Andrea Vedaldi, "3d multi-bodies: Fitting sets of plausible 3d human models to ambiguous image data," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 20496–20507, Curran Associates, Inc.

[29] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla, "Humaniflow: Ancestor-conditioned normalising flows on so(3) manifolds for human pose and shape distribution estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 4779–4789.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks," in *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, Eds., Cham, 2016, pp. 630–645, Springer International Publishing.

[31] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[32] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei, "Integral human pose regression," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[33] 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)*, October 2021, pp. 13309–13318.

[34] 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*, Berlin, Heidelberg, 2020, p. 346–364, Springer-Verlag.

[35] Xiang-Rong Sheng, Jingyue Gao, Yueyao Cheng, Siran Yang, Shuguang Han, Hongbo Deng, Yuning Jiang, Jian Xu, and Bo Zheng, "Joint optimization of ranking and calibration with contextualized hybrid model," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2023, KDD '23, p. 4813–4822, Association for Computing Machinery.

[36] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee, "Accurate 3d hand pose estimation for whole-body 3d human mesh estimation,"

in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2022, pp. 2308–2317.

[37] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee, "Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image," in *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, Eds., Cham, 2020, pp. 548–564, Springer International Publishing.

[38] Ilya Loshchilov and Frank Hutter, "Fixing weight decay regularization in adam," *CoRR*, vol. abs/1711.05101, 2017.

[39] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao, "Diffusion-based 3d human pose estimation with multi-hypothesis aggregation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 14761–14771.

[40] Diederik P Kingma, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[41] Hai Ci, Mickel Liu, Xuehai Pan, fangwei zhong, and Yizhou Wang, "Proactive multi-camera collaboration for 3d human pose estimation," in *The Eleventh International Conference on Learning Representations*, 2023.