Supplementary Materials: Decoupling Heterogeneous Features for Robust 3D Interacting Hand Poses Estimation

Anonymous Authors

1 INTRODUCTION

This supplemental material provides:

- More qualitative results to showcase the effectiveness and practicality of our method. These results consist of in-the-wild images or videos captured using a smartphone or from the internet in Sec. 1.1,
- More detailed qualitative analysis of how the proposed decoupling strategy works in Sec. 1.2,
- Details of the Feature Interaction between Joints and the Image in Sec. 1.3,
- More ablation experiment. in Sec. 1.4.

Please note that all the notation and abbreviations in this supplementary material are consistent with those in the main manuscript.

1.1 More qualitative results

In this section, we provide more qualitative results on images captured by a smartphone Fig. 2 and sourced from the Internet Fig. 3. Our model is only trained on the InterHand2.6M dataset. We compare our method with the previous state-of-the-art model-free work [2]. We obtained the result by utilizing a hand detector to extract the hand region from an in-the-wild image while preserving the aspect ratio. Our method achieves robust results even in cases of severe self-similarity compared to [2] (For more qualitative results, please refer to the video demo).

1.2 More qualitative Analysis

More heatmaps are provided to illustrate the effectiveness of the decoupling strategy further in Fig. 1. Due to the entanglement of the two types of features, the baseline method struggles to precisely focus on the positions of the joints in the presence of severe self-similarity. After incorporating our decoupling strategy, the two types of features mutually enhance by leveraging visual and spatial cues. As a result, both the position and appearance features can localize the positions of the joints in the image.

1.3 Details of the Feature Interaction between Joints and the Image

In this section, we provide more mathematical details about the implementation of feature interaction between joints and the image. In the main paper, we fuse the relationships between position and appearance and use these relationships to guide the enhancement of two types of features:

$$\begin{aligned}\mathbf{Q}_t &= Concat(\mathbf{J}_t^{a'},\mathbf{J}_t^{p'}), \mathbf{K}_t = \mathbf{V}_t = Concat(\mathbf{F}_t^{a},\mathbf{F}_t^{p}),\\ \mathbf{J}_{t+1}^{a},\mathbf{J}_{t+1}^{p} &= Split(Attn(\mathbf{Q}_t,\mathbf{K}_t,\mathbf{V}_t)).\end{aligned} \tag{1}$$

where $Attn(q,k,v)$ denotes multi-head attention [4]. *Split* and *Concat* represent the operations of separating and concatenating along the last dimension, respectively. Assuming the number of heads in the multi-head attention is 1, the attention mechanism first applies linear transformations to the position and appearance features using weight matrices. For brevity of exposition, we assume that the weight matrices are identity matrices similar to the proof in [5]. We keep the notation for the position and appearance features unchanged after the identity mapping. Then, the attention mechanism performs similarity calculations as:

$$\mathbf{A} = Softmax(DP(\left[\mathbf{J}_t^{a'},\mathbf{J}_t^{p'}\right], \left[\mathbf{F}_t^{a},\mathbf{F}_t^{p}\right]^T)). \tag{2}$$

where $DP(\mathbf{M}_1,\mathbf{M}_2)$ denotes Dot Product computation representing the pairwise dot product operation between the row vectors of matrix $\mathbf{M}_1$ and the column vectors of matrix $\mathbf{M}_2$. $\mathbf{A} \in \mathbb{R}^{2J\times(H_tW_t)}$ is the normalised attention map. For simplification, we disregard constants. We can rewrite it as:

$$\begin{aligned}\mathbf{A} &= Softmax(DP(\mathbf{J}_t^{a'},\mathbf{F}_t^{a}) + DP(\mathbf{J}_t^{p'},\mathbf{F}_t^{p}))\\ &= Softmax(\mathbf{A}_a + \mathbf{A}_p)\end{aligned} \tag{3}$$

where the differences in constants are disregarded. This equation represents the extraction of respective relationships from each feature and the fusion of them. Afterwards, the relationships are used to aggregate each feature.

Table 1: Ablation study on InterHand2.6M [3].

ID	MPJPE (mm)↓		
	Single	Two	All
Best model	7.35	9.82	8.67
w/o Multi-scale	7.54	9.84	8.77
w/o Identity Info	7.56	9.86	8.78
w/o All	7.62	9.88	8.82

1.4 Ablation Study

In our best model, we incorporate multi-scale features and part segmentation to improve performance. Multi-scale features allow the model to capture information at different scales. Part segmentation, on the other hand, helps provide identity information for individual pixels, improving the model's ability to classify different hand parts as mentioned in [1]. As indicated in Table 1, both designs resulted in a performance improvement of approximately 0.1mm. If both designs are removed, the MPJPE drops 0.15mm.

REFERENCES

[1] Zicong Fan, Adrian Spurr, Muhammed Kocabas, Siyu Tang, Michael J Black, and Otmar Hilliges. 2021. Learning to disambiguate strongly interacting hands via probabilistic per-pixel part segmentation. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 1–10.

[2] Changlong Jiang, Yang Xiao, Cunlin Wu, Mingyang Zhang, Jinghong Zheng, Zhiguo Cao, and Joey Tianyi Zhou. 2023. A2J-Transformer: Anchor-to-Joint Transformer Network for 3D Interacting Hand Pose Estimation from a Single RGB Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8846–8855.

[3] 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. In *European Conference on Computer Vision*. 548–564.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[5] Frederic Z Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould. 2023. Exploring Predicate Visual Context in Detecting of Human-Object Interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10411–10421.

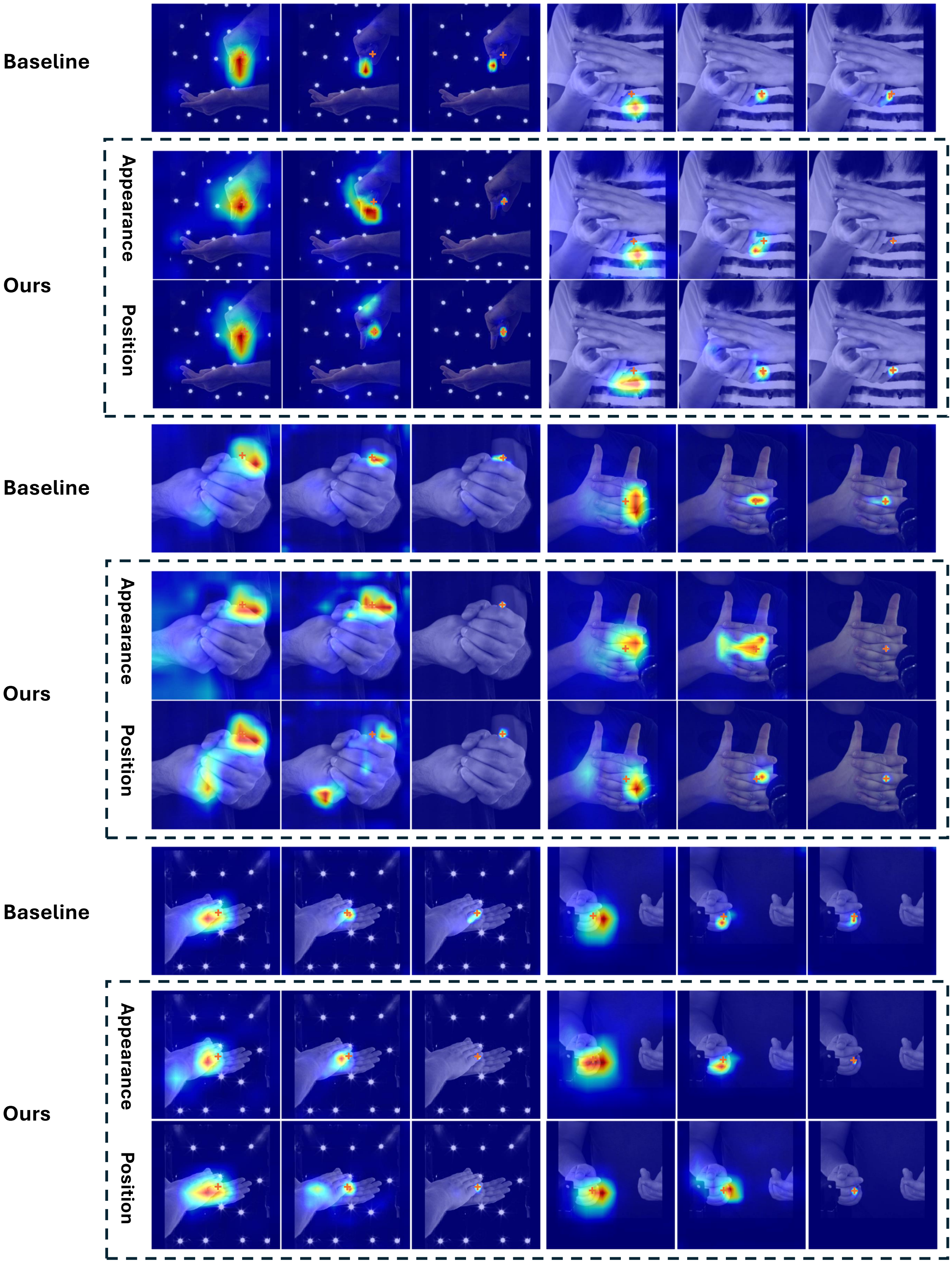

Figure 1: Illustration of the 3-iteration attention maps from the feature interaction between joints and image. The heatmaps from the baseline method without the decoupling strategy are presented first, followed by the heatmaps of the appearance and position from our method. The ground truth joint positions are marked with an orange cross.

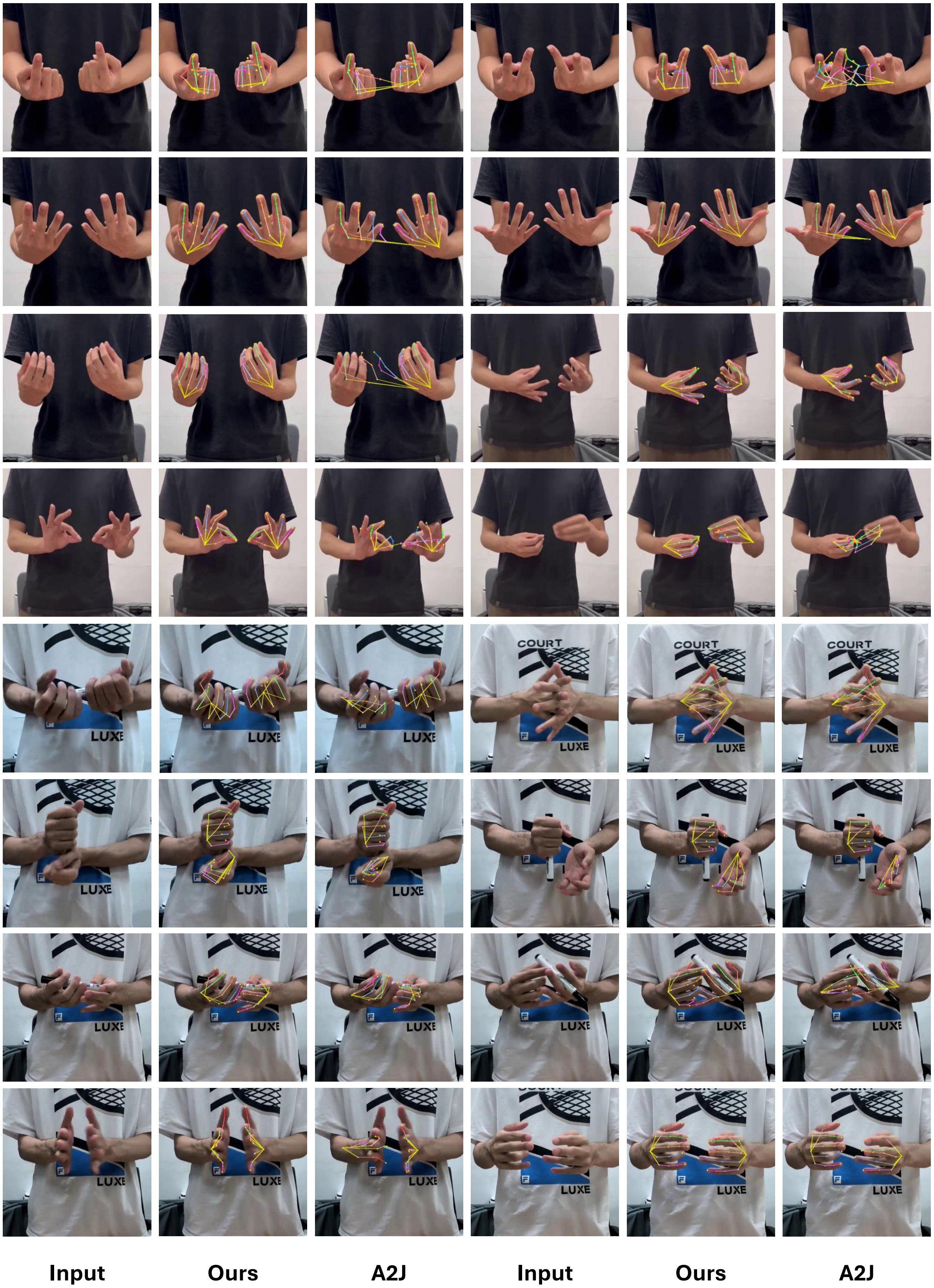

Figure 2: Qualitative results of A2J [2] and ours on images captured using a smartphone.

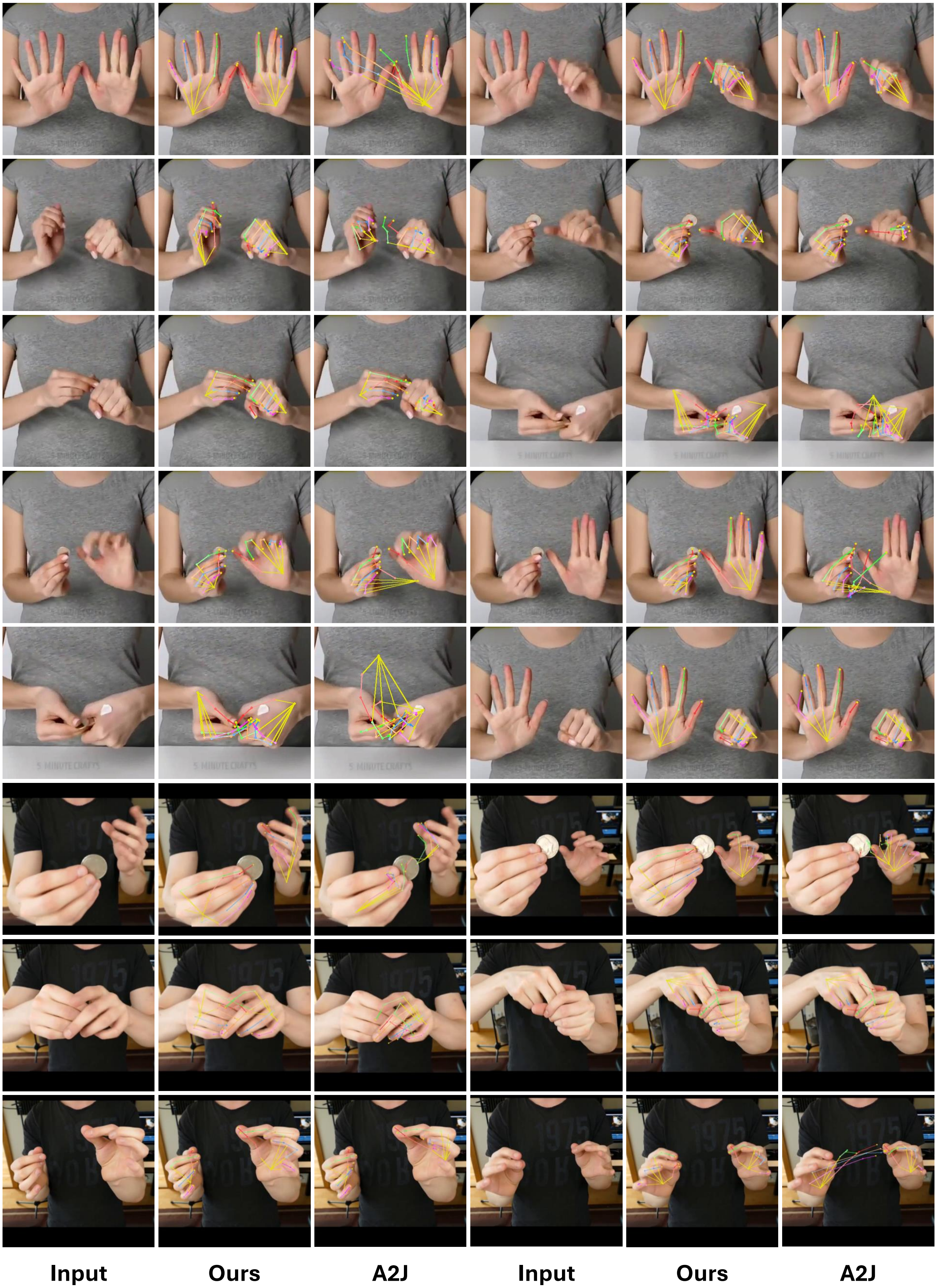

Figure 3: Qualitative results of A2J [2] and ours on images sourced from the Internet.