2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

TransDexNet: An End-to-end Motion Retargeting Network with Transformer for Dexterous Hand Teleoperation from RGB Images

Jiaying Tan, Qing Gao* and Yuanchuan Lai

***Abstract*— Dexterous hand teleoperation is becoming increasingly common, yet existing methods rarely provide both efficiency and convenience. The core challenge is to achieve motion retargeting from the human hand to a dexterous hand. To address this, we introduce TransDexNet, an end-to-end vision-based motion retargeting architecture for dexterous hands. Equipped with a Vision Transformer backbone, it takes a single RGB image of a human hand and directly regresses the joint angles of a dexterous hand without any intermediate pose estimation. The architecture employs dual branches bridged by an alignment layer to close the gaps in degrees of freedom (DoFs), geometry, and kinematics between the human and dexterous hands, enabling domain-invariant latent features. To train TransDexNet, we built a dataset named TransDexData, consisting of 91,000 RGB images of human hands paired with the corresponding dexterous hand RGB images and joint angles. In evaluation, the proposed network achieves an average joint angle error of 0.076 rad. Both simulation and real-world experiments demonstrate accurate and efficient performance. The project page is available at: https://joyyyy-gaint.github.io/TransDexNet.**

I. INTRODUCTION

Robot teleoperation has been widely applied in medical rehabilitation [1], industrial manufacturing [2], and hazardous environment rescue [3]. By leveraging human perception and motor skills, teleoperated systems can accomplish complex tasks. Traditional teleoperation has shown notable success in gripper-level manipulation [4]. However, for multi-fingered dexterous hands, the combination of high-dimensional motion mapping and real-time constraints significantly increases manipulation difficulty, which remains a major challenge.

Most dexterous hand teleoperation solutions rely on specialized hardware [5], [6], such as inertial sensors [7], [8], data gloves [9], or virtual-reality equipment [10], [11]. Although they provide accurate data, these devices are costly, complex to deploy, and environmentally constrained. Recently, vision-based teleoperation has emerged as a low-cost, easy-to-use alternative, yet accurately converting human hand motions into executable robot commands remains challenging.

The central issue in vision-based dexterous hand teleoperation is motion retargeting—mapping human hand motions to a dexterous hand. Conventional pipelines typically adopt a two-stage framework: the pose of the hand is first estimated from images, then mapped to the dexterous hand via optimization [12] or data-driven methods [13]. These methods depend heavily on pose accuracy and often incur considerable latency due to their multi-stage design.

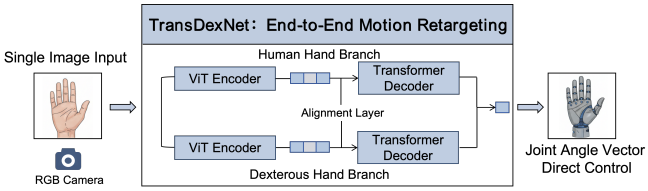


Fig. 1. Overview of the vision-based teleoperation pipeline. The framework is based on the end-to-end retargeting network TransDexNet, which maps a single RGB image of a human hand directly to the joint angles of a dexterous hand.

To address these challenges, recent studies have explored end-to-end vision-based motion retargeting methods [14], [15]. These methods map human hand images to dexterous hand joint angles, reducing error accumulation and enabling real-time performance. However, most approaches rely on depth images from specialized sensors, limiting their deployment in everyday scenarios. In contrast, RGB images are more accessible and widely used in computer vision. Motivated by this, we investigate an RGB-only teleoperation framework to improve deployment feasibility and practical usability.

Compared with depth images, RGB images lack explicit 3D geometric information, making it difficult to directly apply traditional 3D pose-based networks to RGB inputs. To address this limitation, we adopt a 3D mesh-based network for motion retargeting. A 3D mesh provides richer structural information than sparse 3D poses, mitigating performance degradation caused by missing depth cues and improving retargeting accuracy.

Based on the above, we introduce TransDexNet, a vision-based end-to-end motion retargeting network for dexterous hand teleoperation (see Fig. 1). It directly maps a single RGB image of a human hand to the 22 joint angles of a dexterous hand in a single forward pass, thus completely eliminating the intermediate 3D pose estimation stages that typically introduce error accumulation and computational overhead.

Specifically, TransDexNet adopts a dual-branch Transformer with Vision Transformer (ViT) backbones to extract features from human and robotic hand domains. An alignment layer bridges the domain gap by mapping both into a shared embedding space, followed by a shared regression head that directly predicts joint angles, improving parameter efficiency and generalization.

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515011954, 2023A1515110074, in part by the Shenzhen Science and Technology Program under Grant ZDCY20250901100201002.

Jiaying Tan, Qing Gao, and Yuanchuan Lai are with the School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen, China (e-mail: tanjy79@mail2.sysu.edu.cn; gaoqing2@mail.sysu.edu.cn; 2865326869lai@gmail.com).

*Corresponding author: Qing Gao, gaoqing2@mail.sysu.edu.cn.

To train our retargeting network, we constructed TransDexData, a large-scale dataset with 91,000 samples. Each sample contains a human hand RGB image, nine multi-view dexterous hand images, and the corresponding 22-dimensional joint angles. The dataset was generated via a motion optimization pipeline with physical constraints and collision avoidance, ensuring kinematic feasibility and natural motion transfer.

Experimental results show that TransDexNet achieves an average joint angle error of only 0.076 radians with a per-frame inference time of 0.22 seconds. The system also shows excellent generalizability in both simulation environments and in real-world experiments. In particular, the network successfully retargets various complex gestures, including human daily motions, with high accuracy and natural movement dynamics.

The main contributions of this study are as follows:

- We propose TransDexNet, an end-to-end hand motion retargeting network that takes a single RGB image and directly predicts dexterous hand joint angles, using a dual-branch architecture with a consistency loss to align human and robotic hand features.
- We introduce TransDexData, a large-scale paired dataset of 91,000 samples, each containing a human hand RGB image, nine simulated multi-view dexterous hand images, and corresponding joint angles.
- Experiments show that TransDexNet achieves low average joint angle error and demonstrates high accuracy and efficiency in both simulation and real-world experiments.

II. Related Work

A. Vision-Based Teleoperation

In recent years, vision-based teleoperation has attracted widespread attention in robot control [16], [17]. With advantages such as low cost, easy deployment, and non-invasiveness, these methods provide a natural and flexible control approach for robotic systems. In dexterous hand control, vision-based teleoperation shows significant potential, with the main challenge being accurately and robustly mapping human hand movements to multi-fingered robotic hands [18], [19].

Most existing methods adopt a phased processing pipeline [20], which typically first estimates the human hand pose based on visual input [21], [22], and then generates control commands for the dexterous hand through motion mapping algorithms [23]. Although such methods have realized motion transfer to a certain extent, the existing problems of error accumulation and high computational complexity limit their application in real-time scenarios.

Meanwhile, end-to-end control methods have gradually emerged [14], [15]. These methods aim to directly generate robot control commands from raw visual input through models, which alleviates the cascading error problem in traditional frameworks to a certain extent. However, relevant research in the field of dexterous hand control is still in the exploratory stage, especially in the research direction that only uses RGB images and does not require intermediate pose representation.

B. Robot Motion Retargeting

Robot motion retargeting is a core technology in teleoperation, aiming to map human movements to robot systems with significantly different structure, size, and degrees of freedom in real time. Existing approaches fall into three main categories:

Mapping-based methods utilize predefined joint correspondences and inverse kinematics to transfer motion. They are straightforward to implement and computationally efficient, but their generalization is heavily constrained by the structural similarity required between human and robot [24].

Optimization-based methods formulate an objective function and solve it iteratively to achieve motion mapping. Recent improvements have enhanced convergence speed, yet high computational cost still restricts them mainly to offline use [25], [26].

Learning-based methods learn motion policies directly from demonstration data, enabling end-to-end motion generation with strong flexibility and generalization potential. However, their performance is highly sensitive to data quality and often struggles to achieve both real-time performance and practical usability in dexterous hand operation [27], [28].

To address these limitations common to all three paradigms, we propose an end-to-end vision-based architecture for dexterous hand motion retargeting—TransDexNet. Building on the learning-based paradigm, TransDexNet uses data-driven training to deliver an accurate, efficient, and fast retargeting pipeline.

C. RGB-Based 3D Hand Pose Estimation

3D hand pose estimation from RGB images is a fundamental step in vision-based teleoperation. Existing methods can be categorized as parametric or non-parametric. Parametric methods [29]–[32] use a predefined hand model to regress hand shape and pose parameters, forming a single image end-to-end prediction pipeline. Although efficient, their accuracy often decreases when the observed hand posture significantly deviates from the prior model. Non-parametric methods [33], [34] bypass the limitations of parametric models by directly predicting mesh vertex coordinates. This improves alignment with the image but often increases sensitivity to occlusions.

Recent efforts have aimed to improve the practicality of these methods. For example, MobRecon [35] focuses on real-time inference on mobile devices, while HandOccNet [36] improves robustness under occlusions. Despite these advances, both still require costly 3D annotations for training. More recently, HaMeR [37] has demonstrated improved generalization and accuracy by scaling training data and simplifying the network architecture.

Inspired by HaMeR, our approach avoids explicit 3D hand reconstruction. Instead, we adopt a Vision Transformer [38] backbone to directly map a single RGB image to the joint angles of the dexterous hand. This end-to-end design eliminates intermediate 3D pose estimation, thereby reducing error accumulation and improving inference speed and retargeting accuracy.

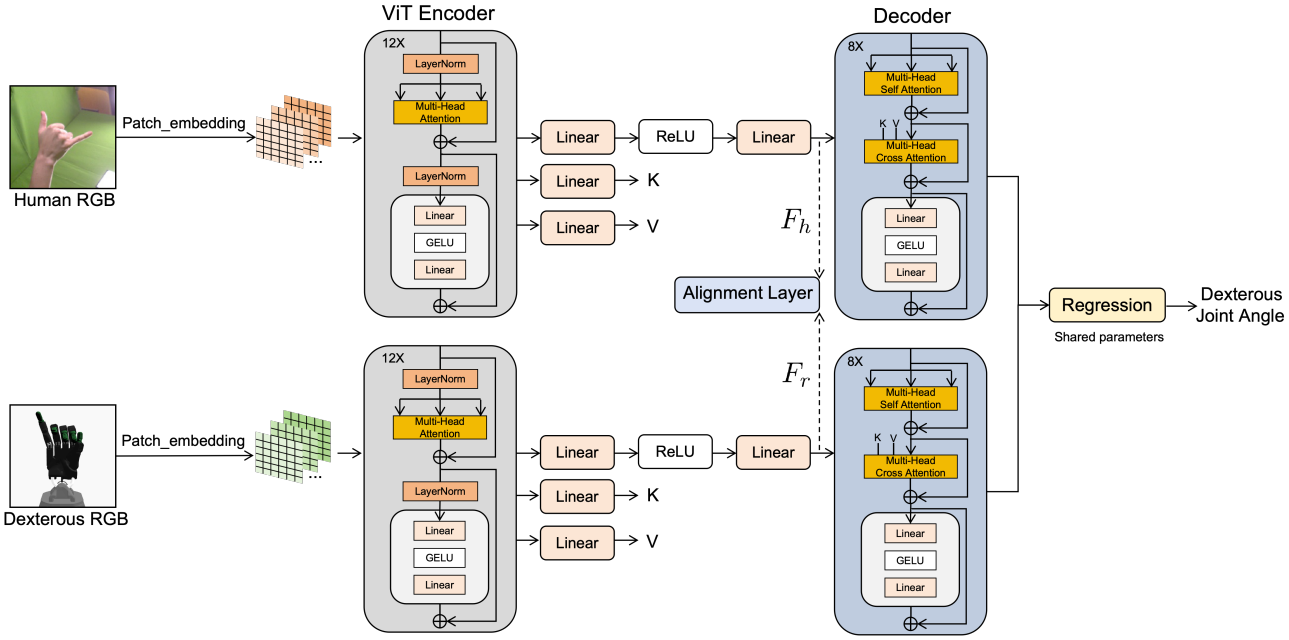


Fig. 2. The architecture of TransDexNet. Two branches employ independent Vision Transformer (ViT) backbones to encode images of the human hand and robot hand respectively. After feature extraction, an alignment layer is used to synchronize their latent features. These aligned features are then fed into a Transformer Decoder for further processing, which refines and models the relationships within the feature representations. Finally, a shared regression head outputs 22-dimensional joint angles for the dexterous hand.

III. TransDexNet Network

Owing to large differences in morphology and kinematics between the human hand and multi-fingered dexterous hands, a single-branch regression network cannot easily bridge the large domain gap and mismatched feature distributions. We therefore propose TransDexNet, a dual-branch motion retargeting network whose key idea is to perform an end-to-end mapping from a human hand image to the joint angles of the dexterous hand through feature alignment and parameter sharing. The network can be formalized as the following mapping:

$$F{:}I_h \rightarrow \hat{\theta}_r, \tag{1}$$

where I_h represents the input RGB image of the human hand, and $\hat{\theta}_r$ denotes the predicted 22-dimensional joint angle vector of the dexterous hand.

A. Network Architecture

TransDexNet adopts a dual-branch architecture consisting of a human hand branch and a dexterous hand branch, as illustrated in Fig. 2. Each branch comprises three modules: a ViT, a Transformer Decoder, and a regression module. ViT is employed to extract high level semantic features from the input image:

$$F_h = \mathrm{ViT}(I_h), F_r = \mathrm{ViT}(I_r), \tag{2}$$

where $F_h, F_r \in \mathbf{R}^{768}$ denote the extracted latent feature vectors, I_h is the input RGB image of the human hand, and I_r represents the RGB image of the dexterous hand.

The first two modules (ViT and Transformer Decoder) of each branch are trained with independent parameters, while the regression module adopts shared parameters to output the 22 joint angles of the dexterous hand. During training, the network takes paired RGB images as input. The human branch receives an RGB image of the human hand, and the robot branch receives an RGB image of the dexterous hand in the same pose. Both branches independently predict the joint angles of the dexterous hand.

B. Feature Alignment Mechanism

Directly regressing the joint angles of a dexterous hand from a human hand image poses significant challenges. To address this, we designed a dual-branch architecture to accomplish the following three key tasks:

- Learning deep latent feature representations of both the human hand and the dexterous hand through the ViT network.
- Spatially aligning the latent features of the two branches to identify common characteristics across domains.
- Decoding and regressing the aligned features to obtain accurate joint angle commands.

An Alignment Layer is introduced between the two branches specifically to align the latent feature spaces of the human hand and the dexterous hand. This design effectively learns shared feature representations between the two domains, thereby better addressing the problem of directly predicting dexterous hand joint angles from human hand images. Compared to a single-branch architecture, the dual-branch architecture more effectively captures both the similarities and differences between human and robotic hands, significantly improving prediction accuracy.

C. Loss Functions

The network training employs three loss functions: joint loss, consistency loss, and physical loss.

Joint loss. The joint loss is applied to both the human hand branch and the dexterous hand branch. It measures the difference between the predicted joint angles and the ground truth angles using the Mean Squared Error (MSE). Joint loss for the human hand branch and the dexterous hand branch:

$$L_{joint}^{h} = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\theta}_{h}^{i} - \theta_{r}^{i}\right)^{2}, \tag{3}$$

$$L_{joint}^{r} = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\theta}_{r}^{i} - \theta_{r}^{i}\right)^{2}, \tag{4}$$

where $\hat{\theta}_h$ denotes the joint angles predicted by the human hand branch, $\hat{\theta}_r$ represents the joint angles predicted by the dexterous hand branch, and θ_r indicates the ground-truth joint angles of the dexterous hand from the dataset.

Consistency loss. This loss aligns the features between the two branches by encouraging the latent features learned by both branches to be as consistent as possible. The consistency loss is computed using MSE:

$$L_{align} = \frac{1}{d}\sum_{j=1}^{d}(F_{h}^{j} - F_{r}^{j})^{2}, \tag{5}$$

where F_h and F_r respectively represent the latent hand features learned by the two branches.

Physical loss. This loss constrains the predicted joint angles within physically plausible ranges, ensuring compliance with actual mechanical limitations of the joints. Physical loss for the human hand branch and the dexterous hand branch:

$$L_{phys}^{h} = \frac{1}{N}\sum_{i=1}^{N}\left[\max\left(0, \hat{\theta}_{h}^{i} - \theta_{\max}^{i}\right) + \max\left(0, \theta_{\min}^{i} - \hat{\theta}_{h}^{i}\right)\right], \tag{6}$$

$$L_{phys}^{r} = \frac{1}{N}\sum_{i=1}^{N}\left[\max\left(0, \hat{\theta}_{r}^{i} - \theta_{\max}^{i}\right) + \max\left(0, \theta_{\min}^{i} - \hat{\theta}_{r}^{i}\right)\right], \tag{7}$$

where N is the number of joints, θ_{max} is the upper mechanical limit of the joint range, and θ_{min} is the lower mechanical limit of the joint range.

The total loss function is defined as the sum of the above three loss functions:

$$L_{total} = L_{joint}^{h} + L_{joint}^{r} + L_{align} + L_{phys}^{h} + L_{phys}^{r}. \tag{8}$$

IV. TransDexData Dataset

To train the TransDexNet network for learning the motion mapping relationship between human hands and dexterous hands, it is necessary to construct a large-scale and accurate paired dataset. Since publicly available data of this kind are currently lacking, we construct a new large-scale paired dataset named TransDexData, based on two RGB hand datasets—FreiHAND [39] and HO-3D [40], both annotated with 3D coordinates of 21 hand joints.

We adopt a motion optimization and physics-based simulation mapping method to transfer human hand motion features to the Shadow C6 dexterous hand model, while simultaneously collecting corresponding RGB images and 22-dimensional joint angles. Eventually, a total of 91,000 data samples were ultimately collected, each comprising one human hand RGB image, multiple dexterous hand RGB images, and the corresponding joint angles.

A. Kinematic Structure of Human and Dexterous Hand

This study employs the Shadow C6 dexterous hand, which integrates BioTac tactile sensors at the tip of each finger. The hand consists of five fingers, each containing four joints (distal, middle, proximal, and metacarpal), with the distal joint being fixed. The thumb and little finger are each equipped with an additional auxiliary joint to enhance motion stability. The hand itself has 17 DoF. Together with the 2-DoF wrist, the system has a total of 19 DoF.

In comparison, the kinematic structure of the human hand is more complex. The human hand models in the FreiHAND and HO-3D datasets comprise 21 joints and exhibit up to 31 DoF. Significant differences exist between the human and robotic hands in terms of joint range of motion and wrist structure. To simplify the mapping problem, the two wrist joints of the dexterous hand are fixed in this study.

B. Motion Optimization Mapping Method

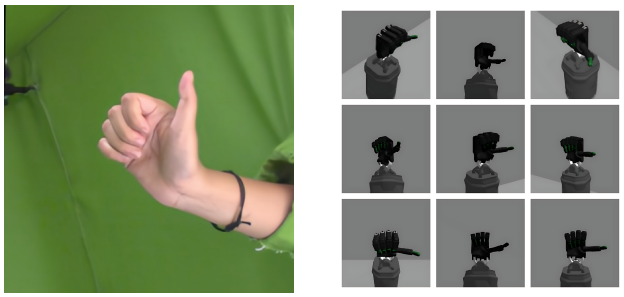
To achieve high quality motion mapping from human hands to dexterous hands, we adopt a motion retargeting approach based on multi-constraint optimization inspired by TeachNet. This method comprehensively considers position constraints, orientation constraints, and collision avoidance to ensure that the generated dexterous hand poses accurately reflect human motion intent while satisfying the physical feasibility of the mechanical structure.

First, a unified reference coordinate frame F is established, with its origin located at the human wrist joint and its z-axis aligned with the wrist coordinate frame of the dexterous hand. This coordinate frame is chosen based on the high kinematic similarity between the human and dexterous hands in the wrist region.

In terms of constraint settings, we apply a primary constraint (weight coefficient ω_{pf}) to the fingertip positions, an auxiliary constraint (weight coefficient ω_{pp}) to the proximal interphalangeal positions, and an orientation constraint (weight coefficient ω_{dir}) to the proximal phalanges and the distal thumb phalanx. The constraint weight configuration adopted in this study is: $\omega_{\text{pf}},\ \omega_{\text{pp}},\ \omega_{\text{dir}} = 1.0,\ 0.5,\ 0.1$.

Based on the above constraints, the joint angles θ of the dexterous hand are solved through an optimization algorithm. The objective function is defined as follows:

$$\begin{aligned}\arg\min_{\theta}\Big[&\omega_{\text{pf}} \cdot L_{\text{pos}}(\mathbf{X}_{\text{tip}}, \hat{\mathbf{X}}_{\text{tip}}) \\ &+ \omega_{\text{pp}} \cdot L_{\text{pos}}(\mathbf{X}_{\text{pip}}, \hat{\mathbf{X}}_{\text{pip}}) + \omega_{\text{dir}} \cdot L_{\text{dir}}(\mathbf{D}, \hat{\mathbf{D}})\Big],\end{aligned} \tag{9}$$



(a) (b)

Fig. 3. TransDexData sample. (a) A human hand image. (b) The corresponding multi-view images of the dexterous hand.

where L_{pos} denotes the position loss function and L_{dir} the direction loss function; $\mathbf{X}$ and $\mathbf{D}$ respectively represent the target position and orientation, while $\hat{\mathbf{X}}$ and $\hat{\mathbf{D}}$ are their current estimates.

To ensure the physical plausibility of the generated poses, we perform motion execution validation in the Gazebo simulation environment and utilize MoveIt for collision detection. If self-collision is detected, the severity of the collision is evaluated by computing the minimum distance between the links:

$$C_{\text{collision}} = \max\left(0, d_{\text{safe}} - \|P_i - P_j\|\right), \tag{10}$$

where P_i and P_j denote two points on the links that may collide, and d_{safe} is the predefined safety distance.

C. Multi-view Data Collection

To enhance the viewpoint diversity of the dataset, we deployed nine virtual RGB cameras at different poses in the Gazebo environment to synchronously capture multi-view images of the same dexterous hand posture (see Fig. 3). This multi-view acquisition strategy significantly increases the data variability, which is crucial for improving the model's generalization capability across different viewpoints.

The high-quality TransDexData dataset, constructed using this methodology, provides essential resources for training and evaluating models for motion retargeting from human hands to dexterous hands.

V. Experiments

A. Experimental Setup

The experiments were conducted using the TransDexData dataset constructed in the TransDexData Dataset chapter, under a hardware environment equipped with an Intel(R) Core(TM) i7-10700K CPU @ 3.80GHz and an NVIDIA GeForce RTX 3080. RGB images of both human and dexterous hands were uniformly cropped to a resolution of 224×224 before being fed into the network. For simulation experiments, the Shadow dexterous hand model consistent with the dataset was used, while in real-world experiments, due to equipment constraints, the Inspire dexterous hand with fewer degrees of freedom served as a physical validation platform.

The evaluation metrics include three measures: average joint angle error, maximum error frame accuracy, and mean error frame accuracy. Here, maximum error frame accuracy is defined as the proportion of frames where the maximum joint angle error is below a set threshold, and mean error frame accuracy refers to the proportion of frames where the mean joint angle error is below a set threshold.

B. Network Experiments

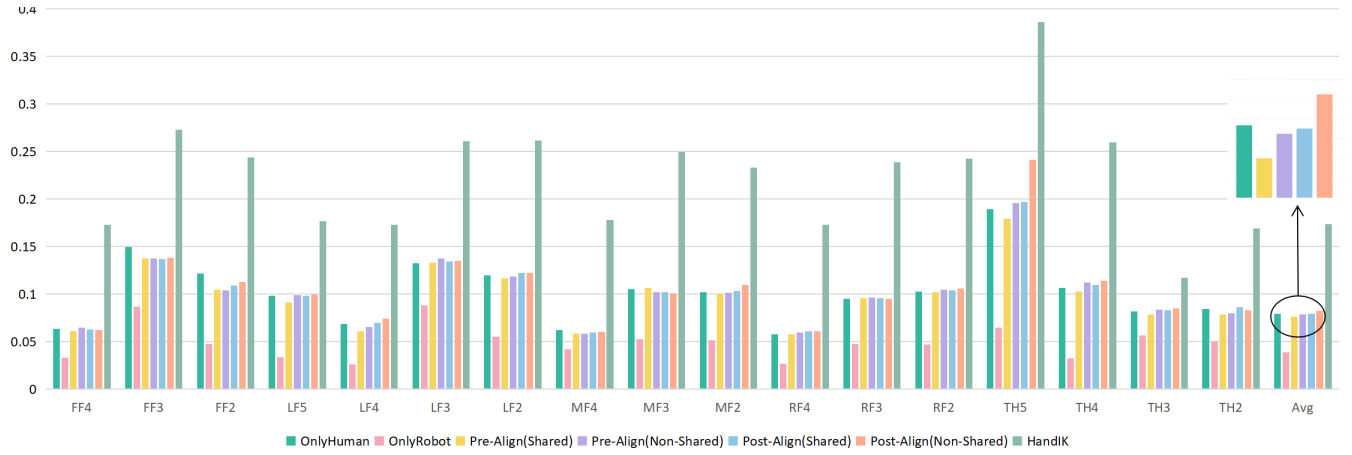
This subsection presents the network comparison and ablation studies. In total, seven motion retargeting network configurations were evaluated. These include four variants of the proposed architecture: Pre-Align (Shared), Pre-Align (Non-Shared), Post-Align (Shared) and Post-Align (Non-Shared); two single-branch architectures: OnlyHuman (trained solely on the human hand branch) and OnlyRobot (trained solely on the dexterous hand branch) and HandIK, which serves as the baseline method for comparison.

The model configurations are as follows: the number of layers and heads in the ViT are both set to 12, the number of layers and heads in the Transformer Decoder are both set to 8. The alignment layer vector dimension is 768. During training, the initial learning rate is set to 1×10^{-4} and the batch size is 32. All models are trained for 200 epochs, during which the training process remains stable.

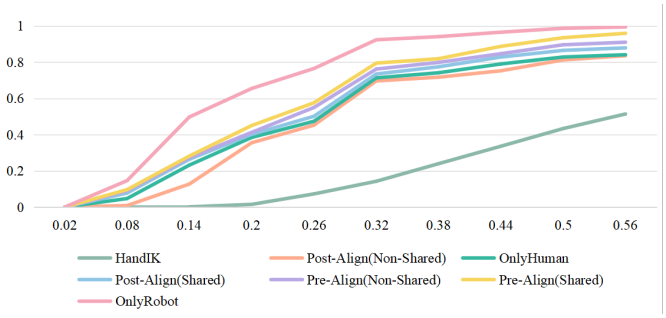
In Pre-Align(Shared) and Pre-Align(Non-Shared), the alignment layer is inserted after the ViT module, with the former sharing regression head parameters and the latter using independent regression heads. Post-Align(Shared) and Post-Align(Non-Shared) place the alignment layer after the Decoder module, following similar parameter-sharing strategies. OnlyHuman is trained exclusively on human hand images, while OnlyRobot is trained only on robotic hand images. Since the actual input consists of human hand images, the OnlyRobot configuration serves only as a reference for comparison. HandIK adopts a phased processing pipeline: it first extracts 21 keypoints from the human hand using MediaPipe [41], then solves inverse kinematics via BioIK to obtain joint angles.

The experimental results of the motion retargeting networks are shown in Fig. 4. In descending order, the average joint angle error of the networks is: HandIK, Post-Align(Non-Shared), OnlyHuman, Post-Align(Shared), Pre-Align(Non-Shared), Pre-Align(Shared) and OnlyRobot. Compared to HandIK, a traditional phased processing method, the end-to-end network reduces the average joint angle error by 0.09 radians, demonstrating significant performance improvement.

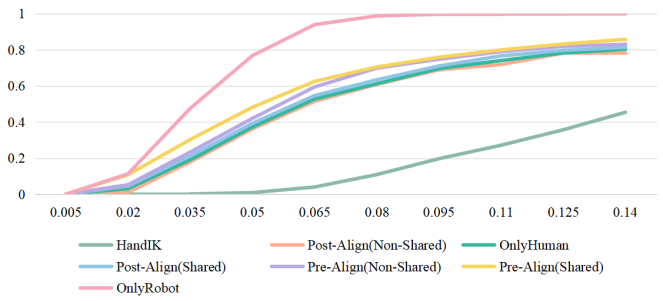
Among the four dual-branch architectures, the Pre-Align series consistently outperforms the Post-Align series, indicating that early-stage cross-domain alignment helps better preserve and share kinematic features, thereby improving mapping accuracy. Parameter sharing also contributes to performance gains: sharing regression heads not only reduces the number of model parameters and mitigates overfitting but also, due to the aligned front-end features, ensures higher consistency in the input features to the regression module, leading to better generalization under shared parameters.



(a) Average joint angle error.



(b) Maximum error frame accuracy.



(c) Mean error frame accuracy.

Fig. 4. Evaluation of motion retargeting networks.

In the ablation study, Post-Align(Shared), Pre-Align(Non-Shared), and Pre-Align(Shared) all significantly outperform the single-branch baseline OnlyHuman, validating the effectiveness of the dual-branch alignment architecture. OnlyRobot achieves the best performance as it performs same-domain regression directly based on robotic hand images. However, since practical tasks require inferring the robotic hand state from human hand images, this configuration lacks practical applicability.

Analysis of joint error distribution shows that the proposed method achieves low overall prediction errors, with higher errors mainly in the thumb (TH5), index finger (FF3), and little finger (LF3) due to their higher degrees of freedom. In contrast, the MCP joints of all fingers (e.g., FF4, LF4, MF4, RF4) exhibit lower errors, indicating more robust prediction for proximal joints. Overall, the method performs satisfactorily, though accuracy for complex joints like the thumb can be further improved.

C. Simulation Experiments

In the motion retargeting simulation experiments, we employed the trained Pre-Align (Shared) network configured in single-branch (human hand branch) inference mode. The network takes an RGB image of a human hand as input and outputs joint angle commands for the dexterous hand in an end-to-end way. The measured average inference time per sample was 0.22 seconds, enabling near real-time interaction.

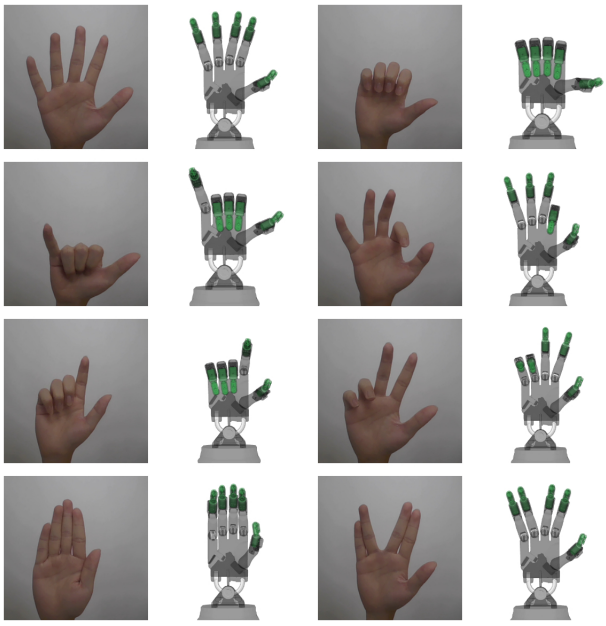
To evaluate performance in practical interactive scenarios, we invited three untrained operators to perform random daily gestures while teleoperating the Shadow Hand in simulation. Using an end-to-end motion retargeting approach, the system enables intuitive human-to-robot hand motion transfer without requiring robotics expertise. As shown in Fig. 5, the system achieves effective retargeting performance in simulation.

The results indicate satisfactory overall retargeting accuracy, with errors mainly concentrated in the thumb. This may result from thumb-specific annotation noise and its higher degrees of freedom and structural complexity, which make generalization more difficult. Nevertheless, the results validate the effectiveness of the proposed method in simulation and demonstrate its feasibility for vision-based teleoperation of multi-degree-of-freedom dexterous hands.

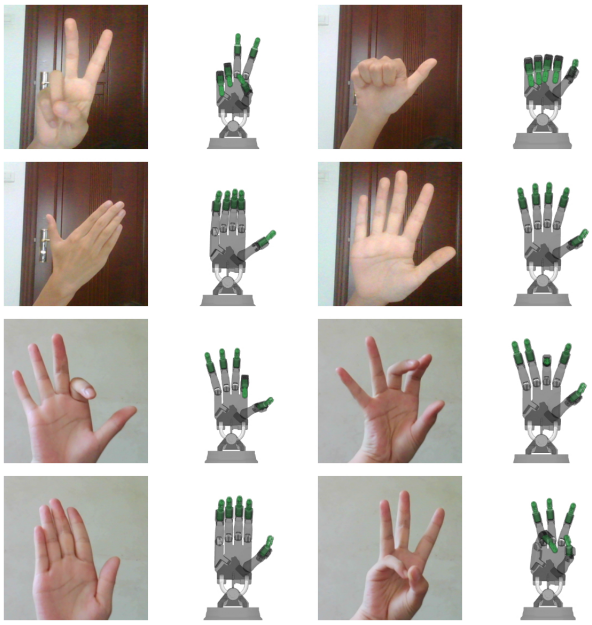
D. Real-world Experiments

In the real-world validation, due to the unavailability of a physical Shadow Hand, the 22 predicted joint angles (corresponding to 17 degrees of freedom) of the Shadow Hand from the network were mapped to the Inspire dexterous hand, which has 12 joints (6 degrees of freedom).

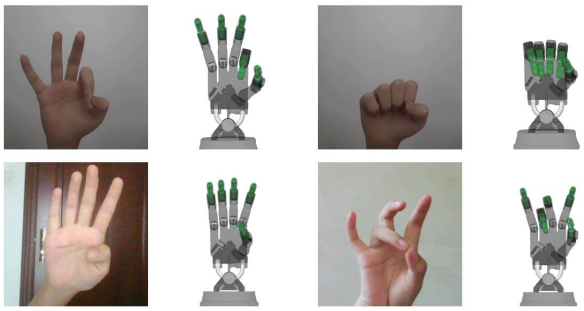
The specific mapping strategy was as follows: the outputs of the FJ2 and FJ3 joints of the index, middle, ring, and little fingers were merged with a weight ratio of 0.8:0.2 to drive the corresponding single joint of the Inspire dexterous hand. For the thumb, TH5 and TH4 were mapped to the first and



(a)



(b)



(c)

Fig. 5. Simulation experimental results. (a) Successful gesture by Operator 1. (b) Successful gesture by Operators 2 and 3. (c) Failure case by three operators.

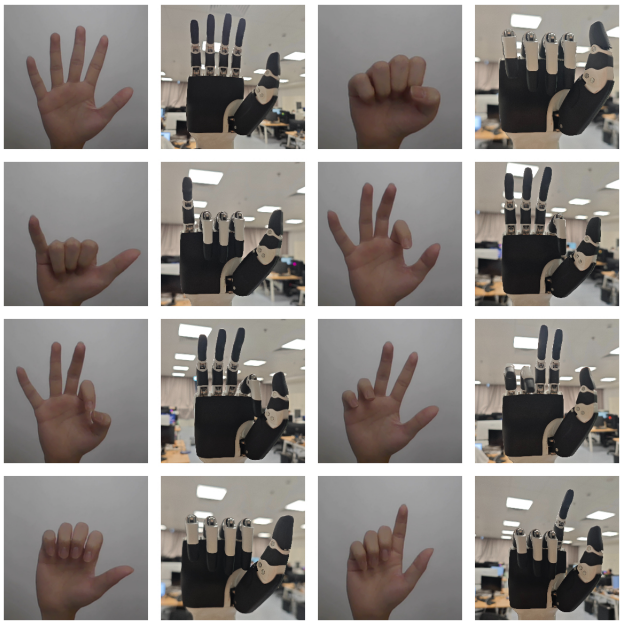


Fig. 6. Real-world Experimental Results. The joint angles applied in the real-world experiments were mapped from the simulation-based results, and the evaluation was performed using the identical gesture set employed in the preceding simulation experiments.

second joints of the Inspire thumb. The gestures used in the real-world experiments remained consistent with those in the simulation, covering a variety of daily hand poses.

As shown in Fig. 6, the motion retargeting results in the real-world environment indicate that the Inspire dexterous hand successfully replicated most target gestures, and it demonstrates that the proposed method effectively maintains accuracy and stability in motion mapping even in physical systems.

VI. Conclusion

This paper presents TransDexNet, a Transformer-based end-to-end motion retargeting network for direct teleoperation from a single RGB image to a multi-degree-of-freedom dexterous hand. The network adopts a dual-branch architecture and incorporates a learnable alignment layer to bridge the kinematic, morphological, and structural differences between the human and robotic hands at the feature level, achieving cross-domain unified representation. Experimental results demonstrate that the proposed method achieves an average joint angle error of 0.076 radians while maintaining low inference latency (0.22 seconds per frame), significantly outperforming traditional phased processing methods and single-branch network structures. Both simulation and real-world experiments further validate the accuracy of the system across various daily gesture tasks.

Future work will focus on designing lightweight network architectures to improve inference speed and developing more robust data augmentation methods to reduce dependency on precise annotations.

REFERENCES

[1] Toedtheide A, Chen X, Sadeghian H, et al. A force-sensitive exoskeleton for teleoperation: An application in elderly care robotics[C]//2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023: 12624-12630.

[2] Mai X, Chen J, Wang Y, et al. A teleoperation framework of hot line work robot[C]//2018 IEEE International Conference on Mechatronics and Automation (ICMA). IEEE, 2018: 1872-1876.

[3] Asami R, Sawai Y, Sato N, et al. Teleoperation system with virtual 3D diorama for moving operation of a tracked rescue robot[C]//2016 International Conference on advanced Mechatronic Systems (ICAMechS). IEEE, 2016: 90-95.

[4] Khurshid R P, Fitter N T, Fedalei E A, et al. Effects of grip-force, contact, and acceleration feedback on a teleoperated pick-and-place task[J]. IEEE transactions on haptics, 2016, 10(1): 40-53.

[5] Wu P, Shentu Y, Yi Z, et al. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators[C]//2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2024: 12156-12163.

[6] Yang S. ACE: A Cross-platform Visual-Exoskeleton System for Low-Cost Dexterous Teleoperation[D]. University of California, San Diego, 2025.

[7] Zhang H, Zhao Z, Yu Y, et al. A feasibility study on an intuitive teleoperation system combining IMU with sEMG sensors[C]//International Conference on Intelligent Robotics and Applications. Cham: Springer International Publishing, 2018: 465-474.

[8] Li S, Jiang J, Ruppel P, et al. A mobile robot hand-arm teleoperation system by vision and imu[C]//2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020: 10900-10906.

[9] Liu H, Xie X, Millar M, et al. A glove-based system for studying hand-object manipulation via joint pose and force sensing[C]//2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2017: 6617-6624.

[10] Arunachalam S P, Güzey I, Chintala S, et al. Holo-dex: Teaching dexterity with immersive mixed reality[J]. arXiv preprint arXiv:2210.06463, 2022.

[11] Ding R, Qin Y, Zhu J, et al. Bunny-visionpro: Real-time bimanual dexterous teleoperation for imitation learning[C]//2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2025: 12248-12255.

[12] Qin Y, Yang W, Huang B, et al. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system[J]. arXiv preprint arXiv:2307.04577, 2023.

[13] Handa A, Van Wyk K, Yang W, et al. Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system[C]//2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020: 9164-9170.

[14] Li S, Hendrich N, Liang H, et al. A dexterous hand-arm teleoperation system based on hand pose estimation and active vision[J]. IEEE Transactions on Cybernetics, 2022, 54(3): 1417-1428.

[15] Li S, Ma X, Liang H, et al. Vision-based teleoperation of shadow dexterous hand using end-to-end deep neural network[C]//2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019: 416-422.

[16] Weiming Q, Xiaomei Z, Jiwei H, et al. Real-time virtual UR5 robot imitation of human motion based on 3D camera[C]//2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE). IEEE, 2020: 387-391.

[17] Ajili I, Mallem M, Didier J Y. Gesture recognition for humanoid robot teleoperation[C]//2017 26Th IEEE international symposium on robot and human interactive communication (RO-MAN). IEEE, 2017: 1115-1120.

[18] Qin Y, Su H, Wang X. From one hand to multiple hands: Imitation learning for dexterous manipulation from single-camera teleoperation[J]. IEEE Robotics and Automation Letters, 2022, 7(4): 10873-10881.

[19] Gao Q, Li J, Zhu Y, et al. Hand gesture teleoperation for dexterous manipulators in space station by using monocular hand motion capture[J]. Acta Astronautica, 2023, 204: 630-639.

[20] Gao Q, Ju Z, Chen Y, et al. An efficient RGB-D hand gesture detection framework for dexterous robot hand-arm teleoperation system[J]. IEEE Transactions on Human-Machine Systems, 2022, 53(1): 13-23.

[21] Zhang M, Gao Q, Lai Y, et al. 3D Whole-Body Pose Estimation Using Graph High-Resolution Network for Humanoid Robot Teleoperation[C]//2025 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2025: 1090-1096.

[22] Pang W, Gao Q, Zhao Y, et al. BaSICNet: Lightweight 3-D hand pose estimation network based on biomechanical structure information for dexterous manipulator teleoperation[J]. IEEE Transactions on Cognitive and Developmental Systems, 2022, 16(2): 448-457.

[23] Lai Y, Ju Z, Gao Q. Motion Retargeting Using Graph Neural Network for Vision-Guided Dexterous Robot Teleoperation[C]//2024 17th International Convention on Rehabilitation Engineering and Assistive Technology (i-CREATe). IEEE, 2024: 1-6.

[24] Penco L, Clément B, Modugno V, et al. Robust real-time whole-body motion retargeting from human to humanoid[C]//2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids). IEEE, 2018: 425-432.

[25] Liang Y, Li W, Wang Y, et al. Dynamic movement primitive based motion retargeting for dual-arm sign language motions[C]//2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021: 8195-8201.

[26] Zhang H, Li W, Liu J, et al. Kinematic motion retargeting via neural latent optimization for learning sign language[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 4582-4589.

[27] Kim T, Lee J H. C-3po: Cyclic-three-phase optimization for human-robot motion retargeting based on reinforcement learning[C]//2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020: 8425-8432.

[28] Yin H, Melo F, Billard A, et al. Associate latent encodings in learning from demonstrations[C]//Proceedings of the AAAI conference on artificial intelligence. 2017, 31(1).

[29] Kanazawa A, Black M J, Jacobs D W, et al. End-to-end recovery of human shape and pose[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7122-7131.

[30] Baek S, Kim K I, Kim T K. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 1067-1076.

[31] Boukhayma A, Bem R, Torr P H S. 3d hand shape and pose from images in the wild[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 10843-10852.

[32] Rong Y, Shiratori T, Joo H. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 1749-1759.

[33] Choi H, Moon G, Lee K M. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose[C]//European Conference on Computer Vision. Cham: Springer International Publishing, 2020: 769-787.

[34] Ge L, Ren Z, Li Y, et al. 3d hand shape and pose estimation from a single rgb image[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 10833-10842.

[35] Chen X, Liu Y, Dong Y, et al. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 20544-20554.

[36] Park J K, Oh Y, Moon G, et al. Handoccnet: Occlusion-robust 3d hand mesh estimation network[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 1496-1505.

[37] Pavlakos G, Shan D, Radosavovic I, et al. Reconstructing hands in 3d with transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 9826-9836.

[38] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[39] Zimmermann C, Ceylan D, Yang J, et al. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 813-822.

[40] Hampali S, Sarkar S D, Lepetit V. `Ho-3d_v3`: Improving the accuracy of hand-object annotations of the ho-3d dataset[J]. arXiv preprint arXiv:2107.00887, 2021.

[41] Lugaresi C, Tang J, Nash H, et al. Mediapipe: A framework for building perception pipelines[J]. arXiv preprint arXiv:1906.08172, 2019.