HandDGCL: Two-hand 3D reconstruction based Disturbing Graph Contrastive Learning

Bing Han[1,3] Chao Yao[2*] Xiaokun Wang[2,4]
Jian Chang[5] Xiaojuan Ban[1,6*]
[1]Beijing Advanced Innovation Center for Materials Genome Engineering,
Institute of Artificial Intelligence,
University of Science and Technology Beijing, Beijing 100083, China
[2]School of Computer and Communication Engineering,
University of Science and Technology Beijing, Beijing 100083, China
[3]Shunde Innovation School,
University of Science and Technology Beijing, Foshan 528300, China
[4]School of Intelligence Science and Technology,
University of Science and Technology Beijing, Beijing 100083, China
[5]Bournemouth University, Fern Barrow, Poole, Dorset, BH12 5BB, UK
[6]Key Laboratory of Intelligent Bionic Unmanned Systems,
Ministry of Education,
University of Science and Technology Beijing, Beijing 100083, China
yaochao@ustb.edu.cn, banxj@ustb.edu.cn

Abstract

Virtual Reality (VR) and Augmented Reality (AR) applications are becoming increasingly prevalent. However, constructing realistic 3D hands, especially when two hands are interacting, from a single RGB image remains a major challenge due to severe mutual occlusion and the enormous diversity of hand poses. In this paper, we propose a Disturbing Graph Contrastive Learning strategy for two-hand 3D reconstruction. This involves a graph disturbance network designed to generate graph feature pairs to enhance the consistency of the two-hand pose features. A contrastive learning module leverages high-quality generative features for a strong feature expression. We further propose a similarity distinguish method to divide positive and negative features for accelerating the model convergence. Additionally, a multi-term loss is designed to balance the relation among the hand pose, the visual scale and the viewpoint position. Our model has achieved State-of-the-Art results in the InterHand2.6M benchmark. Ablation studies show the model's great ability to correct unreasonable hand movements. In subjective assessments, our Graph Disturbance Learning method significantly improves the construction of realistic 3D hands, especially when two hands are interacting.

Keywords: hand shape reconstruction, graph contrastive learning, hand pose estimation

1 Introduction

With the development of virtual reality and augmented reality (VR/AR), hand poses are widely

*Corresponding Author

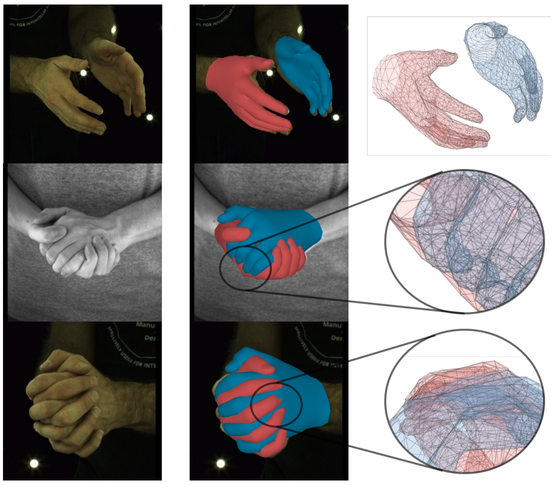

Figure 1: **Interactive & non-interactive hand shape reconstruction from an RGB image.** Geometric deformation occurs in case of the two-hand interaction, which limits visual quality of 3D hand reconstruction.

used for virtual-real interaction in many scenarios. Recently, hand shape reconstruction, which is considered an extension of hand pose estimation, has drawn a lot of attention due to its realistic demonstration. Previous studies [1, 2] have successfully employed 3D hand techniques to visualize hand joints. However, visualizing two interacting hands still presents significant challenges.

Many existing studies have concentrated on single-hand reconstruction from RGB [1, 2, 3, 4], depth [5], or sparse keypoints [6, 7]. When dealing with interacting two-hand reconstruction, these single hand-based methods have poor performance, since it increases the difficulty including inter-hand collisions and mutual occlusions. Recently, some large-scale interacting hand datasets are released to support the two-hand shape reconstruction. Two-Hand-Shape-Pose [8] and IHMR [9] reconstruct two-hands by estimating MANO [10] parameters, which are later mapped to triangular hand meshes using a pre-defined statistical model (i.e.,MANO). IntagHand [11] directly regresses a fixed number of mesh vertex coordinates using a graph convolutional network (GCN). Nevertheless, few effective methods are proposed to resolve the problem of two-hand 3D reconstruction, as the complex occlusion and deformation occur from the interaction of two hands.

In this paper, we present a graph contrastive learning based approach to reconstruct two interacting hands, where an explicit graph disturbing strategy is performed at the feature level to augment the hand feature. The explicit disturbances change the external properties of the hand graph, such as graph dimension, graph edge and some feature noise. To effectively utilize these features, we incorporate a graph contrastive learning for generating strong hand-feature expression. We also propose a similarity distinguish approach to divide positive and negative features for accelerating the model convergence. Finally, a new loss function is designed to enhance 3D visual quality by balancing relation among the hand pose, the visual scale and the viewpoint position. Experimental results prove that our approach achieves a qualitative improvement compared to existing methods on subjective and objective evaluation.

Overall, our contributions are summarized as follows:

- We propose a novel graph disturbance learning to solve the problem of 3D two hands reconstruction by generating graph feature pairs without compromising consistency to hand pose. Explicit as well as implicit graph disturbance is added, which improves the expression of the features.

- We incorporate a scheme of similarity control in order to smooth the learning curve and reduce divergence within the contrastive learning structure.

- We design a multi-term loss function that deals with the pose, the scale and the camera parameter separately so that the model can balance the weight of each task.

2 Related works

According to different downstream task, 3D hand field can be categorized into three: 1) 3D hand pose estimation; 2) 3D hand shape reconstruction, 3) hand-object interactions. There are numerous datasets [12, 13, 14, 15, 16, 17, 18] proposed these years of above three, which greatly boost the machine understanding of human hands.

3D hand pose estimation. Previous works mainly focus on the depth domain [19, 20, 21, 22] and the RGB domain [1, 3, 15, 22, 23, 24, 25]. Compared with 2D hand, 3D hand pose has more complexity and lack one dimension of information for single RGB naturally. Recently, 3D hand pose estimation from a monocular RGB image has achieved great progress [18]. To alleviate the confusion of uniform appearance in hands, the method in [26] uses ResGCN based refinement and conditional adversarial learning scheme to fully exploit hand features. [27] associates keypoints in the heatmap with hand joints using multi-head self-attention to predict complex hand interactions. The work of [28] shows that contrastive training scheme extracts better hand feature representation especially for those vague or easily confused cases.

Graph CNN based hand shape reconstruction. GCN [29] has been proved that it is stable to convey 3D hand shape information with less computational expense. [2] proposed a pipeline to make mesh generation from coarse to precise via GCN in hand shape reconstruction. In [30], RGB images are encoded into embedings of a graph morphable model, which helps the reconstruction of 3D hand from RGB space. These two works mentioned above are the first to construct 3D hand mesh by GCN, and both stand for the great ability of GCN in hand shape reconstruction. After them, GCN is widely used in this aera. [31] designs two transformer-based modules to predict the shape of hands at the occlusion part. [11] adds cross-hand attention modules behind the GCN operation within three blocks to refine the representation of two-hand interactions.

Contrastive learning. Self-supervised methods have been intensively studied in recent years. As one of them, contrastive learning is competitive due to its simple structure as well as high performance. [32] makes it possible to catch up supervised methods in image classification using a simple framework of contrastive learning. The work of [33] shows that pretraining with contrastive learning can also boost the performance of fine-tuning on a few labeled dataset. [34] builds a dynamic dictionary with a queue and a moving-averaged encoder to enable a large and consistent dictionary that facilitates contrastive unsupervised representations learning. While these works rely on hard negative samples, [35] cast contrastive pairs only using positive samples via a momentum encoder trained by the parameter update strategy of moving average, which makes pretraining easier and more friendly to low memory devices. These classical contrastive learning methods above all focus on the encoding of original data to extract better representation while our point is on the decoder. Before this paper, [36] builds hand representation by making expressive positive pairs on multiview images. They take the images of one hand in multiviews as positive samples that share the identical inherent representation, and learn these representations by the similar method of [34]. However, the geometric property of hands is not fully extracted and the contrastive pretraining can not get regular performance when deals with complex hand poses or two-hand interactions. To solve these problems, we introduce graph contrastive learning to the field of 3D hand pose estimation and shape reconstruction in this paper. As for graph contrastive learning, existing methods [37, 38, 39, 40, 41] excavate inherent features on graph in contrastive manners that inspire us utilize it in hand graph domain. We use a simple but powerful framework named SimSiam [42] that combines the advantages of classical contrastive learning, and mixes augmentations in hand graph domain to learn geometry-structured hand representation deeply.

3 Methodology

3.1 Overview

An overall structure of our model is illustrated in Fig.2. The main pipeline at the top of the figure shows the main workflow of the hand shape reconstruction task, which can be summarized as image feature extraction, graph feature conversion, disturbing graph contrastive learning and MANO-hand construction. The disturbing graph contrastive learning network (DGCLNet) module utilizes graph contrastive learning to generate pairs with consistent hand pose information, which is specially designed for graph-level visual contrastive learning task. Note that DGCLNet is only used to train the GCN, which means that DGCLNet are removed except the inside GCN during evaluation. The similarity

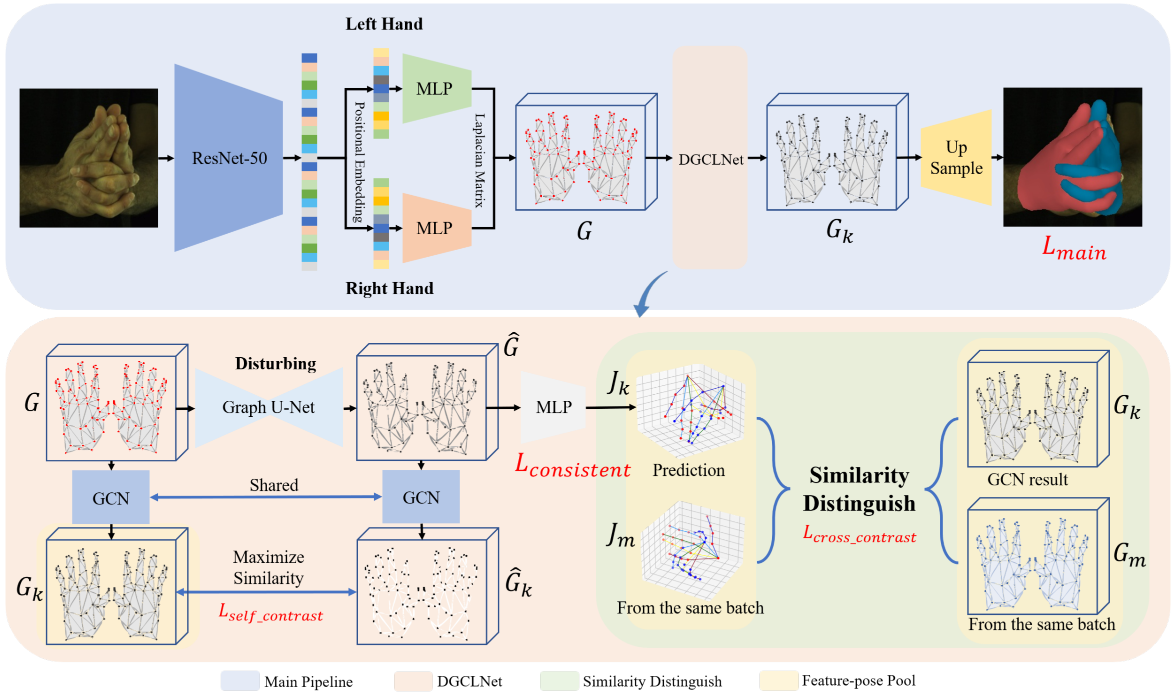

Figure 2: **Architecture overview.** Given an RGB image, our model uses ResNet-50 for feature extraction and then the feature is divided into left hand and right hand via positional embedding. Both of the hand feature are further processed by two independent MLP layers, and build up the graph feature using the knowledge of Laplacian matrix. Next our DGCLNet makes disturbance by regenerating the graph feature through a graph u-net [43]. The pair of graph feature is passed through the same GCN, and the similarity between them is maximized. Finally the graph feature of original branch is thrown into the up sample layer to get the position of 778 vertices in standard MANO [10] way.

distinguish is an extended module of DGCLNet to calculate similarity of a pair precisely from a feature-pose pool, and makes the model converge faster and better. Detailed information of the pipeline will be discussed in this section.

3.2 *Disturbing Graph Contrastive Learning*

Our model tackles the challenging visual task of two-hand shape reconstruction in a simply constructed end-to-end framework. The graph contrastive learning modules are proved to well decode image embedings to hand mesh in the form of graph representation.

Explicit Graph Disturbing (EGD) Strategy. Naturally, we consider using basic image-to-image data augmentation methods to perturb the target image features. In the location of the graph u-net in DGCLNet depicted in Fig.2, we applied the following four types of disturbance operations to the model. (a)Random mask on feature dimension. (b)Random noise on feature space. (c)Add random edges on graph. (d)Flip the edges on graph. After disturbance is added, the shared GCN processes the feature pair and a loss function is used to maximize the similarity of them.

However, there is an essential problem on steps above that is whether the intrinsic essence is changed by disturbance. It is well known that classical contrastive method does not change the intrinsic essence because the disturbance is mainly pixel-level and the effect of the disturbance makes slightly different on feature level. Therefore, the feature vector extracted by encoder maintain its consistency and contrastive method works out. But now we have the graph features distorted which makes it possible that the consistency of different branches is broken.

Disturbing Graph Contrastive Learning Network. To avoid this problem, we design a new contrastive structure to generate pairs of disturbance as Fig.2 described, which corresponds to the $L_{self\_contrast}$. The core of DGCLNet is a generative neural network essentially, which can be noted as F_{gen}. And we use graph u-net [43] in this paper. The input of DGCLNet $\mathbf{G}$ is the original graph features which built by the previous network, and $\hat{\mathbf{G}}$ is the graph features after disturbance. All $\mathbf{G}$ are normalized before use, but for the sake of brevity, this step is omitted in the formula. Simply, we have

$$\hat{G} = F_{gen}(G), \tag{1}$$

The two graph features $\mathbf{G}$, $\hat{\mathbf{G}}$ are considered as a contrastive pair just like other contrastive learning method. F_{gen} replaces those 4 kinds of simple graph-level operations above and the other setting remains the same. Therefore, the model feeds the contrastive pairs to the main GCN decoder and get graph features with lower feature dimension noted as $\mathbf{G_k}$, $\hat{\mathbf{G}}_\mathbf{k}$.

$$G_k = GCN(G), \hat{G}_k = GCN(\hat{G}), \tag{2}$$

$\mathbf{G_k}$, $\hat{\mathbf{G}}_\mathbf{k}$ are actually in a pool of features and poses, which accounts for their subscript. The similarity can be predicted as

$$S = G_k \cdot \hat{G}_k^\top, \tag{3}$$

The $\mathbf{S}$ here describes the similarity between one graph feature and its generative copy. $\mathbf{G_k}$ is passed to next part of the model while all other features in DGCLNet are abandoned.

But how to train DGCLNet? DGCLNet is also trained in end-to-end manner. In order to preserve pose information during training, we innovatively add a linear MLP after $\hat{\mathbf{G}}$ to predict hand pose labels which is taken as a subtask of hand reconstruction.

$$J_k = MLP(\hat{G}). \tag{4}$$

It is because of the subtask that $\mathbf{G}$, $\hat{\mathbf{G}}$ are forced to maintain sufficient supervised information while $\mathbf{G}$ cannot be identical with $\hat{\mathbf{G}}$ due to the bottleneck structure of DGCLNet. Then we get a disturbance reproduction without losing the intrinsic essence of the hand pose naturally, which ensures the consistency that the contrastive learning scheme needs. Besides, the feature output by the graph u-net will not be identical with the input because of the bottleneck structure. Therefore this operation can be considered as disturbing. Note that the DGCLNet is only used during training, our goal is to get a great GCN decoder. The simplicity of the model during inference is also one of our advantages.

3.3 Similarity Distinguish

Based on the pre-estimate relative hand joints location, we intend to extend to the contrastive space of positive and negative samples, which can improve the cross-pose contrastive learning performance. According to the general contrastive learning paradigm, the ground truth of similarity between a positive sample pair is set to 1, while for a negative sample pair it is set to 0. To express it in a neat form, assuming similarity between a random pair from a feature-pose pool is noted by $\mathbf{S_p}$. The pool is defined as a dictionary of extracted features and their corresponding 3D poses as [34] did.

$$S_p[m, n] = G_m \cdot G_n^\top, \tag{5}$$

And we need it equals to 1 when the two of the pair are positive, equals to 0 when negative, which means the quantity of the similarity is discrete. However, we can make it continuous by defining a ground truth of similarity, which is a real number between 0 and 1. The pre-estimate relative hand joints location, noted $\mathbf{J}$, can be used to build the ground truth of similarity. Suppose we have J_m and J_n, the

$$S_g[m, n] = J_m \cdot J_n^\top. \tag{6}$$

where $\mathbf{S_g}$ is the ground truth similarity of $\mathbf{S_p}$.

3.4 Loss Functions

The loss functions we implement in this paper can be divided into three categories, main training objective, auxiliary training objective and contrastive training objective. The λs noted in the following formulas are hyper-parameters to balance the losses.

$$L = L_{main} + L_{auxiliary} + L_{DGCLNet}, \tag{7}$$

Main training objective. The main training objective contains three parts, 3D hand joints loss, 3D hand mesh vertices loss and 2D hand joints loss. These three losses make sure the model performs a basical level of hand estimation. This part of the loss is calculated by minimizing the distance between the results of the main pipeline and the ground truth label.

$$L_{main} = \lambda_1 L_{3djoints} + \lambda_2 L_{3dverts} + \lambda_3 L_{2djoints}, \tag{8}$$

Note that the location of 3D hand joints and 3D hand mesh vertices is aligned with a root joint and is scaled by a length parameter. Therefore the two loss only describe the pose information of hand. The absolute position information and the scale information are learned by 2D hand joints loss. Besides, the camera parameter and rotation matrix are also learned by 2D hand joints loss, which can make projections to the 2D plane of images for the use of demonstration. We use L1 loss for all the three losses of the main training objective. In detail, it is L1 distance between estimated results and ground truth labels.

Auxiliary training objective. The auxiliary training objective contains three parts, length loss, consistent loss and transition loss. These three parts improve the model performance in some aspects.

$$\begin{split} L_{auxiliary} = \lambda_4 L_{length} + \lambda_5 L_{consistent} + \\ \lambda_6 L_{transition}, \end{split} \tag{9}$$

$$L_{consistent} = \|J_k - \bar{J}_k\|_1, \tag{10}$$

The length loss here refers to the distance between joints. It can not only learn the scale of the hands but also can regularize the geometric stucture of the hands. The consistent loss means the consistency between the graph features and the hand pose that the feature represents. With the help of the consistent loss, the model will be more stable and interpretable. The difference between the consistent loss and the 3D joints loss mentioned above is whether the GCN module is used. Consistent loss is calculated linearly from the feature generated by graph u-net while 3D joints loss is from MANO upsample layer. The former only contains pose information while the latter decodes pose information to precise 3D location. It can be calculated by the L1 distance between joint locations J_k and ground truth labels $\bar{J}_k$. And the transition loss is the relative translation of each pair of vertices from left and right hand, which is a widely used restriction for two hand reconstruction mission.

Contrastive training objective.

$$L_{DGCLNet} = -\lambda_7 (L_{cross\_contrast} + L_{self\_contrast}), \tag{11}$$

$$L_{cross\_contrast} = \|S_p - S_g\|_2, \tag{12}$$

$$L_{self\_contrast} = log \frac{exp(S/\tau)}{\sum exp(S/\tau)}. \tag{13}$$

Note that $L_{cross\_contrast}$ and $L_{self\_contrast}$ describe cross-pose contrastive loss and self-pose contrastive loss. The cross-pose loss is evaluated by the similarity distinguish module, which contains $\mathbf{S_g}$ and $\mathbf{S_p}$. The self-pose contrastive loss is calculated by maximizing the similarity of input and output of the graph u-net. In order to utilize large enough batch size in our implementation, we adopt MoCo-like[34] dictionary settings to save the graph features of recent batches.

4 Experiments

4.1 Experimental Settings

Implementation Details. Our network is implemented using Pytorch. And the main model structure as well as the training pipeline is based on Pytorch lightning. All experiments are conducted on 4 NVIDIA RTX 2080ti GPUs.Training minibatch size is set as 32 and Adam is used for model optimization. The learning rate is initialized with 1×10^{-4}, and decays by half when the training loss decreases less than 10^{-4} continuously in 4 epoch. The minimum of learning rate is set by 10^{-6}. The whole training takes 50 epochs, which takes about 2 days. We take input image of 256×256. The encoder is ResNet50 pretrained on ImageNet. And the GCN decoder is Chebyshev spectral graph CNN that is initialized randomly. $\lambda_1 = \lambda_2 = 10$ for 3d joints loss and 3d vertices loss. $\lambda_3 = \lambda_4 = 10^{-3}$ for 2d joints loss and length loss. $\lambda_5 = \lambda_6 = 1$ for consistent loss and transition loss. And $\lambda_7 = 10^{-2}$ for contrastive loss. The MoCo-like dictionary is implemented by a queue of 2048 and the model updates momentumly.

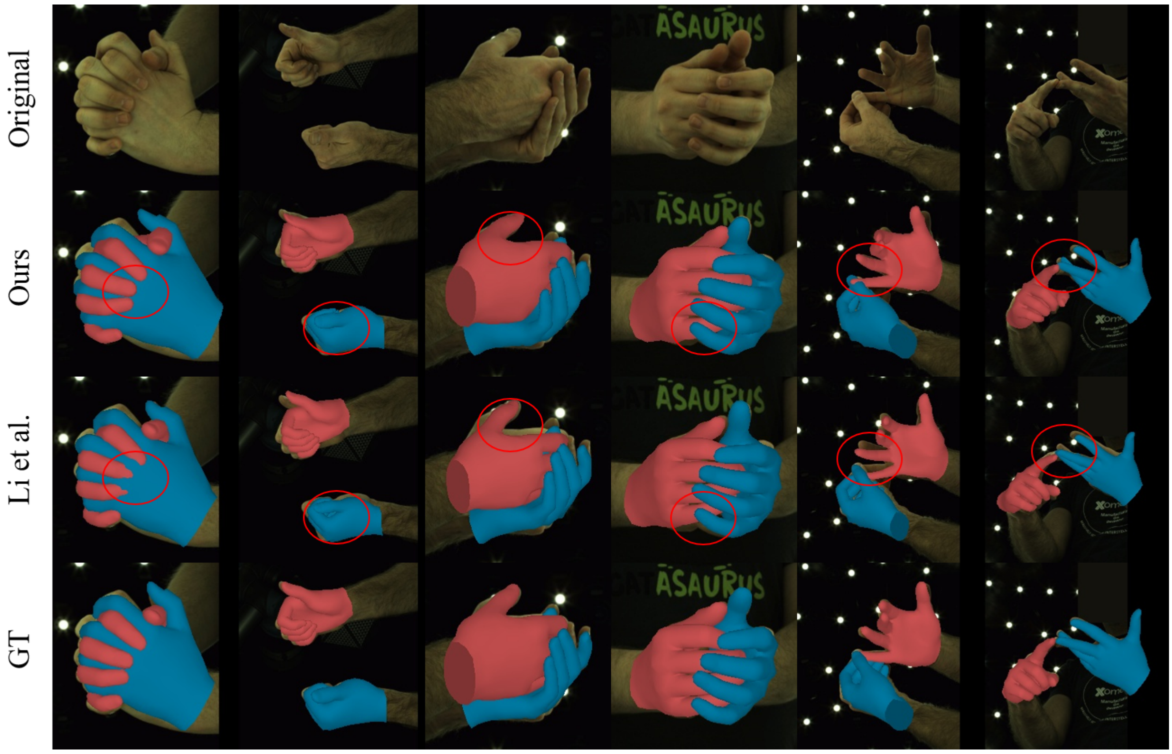

Figure 3: Qualitative results of our method on InterHand2.6M test dataset and comparison with SOTA method proposed by Li et al.[11].

Evaluation Metrics. We used the Mean Per Joint Position Error (MPJPE) to evaluate the hand pose estimation performance. MPJPE is defined as the mean Euclidean distance between ground truth and predicted 3D joint positions. Similarly the Mean Per Vertex Position Error (MPVPE) is used to evaluatethe hand shape reconstruction performance. Note that MPJPE and MPVPE in this paper are reported after aligning the root joint and scaling the length of the middle metacarpal for fair comparison. All the measures mentioned above are calculate in millimeters.

4.2 *Dataset*

Our model is trained and tested on InterHand2.6M [18] following the dataset settings of [11]. 366K training samples and 261K testing samples are picked out from the original InterHand2.6M dataset. And these samples are all the interacting two-hand data with annotation by both human and machine. The image samples are cropped and resize to 256 × 256 resolution.

Methods	MPJPE↓	MPVPE↓
Boukhayma et al.[1]	16.63	17.98
Zhang et al.[8]	13.48	13.95
Rong et al.[9]	13.56	-
Kim et al.[44]	12.08	-
Meng et al.[45]	10.97	-
Li et al.[11]	**8.79**	**9.03**
Ours	**8.86**	**9.14**

Table 1: Comparison with state-of-the-art methods on InterHand2.6M. The MPJPE and MPVPE are reported in mm.

4.3 *Qualitative Results*

Fig.3 shows our qualitative results for hand shape reconstruction. We compare the performance of our model to state-of-the-art reconstruction methods, which can be integrated with virtual reality (VR) or augmented reality (AR) applications. It is clear that our model offers a much better result in the details of processing double hands, especially the interactions among them, which makes the hand mesh closer to the natural state. Our model demonstrates the ability to successfully capture and reconstruct

	MPJPE↓	MPVPE↓
GCN baseline	9.97	10.63
GCN + random mask	10.35	10.65
GCN + random noise	10.56	10.89
GCN + random edge	10.59	10.93
GCN + flip edge	10.92	11.30
GCN + DGCLNet	**9.02**	**9.31**
GCN + DGCLNet*(**Ours**)	**8.86**	**9.14**

Table 2: Ablation study of module choice on InterHand2.6M. * represents similarity distinguish strategy is used.

complex interactions of two hands. The first column of our qualitative results indicates that our model has fewer collisions. The results of Columns 2 and 4 indicate that the performance of our model is superior when the two hand meshes are close to each other, yet not making contact. Columns 3 and 5 indicate that our model results in a better matching between the hand mesh and the hand mask, which implies enhanced location accuracy. The last column of our results highlights that our model has a deep understanding of hand occlusion, and is able to reconstruct detailed hand poses competently.

4.4 *Ablation study*

Baseline GCN. Before implementing our contrastive method, we run a GCN decoder as a baseline for following comparison. The result is shown in Tab. 2. The GCN is built on the base of [2] and [11]. Moreover, the graph laplacian matrix obtained from MANO-pretraining was fixed during training. We observe that the baseline is strong, but may struggle with fingertip alignment and encountering obstacles.

Adding contrastive module. We tested the contrastive module with both EGD and DGCLNet based on GCN baseline. The result is shown in Tab.2. As discussed in Section 3.4, explicit graph disturbing methods may lead to a loss of consistency for graph features which accounts for their poor performance. We see that DGCLNet module reduces the error by 1.6 mm. Thanks to the disturbance network, the model is able to learn from slightly different graph features without compromising the consistency of the pose and feature. The qualitative results are given in Fig.4. It is clear that our DGCLNet module improves the baseline by reducing collisions, resulting in a hand mesh that is more realistic and lifelike. Besides, the overlap of hand mask is more precise, which indicates the higher performance in visual comprehension.

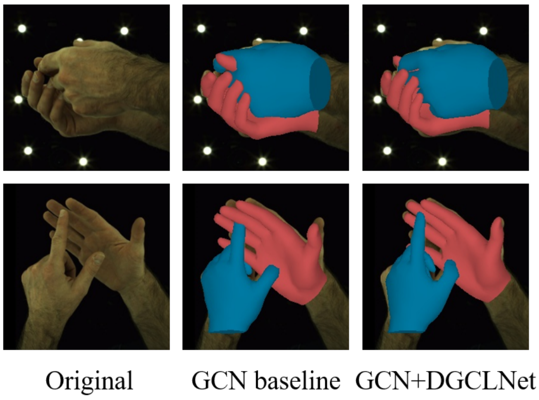

Figure 4: Qualitative ablation study on InterHand2.6M.

Similarity distinguish. Our similarity distinguish method makes the target of contrastive pairs more reliable and yields better robustness while testing. Contrastive learning can be notoriously difficult to converge and can easily become trapped in severe oscillations. We manage to solve this problem by making the similarity of the pose of embedded feature the target similarity of embedded graph feature. The ablation result is shown in Tab. 2.

5 Conclusion

In conclusion, this paper introduces a novel graph contrastive method for two-hands 3D reconstruction from a single image. To deal with challenges such as hand-hand occlusion and the homogeneous and self-similar appearance of hands, we introduce novel graph contrastive method to enhance model ability to rebulid 3D topographical structure. More specifically, an explicit graph disturbing is tried but does not perform well due to its inconsistency. To resolve that issue, a DGCLNet was proposed to force the features of disturbed data to remain consistent. A similarity distinguish strategy was further used instead of the discrete settings which made the model more robust and yielded better results. Moreover, a multi-term loss is designed to balance the relation among the hand pose, the visual scale and the viewpoint position. Experiments validate the superiority of our proposed method over the state-of-the-art methods.

6 Acknowledgment

This paper is sponsored by National Key R&D Program of China (2022ZD0118001), National Natural Science Foundation of China under Grant U22A2022 and 61972028, Guangdong Basic and Applied Basic Research Foundation (2023A1515030177), and Scientific and Technological Innovation Foundation of Shunde Graduate School, USTB under Grant BK21BF002.

References

[1] Adnane Boukhayma, Rodrigo de Bem, and Philip H.S. Torr. 3d hand shape and pose from images in the wild. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10835–10844, 2019. 2, 3, 7

[2] 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10825–10834, 2019. 2, 3, 8

[3] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1067–1076, 2019. 2, 3

[4] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Weakly-supervised domain adaptation via gan and mesh model for estimating 3d hand poses interacting objects. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6120–6130, 2020. 2

[5] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Dual grid net: Hand mesh vertex regression from single depth maps. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, page 442–459, Berlin, Heidelberg, 2020. Springer-Verlag. 2

[6] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multimodal data. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5345–5354, 2020. 2

[7] Korrawe Karunratanakul, Adrian Spurr, Zicong Fan, Otmar Hilliges, and Siyu Tang. A skeleton-driven neural occupancy representation for articulated hands. In *2021 International Conference on 3D Vision (3DV)*, pages 11–21, 2021. 2

[8] Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. Interacting two-hand 3d pose and shape reconstruction from single color image. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11334–11343, 2021. 2, 7

[9] Yu Rong, Jingbo Wang, Ziwei Liu, and Chen Change Loy. Monocular 3d reconstruction of interacting hands via collision-aware factorized refinements. In *2021 International Conference on 3D Vision (3DV)*, pages 432–441, 2021. 2, 7

[10] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 36(6), nov 2017. 2, 4

[11] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, jun 2022. 2, 3, 7, 8

[12] Srinath Sridhar, Franziska Mueller, Michael Zollhöfer, Dan Casas, Antti

Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 294–310, Cham, 2016. Springer International Publishing. 2

[13] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. A hand pose tracking benchmark from stereo matching. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 982–986, 2017. 2

[14] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1163–1172, 2017. 2

[15] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4913–4921, 2017. 2, 3

[16] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):190–204, 2019. 2

[17] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max J. Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 813–822, 2019. 2

[18] 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 *European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 7

[19] Ayan Sinha, Chiho Choi, and Karthik Ramani. Deephand: Robust hand pose estimation by completing a matrix imputed with deep features. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4150–4158, 2016. 3

[20] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. Robust 3d hand pose estimation in single depth images: From single-view cnn to multi-view cnns. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3593–3601, 2016. 3

[21] Jameel Malik, Ahmed Elhayek, Fabrizio Nunnari, Kiran Varanasi, Kiarash Tamaddon, Alexis Heloir, and Didier Stricker. Deephps: End-to-end estimation of 3d hand pose and shape by learning from synthetic depth. In *2018 International Conference on 3D Vision (3DV)*, pages 110–119, 2018. 3

[22] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 89–98, 2018. 3

[23] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4645–4653, 2017. 3

[24] Umar Iqbal, Pavlo Molchanov, Thomas Breuel, Juergen Gall, and Jan Kautz. Hand Pose Estimation via Latent 2.5D Heatmap Regression. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 125–143, Cham, 2018. Springer International Publishing. 3

[25] Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A. Otaduy, Dan Casas, and Christian Theobalt. Rgb2hands: Real-time tracking of 3d hand interactions from monocular rgb video. volume 39, New York, NY, USA, nov 2020. Association for Computing Machinery. 3

[26] Yiming He and Wei Hu. 3d hand pose estimation via regularized graph representation learning. In Lu Fang, Yiran Chen, Guangtao Zhai, Jane Wang, Ruiping Wang, and Weisheng Dong, editors, *Artificial Intelligence*, pages 540–552, Cham, 2021. Springer International Publishing. 3

[27] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *IEEE Computer Vision and Pattern Recognition Conference*, 2022. 3

[28] Adrian Spurr, Aneesh Dahiya, Xi Wang, Xucong Zhang, and Otmar Hilliges. Self-supervised 3d hand pose estimation from monocular rgb via contrastive learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11210–11219, 2021. 3

[29] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3844–3852, Red Hook, NY, USA, 2016. Curran Associates Inc. 3

[30] Dominik Kulon, Haoyang Wang, Riza Alp Güler, Michael M. Bronstein, and Stefanos Zafeiriou. Single image 3d hand reconstruction with mesh convolutions. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2019. 3

[31] 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)*, pages 1496–1505, June 2022. 3

[32] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020. 3

[33] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22243–22255. Curran Associates, Inc., 2020. 3

[34] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. 3, 5, 6

[35] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. 3

[36] Christian Zimmermann, Max Argus, and Thomas Brox. Contrastive representation learning for hand shape estimation. In Christian Bauckhage, Juergen Gall, and Alexander Schwing, editors, *Pattern Recognition*, pages 250–264, Cham, 2021. Springer International Publishing. 3

[37] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5812–5823. Curran Associates, Inc., 2020. 3

[38] Guanyi Chu, Xiao Wang, Chuan Shi, and Xunqiang Jiang. Cuco: Graph representation with curriculum contrastive learning. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2300–2306. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track. 3

[39] Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, Mehdi Azabou, Eva L Dyer, Remi Munos, Petar Veličković, and Michal Valko. Large-scale representation learning on graphs via bootstrapping. In *International Conference on Learning Representations*, 2022. 3

[40] Namkyeong Lee, Junseok Lee, and Chanyoung Park. Augmentation-free self-supervised learning on graphs. 36(7):7372–7380, 2022. 3

[41] Zekun Tong, Yuxuan Liang, Henghui Ding, Yongxing Dai, Xinke Li, and Changhu Wang. Directed graph contrastive learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 19580–19593. Curran Associates, Inc., 2021. 3

[42] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753, 2021. 3

[43] Hongyang Gao and Shuiwang Ji. Graph u-nets. In *International Conference on Machine Learning*, pages 2083–2092, 2019. 4, 5

[44] Dong Uk Kim, Kwang In Kim, and Seungryul Baek. End-to-end detection and pose estimation of two interacting hands. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11169–11178, 2021. 7

[45] Hao Meng, Sheng Jin, Wentao Liu, Chen Qian, Mengxiang Lin, Wanli Ouyang, and Ping Luo. 3d interacting hand pose estimation by hand de-occlusion and removal. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, page 380–397, Berlin, Heidelberg, 2022. Springer-Verlag. 7

7 Author Biography

Bing Han received B.S. degree from University of Science and Technology Beijing (USTB), Beijing, China, in 2021. He is currently pursuing a Master's degree in Computer Science at USTB. His research interests mainly focus on hand pose estimation and hand mesh reconstruction.

Chao Yao received the B.S. degree in computer science from Beijing Jiaotong University (BJTU), Beijing, China, in 2009. He received the Ph.D. degree from the Institute of Information Science at BJTU in 2016. From 2014 to 2015, he served as a Visiting Ph.Dad student at Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland. Since July 2016, he served as a Postdoc in Beijing University of Posts and Telecommunications (BUPT), Beijing, China. His current research interests include image and video processing and computer vision.

Xiaokun Wang is an associate professor in School of Intelligence Science and Technology, University of Science and Technology Beijing, China. He received the Ph.D. degree in Computer Science and Technology from the University of Science and Technology Beijing, in 2017. He is currently working at the National Centre for Computer Animation at Bournemouth University funded by the EU's Horizon 2020 Marie Curie Individual Fellowship. His research interests include computer graphics, virtual reality and human-computer interaction.

Jian Chang is a professor in National Centre Computer Animation at Bournemouth University, United Kindom. He received his Ph.D. degree in Computer Graphics from Bournemouth University, in 2007. His research interests include physics based modelling (deformation & fluid), motion synthesis, virtual reality (surgery simulation), and novel HCI (eye tracking, gesture control and haptic).

Xiaojuan Ban received the Ph.D. degree from the University of Science and Technology Beijing, Beijing, in 2003. She is currently a Ph.D. Supervisor with the University of Science and Technology Beijing. She is the Managing Director of the Chinese Association for Artificial Intelligence (CAAI). She has received the New Century Excellent Talent of the Ministry of Education. Her current research interests are artificial intelligence, natural human–computer interactions, and 3D visualization. Co-corresponding author of this paper