MGDHand: Multi-Granularity Prior-to-Inertial Distillation Framework for Sequential 3D Hand Pose Estimation from Sparse IMUs

Supplementary Material

In the supplementary material, we provide:

- More details of network structure and computational requirements in Sec. 1,
- The details of hand pose reconstruction loss in Sec. 2,
- The effectiveness of prior decoupling in Sec. 3,
- More ablation experiments in Sec. 4,
- More qualitative results in Sec. 5,

Note that all the notation and abbreviations here are consistent with those in the main paper.

1. Details of Network Structure and Computational Requirements

Given a MANO parameter sequence $\mathbf{\Theta} = \{(\boldsymbol{\theta}_t, \boldsymbol{\beta})\}_{t=1}^{T}$, we introduce random joint-wise and frame-wise masking on the pose parameters $\boldsymbol{\theta}$ to generate a self-supervised signal for cross-modal completion. The pose parameters $\boldsymbol{\theta}_t \in \mathbb{R}^{J\times d_\theta}$ describe joint rotations, and the sequence-wise shape parameter $\boldsymbol{\beta} \in \mathbb{R}^{d_\beta}$ controls hand morphology. Let $\mathbf{M}^{\text{joint}} \in \{0,1\}^{T\times J}$ be a joint-level binary mask from Bernoulli(p_m) sampling. We apply this mask to $\boldsymbol{\theta}$, replacing masked positions with a random noise vector $\mathbf{t}_{\text{mask}} \in \mathbb{R}^{d_\theta}$, yielding the corrupted pose sequence $\tilde{\boldsymbol{\theta}}$. The resulting masked MANO sequence is $\tilde{\mathbf{\Theta}} = \{(\tilde{\boldsymbol{\theta}}_t, \boldsymbol{\beta})\}_{t=1}^{T}$. By forcing the teacher to recover the full MANO sequence from $\tilde{\mathbf{\Theta}}$ and IMU signals, we encourage it to capture robust hand pose, shape and motion priors.

We train the MGDistill framework with a single GPU (NVIDIA 4090) and a batch size of 32. We adopt the DSTformer [4] backbone with depth N = 5, number of heads h = 8, feature size C = 512, kernel size k = 4. For the teacher network, the FLOPs is 9.32 G, and the model parameters are 12.52 M. For the student network, the FLOPs is 2.22 G, and the model parameters are 9.62 M. The FLOPs are measured with sequence length T = 32 and the 7-IMU setting.

2. Hand Pose Reconstruction Loss

The predicted parameters $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$ are fed into the MANO layer $\mathcal{M}$ to explicitly reconstruct 3D joints $\hat{\mathbf{J}} \in \mathbb{R}^{T\times J\times 3}$ and mesh $\hat{\mathbf{V}} \in \mathbb{R}^{T\times N_v\times 3}$. Similar to previous methods [1–3], we supervise the joints and mesh vertices with the smooth L1 loss. Meanwhile, we use smooth L1 loss to supervise the pose parameters θ and shape parameters β. We denote the mesh, joint, pose, and shape losses by $\mathcal{L}_{mesh}$, $\mathcal{L}_{\text{joint}}$, $\mathcal{L}_\theta$, and $\mathcal{L}_\beta$ respectively, and define the overall task loss as:

$$\mathcal{L}_{\text{recon}} = \lambda_{\text{joint}}\mathcal{L}_{\text{joint}} + \lambda_{\text{mesh}}\mathcal{L}_{\text{mesh}} + \lambda_\theta\mathcal{L}_\theta + \lambda_\beta\mathcal{L}_\beta, \quad (1)$$

where λ_{joint}, λ_{mesh}, λ_θ, and λ_β are scalar weights that balance the different supervision terms.

3. Effectiveness of Prior Decoupling

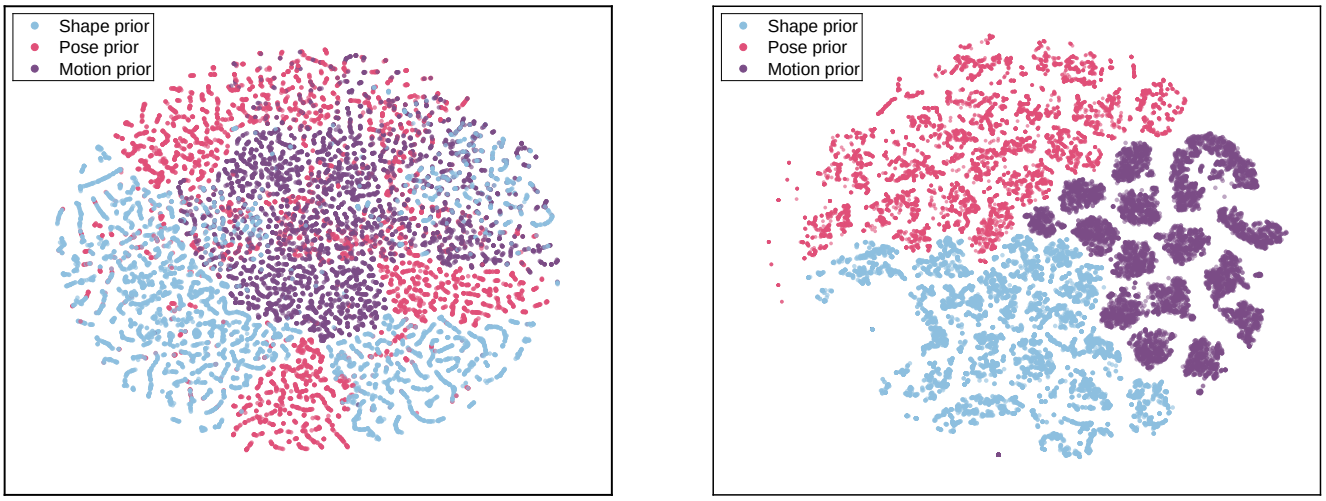
To verify the effectiveness of the proposed prior decoupling, we further analyze the teacher representations before and after applying MGDistill. Specifically, we collect the fused global features from the MANO–IMU teacher, as well as the decoupled pose, shape, and motion features, and visualize them using t-SNE. As shown in Fig. 1 (a), the fused features space before disentanglement exhibits strong semantic entanglement. Pose, shape, and motion features are heavily overlapped in the same latent region, which indicates that different types of priors are encoded in a single global representation. Directly distilling the entangled features to the IMU-based student leads to optimization difficulty of the student model.

After decoupling, the priors become much more structured and interpretable, as illustrated in Fig. 1 (b). The static shape, dynamic pose, and temporal motion features now occupy clearly different areas in the latent space. This observation demonstrates that MGDistill successfully factorizes the original complex coupled features into three complementary priors with clearer semantics, which in turn facilitates more targeted and stable knowledge transfer to the student model.

4. Ablation Study

To investigate how each type of teacher prior contributes to the student model, we conducted an ablation study by selectively enabling different distillation modules. We first contrast the effect of multi-granularity decoupled distillation with that of direct distill coupling fused features. Then, we construct three variants of MGDistill that only the Static Shape Distillation (SSD), only the Dynamic Pose Distillation (DPD), or only the Temporal Motion Distillation (TMD), respectively. The results are summarized in Tab. 1.

Compared with the student model that directly regresses hand poses from IMU signals without distillation, distilling the teacher's globally entangled features into the student reduces the MPJPE error by 2.48 mm. However, the global features simultaneously mixes shape, pose, and motion semantics and differs markedly from the student representation in both information density and semantic granularity. As a result, directly distilling such entangled priors still makes optimization difficult, and its performance is clearly



(a) Before decoupling **(b) After decoupling**

Figure 1. T-SNE visualization of teacher feature distribution before and after prior disentanglement. (a) Fused priors features before disentanglement. (b) Shape, pose and motion priors after decoupling.

Table 1. Ablation study of different distillation configurations on VIHand with the 7-IMU setting. Baseline denotes the student model without distillation. w/o decouple denotes directly distills the fused features from the teacher.

Methods	Distill	L_{ss}	L_{ps}	L_{ts}	MPJPE	MPVPE
baseline	✗	✗	✗	✗	15.40	17.15
w/o decouple	✓	✗	✗	✗	12.92	14.43
SSD-only	✓	✓	✗	✗	14.72	16.97
DPD-only	✓	✗	✓	✗	10.36	12.52
TMD-only	✓	✗	✗	✓	12.68	14.71
MGDistill	✓	✓	✓	✓	**9.13**	**10.46**

inferior to our proposed multi-granularity decoupled distillation MGDistill. On top of distillation without decouple, MGDistill further reduces MPJPE by 3.79 mm, achieving a total reduction of 6.27 mm over the non-distilled baseline and thus confirming the necessity of decoupling the teacher priors and transferring them at multiple granularities. We further analyze the effect of distilling each prior individually. When only the static shape prior is distilled, the MPJPE error is reduced by just 0.68 mm, mainly improving the global hand morphology and alleviating shape distortion, but remaining insufficient to regularize fine-grained joint kinematics and temporal consistency. Distilling only the dynamic pose prior reduces MPJPE by 5.04 mm, significantly improving joint position accuracy and pose plausibility, yet it lacks constraints on hand shape and bone length and still exhibits a certain degree of motion jitter in the sequence. When only the temporal motion prior is distilled, MPJPE decreases by 2.72 mm, which mainly enhances dynamic smoothness and motion continuity, but the per-frame pose accuracy and hand morphology are still suboptimal.

Overall, these three priors are clearly complementary: static shape distillation regularizes subject-specific hand morphology, dynamic pose distillation constrains fine-grained joint relations, and temporal motion distillation reinforces global motion smoothness and dynamical consistency. The ablation results show that each prior brings non-negligible gains on its targeted aspect, while jointly distilling all of them in MGDistill yields the best overall performance and robustness under sparse IMU configurations.

5. Qualitative Results

We provide more qualitative results in Fig. 2. We can observe that directly regressing dense hand poses from the sparse IMUs often leads to pose misalignment (row 5) and morphological distortion (row 3). In addition, there is also a temporal prediction lag when rapidly changing gestures (row 1, columns 6 and 7). These results highlight the inherent ambiguity and the difficulty of learning robust sparse-to-dense mappings. Overall, our MGDistill strategy significantly optimizes the dense pose misalignment, geometric shape deformation, and temporal motion bias after multi-granularity decoupled distillation, resulting in consistently accurate joint positions and geometries.

References

[1] Xinghao Chen, Guijin Wang, Hengkai Guo, and Cairong Zhang. Pose guided structured region ensemble network for cascaded hand pose estimation. *Neurocomputing*, 395:138–149, 2020. 1

[2] Pengfei Ren, Haifeng Sun, Qi Qi, Jingyu Wang, and Weiting Huang. Srn: Stacked regression network for real-time 3d hand pose estimation. In *BMVC*, 2019.

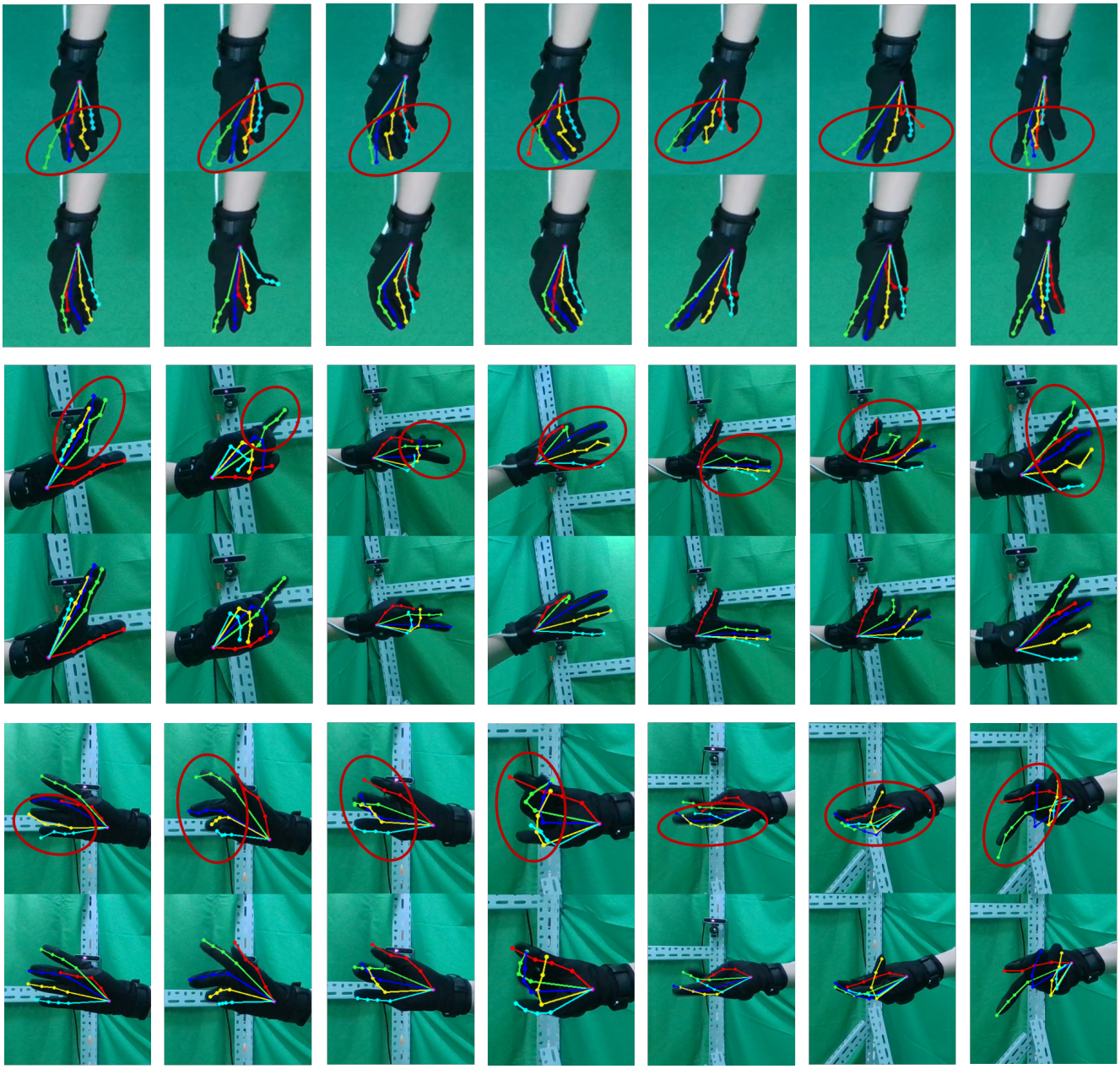


Figure 2. Qualitative results. For each sequence, the top row shows the baseline pose (before distillation), and the bottom row shows the pose after applying our MGDistill. We only use sparse IMUs in the data glove as input; RGB images are for visualization only.

[3] Pengfei Ren, Chao Wen, Xiaozheng Zheng, Zhou Xue, Haifeng Sun, Qi Qi, Jingyu Wang, and Jianxin Liao. Decoupled iterative refinement framework for interacting hands reconstruction from a single rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8014–8025, 2023. 1

[4] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15085–15099, 2023. 1