Knowledge as Priors: Cross-Modal Knowledge Generalization for Datasets without Superior Knowledge (Supplementary Material)

Long Zhao[1] Xi Peng[2] Yuxiao Chen[1] Mubbasir Kapadia[1] Dimitris N. Metaxas[1]
[1]Rutgers University [2]University of Delaware
{lz311,yc984,mk1353,dnm}@cs.rutgers.edu, xipeng@udel.edu

Abstract

This supplementary material provides additional results supporting the claims of the main paper. First, the proof of Proposition 1 of the main paper is given in Sect. 1. Second, we provide an efficient way to implement the meta-training method (Algorithm 1 of the main paper) in Sect. 2. Third, Sect. 4 describes the detailed experimental setup on FreiHAND [5], which is supplementary to the discussion section in the experiments of the main paper. Finally, we show more visual results of our method on RHD [4], STB [2] and synthetic dataset in Sect. 4.

1. Proof of Proposition 1

Proposition 1. *Let q be any posterior distribution function over the latent variables $\boldsymbol{\theta}$ given the evidence $\mathcal{D}_S$. Then, the marginal log-likelihood can be lower bounded:*

$$\log P(\mathcal{D}_S|\boldsymbol{\phi}) = \log \int P(\mathcal{D}_S, \boldsymbol{\theta}|\boldsymbol{\phi}) d\boldsymbol{\theta} \geq \mathcal{E}(q, \boldsymbol{\phi}),$$

where $\mathcal{E}$ is the evidence lower-bound (ELBO) defined as:

$$\mathcal{E}(q, \boldsymbol{\phi}) \triangleq \mathbb{E}_q[\log P(\mathcal{D}_S|\boldsymbol{\theta})] - \mathrm{KL}[q(\boldsymbol{\theta}|\mathcal{D}_S) \| P(\boldsymbol{\theta}|\boldsymbol{\phi})].$$

Proof. The proposed meta-training as described in Algorithm 1 of the main paper makes a posterior inference based on the graphical model in Fig. 1. Given the evidence $\mathcal{D}_S$, learning the parameters $\boldsymbol{\phi}$ leads to maximize the likelihood $P(\mathcal{D}_S|\boldsymbol{\phi})$:

$$\begin{aligned}\log P(\mathcal{D}_S|\boldsymbol{\phi}) &= \log \int P(\mathcal{D}_S, \boldsymbol{\theta}|\boldsymbol{\phi}) d\boldsymbol{\theta} \\ &= \log \int P(\mathcal{D}_S|\boldsymbol{\theta}, \boldsymbol{\phi}) P(\boldsymbol{\theta}|\boldsymbol{\phi}) d\boldsymbol{\theta} \\ &= \log \int P(\mathcal{D}_S|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\boldsymbol{\phi}) d\boldsymbol{\theta} \\ &= \log \int q(\boldsymbol{\theta}|\mathcal{D}_S) \frac{P(\mathcal{D}_S|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\boldsymbol{\phi})}{q(\boldsymbol{\theta}|\mathcal{D}_S)} d\boldsymbol{\theta}.\end{aligned}$$

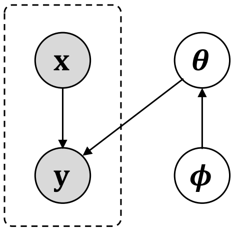

Figure 1. Graphical models for meta-training algorithm.

By Jensen's inequality, we have:

$$\begin{aligned}\log P(\mathcal{D}_S|\boldsymbol{\phi}) &= \log \int q(\boldsymbol{\theta}|\mathcal{D}_S) \frac{P(\mathcal{D}_S|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\boldsymbol{\phi})}{q(\boldsymbol{\theta}|\mathcal{D}_S)} d\boldsymbol{\theta} \\ &\geq \int q(\boldsymbol{\theta}|\mathcal{D}_S) \log \frac{P(\mathcal{D}_S|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\boldsymbol{\phi})}{q(\boldsymbol{\theta}|\mathcal{D}_S)} d\boldsymbol{\theta} \\ &\triangleq \mathcal{E}(q, \boldsymbol{\phi}),\end{aligned}$$

where $\mathcal{E}(q, \boldsymbol{\phi})$ is the evidence lower-bound (ELBO) of the likelihood $\log P(\mathcal{D}_S|\boldsymbol{\phi})$. Then, we further have:

$$\begin{aligned}\mathcal{E}(q, \boldsymbol{\phi}) &= \int q(\boldsymbol{\theta}|\mathcal{D}_S) \log \frac{P(\mathcal{D}_S|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\boldsymbol{\phi})}{q(\boldsymbol{\theta}|\mathcal{D}_S)} d\boldsymbol{\theta} \\ &= \int q(\boldsymbol{\theta}|\mathcal{D}_S) \log P(\mathcal{D}_S|\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &\qquad + \int q(\boldsymbol{\theta}|\mathcal{D}_S) \log \frac{P(\boldsymbol{\theta}|\boldsymbol{\phi})}{q(\boldsymbol{\theta}|\mathcal{D}_S)} d\boldsymbol{\theta} \\ &= \int q(\boldsymbol{\theta}|\mathcal{D}_S) \log P(\mathcal{D}_S|\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &\qquad - \int q(\boldsymbol{\theta}|\mathcal{D}_S) \log \frac{q(\boldsymbol{\theta}|\mathcal{D}_S)}{P(\boldsymbol{\theta}|\boldsymbol{\phi})} d\boldsymbol{\theta} \\ &= \mathbb{E}_{\boldsymbol{\theta} \sim q(\boldsymbol{\theta}|\mathcal{D}_S)} \left[\log P(\mathcal{D}_S|\boldsymbol{\theta})\right] \\ &\qquad - \mathrm{KL}\left[q(\boldsymbol{\theta}|\mathcal{D}_S) \| P(\boldsymbol{\theta}|\boldsymbol{\phi})\right].\end{aligned}$$

We have thus proven Proposition 1. □

2. Efficient Implementation of Algorithm 1

In order to implement Algorithm 1 of the main paper, we need to compute the second order derivative of the network parameters when a set of $\boldsymbol{\phi}$ are updated by gradient descent. This is computational expensive especially when the scale of the backbone network becomes very large. In this section, we provide an efficient implementation of Algorithm 1 when the derivative w.r.t. $\boldsymbol{\phi}$ of the regularizer $\mathcal{R}$ can be calculated directly.

As described in the main paper, we implement $\mathcal{R}$ by a weighted ℓ^2 regularizer in this work. Therefore, the regularized objective function of Eq. (9) in the main paper can be rewritten by:

$$\mathcal{F}(\mathbf{x}_i, \mathbf{y}_i; \boldsymbol{\theta}, \boldsymbol{\phi}) = \mathcal{L}_{\text{REG}}(\mathbf{x}_i, \mathbf{y}_i; \boldsymbol{\theta}) + \sum_i \phi_i \|\theta_i\|^2, \quad (1)$$

where ϕ_i is the i-th weight of the regularizer and θ_i is the i-th parameter of the student network. Then, the k-th gradient descent step of the network parameter θ_i^k is:

$$\begin{aligned} \theta_i^{k+1} &= \theta_i^k - \alpha \frac{\partial \mathcal{F}}{\partial \theta_i^k} = \theta_i^k - \alpha \frac{\partial \left(\mathcal{L}_{\text{REG}} + \sum_i \phi_i \|\theta_i^k\|^2 \right)}{\partial \theta_i^k} \\ &= \theta_i^k - \alpha \frac{\partial \mathcal{L}_{\text{REG}}}{\partial \theta_i} - 2\alpha\phi_i\theta_i^k \\ &= \theta_i^k(1 - 2\alpha\phi_i) - \alpha \frac{\partial \mathcal{L}_{\text{REG}}}{\partial \theta_i^k}, \end{aligned} \quad (2)$$

where α is the learning rate of θ_i. We can see that Eq. (2) converts our regularizer formulation into the weight decay mechanism, where $2\phi_i$ turns into the decay rate. Since the second term of Eq. (2) is independent with ϕ_i, we only need to compute the first order derivative when updating ϕ_i of the regularizer $\mathcal{R}$. The modified meta-training approach is illustrated in Algorithm 1.

Algorithm 1 Efficient implementation of meta-training.

```
Input: Batch size $N$, # of iterations $K$, learning rate $\alpha$.
Input: # of inner iterations $l$, meta learning rate $\beta$.
  Initialize $\boldsymbol{\theta}_0$, $\boldsymbol{\phi}_0$
  for $k = 0$ to $K - 1$ do
      Sample $N$ examples $\{(\mathbf{x}_n^S, \tilde{\mathbf{x}}_n^S, \mathbf{y}_n^S) \sim \mathcal{D}_S\}_{n=1}^N$
      $\ddot{\boldsymbol{\theta}}_0 \leftarrow \boldsymbol{\theta}_k$
      for $i = 0$ to $l - 1$ do
          $\ddot{\boldsymbol{\theta}}_{i+1} \leftarrow \ddot{\boldsymbol{\theta}}_i(1 - 2\alpha\boldsymbol{\phi}_k) - \alpha\nabla_{\ddot{\boldsymbol{\theta}}_i}\mathcal{L}_{\text{REG}}(\mathbf{x}_n^S, \mathbf{y}_n^S; \ddot{\boldsymbol{\theta}}_i)$
      end for
      $\ddot{\boldsymbol{\theta}}_k \leftarrow \ddot{\boldsymbol{\theta}}_l$
      $\boldsymbol{\phi}_{k+1} \leftarrow \boldsymbol{\phi}_k - \beta\nabla_{\boldsymbol{\phi}_k}\mathcal{G}(\mathbf{x}_n^S, \tilde{\mathbf{x}}_n^S, \mathbf{y}_n^S; \ddot{\boldsymbol{\theta}}_k)$
      $\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k - \alpha\nabla_{\boldsymbol{\theta}_k}\mathcal{G}(\mathbf{x}_n^S, \tilde{\mathbf{x}}_n^S, \mathbf{y}_n^S; \boldsymbol{\theta}_k)$
  end for
  $\boldsymbol{\phi}_{\text{META}} \leftarrow \boldsymbol{\phi}_K$
```

3. Experimental Setup on FreiHAND

FreiHAND [5] is a 3D hand pose dataset which records different hand actions performed by 32 people. For each hand image, MANO-based 3D hand pose annotations are provided. It currently contains 32,560 unique training samples and 3960 unique samples for evaluation. The training samples are recorded with a green screen background allowing for background removal. In addition, it applies three different post processing strategies to training samples for data augmentation. However, these post processing strategies are not applied to evaluation samples.

In Sect. 5.5 of the main paper, we conduct the experiment to evaluate the performance of the learned regularizer when it is applied to different target datasets (domains). In this experiment, we treat the original images collected with the green screen background ($\mathcal{G}$) in FreiHAND, together with their post-processed results using three different strategies: harmonization [1] ($\mathcal{H}$), colorization auto [3] ($\mathcal{A}$), colorization sample [3] ($\mathcal{S}$), as three different domains contained by FreiHAND. However, since the domains of $\mathcal{H}$, $\mathcal{A}$ and $\mathcal{S}$ are not provided for the original evaluation samples, we create new training and evaluation splits from the original training data of FreiHAND. Therefore, for each domain, the first 30,000 training samples are used for network training while the rest 2,560 samples are leveraged for evaluation. We use the same setting as described in Sect. 5.1 of the main paper to train the network in this dataset.

4. Additional Visual Results

In Figs. 2 to 4, we show additional visual results predicted by our method on RHD [4], STB [2] and the synthetic dataset. We can see that our method is able to accurately estimate 3D hand poses across different datasets.

References

[1] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *CVPR*, pages 3789–3797, 2017.

[2] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. A hand pose tracking benchmark from stereo matching. In *ICIP*, pages 982–986, 2017.

[3] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S. Lin, Tianhe Yu, and Alexei A. Efros. Real-time user-guided image colorization with learned deep priors. In *SIGGRAPH*, 2017.

[4] Christian Zimmermann and Thomas Brox. Learning to estimate 3D hand pose from single RGB images. In *ICCV*, pages 4903–4911, 2017.

[5] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russel, Max Argus, and Thomas Brox. FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In *ICCV*, pages 813–822, 2019.

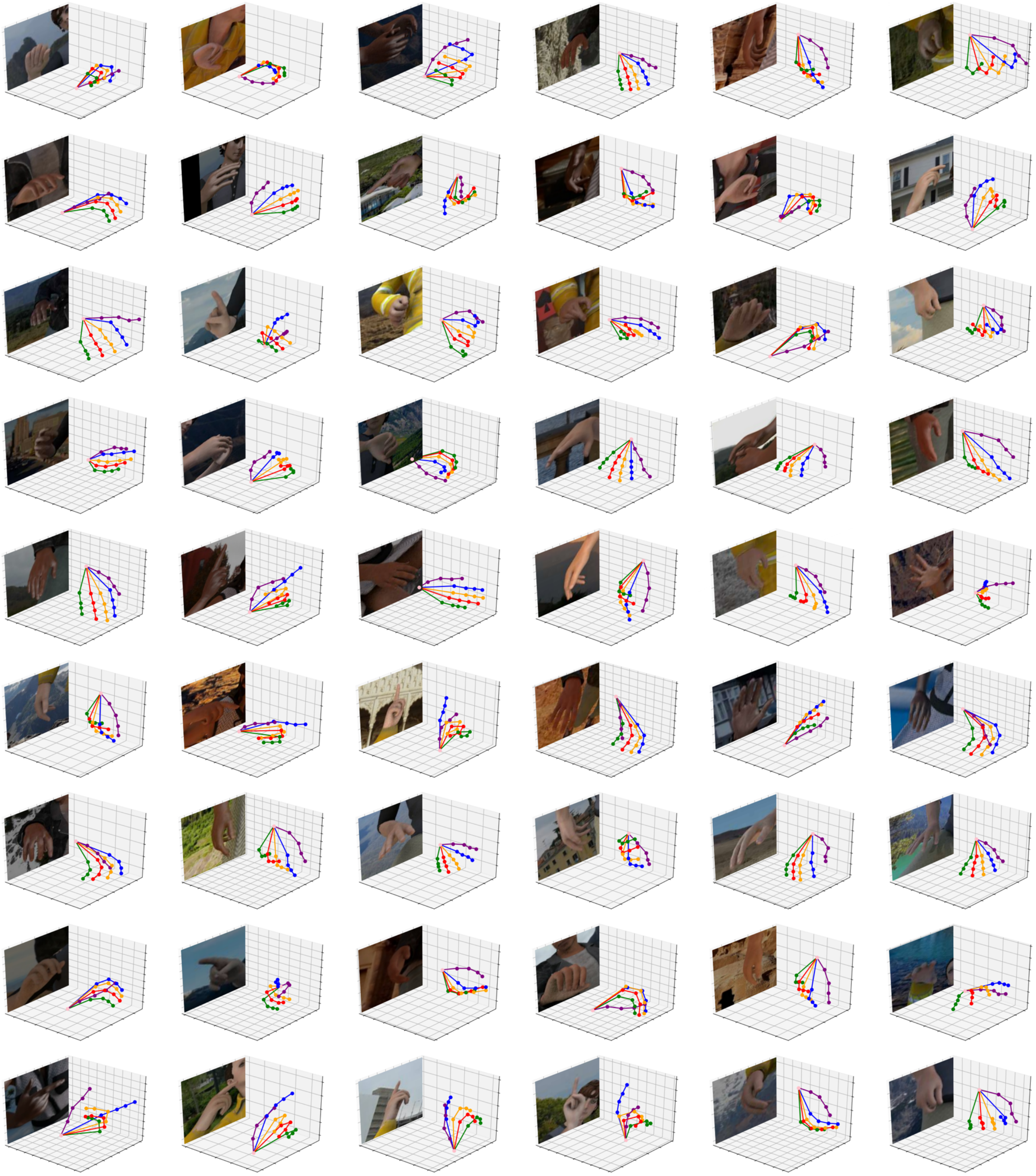

Figure 2. Additional visual results of our approach on RHD [4] dataset.

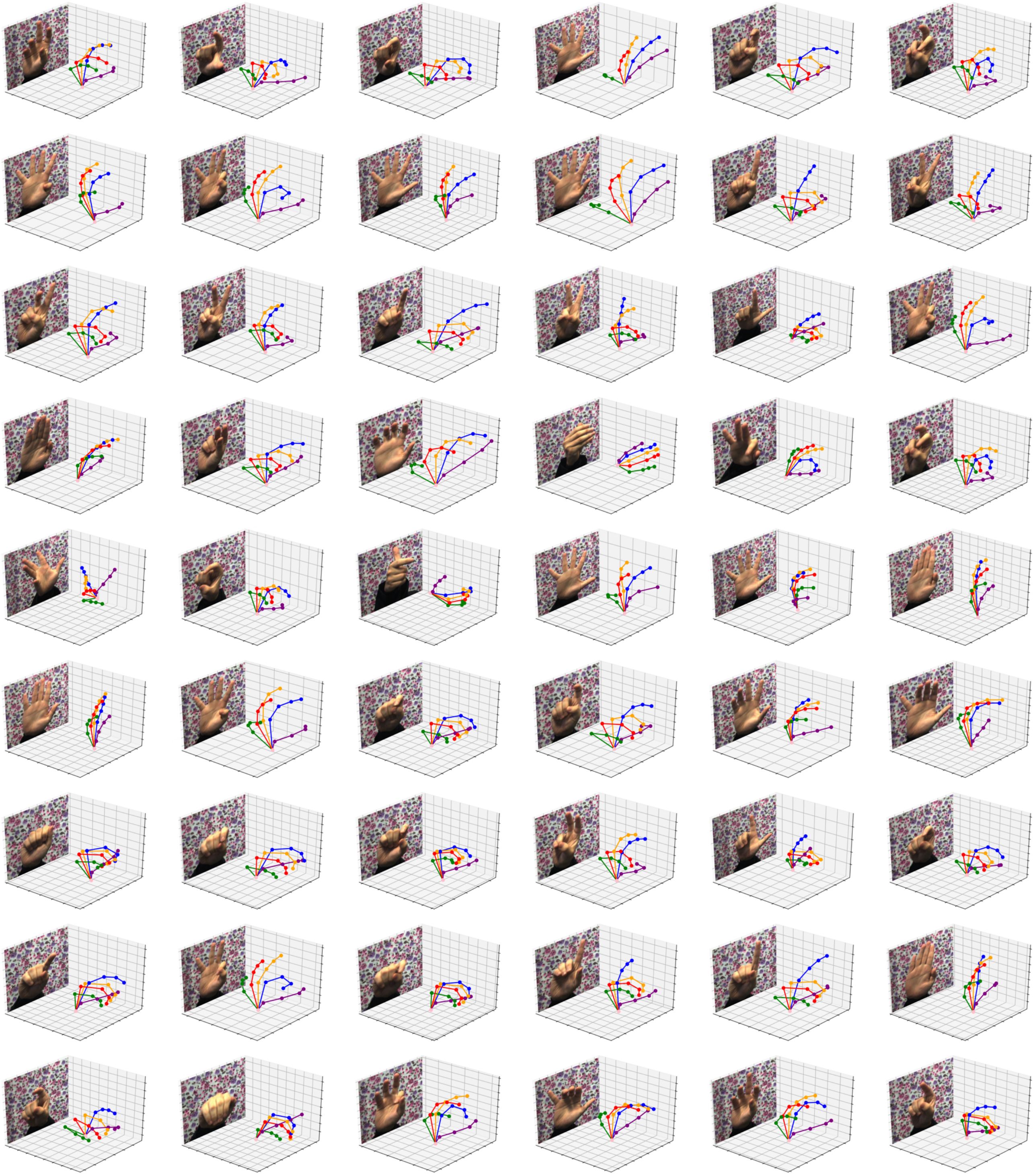

Figure 3. Additional visual results of our approach on STB [2] dataset.

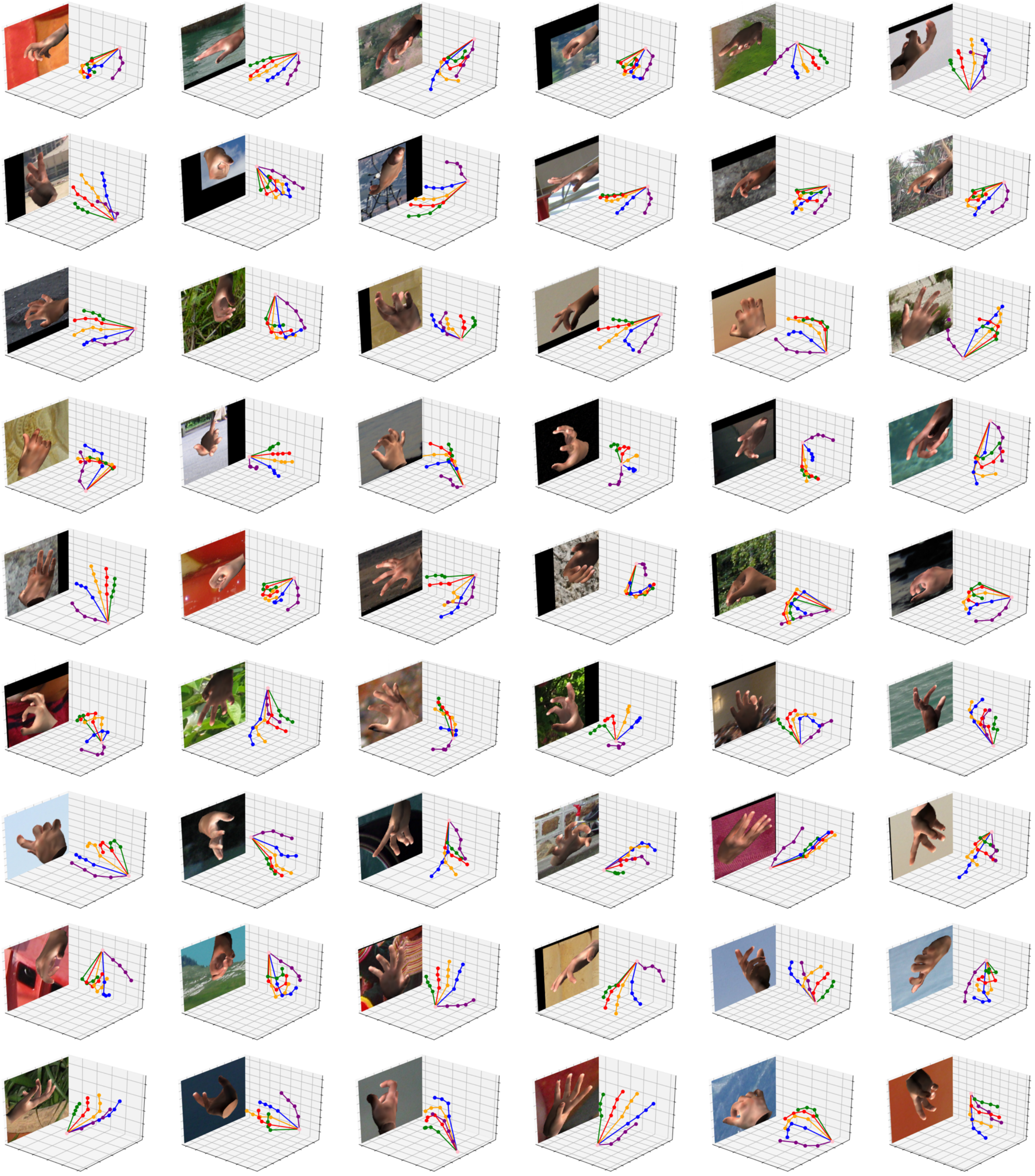

Figure 4. Additional visual results of our approach on synthetic dataset.