DOI: 10.19884/j.1672-5220.202411014

Multi-Modal Multi-View 3D Hand Pose Estimation

WANG Hao, WANG Ping*, YU Haoran, DING Dong, XIANG Weiming
College of Information Science and Technology, Donghua University, Shanghai 201620, China

Abstract: With the rapid progress of the artificial intelligence (AI) technology and mobile internet, 3D hand pose estimation has become critical to various intelligent application areas, e. g., human-computer interaction. To avoid the low accuracy of single-modal estimation and the high complexity of traditional multi-modal 3D estimation, this paper proposes a novel multi-modal multi-view (MMV) 3D hand pose estimation system, which introduces a registration before translation (RT)-translation before registration (TR) jointed conditional generative adversarial network (cGAN) to train a multi-modal registration network, and then employs the multi-modal feature fusion to achieve high-quality estimation, with low hardware and software costs both in data acquisition and processing. Experimental results demonstrate that the MMV system is effective and feasible in various scenarios. It is promising for the MMV system to be used in broad intelligent application areas.

Keywords: 3D hand pose estimation; registration network; multi-modal; multi-view; conditional generative adversarial network (cGAN)
CLC number: TP391.4 **Document code**: A
Article ID: 1672-5220(2025)06-0673-10

Open Science Identity (OSID)

0 Introduction

With the rapid development of information technologies, such as artificial intelligence (AI), virtual reality (VR) and digital twins, advanced human-computer interaction (HMI) technologies have been successfully applied across various fields. As a primary means for human beings to interact with the physical world, accurate hand pose estimation results have broad applications in intelligent areas like robotic dexterous hands, remote healthcare and automated manufacturing. Consequently, 3D hand pose estimation has drawn significant attention from both domestic and international researchers.

Recently, significant developments have been made in 3D hand pose estimation based on visual data, driven by the proliferation of low-cost and high-quality binocular commercial cameras. To meet the demands of refined scenarios, it is essential to enhance the qualities of 3D hand pose estimation, including multi-view adaptability, on-site performance and high accuracy. However, existing methods based on monocular RGB cameras still face limitations, since they can only provide insufficient depth data from a single modality image[1]. Besides, other 3D object modeling methods rely on deep learning techniques to extract both RGB and depth features from dense point clouds captured by binocular cameras[2-3]. However, the spatial accuracy of these methods is easily influenced by environmental interference, leading to increased complexity and latency during model training, especially in dynamic hand pose estimation. This makes them unsuitable for hand pose estimation applications. Therefore, achieving an effective fusion of RGB and depth data without increasing complexity and cost remains a significant challenge.

To overcome these challenges, this paper proposes a multi-modal multi-view (MMV) 3D hand pose estimation system, hereafter referred to as the MMV system. By introducing a multi-modal fusion approach which can fuse multiple data features to enhance the sensing accuracy, the MMV system achieves highly accurate multi-modal registration between RGB and depth data for each feature point.

There are three main types of 3D hand pose estimation methods: depth-based, RGB image-based, and multi-modal methods.

Firstly, depth-based methods, such as PosNet[4], PointNet[5] and weak supervision-based Pose-REN[6], use the depth information from point cloud as the input to the network, allowing for the direct estimation of spatial coordinates. However, these methods suffer from serious self-occlusion effects in the spatial domain and have low noise immunity.

Secondly, in RGB image-based 3D hand pose estimation, some studies have improved deep learning-based techniques by parameterization[7], introducing generative adversarial networks (GANs)[8], evolving the Transformer's nonlocal self-attention mechanism with adaptive local feature learning[9], or incorporating multi-view self-supervision[10], achieving promising results. Especially, Zhang et al.[11] proposed a real-time and hardware-free hand-tracking solution, MediaPipe, with a two-stage pipeline and 2.5D pose estimation using only RGB input. By the way, some non-parametric methods

Received date: 2024-11-19
* Correspondence should be addressed to WANG Ping, email: pingwang@dhu.edu.cn
Citation: WANG H, WANG P, YU H R, et al. Multi-modal multi-view 3D hand pose estimation[J]. *Journal of Donghua University (English Edition)*, 2025, 42(6): 673-682.

directly estimate 3D coordinates or heatmaps of the hand key points by using a graph convolutional network (GCN)[12], spiral convolutions, and Transformer[13], to achieve high-precision 3D pose estimation.

As discussed above, both single depth-based and RGB image-based learning methods lack sufficient 3D information, leading to an ill-posed problem in the 3D coordinate regression process[4, 14-15]. In order to estimate the real depth, previous work has included the estimation of wrist-relative hand representation[16] with strong assumptions, and reconstruction of the absolute 3D hand pose close to a scaling factor[17]. However, these methods cannot always be effective in resolving ill-posed problems.

Thirdly, since the accurate spatial relationship between the fingers and the environment dominates hand poses, multi-modal methods[18-20] that combine depth data with RGB images are used to enrich the 3D information of the hand, thereby improving the accuracy and robustness of hand pose estimation. Tu et al.[21] integrated tactile information into the RGB-D data, especially to overcome issues such as hand occlusion. However, the use of tactile sensors also adds to the complexity and cost of the method to some extent. In order to fulfill the feature fusion between multi-modal data, multi-modal image registration (MIR) is employed. It transfers original pixels to target pixels of the same object in different images through a deformation field.

Deep learning-based image registration methods are typically categorized into supervised[22-24] and unsupervised[25-27] methods. Supervised registration methods rely on prior information of the landmark points as ground truth. In contrast, unsupervised registration methods use the similarity between reference and deformed images as a loss function, thereby avoiding the need for prior information during network training. These unsupervised registration methods optimize cross-modal similarity metrics by techniques such as spatial disentangling[25] and the introduction of modality translation networks[27], achieving better registration results. Due to the information differences between modalities, the registration process in the unsupervised multi-modal is more complex than the traditional single-modal registration. To simplify the similarity evaluation, Arar et al.[26] introduced an unsupervised multi-modal registration method by the registration before translation (RT)-translation before registration (TR) jointed conditional generative adversarial network (cGAN), and the method converted cross-modal similarity metrics into single-modal similarity metrics. There have been few publications addressing the issue of MIR for 3D multi-modal hand pose estimation up to now.

Motivated by the existing research achievements mentioned above, to realize multi-view accurate 3D hand pose estimation, this paper proposes a multi-modal cGAN-based network, MMV, with MIR, which is crucial for feature aligning across different modalities between both depth and RGB images, to obtain accurate spatial coordinates of the 3D hand poses. The up-down refined registration network trained by a multi-modal cGAN can effectively generate a refined deformation field to extract the depth and RGB features. In addition, the texture loss of the hand pose can be suppressed by optimizing the similarity loss function. Then, the multi-modal feature fusion from multi-view is achieved by the feature distribution consistency. Based on the performance evaluation results in terms of the registration accuracy and training time, as well as the multiple hand pose experiment results on real data, it can be seen that the proposed MMV 3D hand pose estimation method demonstrates advantages such as high feasibility, enhanced accuracy and low cost compared with existing counterparts, and a promising opportunity in applications.

1 Methods

1.1 MMV 3D hand pose estimation network architecture

Figure 1 illustrates the MMV 3D hand pose estimation network architecture. The MMV system consists of three key components: the modality registration network R, the 2.5D hand pose estimation network module (MediaPipe)[11] and the multi-modal feature fusion module. Compared to existing multi-modal methods which require 3D point cloud processing in the multi-modal feature fusion stage and lead to high computational complexity, the proposed MMV system, through the MIR method based on 2D spatial feature point matching, effectively reduces the computational complexity, and then reduces the cost of implementation.

The modality registration network aligns the independent depth modality I_{Depth} with the RGB modality I_{RGB} of input images (pixels by pixels), with the aid of a dedicated modality deformation field prediction network. This network is trained through modality learning in a cGAN framework, producing registered depth modality $I_{\text{Depth(reg)}}$. Meanwhile, MediaPipe[11] generates a single-view hand pose estimation result V_{Hand} at the output. Leveraging the consistency in the hand skeleton point distribution between $I_{\text{Depth(reg)}}$ and I_{RGB}, the multi-modal feature fusion module maps the points from both modalities to the corresponding 3D hand pose $V_{\text{3D-hand}}$ with high accuracy in the spatial domain.

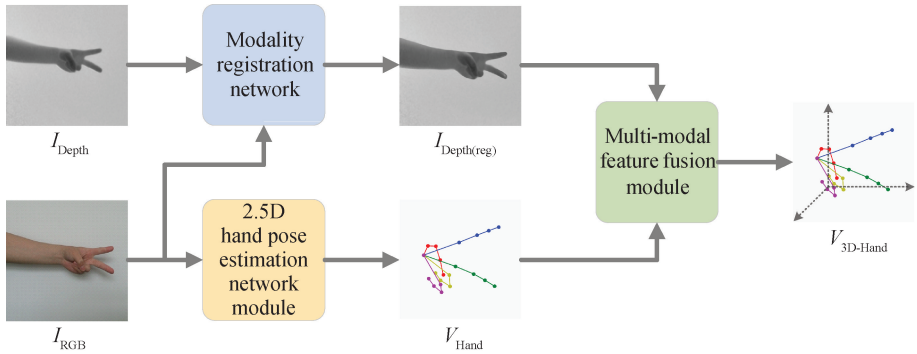

Fig. 1 MMV 3D hand pose estimation network architecture

1.2 Modality registration network

The proposed modality registration network R consists of two key components: the deformation field prediction network R_{Φ} and the spatial transformation function R_{S}. The deformation field prediction network which is trained through cGAN modality learning, predicts the deformation field Φ corresponding to both I_{Depth} and I_{RGB}. Utilizing the deformation vector associated with Φ, R_{S} maps the original input pixels to their target positions, yielding $I_{\text{Depth(reg)}}$ as the output.

1.2.1 Deformation field prediction network based on U-Net with up-down refinement

As shown in Fig. 2(a), R_{Φ} is based on a U-Net architecture, primarily consisting of a series of multi-dimensional feature encoders and decoders, and the detailed structure of each layer in the network is shown in Fig. 2(b). The encoder layers include a convolutional layer followed by ReLU activation functions and a residual (Res) block containing two 3×3 convolutional layers. The Res blocks help mitigate the vanishing gradient problem, while the skip connections between the encoder and decoder layers facilitate feature propagation across different layers. In the decoder, the feature dimensionality mismatch between the down-refined features and the skip-connected features is addressed by using bilinear interpolation.

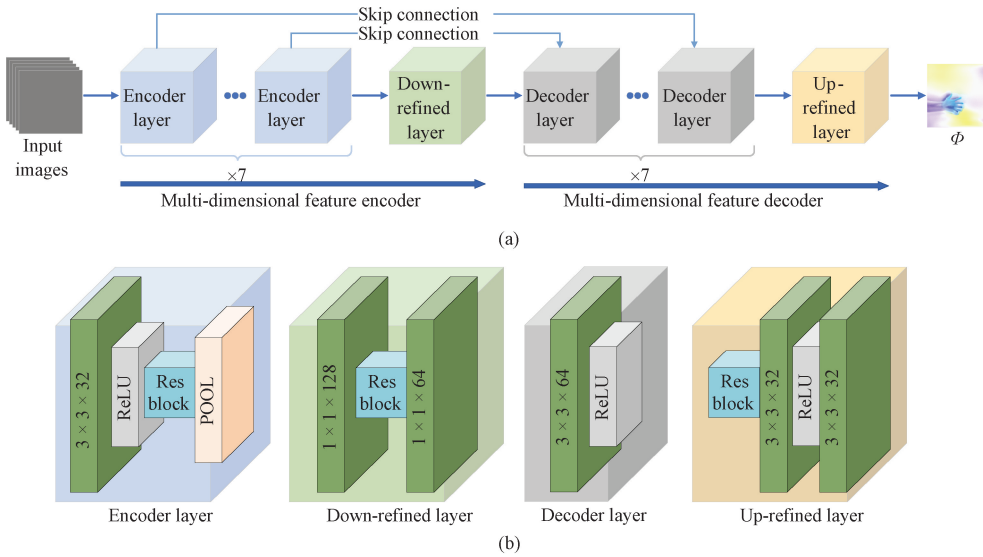

Fig. 2 R_{Φ} based on U-Net with up-down refinement: (a) registration network; (b) details of each module in network architecture

Furthermore, down-refined and up-refined layers are specifically added at the ends of the encoder and decoder, respectively, to enhance the extraction of low-dimensional gray features, thereby enabling more effective refinement of Φ at the output.

Generally, the modality deformation field before optimization Φ_i at the output can be calculated as[26]

$$\Phi_i = \text{Conv}_i(\text{up}(f_{\text{in}}(i), f_{\text{skip}}(i)) + f_{\text{skip}}(i)), \quad (1)$$

where $f_{\text{in}}(i)$ represents the input features of the i_{th} decoder layer; $f_{\text{skip}}(i)$ denotes the encoded features at the i_{th} layer before pooling; up $(f_{\text{in}}(i), f_{\text{skip}}(i))$ signifies the up-sampling of $f_{\text{in}}(i)$ through the bilinear interpolation to match the dimensions of $f_{\text{skip}}(i)$; Conv_i refers to the tensor convolution process applied to the summation of up $(f_{\text{in}}(i), f_{\text{skip}}(i))$ and $f_{\text{in}}(i)$ at each decoder layer.

1.2.2 RT-TR jointed cGAN training scheme

In practice, there may be overlap or occlusion in different hand poses. It is essential for the spatial transformation field to leverage multi-modal features in

order to ensure both the accuracy and quality, thereby preventing any degradation in registration performance. To mitigate the effects of modality discrepancies in the registration network, the RT-TR jointed cGAN training scheme is employed, aiming to produce a high-quality deformation field.

Figure 3 illustrates the RT-TR jointed cGAN training scheme. This scheme consists of three task branches: deformation field prediction based on R_Φ, RT and TR both of which are based on the cGAN principle. In parallel, the distinction between these two branches lies in whether the generator is RT or TR within the overall cGAN structure. The fake images generated by RT or TR are input into the discriminator D, referring to I_{RGB}, to minimize the adversarial loss L_{cGAN} and produce a well-trained and independent R. In each RT or TR, the generator and the discriminator jointly generate fake images, with the discriminator trying to classify the generator's output as false, thus minimizing L_{cGAN}. The generator is optimized by minimizing L_{cGAN} in the game between D and RT, or between D and TR. The modality translation network T transfers the registered results from the depth modality to the RGB modality. Furthermore, the reconstruction loss L_{recon} is incorporated in both training branches to encourage the generator to generate fake images that closely resemble the target image, thereby making R and T maintain task independence. To encourage R_Φ to generate a smooth deformation field, we introduce the smooth loss L_{smooth}, which serves as a regularization term, encouraging adjacent pixels to deform similarly. L_{cGAN}, L_{recon} and L_{smooth} are fed back to their corresponding networks to form a closed loop, minimizing the losses for network training. In detail, L_{cGAN} is output by D and fed back to both T and R_Φ, while L_{recon} and L_{smooth} are both fed back to R_Φ.

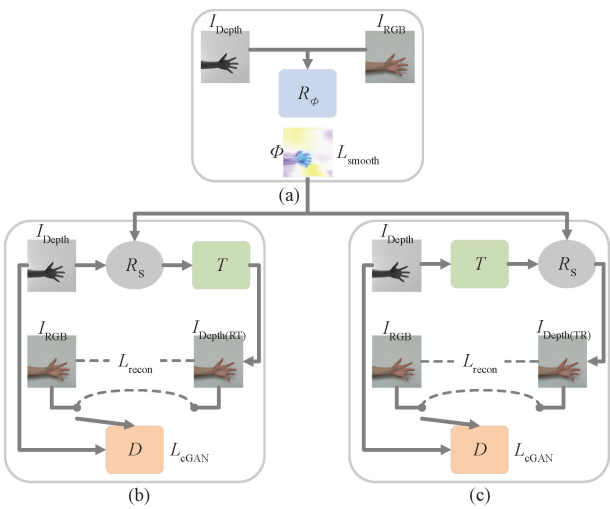

Fig. 3 RT-TR jointed cGAN training scheme: (a) deformation field prediction; (b) RT; (c) TR

In the RT branch, I_{Depth} is first registered into $I_{\mathrm{Depth(reg)}}$ by the differential deformation function based on a sampled point grid and interpolation transformations R_{S} under the control of Φ, and is then translated into the RGB modality $I_{\mathrm{Depth(RT)}}$ image by T. Similarly, in the TR branch, I_{Depth} is first translated to the RGB modality $I_{\mathrm{Depth(fake)}}$ image, and is then registered into the depth modality $I_{\mathrm{Depth(TR)}}$ image.

According to the above principle, $I_{\mathrm{Depth(RT)}}$ can be given as

$$I_{\mathrm{Depth(RT)}} = T(I_{\mathrm{Depth(reg)}}) = T(R_{\mathrm{S}}(R_\Phi(I_{\mathrm{Depth}}, I_{\mathrm{RGB}}))). \quad (2)$$

$I_{\mathrm{Depth(TR)}}$ can be given as

$$I_{\mathrm{Depth(TR)}} = R_{\mathrm{S}}(I_{\mathrm{Depth(fake)}}, \Phi) = R_{\mathrm{S}}(T(I_{\mathrm{Depth}}), R_\Phi(I_{\mathrm{Depth}}, I_{\mathrm{RGB}})). \quad (3)$$

Ensuring consistency between $I_{\mathrm{Depth(TR)}}$ and $I_{\mathrm{Depth(RT)}}$ encourages T and R to maintain task independence, thereby resulting in a well-trained registration network.

1.2.3 Training losses

The calculation of the above loss functions is as follows.

L_{recon} can be expressed as

$$L_{\mathrm{recon}} = \|I_{\mathrm{Depth(TR)}} - I_{\mathrm{RGB}}\|_1 + \|I_{\mathrm{Depth(RT)}} - I_{\mathrm{RGB}}\|_1. \quad (4)$$

By making $I_{\mathrm{Depth(TR)}} \approx I_{\mathrm{Depth(RT)}}$, L_{recon} is minimized.

L_{cGAN} for the two task branches is given as

$$L_{\mathrm{cGAN}}(T, R, D) = E[(D(I_{\mathrm{Depth}}, I_{\mathrm{RGB}}))^2] + E[(D(I_{\mathrm{Depth(TR)}}, I_{\mathrm{RGB}}) - 1)^2] + E[(D(I_{\mathrm{Depth(RT)}}, I_{\mathrm{RGB}}) - 1)^2], \quad (5)$$

where E denotes the mathematical expectation.

L_{smooth} can be constructed at a pixel $P=(x, y)$:

$$L_{\mathrm{smooth}} = \sum_{P\subseteq\Omega} \left\| \frac{\partial\Phi(P)}{\partial x}, \frac{\partial\Phi(P)}{\partial y} \right\|_2^2, \quad (6)$$

where the partial derivatives $\partial\Phi(P)/\partial x$ and $\partial\Phi(P)/\partial y$ represent the gradients in the x- and y- directions at a point P within Φ; $\sum_{P\subseteq\Omega} \|\cdot\|_2^2$ denotes the L2 norm of the gradients at all points within Φ, summed across the entire domain Ω. Ω represents the pixel space of the image.

The overall loss L_{total} for the network training can be described as

$$L_{\mathrm{total}} = \min_{(T,R)} (\max_{D} L_{\mathrm{cGAN}} + \lambda_1 L_{\mathrm{recon}} + \lambda_2 L_{\mathrm{smooth}}), \quad (7)$$

where λ_1 and λ_2 are the weights for L_{recon} and L_{smooth}, respectively. The final goal is to optimize R and T to minimize this loss.

2 Results and Discussion

In this section, the performance for R is evaluated, followed by an ablation study of R_Φ. Subsequently, real-world testing of the MMV system is conducted, with its performance being compared to the existing 2.5D MediaPipe across different hand pose scenarios.

2.1 Experimental setup

Figure 4 illustrates the experimental setup flowchart

of the MMV system. Firstly, the MMV system acquires data from a multi-modal camera (Kinect v2, Microsoft, USA) and constructs a dataset. Secondly, the training set from the dataset is used to train the network, and the trained network is applied to perform modality registration on the test set. Thirdly, 3D hand pose estimation is performed based on the registered images. Finally, the 3D hand pose estimation results are visualized on the computer. The MMV system is evaluated in terms of performance evaluation, ablation study and real-world testing.

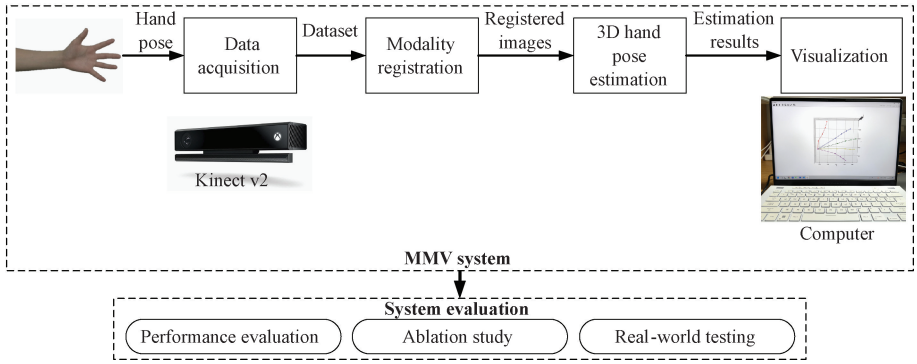

Fig. 4 Experimental setup flowchart of MMV system

This study constructs a dataset by collecting pairs of depth and RGB images of various real-world hand poses using the Kinect v2 camera. The dataset consists of 550 image pairs, with each pair containing depth and RGB images of the same hand pose, including both self-occluded (e.g., fist) and non-self-occluded (e.g., open hand) common hand poses. The entire dataset is divided into a training set (450 image pairs) and a test set (100 image pairs). Specifically, the test set includes RGB images of 50 representative hand poses from three mutually orthogonal viewpoints, with two of these specific viewpoint images used as references to evaluate the multi-view performance of our method. Specific experimental environment configurations, such as computer parameters, network training environment and image acquisition sizes, are shown in Table 1.

Table 1 Experimental environment configurations

Configuration	Parameter
CPU	AMD Ryzen™ 9 7940HS 5.2 GHz
GPU	NVDIA® GeForce RTX™ 4090 (16 GB)
Memory	32 GB
Operating system	Windows11 64 bit
Deep learning framework	Pytorch 2.3.0, Cuda 12.1
Development language	Python 3.9.13
3D visualization tool	Python-Matplotlib
Image resolution (in pixel)	1 920×1 080 (RGB) 512×424 (depth)

The MMV system does not depend on a huge-volume 3D point cloud. For hardware, the MMV system uses the usual Kinect v2 instead of expensive high-precision 3D point cloud collection devices, nor does it rely on expensive AI training hardware or data storage. For software, the MMV system leverages standard machine learning frameworks, together with the lightweight MediaPipe components, effectively reducing development costs and deployment requirements.

2.2 Performance evaluation of multi-modal registration network

2.2.1 Performance metrics

Specifically, the structural similarity index (SSIM), mean squared error (MSE), mutual information (MI), average Euclidean distance (AED), preservation of hand depth information (DIP-Hands), and the training time per epoch (TTE) are used to assess the registration accuracy and the learning time cost. Typically, SSIM, MSE and MI are employed to evaluate the similarity between two arbitrary images. Particularly, AED is introduced in this paper to evaluate the registration accuracy pixel-to-pixel over hand feature points on average. Referring to Fig. 1, perfect modality registration occurs when both the registered depth image and target RGB image align at the same pixel position, i.e., AED is 0. Otherwise, the distance error between these two modalities can be defined as AED. DIP-Hands is also introduced to indicate the depth difference of the hand feature point before and after registration.

2.2.2 Comparison of registration accuracy

Figure 5 shows an example of AED calculation for modality registration with the critical points in the hand image samples, including wrist joint points, fingertip points and finger joint points.

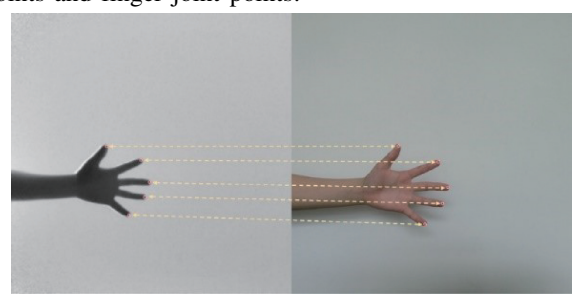

Fig. 5 Modality registration with critical points (the left being input image and the right being target image)

The typical existing registration networks, such as U-Net[26] and Affined[28], are also evaluated based on the same dataset during training. These two methods have no multi-dimensional refined feature layer in their encoder/decoder structures for modality registration, as the counterpart of the MMV system.

Table 2 gives the comparison results on the performance metrics. The higher SSIM and MI are, and the lower MSE, AED and TTE are, the higher the accuracy is. It can be observed that among the existing registration methods, Affined gets higher accuracy than U-Net with slightly higher SSIM and MI, lower MSE and AED, and a similar TTE. The MMV system is superior to Affined in terms of SSIM, MSE, AED and TTE except MI. The low MI value indicates less correlation between the depth and RGB modalities. Moreover, the MMV system demonstrates an advantage of relatively low AED and TTE. Therefore, the registration accuracy has been improved with fast network convergence.

Table 2 Comparison results on the performance metrics

Network	SSIM	MSE	MI	AED	TTE
Affined	0.910 8	7.379 3	0.492 4	5.437	50.7
U-Net	0.896 0	8.600 5	0.486 9	7.035	50.4
MMV system	0.911 2	7.210 4	0.479 9	5.282	47.4

2.2.3 Depth error before and after registration on average

Table 3 shows the depth error before and after registration on average. There are 10 sets of hand pose samples, each including different numbers of critical feature points. The average DIP-Hands value for each set is listed. The overall average value of DIP-Hands is 22.478 mm. Since the observation distance between the hand and Kinect v2 camera ranges from 850 to 1 100 mm during testing, the error of DIP-Hands relative to the observation distance, around 2% on average, can be tolerable.

Table 3 Depth error before and after registration on average

Hand pose sample	Number of wrist joint points	Number of fingertip points	Number of finger joint points	DIP-Hands/mm
1	1	5	5	25.909
2	1	—	3	12.031
3	1	2	3	15.292
4	1	2	3	21.604
5	1	1	4	22.000
6	1	5	5	8.807
7	1	5	5	26.545
8	1	5	5	29.795
9	1	5	5	31.863
10	1	1	4	30.937
Average	—	—	—	**22.478**

2.2.4 Ablation study

Figure 6 shows the ablation study results for R. Considering two kinds of hand poses, six groups of results for each hand pose are listed, including the input depth images, target RGB images, registered depth images, Φ with optical flow visualization, and the pre-registered and post-registered depth images.

Compared to Fig. 6 (a), the visualized deformation field Φ in Fig. 6 (b) shows observable characteristics with denser and smoother optical flow inside the hand palm. The registered depth image also shows a better-aligned depth of the hand palm, referring to the depth image. From the pre-registered depth image to the post-registered depth image, Fig. 6 (b) gives a richer texture feature in addition to the same well-registered edge feature, which fits the red mask of the hand pose. As a result, since the proposed encoder-decoder structure with the refined layers can generate an optimal deformation field superior to the traditional U-Net, the proposed registration network can significantly enhance the quality of the multi-modal registration in 3D hand pose estimation.

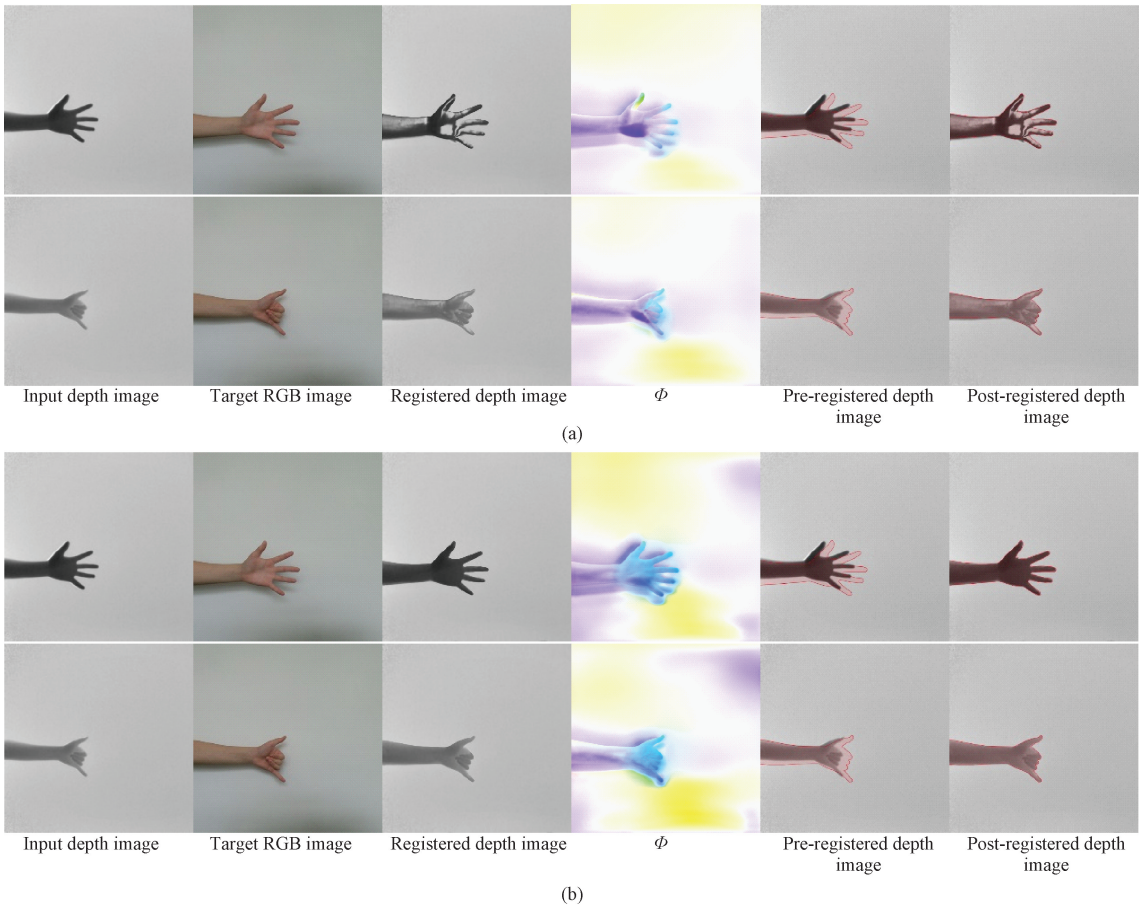

Fig. 6 Ablation study results for R: (a) registered results using an original encoder-decoder structure without refinement layers; (b) registered results using the proposed encoder-decoder structure with refinement layers

2.3 Real-world testing of MMV 3D hand pose estimation system

In this section, the real-world testing of the MMV 3D hand pose estimation system has been conducted and compared with the existing 2.5D MediaPipe in two hand pose scenarios: Pose 1, where both an arm and a hand are extended in front of each other with fingers being held straight; Pose 2, where both an arm and a hand are extended in front of each other with fingers being slightly bent. The slight pose difference between these two hand poses lies in the finger gesture, while both the hand palm and fingers remain unchanged within the same pose. Take three orthogonal observation views in the 3D domain as examples. View 1 is defined as the view from the direction perpendicular to the inner palm; View 2 is the view rotated 90° around the forearm relative to View 1; View 3 is the view rotated 90° around the elbow joint relative to View 2, as shown in Fig. 7.

Figure 7 gives the real-world testing results of the 3D MMV system. Figure 7(a) shows the RGB images corresponding to the ground truth. During the testing, View 1 from Fig. 7(a) is used as the active input for the hand pose estimation systems. Figures 7(b) and 7(c) show the corresponding visualized results of the 3D hand skeleton point model at View 1 by using 2.5D MediaPipe and the proposed 3D MMV system, respectively. Next, the 3D hand skeleton point model at View 1 is rotated by 90° and 180° to obtain the updated 3D hand skeleton point models at View 2 and View 3. Figure 7(b) fails in both cases of View 2 and View 3 but Fig. 7(c) can fit well to the ground truth over all cases of different views. Thus, Fig. 7(c) aligns better with Fig. 7(a) than Fig. 7(b) in the given view. Similarly, Pose 2 shows the same comparison results as Pose 1. This improved estimation accuracy of the proposed 3D MMV system lies in the fact that the deformation field optimized by the RT-TR jointed cGAN training scheme in the registration process can leverage multi-modal features to ensure accurate registration without much registration performance loss.

Therefore, based on the above analysis of the real-world testing results, it can be concluded that the MMV system is highly competent in recognizing slight changes in hand poses from different views in the 3D domain, thanks to the enhanced multi-modal registration quality. Compared with traditional single-modal hand pose estimation methods such as MediaPipe, the MMV system can obtain more abundant spatial information and can effectively alleviate the multi-view ill-posed problem.

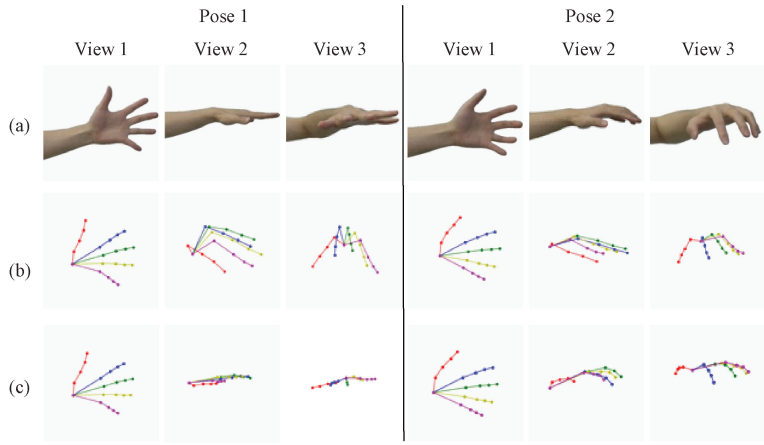

Fig. 7 Comparison results of real-world hand pose estimation: (a) RGB (ground truth); (b) 2.5D MediaPipe; (c) 3D MMV system

3 Conclusions

This paper has proposed a novel MMV system based on the multi-modal registration network trained by an RT-TR jointed cGAN scheme to fulfill high-accuracy and multi-view well-posed 3D hand pose estimation effectively. The MMV system adopts a modular design, allows for flexible adjustments and optimizations of the registration network in different application scenarios, and avoids extensive calculations. The method has low computational complexity and low cost. Experimental results demonstrate superior 3D estimation performances in various hand pose scenarios by fusing the multi-modal data features between RGB and depth. In short, the proposed MMV system is promising to be used in rich application fields such as HMI and remote medical at low cost. Future research will further study the network lightweight and dynamic hand tracking based on the currently proposed method.

References

[1] GAO C Y, YANG Y J, LI W S. 3D interacting hand pose and shape estimation from a single RGB image[J]. *Neurocomputing*, 2022, 474: 25-36.

[2] CHEN J Y, YAN M, ZHANG J Z. Tracking and reconstructing hand object interactions from point cloud sequences in the wild[C]//37th Annual AAAI Conference on Artificial Intelligence (AAAI-24). Washington DC: AAAI Press, 2023: 304-312.

[3] YU Z W, YANG L L, CHEN S C. Local and global point cloud reconstruction for 3D hand pose estimation[EB/OL]. (2021-12-13)[2024-11-02]. https://arxiv.org/abs/2112.06389.

[4] CHANG J Y, MOON G, LEE K M. V_2V-PoseNet: voxel-to-voxel prediction network for accurate 3D hand and human pose estimation from a single depth map[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 5079-5088.

[5] CHARLES R Q, HAO S, MO K C, et al. PointNet: deep learning on point sets for 3D classification and segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2017: 77-85.

[6] CHEN X H, WANG G J, GUO H K, et al. Pose guided structured region ensemble network for cascaded hand pose estimation [J]. *Neurocomputing*, 2020, 395: 138-149.

[7] ROMERO J, TZIONAS D, BLACK M J. Embodied hands: modeling and capturing hands and bodies together [J]. *ACM Transactions on Graphics (TOG)*, 2022, 36(6): 1-17.

[8] MUELLER F, BERNARD F, SOTNYCHENKO O, et al. Generated hands for real-time 3D hand tracking from monocular RGB[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2018: 49-59.

[9] JIANG C L, XIAO Y, WU C L, et al. A2J-transformer: anchor-to-joint transformer network for 3D interacting hand pose estimation from a single RGB image [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2023: 8846-8855.

[10] GAO D H, ZHANG X D, CHEN X Y, et al. CycleHand: increasing 3D pose estimation ability on in-the-wild monocular image through cyclic flow [C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM, 2022: 2452-2463.

[11] ZHANG F, BAZAREVSKY V, VAKUNOV A,

et al. MediaPipe hands: on-device real-time hand tracking [EB/OL]. (2020-06-18) [2024-11-02]. https://arxiv.org/abs/2006.10214.

[12] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks [EB/OL]. (2016-09-09) [2024-11-02]. https://arxiv.org/abs/1609.02907.

[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//31st Conference on Neural Information Processing Systems (NIPS 2017). Long Beach: NIPS, 2017.

[14] FAN Z C, SPURR A, KOCABAS M, et al. Learning to disambiguate strongly interacting hands via probabilistic per-pixel part segmentation [C]//2021 International Conference on 3D Vision (3DV). New York: IEEE, 2021: 1-10.

[15] HUANG L, ZHANG B S, GUO Z L, et al. Survey on depth and RGB image-based 3D hand shape and pose estimation[J]. *Virtual Reality & Intelligent Hardware*, 2021, 3(3): 207-234.

[16] GE L H, REN Z, LI Y C, et al. 3D hand shape and pose estimation from a single RGB image [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2019: 10825-10834.

[17] SPURR A, IQBAL U, MOLCHANOV P, et al. Weakly supervised 3D hand pose estimation via biomechanical constraints[EB/OL]. (2016-09-09)[2024-11-02]. https://arxiv.org/pdf/2003.09282.

[18] SPURR A, SONG J, PARK S, et al. Cross-modal deep variational hand pose estimation [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 89-98.

[19] CHEN L J, LIN S Y, XIE Y S, et al. DGGAN: depth-image guided generative adversarial networks for disentangling RGB and depth images in 3D hand pose estimation [C]//2020 IEEE Winter Conference on Applications of Computer Vision (WACV). New York: IEEE, 2020: 400-408.

[20] HOANG D C, TAN P X, NGUYEN A N, et al. Multi-modal hand-object pose estimation with adaptive fusion and interaction learning [J]. *IEEE Access*, 2024, 12: 54339-54351.

[21] TU Y Y, JIANG J N, LI S, et al. PoseFusion: robust object-in-hand pose estimation with SelectLSTM[C]//2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). New York: IEEE, 2023: 6839-6846.

[22] HU J, LUO Z W, WANG X, et al. End-to-end multimodal image registration via reinforcement learning[J]. *Medical Image Analysis*, 2021, 68: 101878.

[23] HU Y P, MODAT M, GIBSON E, et al. Weakly-supervised convolutional neural networks for multimodal image registration[J]. *Medical Image Analysis*, 2018, 49: 1-13.

[24] FAN J F, CAO X H, YAP P T, et al. BIRNet: brain image registration using dual-supervised fully convolutional networks[J]. *Medical Image Analysis*, 2019, 54: 193-206.

[25] QIN C, SHI B B, LIAO R, et al. Unsupervised deformable registration for multi-modal images via disentangled representations [C]// International Conference on Information Processing in Medical Imaging. Berlin: Springer, 2019: 249-261.

[26] ARAR M, GINGER Y, DANON D, et al. Unsupervised multi-modal image registration via geometry preserving image-to-image translation [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2020: 13407-13416.

[27] DAI X R, MA T, CAI H B, et al. Unsupervised hierarchical translation-based model for multi-modal medical image registration [C]//2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New York: IEEE, 2022: 1261-1265.

[28] EVAN M Y, WANG A Q, DALCA A V, et al. Keymorph: robust multi-modal affine registration via unsupervised keypoint detection [C]// Proceedings of the 5th International Conference on Medical Imaging with Deep Learning. New York: PMLR, 2022: 172: 1482-1503.

多模态多视角 3D 手部姿态估计

王　浩，王　萍*，于昊冉，丁　东，向未名
东华大学 信息科学与技术学院，上海 201620

摘　要：随着人工智能技术和移动互联网的飞速发展，3D 手部姿态估计已成为人机交互等各种智能应用领域的关键技术之一。为了避免单一模态估计精度低，以及传统多模态 3D 估计复杂度高的问题，该文提出了一种新型的多模态多视角（multi-modal multi-view，MMV）3D 手部姿态估计系统：引入 RT-TR 联合的条件式生成对抗网络（conditional generative adversarial network，cGAN），以训练多模态配准网络，并利用多模态特征融合实现高质量估计，在数据采集和处理过程中具有较低的硬件和软件成本。实验结果表明，MMV 系统在各种场景下都显示出其有效性和可行性，有望广泛应用于智能领域。

关键词：3D 手部姿态估计；配准网络；多模态；多视角；条件式生成对抗网络（cGAN）