Semantics2Hands: Transferring Hand Motion Semantics between Avatars

Zijie Ye
yzjscwy@gmail.com
Department of Computer Science and Technology, Tsinghua University
Beijing 100084, China

Jia Jia*
jjia@tsinghua.edu.cn
Department of Computer Science and Technology, Tsinghua University
Beijing National Research Center for Information Science and Technology
Beijing 100084, China

Junliang Xing
jlxing@tsinghua.edu.cn
Department of Computer Science and Technology, Tsinghua University
Beijing 100084, China

ABSTRACT

Human hands, the primary means of non-verbal communication, convey intricate semantics in various scenarios. Due to the high sensitivity of individuals to hand motions, even minor errors in hand motions can significantly impact the user experience. Real applications often involve multiple avatars with varying hand shapes, highlighting the importance of maintaining the intricate semantics of hand motions across the avatars. Therefore, this paper aims to transfer the hand motion semantics between diverse avatars based on their respective hand models. To address this problem, we introduce a novel anatomy-based semantic matrix (ASM) that encodes the semantics of hand motions. The ASM quantifies the positions of the palm and other joints relative to the local frame of the corresponding joint, enabling precise retargeting of hand motions. Subsequently, we obtain a mapping function from the source ASM to the target hand joint rotations by employing an anatomy-based semantics reconstruction network (ASRN). We train the ASRN using a semi-supervised learning strategy on the Mixamo and InterHand2.6M datasets. We evaluate our method in intra-domain and cross-domain hand motion retargeting tasks. The qualitative and quantitative results demonstrate the significant superiority of our ASRN over the state-of-the-arts. Code available at *Semantics2Hands*.

CCS CONCEPTS

• **Computing methodologies** → **Motion processing**; • **Applied computing** → **Media arts**.

KEYWORDS

Hand Motion Retargeting, Neural Motion Processing

ACM Reference Format:
Zijie Ye, Jia Jia, and Junliang Xing. 2023. Semantics2Hands: Transferring Hand Motion Semantics between Avatars. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3612703

*Corresponding author.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3612703

1 INTRODUCTION

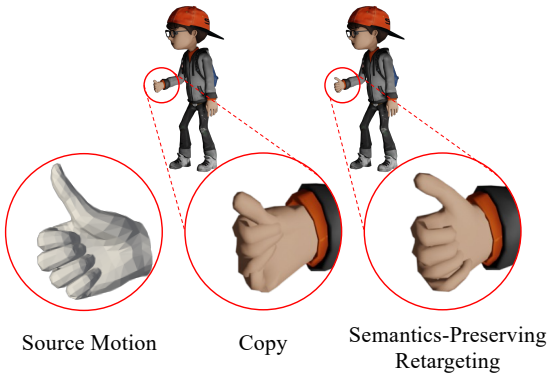

Figure 1: Despite the accurate body motions, errors introduced by copying finger joint rotations make the "thumb-up" gesture illegible.

The generation of realistic hand motions has demonstrated promising potential in diverse virtual avatar scenarios, including co-speech gesture synthesis [25, 27, 38] and sign language synthesis [14, 30, 39]. Human hands, being the primary means of non-verbal communication [31], convey subtle nuances during the execution of particular hand gestures. Given people's high sensitivity to hand motions, even slight errors can significantly impact the user experience in virtual avatar applications. Consequently, maintaining consistent hand motion semantics across various virtual avatar hands is paramount. However, due to the highly articulated nature of the human hand with multiple degrees of freedom (DoFs) and the varying hand shapes and proportions of different avatars, directly copying joint rotations would significantly compromise the intricate semantics of hand motions, as shown in Figure 1. Consequently, developing a methodology that can preserve the semantics of hand motions when retargeting them to diverse avatars is essential.

Previous research has focused on motion retargeting and hand-object interaction. Motion retargeting, pioneered by Gleicher [11], aims to identify the characteristics of source motions and transfer them to target motions on different characters. Early work [3, 9, 19] focused on optimization-based approaches. Recently, researchers have proposed data-driven approaches [1, 35, 41] using various network architectures and semantic measurements. These approaches can successfully retarget realistic body motions but do not apply to dexterous hand motion retargeting. Ge et al. [10] proposed a

rule-based approach for retargeting sign language motions; however, their method is limited to a specific set of pre-defined hand movements and lacks sufficient testing. Hand-object interaction is a research area that focuses on synthesizing realistic hand motions during interactions with objects, including static grasp synthesis [12, 34, 45] and manipulation motion synthesis [24, 37, 40, 43]. However, these methods fail to preserve the semantics of hand motions in communication scenarios. Furthermore, they do not apply to diverse hand models with varying shapes and proportions. Despite the existing methods, it remains a challenge: retarget realistic hand motions with high fidelity across different hand models while preserving intricate motion semantics.

This paper focuses on retargeting dexterous hand motions across different hand models while preserving the semantics of the source hand motions. Hand motion retargeting requires a higher level of semantic measurement precision than body motion retargeting, making this idea novel. The semantic measurements previously employed in motion retargeting, including cycle consistency [1, 35] and distance matrix [41], are inadequate due to the high density of hand joints within a limited space, which results in significant spatial interactions between finger joints and the palm.

Therefore, our central insight is that the spatial relationships between the finger joints and the palm are crucial for preserving hand motion semantics. Consequently, we encode the spatial relationships into a novel anatomy-based semantic matrix (ASM). We utilize ASM as the semantic measurement for precise hand motion retargeting. In particular, we first build anatomical local coordinate frames for finger joints on different hand models. Then we construct ASM based on the anatomical local coordinate frames. ASM quantifies the positions of the palm and other joints relative to the local frame of the given finger joint. Next, we acquire a mapping function from the source motion ASM to the target motion rotations using an anatomy-based semantics reconstruction network (ASRN). We train ASRN on two heterogeneous hand motion datasets [2, 23]. Unlike template mesh-based methods [40, 43] for semantic correspondence, our approach is not dependent on template meshes and can be applied to various hand models.

We conducted comprehensive experiments to assess the quality of the hand motions generated by our ASRN. These experiments encompassed both intra-domain and cross-domain hand motion retargeting scenarios involving intricate hand motion sequences and a diverse range of hand shapes. The qualitative and quantitative results show that our ASRN outperforms existing motion retargeting methods by a large margin.

To summarize, our contributions are as three-fold:

- We propose a novel task: semantics-preserving retargeting of dexterous hand motions across diverse hand models.
- We introduce an anatomy-based semantic matrix (ASM) that quantifies hand motion semantics without relying on any template mesh, making it applicable to various hand models.
- We propose a novel framework for semantics-preserving hand motion retargeting, leveraging the ASM. Experimental results on both intra-domain and cross-domain hand motion retargeting tasks validate the superior performance of our framework over existing methods.

2 RELATED WORK

2.1 Motion Retargeting

Motion retargeting aims to identify the features of the source motions and transfer them to the target motions on a different character. The pioneering work by Gleicher [11] addresses motion retargeting as a spatial-temporal optimization problem with the source motion features as kinematic constraints. Subsequent studies propose solutions to this optimization problem with various constraints [3, 5, 19, 33].

Recently, data-driven methods [1, 7, 15, 20, 35, 41, 44] have become increasingly appealing due to the growing availability of motion capture data. Delhaisse et al. [7] and Jang et al. [15] train neural networks for retargeting using paired training data. Subsequently, Villegas et al. [35] develop an adversarial neural network trained with cycle consistency [44], eliminating the need for paired ground truth. Aberman et al. [1] propose a skeleton-aware network for retargeting motions between skeletons with varying topologies. Zhang et al. [41] also introduces the Distance Matrix for measuring body motion semantics.

However, all the methods above either truncate finger movements or merely replicate finger joint rotations during retargeting, resulting in the loss of intricate semantics in dexterous hand motions. In contrast, our framework carefully measures the hand motion semantics with an anatomy-based semantic matrix (ASM), and transfers these semantics to the target hand motion through a novel anatomy-based semantics reconstruction network (ASRN).

2.2 Hand-object Interaction Synthesis

The synthesis of hand grasping given an object has been extensively studied in robotics [4, 8, 29]. Recently, several data-driven methods have been proposed [6, 12, 32, 45]. Among these methods, Karunratanakul et al. [17] and Jiang et al. [16] propose to represent the proximity between the hand and the object as an implicit function.

Object manipulation synthesis involves dynamic hand and object interaction, which makes it more relevant to our research. Previous researchers have tackled this issue by optimizing hand poses to meet different constraints [22, 24, 37, 42]. In a recent study, Zhang et al. [40] employed hand-object spatial representations to learn object manipulation using motion capture data. Subsequently, Zhou et al. [43] devised a different object-centric spatiotemporal representation.

However, these representations cannot capture the semantics of hand motion as they neglect the interaction between the palm and the fingers. Furthermore, these representations are explicitly designed for a given template hand mesh, which restricts their applicability to different hand models. In contrast, our ASM quantifies hand motion semantics without depending on a template mesh, allowing its application to diverse hand models.

3 PROBLEM FORMULATION

This paper aims to learn a mapping function f that transfers the source hand motion to the target hand while preserving the semantics of the source hand motion. The inputs to the function are the source joint rotation sequence $\mathbf{Q}_\mathrm{A}$, the source hand shape parameter $\mathbf{H}_\mathrm{A}$, the source hand anatomical parameter $\mathbf{M}_\mathrm{A}^\mathrm{rest}$, the target hand

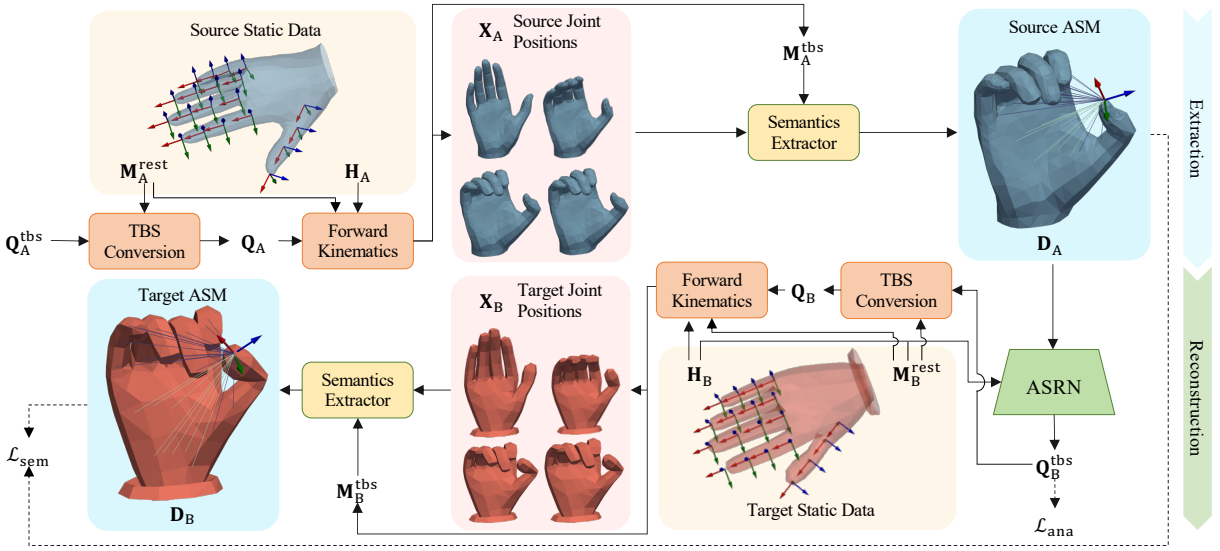

Figure 2: The figure presents an overview of the proposed pipeline consisting of two stages. The extraction stage involves the retrieval of ASM from the source hand motion. The reconstruction stage utilizes the source ASM, target hand shape parameter, and target hand anatomical parameter to reconstruct the target hand motion.

shape parameter $\mathbf{H}_B$, and the target hand anatomical parameter $\mathbf{M}_B^{rest}$. The mapping function can be formulated as follows:

$$f(\mathbf{Q}_A, \mathbf{H}_A, \mathbf{M}_A^{rest}, \mathbf{H}_B, \mathbf{M}_B^{rest}) \implies \mathbf{Q}_B, \tag{1}$$

where $\mathbf{Q}_B$ is the target joint rotation sequence.

4 METHODOLOGY

Based on the formulation in Section 3, we have developed a framework for retargeting hand movements, as depicted in Figure 2. We introduce a novel anatomy-based semantic matrix (ASM) based on the finger anatomical coordinate frame. By utilizing the ASM, we train an anatomy-based semantics reconstruction network (ASRN) to predict the target joint rotation sequence using the source ASM, target hand shape parameter, and target hand anatomical parameter.

In the subsequent subsections, we briefly introduce the anatomical coordinate frame of finger movements, as outlined in Section 4.1. Next, we elaborate on the definition of the ASM in Section 4.2. Finally, we describe the framework pipeline and training details in Section 4.3.

4.1 *Twist-bend-splay* Frame

The human hand exhibits a high degree of articulation. Directly predicting rotations of all 15 finger joints can lead to abnormal hand postures. Previous works [21, 36] suggest that constraints can be applied to the finger joint rotations to prevent abnormal hand movements. Yang et al. [36] extended MANO [28] to develop a hand model called A-MANO incorporating anatomical constraints. A-MANO assigns a Cartesian coordinate frame, known as the *Twist-bend-splay* frame, to each joint in the hand's kinematic tree. The frame's *x*, *y*, and *z* axes align with the three revolute directions:

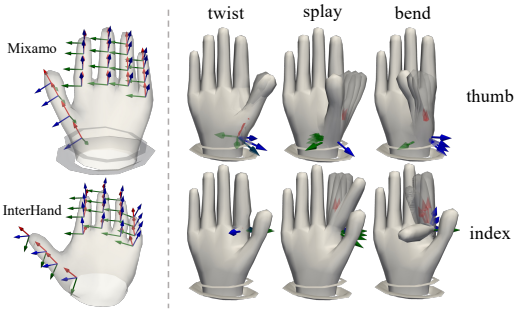

Figure 3: Left: *Twist-bend-splay* frames obtained from different hand models using our annotation tool. Right: Finger movements in the *twist*, *splay*, and *bend* directions. Note that the *bend* and *splay* directions of the thumb joints differ significantly from those of the other four fingers.

twist, *bend*, and *splay*, based on hand anatomy. Most finger joints have only one degree of freedom (DoF) along the *bend* axis.

While A-MANO shows promise in estimating MANO pose during hand-object interaction, it does not apply to hand models from external sources, such as the hands of Mixamo [2] characters. To mitigate this problem, we develop a tool for annotating the *Twist-bend-splay* frames of different hand models. Figure 3 demonstrates that our tool can readily provide the *Twist-bend-splay* frames for hands obtained from both InterHand2.6M [23] and Mixamo [2]. Details of our annotation tool can be found in Appendix A.

4.2 Anatomy-based Semantic Matrix

Our framework aims to preserve the intricate semantics while retargeting hand motions between hand models from different sources.

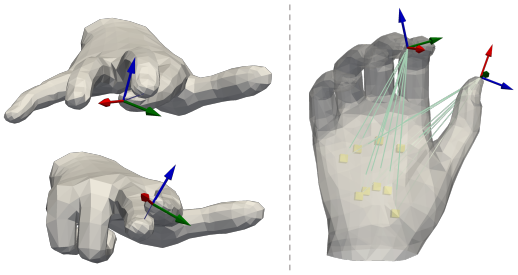

Figure 4: Left: The inter-finger semantic features capture the subtle semantics of finger movements. Right: The palm-finger semantic features capture the overall hand posture. Yellow cubes represent the palm anchors.

This paper defines hand motion semantics as the spatial relationships between the fingers and the palm. Due to the absence of paired ground truth with intense semantic supervision, we introduce a novel anatomy-based semantic matrix (ASM) as a semantic measurement for hand motion retargeting. Compared to existing semantic measurements in body motion retargeting [1, 35, 41] and object manipulation synthesis [40, 43], the proposed ASM captures the intricate semantics of hand motions and can be applied to hand models from different sources without any additional cost.

Our ASM is constructed based on the *twist-bend-splay* frame introduced in Section 4.1. The crucial insight behind constructing the ASM lies in that the orientation of the *twist-bend-splay* frame reveals the finger's structure. As shown in Figure 4, the *splay* axis (blue axis) extends from the finger pulp to the back surface of the finger, while the *bend* axis (green axis) stretches from the right side to the left side of the finger. The *twist* axis also aligns with the finger bone. In this scenario, we can deduce the spatial relationships between the middle fingertip and the index fingertip based on the coordinates of the middle fingertip within the local *twist-bend-splay* frame of the index fingertip.

The proposed ASM applies to hand models composed of five fingers, each consisting of four joints (including a dummy fingertip joint). The semantic matrix comprises two components: inter-finger semantic features and palm-finger semantic features. Formally, at time t, the coordinates of the k-th finger joint within the global frame are represented as ${}^{\mathrm{g}}\mathrm{x}_k \in \mathbb{R}^3$. ${}^{\mathrm{g}}\mathbf{M}_k$ represents the rotation matrix of the *twist-bend-splay* frame of joint k within the global frame. The coordinates of another joint m within the local frame of joint k are given by ${}^{k}\mathrm{x}_m = {}^{\mathrm{g}}\mathbf{M}_k^{\mathrm{T}}({}^{\mathrm{g}}\mathrm{x}_m - {}^{\mathrm{g}}\mathrm{x}_k)$. We define ${}^{k}\mathrm{x}_m$ as the inter-finger semantic feature of joint m concerning joint k. Additionally, we introduce the palm-finger semantic feature to capture the overall hand posture, as depicted in Figure 4. Inspired by Yang et al. [36], we define nine palm anchors along the line connecting the metacarpophalangeal and wrist joints. We denote the palm-finger semantic feature of the n-th anchor with respect to joint k as ${}^{k}\mathrm{x}_{\mathrm{p}_n} = {}^{\mathrm{g}}\mathbf{M}_k^{\mathrm{T}}({}^{\mathrm{g}}\mathrm{x}_{\mathrm{p}_n} - {}^{\mathrm{g}}\mathrm{x}_k)$, where ${}^{\mathrm{g}}\mathrm{x}_{\mathrm{p}_n}$ represents the coordinates of the n-th anchor within the global frame. By combining the inter-finger semantic features and the palm-finger semantic features, we can construct the semantic matrix for joint k as:

$$ {}^{k}\mathbf{D} = [{}^{k}\mathrm{x}_1, {}^{k}\mathrm{x}_2, \ldots, {}^{k}\mathrm{x}_{20}, {}^{k}\mathrm{x}_{\mathrm{p}_1}, {}^{k}\mathrm{x}_{\mathrm{p}_2}, \ldots, {}^{k}\mathrm{x}_{\mathrm{p}_9}] \in \mathbb{R}^{29\times 3}. \quad (2) $$

By having semantic matrices for all 20 finger joints, we obtain the semantic measurement of the entire hand model without relying on any standard mesh template.

4.3 Semantics-Preserving Retargeting

The hand retargeting pipeline comprises two stages: semantic feature extraction and semantics-preserving reconstruction. We extract semantic matrices from the source hand motion during the first stage. In the second stage, we employ the anatomy-based semantics reconstruction network (ASRN) to reconstruct hand motion on the target hand model from the source ASM while preserving the source semantics. The overall pipeline is depicted in Figure 2.

In the semantic feature extraction stage, the T-frame hand motion sequence in the *twist-bend-splay* frame, represented as quaternions of the 15 finger joints, is denoted as $\mathbf{Q}_{\mathrm{A}}^{\mathrm{tbs}} \in \mathbb{R}^{T\times 15\times 4}$. After converting $\mathbf{Q}_{\mathrm{A}}^{\mathrm{tbs}}$ to the global frame using the rest orientation of the joint *twist-bend-splay* frames $\mathbf{M}_{\mathrm{A}}^{\mathrm{rest}} \in \mathbb{R}^{15\times 3\times 3}$, we obtain $\mathbf{Q}_{\mathrm{A}} \in \mathbb{R}^{T\times 15\times 4}$. We then perform forward kinematics (FK) to derive the global coordinates of the finger joints $\mathbf{X}_{\mathrm{A}} \in \mathbb{R}^{T\times 20\times 3}$ and the global orientation of the *twist-bend-splay* frames $\mathbf{M}_{\mathrm{A}}^{\mathrm{tbs}} \in \mathbb{R}^{T\times 20\times 3\times 3}$. It is important to note that the FK results include the dummy fingertip joints. Additionally, the shape parameter $\mathbf{H}_{\mathrm{A}} \in \mathbb{R}^{h_{\mathrm{A}}}$ takes different forms depending on the model type. In the case of MANO models, $\mathbf{H}_{\mathrm{A}}$ represents shape PCA coefficients published by Romero et al. [28], while for Mixamo models, $\mathbf{H}_{\mathrm{A}}$ corresponds to the normalized finger joint offsets. Finally, we extract the semantic matrices $\mathbf{D}_{\mathrm{A}} = [{}^{1}\mathbf{D}_{\mathrm{A}}, {}^{2}\mathbf{D}_{\mathrm{A}}, \ldots, {}^{20}\mathbf{D}_{\mathrm{A}}] \in \mathbb{R}^{20\times T\times 29\times 3}$ from $\mathbf{X}_{\mathrm{A}}$ and $\mathbf{M}_{\mathrm{A}}^{\mathrm{tbs}}$ using Equation 2, where ${}^{k}\mathbf{D}_{\mathrm{A}}$ is the concatenation of ${}^{k}\mathbf{D}$ in T frames.

Having obtained the semantic matrices $\mathbf{D}_{\mathrm{A}}$ from the source hand motion, we utilize our ASRN to reconstruct the target hand motion $\mathbf{Q}_{\mathrm{B}}^{\mathrm{tbs}} \in \mathbb{R}^{T\times 15\times 4}$ on the target hand model B. A ResNet-like [13] architecture is employed. Consecutive 1D ResNet layers process the source ASM $\mathbf{D}_{\mathrm{A}}$. Additionally, ASRN receives the target hand shape parameter $\mathbf{H}_{\mathrm{B}}$ and the target hand local frame rest orientation $\mathbf{M}_{\mathrm{B}}^{\mathrm{rest}}$ as inputs. An MLP encodes $\mathbf{H}_{\mathrm{B}}$ and $\mathbf{M}_{\mathrm{B}}^{\mathrm{rest}}$ initially, followed by concatenation with the input of each ResNet layer. The output of the final ResNet layer is used as input for a fully-connected layer, which predicts the target hand joint rotation $\mathbf{Q}_{\mathrm{B}}^{\mathrm{tbs}}$ in target hand *twist-bend-splay* frames. Next, we extract semantic matrices $\mathbf{D}_{\mathrm{B}}$ from the generated hand motion. In this work, hand motion semantics preservation is modeled as preserving spatial relationships between the fingers and the palm. This design defines the semantic loss $\mathcal{L}_{\mathrm{sem}}$ as the weighted cosine similarity between the source and target semantic matrices:

$$ \mathcal{L}_{\mathrm{sem}} = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{20}\sum_{k=1}^{29}\omega_{jk}\frac{\mathbf{D}_{\mathrm{A}}^{j,t,k}\cdot\mathbf{D}_{\mathrm{B}}^{j,t,k}}{||\mathbf{D}_{\mathrm{A}}^{j,t,k}||_2||\mathbf{D}_{\mathrm{B}}^{j,t,k}||_2}, \quad (3) $$

where the weight ω_{jk} is defined as:

$$ \omega_{jk} = \begin{cases} 1 + \frac{\exp(-||\mathbf{D}_{\mathrm{A}}^{j,t,k}||_2)}{\sum_{m=1}^{20}\exp(-||\mathbf{D}_{\mathrm{A}}^{j,t,m}||_2)} & \text{if } k \in \{1,2,\ldots,20\} \\ 1 & \text{if } k \in \{21,22,\ldots,29\}. \end{cases} \quad (4) $$

This weighting scheme encourages the network to focus on close-finger interactions.

To mitigate abnormal hand postures generated by our network, we propose an anatomical loss, denoted as $\mathcal{L}_{\text{ana}}$. $\mathbf{Q}_{\text{B}}^{\text{tbs}}$ is decomposed into three Euler angles: ϕ_{twist}, ϕ_{bend}, and ϕ_{splay}, aligned with the local *twist-bend-splay* frame axes. Initially, we apply a penalty to ϕ_{twist} for all the joints along the hand's kinematic tree. Additionally, a penalty is imposed on ϕ_{splay} if it exceeds the acceptable range. Finally, we penalize the rotation angle ϕ_{bend} if it exceeds $\pi/2$ or falls below 0. The anatomical loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{ana}} =& \frac{1}{T}\sum_{t=1}^{T}\Big(\sum_{j\in\text{all}}|\phi_{\text{twist}}^{t,j}|^2+\sum_{j\notin\text{knuckle}}|\phi_{\text{splay}}^{t,j}|^2\\ &+\sum_{j\in\text{knuckle}}\max(|\phi_{\text{splay}}^{t,j}|-\pi/18,0)^2\\ &+\sum_{j\in\text{all}}\max(\phi_{\text{bend}}^{t,j}-\pi/2,0)^2+\sum_{j\in\text{all}}\min(\phi_{\text{bend}}^{t,j},0)^2\Big).\end{aligned} \tag{5}$$

Since our network is trained on hand motion data from different hand models, the self-reconstruction supervision signals are only available when A and B belong to the same character. Therefore, ASRN is trained by minimizing the following loss function:

$$\mathcal{L}_{\text{total}} = \mathbb{1}_{\text{A=B}}\cdot\text{MSE}(\mathbf{Q}_{\text{A}}^{\text{tbs}},\mathbf{Q}_{\text{B}}^{\text{tbs}})-\lambda_{\text{sem}}\mathcal{L}_{\text{sem}}+\lambda_{\text{ana}}\mathcal{L}_{\text{ana}}, \tag{6}$$

where λ_{sem} and λ_{ana} are hyper-parameters. The indicator function $\mathbb{1}_{\text{A=B}}$ takes the value 1 if A and B belong to the same character, and 0 otherwise.

5 EXPERIMENTS

5.1 Datasets

The evaluation of our framework encompasses both the Mixamo dataset [2] and the InterHand2.6M dataset [23]. The Mixamo dataset comprises animations performed by various virtual characters with different shapes; however, the dataset does not guarantee consistent hand motion quality and diversity. The InterHand2.6M dataset is a comprehensive collection of hand motion data captured using a multi-view camera system and supplemented with MANO [28] hand pose annotations. While the InterHand2.6M dataset offers high-quality hand motion data with considerable diversity, it has limitations regarding hand shape variations. During the training phase, we gathered 40,903 frames of hand motion data from nine distinct characters. In the testing phase, we obtained 14,316 frames of hand motion data from four different characters, ensuring that none of the testing characters were present during the network's training.

5.2 Implementation Details

The hyper-parameters λ_{sem} and λ_{ana} are set to 1.0 and 0.1 respectively. The network is trained for 100 epochs with a batch size of 64. We use the Adam optimizer [18] with the learning rate set to 10^{-4} to train the network. The input to the network is a sequence of 8 frames with a frame rate of 5 fps. The network is implemented in PyTorch [26] and trained on a single NVIDIA RTX 2080 Ti GPU. Further details can be found in Appendix B.

5.3 Evaluation Metrics

For hand motions with paired ground truth (GT) on different characters, we use Mean Square Error (MSE) to measure how close the retargeted joint positions are to the paired GT. In the absence of paired GT, the following metrics are used to evaluate the quality of the retargeted hand motions:

$$\begin{aligned}S_{\text{palm}} &= \frac{1}{20\times 9\times T}\sum_{t=1}^{T}\sum_{j=1}^{20}\sum_{k=21}^{29}\frac{\mathbf{D}_{\text{A}}^{j,t,k}\cdot\mathbf{D}_{\text{B}}^{j,t,k}}{||\mathbf{D}_{\text{A}}^{j,t,k}||_2||\mathbf{D}_{\text{B}}^{j,t,k}||_2},\\ S_{\text{finger}} &= \frac{1}{20\times 20\times T}\sum_{t=1}^{T}\sum_{j=1}^{20}\sum_{k=1}^{20}\frac{\mathbf{D}_{\text{A}}^{j,t,k}\cdot\mathbf{D}_{\text{B}}^{j,t,k}}{||\mathbf{D}_{\text{A}}^{j,t,k}||_2||\mathbf{D}_{\text{B}}^{j,t,k}||_2}.\end{aligned} \tag{7}$$

S_{palm} and S_{finger} represent the average cosine similarity between the retargeted hand motion and the GT hand motion. Higher values indicate better preservation of the original spatial relationships between the fingers and the palm in the retargeted hand motion.

5.4 Qualitative Results

The results of hand motion retargeting among hands with various shapes are depicted in Figure 5. The TBS Copy method copies $\mathbf{Q}_{\text{A}}^{\text{tbs}}$ to $\mathbf{Q}_{\text{B}}^{\text{tbs}}$, while the Copy method copies $\mathbf{Q}_{\text{A}}$ to $\mathbf{Q}_{\text{B}}$. The DM method replaces our proposed ASM with the distance matrices proposed by Zhang et al. [41]. During training, the network did not encounter any of the source or target hands in the last row. Existing methods barely account for the intricate spatial relationships between the fingers and the palm, leading to inconsistent and unnatural hand motions. In contrast, our method effectively preserves the spatial relationships between the fingers and the palm, resulting in hand motions that are more natural and preserve semantics. Figure 10 shows the detailed spatial relationships in the results of our method.

5.5 Quantitative Results

Table 1 shows comparison between our method and existing body motion retargeting techniques. We compare the methods across three tasks with different sources and targets: Mixamo to Mixamo (**MX2MX**), InterHand to Mixamo (**IH2MX**), and Mixamo to InterHand (**MX2IH**). Because the Mixamo dataset provides paired GT, we use MSE to assess the quality of the retargeted hand motions for the **MX2MX** task. For the other two cross-domain tasks, we utilize S_{palm} and S_{finger} as metrics for the quality of the retargeted hand motions.

Table 1: Comparison with the state-of-the-arts. Ours$_{w/o\,\mathcal{L}_{ana}}$ is the model without anatomical loss in Equation 5. Ours$_{w/o\text{weight}}$ is the model without the weight scheme in Equation 4.

Methods	MX2MX	IH2MX		MX2IH	
	MSE↓	S_{palm}↑	S_{finger}↑	S_{palm}↑	S_{finger}↑
Copy	**4.76e-12**	0.923	0.851	0.941	0.872
TBS Copy	0.155	0.960	0.883	0.968	0.891
SAN [1]	3.134	0.866	0.820	0.034	0.475
DM [41]	2.788	0.888	0.832	0.891	0.878
Ours$_{w/o\,\mathcal{L}_{ana}}$	0.276	**0.983**	**0.932**	**0.985**	**0.935**
Ours$_{w/o\text{weight}}$	0.420	0.972	0.922	0.980	0.927
Ours	0.452	0.971	0.925	0.978	0.929

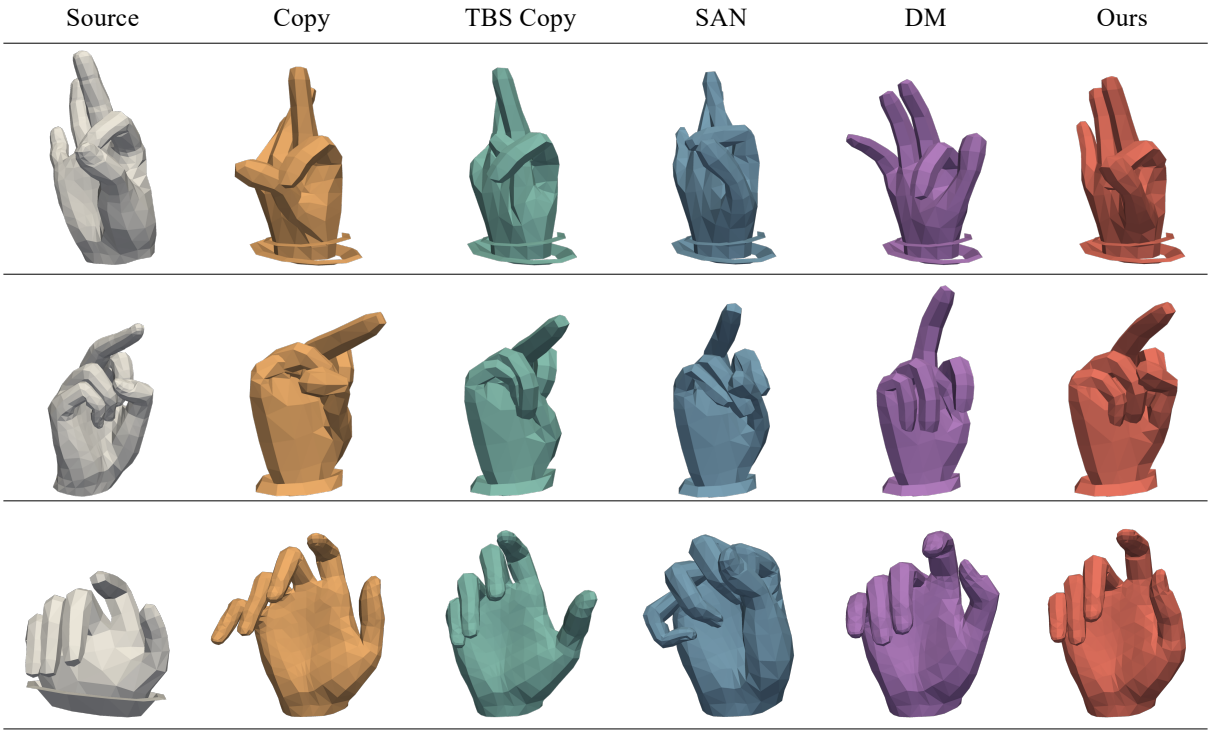

Figure 5: Qualitative comparison between the proposed framework and the state-of-the-art methods.

Because the Mixamo dataset may create a new character with an archived motion by using motion copy, the Copy method has the lowest MSE. However, as the qualitative results reveal, this does not mean the motion copy is optimal. Our method achieves a reduction in MSE of 85.6% and 83.8% compared to SAN [1] and DM [41], which utilize distinct semantic measurements. Additionally, our method achieves the highest S_{palm} and S_{finger} in the **IH2MX** and **MX2IH** tasks, indicating that its superior ability to preserve the original spatial relationships between the fingers and the palm during the retargeted hand motion. This observation suggests that our proposed ASM outperforms the distance matrices [41] and the implicit measurement learned in SAN [1].

5.6 User Study

We conduct a user study to evaluate the performance of our framework against Copy, SAN [1], and DM [41]. We invited 26 participants and showed them six static hand posture pictures and six hand motion videos. Each picture and video contains one source motion and four anonymous results. Participants were instructed to rank the pictures and videos based on three aspects: preservation of static posture semantics (PS), preservation of motion semantics (MS), and motion quality (MQ), from best to worst. The average rankings are presented in Table 2. Overall, our method achieved the best performance in all three aspects.

Table 2: Ranking results of the user study. We invite 26 participants to compare the retargeting results from three aspects: static posture semantics (PS), motion semantics (MS), and motion quality (MQ).

Methods	Ranking		
	PS ↓	MS ↓	MQ ↓
Copy	3.17	3.31	3.33
SAN [1]	2.95	2.99	3.05
DM [41]	2.70	2.53	2.46
Ours	**1.17**	**1.17**	**1.16**

6 CONCLUSION

In this paper, we propose the problem of semantics-preserving hand motion retargeting. We encode the spatial relationships between the fingers and the palm using anatomy-based semantic matrices (ASM). We train an anatomy-based semantics reconstruction network (ASRN) to retarget the motion semantics of the source hand onto the target hand, utilizing the source ASM. We evaluate our framework on both intra-domain and cross-domain retargeting tasks. Our method demonstrates superior performance to existing motion retargeting methods, both qualitatively and quantitatively.

ACKNOWLEDGEMENTS

This work is supported by the National Key R&D Program of China under Grant No. 2021QY1500, the State Key Program of the National Natural Science Foundation of China (NSFC) (No.61831022). It is also supported in part by the NSFC under Grant No. 62222606 and 62076238. Thank Cuiwen for her support and encouragement.

REFERENCES

[1] Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. 2020. Skeleton-aware networks for deep motion retargeting. *ACM Trans. Graph.* 39, 4 (2020), 62. https://doi.org/10.1145/3386569.3392462

[2] Adobe. 2018. Mixamo. https://www.mixamo.com/.

[3] Antonin Bernardin, Ludovic Hoyet, Antonio Mucherino, Douglas Gonçalves, and Franck Multon. 2017. Normalized Euclidean distance matrices for human motion retargeting. In *Proceedings of the 10th International Conference on Motion in Games*. 1–6.

[4] Jeannette Bohg, Antonio Morales, Tamim Asfour, and Danica Kragic. 2013. Data-driven grasp synthesis—a survey. *IEEE Transactions on robotics* 30, 2 (2013), 289–309.

[5] Kwang-Jin Choi and Hyeong-Seok Ko. 2000. Online motion retargetting. *The Journal of Visualization and Computer Animation* 11, 5 (2000), 223–235.

[6] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. 2020. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5031–5041.

[7] Brian Delhaisse, Domingo Esteban, Leonel Rozo, and Darwin Caldwell. 2017. Transfer learning of shared latent spaces between robots with similar kinematic structure. In *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 4142–4149.

[8] S El-Khoury, A Sahbani, and P Bidaud. 2011. 3d objects grasps synthesis: A survey. In *13th World Congress in Mechanism and Machine Science*. 573–583.

[9] Andrew Feng, Yazhou Huang, Yuyu Xu, and Ari Shapiro. 2012. Automating the transfer of a generic set of behaviors onto a virtual character. In *Motion in Games: 5th International Conference, MIG 2012, Rennes, France, November 15-17, 2012. Proceedings 5*. Springer, 134–145.

[10] Chunbao Ge, Yiqiang Chen, Changshui Yang, Baocai Yin, and Wen Gao. 2005. Motion Retargeting for the Hand Gesture. (2005).

[11] Michael Gleicher. 1998. Retargetting motion to new characters. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*. 33–42.

[12] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. 2019. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11807–11816.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[14] Wencan Huang, Wenwen Pan, Zhou Zhao, and Qi Tian. 2021. Towards fast and high-quality sign language production. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3172–3181.

[15] Hanyoung Jang, Byungjun Kwon, Moonwon Yu, Seong Uk Kim, and Jongmin Kim. 2018. A variational u-net for motion retargeting. In *SIGGRAPH Asia 2018 Posters*. 1–2.

[16] Zhenyu Jiang, Yifeng Zhu, Maxwell Svetlik, Kuan Fang, and Yuke Zhu. 2021. Synergies Between Affordance and Geometry: 6-DoF Grasp Detection via Implicit Representations. In *Robotics: Science and Systems XVII, Virtual Event, July 12-16, 2021*, Dylan A. Shell, Marc Toussaint, and M. Ani Hsieh (Eds.). https://doi.org/10.15607/RSS.2021.XVII.024

[17] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. 2020. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*. IEEE, 333–344.

[18] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[19] Jehee Lee and Sung Yong Shin. 1999. A hierarchical approach to interactive motion editing for human-like figures. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. 39–48.

[20] Jongin Lim, Hyung Jin Chang, and Jin Young Choi. 2019. PMnet: Learning of Disentangled Pose and Movement for Unsupervised Motion Retargeting.. In *BMVC*, Vol. 2. 7.

[21] John Lin, Ying Wu, and Thomas S Huang. 2000. Modeling the constraints of human hand motion. In *Proceedings workshop on human motion*. IEEE, 121–126.

[22] C Karen Liu. 2009. Dextrous manipulation from a grasping pose. In *ACM SIGGRAPH 2009 papers*. 1–6.

[23] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. 2020. InterHand2.6M: A Dataset and Baseline for 3D Interacting Hand Pose Estimation from a Single RGB Image. In *European Conference on Computer Vision (ECCV)*.

[24] Igor Mordatch, Zoran Popović, and Emanuel Todorov. 2012. Contact-invariant optimization for hand manipulation. In *Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation*. 137–144.

[25] Evonne Ng, Shiry Ginosar, Trevor Darrell, and Hanbyul Joo. 2021. Body2hands: Learning to infer 3d hands from conversational gesture body dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11865–11874.

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[27] Xingqun Qi, Chen Liu, Muyi Sun, Lincheng Li, Changjie Fan, and Xin Yu. 2023. Diverse 3D Hand Gesture Prediction from Body Dynamics by Bilateral Hand Disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4616–4626.

[28] Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied hands: modeling and capturing hands and bodies together. *ACM Trans. Graph.* 36, 6 (2017), 245:1–245:17. https://doi.org/10.1145/3130800.3130883

[29] Anis Sahbani, Sahar El-Khoury, and Philippe Bidaud. 2012. An overview of 3D object grasp synthesis algorithms. *Robotics and Autonomous Systems* 60, 3 (2012), 326–336.

[30] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2022. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5141–5151.

[31] Michael Studdert-Kennedy. 1994. Hand and Mind: What Gestures Reveal About Thought. *Language and Speech* 37, 2 (1994), 203–209.

[32] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. 2020. GRAB: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 581–600.

[33] Seyoon Tak and Hyeong-Seok Ko. 2005. A physically-based motion retargeting filter. *ACM Transactions on Graphics (TOG)* 24, 1 (2005), 98–117.

[34] Jonathan Taylor, Lucas Bordeaux, Thomas Cashman, Bob Corish, Cem Keskin, Toby Sharp, Eduardo Soto, David Sweeney, Julien Valentin, Benjamin Luff, et al. 2016. Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. *ACM Transactions on Graphics (ToG)* 35, 4 (2016), 1–12.

[35] Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. 2018. Neural kinematic networks for unsupervised motion retargetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8639–8648.

[36] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. 2021. CPF: Learning a contact potential field to model the hand-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11097–11106.

[37] Yuting Ye and C Karen Liu. 2012. Synthesis of detailed hand manipulations using contact sampling. *ACM Transactions on Graphics (ToG)* 31, 4 (2012), 1–10.

[38] Zijie Ye, Jia Jia, Haozhe Wu, Shuo Huang, Shikun Sun, and Junliang Xing. 2023. Salient Co-Speech Gesture Synthesizing with Discrete Motion Representation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[39] Jan Zelinka and Jakub Kanis. 2020. Neural sign language synthesis: Words are our glosses. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3395–3403.

[40] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. 2021. ManipNet: neural manipulation synthesis with a hand-object spatial representation. *ACM Trans. Graph.* 40, 4 (2021), 121:1–121:14. https://doi.org/10.1145/3450626.3459830

[41] Jiaxu Zhang, Junwu Weng, Di Kang, Fang Zhao, Shaoli Huang, Xuefei Zhe, Linchao Bao, Ying Shan, Jue Wang, and Zhigang Tu. 2023. Skinned Motion Retargeting with Residual Perception of Motion Semantics & Geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13864–13872.

[42] Wenping Zhao, Jianjie Zhang, Jianyuan Min, and Jinxiang Chai. 2013. Robust realtime physics-based motion control for human grasping. *ACM Transactions on Graphics (TOG)* 32, 6 (2013), 1–12.

[43] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. 2022. TOCH: Spatio-Temporal Object-to-Hand Correspondence for Motion Refinement. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 13663)*, Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer, 1–19. https://doi.org/10.1007/978-3-031-20062-5_1

[44] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.

[45] Tianqiang Zhu, Rina Wu, Xiangbo Lin, and Yi Sun. 2021. Toward human-like grasp: Dexterous grasping via semantic representation of object-hand. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15741–15751.

A *TWIST-BEND-SPLAY* FRAME ANNOTATION

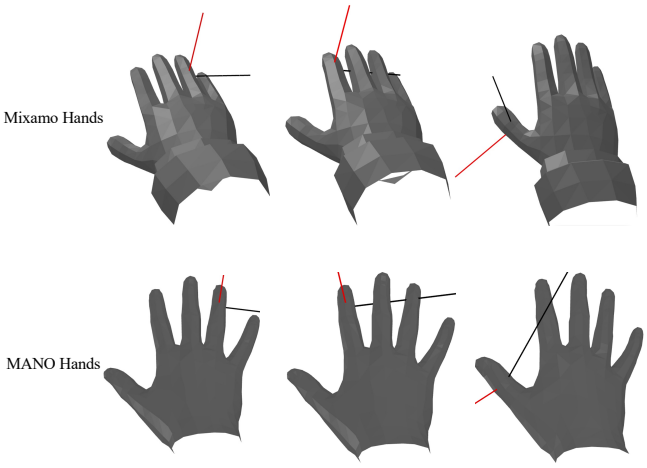

Figure 6: Our annotation tool allows the user to adjust the *splay* axis (red axis) and *bend* axis (black axis) directions for Mixamo and MANO hands.

This section presents our frame annotation tool for *Twist-bend-splay*. A previous study by Yang et al. [36] introduced A-MANO, a hand model that incorporates *Twist-bend-splay* frames. A-MANO, an extension of MANO, is limited in its applicability to other hand models. This paper presents the implementation of a versatile frame annotation tool for *Twist-bend-splay*, applicable to any hand model with five fingers and 15-finger joints. The annotation tool can semi-automatically derive the frame orientation of finger joints for *Twist-bend-splay* from the model's kinematic tree and mesh information.

Specifically, our annotation tool first computes the twist axis $\mathbf{n}_{\text{twist}}$ as the vector from the child of the current joint to the joint itself. Next, we project rays onto the normal plane defined by $\mathbf{n}_{\text{twist}}$ and perform ray-mesh queries. The ray-mesh hit locations on the mutually perpendicular axes $\mathbf{n}_{\text{splay}}$ and $\mathbf{n}_{\text{bend}}$ are denoted as $\mathbf{p}_{\text{splay}}$ and $\mathbf{p}_{\text{bend}}$, respectively. $\mathbf{m}_{\text{splay}}$ and $\mathbf{m}_{\text{bend}}$ represent the normal vectors of the mesh at $\mathbf{p}_{\text{splay}}$ and $\mathbf{p}_{\text{bend}}$, respectively. We iterate through all the possible axis directions and minimize the following loss function:

$$\mathcal{L}_{\text{annotate}} = -\mathbf{n}_{\text{spaly}} \cdot \mathbf{m}_{\text{splay}} - \mathbf{n}_{\text{bend}} \cdot \mathbf{m}_{\text{bend}} + \frac{||\mathbf{p}_{\text{splay}} - \mathbf{o}||_2}{||\mathbf{p}_{\text{bend}} - \mathbf{o}||_2}, \quad (8)$$

where $\mathbf{o}$ is the location of the corresponding finger joint. The underlying insight of $\mathcal{L}_{\text{annotate}}$ is that the fingers are narrower from top to bottom but wider from left to right. Therefore, we minimize $\frac{||\mathbf{p}_{\text{splay}}-\mathbf{o}||_2}{||\mathbf{p}_{\text{bend}}-\mathbf{o}||_2}$. Moreover, we aim to align the axes with the mesh normals, thus maximizing $\mathbf{n}_{\text{spaly}} \cdot \mathbf{m}_{\text{splay}} + \mathbf{n}_{\text{bend}} \cdot \mathbf{m}_{\text{bend}}$. Finally, our annotation tool displays the frames of *Twist-bend-splay* on the hand model, as depicted in Figure 6. If needed, the user can manually adjust the orientation of the *splay* and *bend* axes.

B NETWORK ARCHITECTURE AND TRAINING DETAILS

As depicted in Figure 7, the proposed Action Sequence Reconstruction Network (ASRN) architecture comprises two main components:

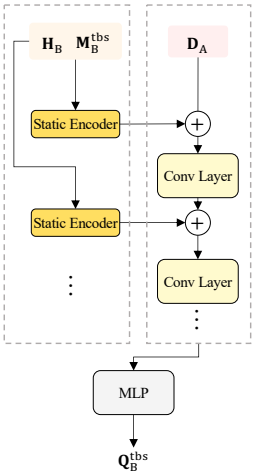

Figure 7: The network architecture of the proposed ASRN.

the static encoders and the motion reconstruction convolutional network. Each static encoder consists of one MLP layer and two ResNet-like convolutional layers. The motion reconstruction convolutional network is composed of four ResNet-like convolutional layers. The input to each layer concatenates the output from the previous layer and the output from the corresponding static encoder. The ASRN takes the source ASM denoted as $\mathbf{D}_\text{A}$ as input and generates the target joint rotation denoted as $\mathbf{Q}_\text{B}^\text{tbs}$ as output. To train the ASRN, we employ the Adam optimizer [18] with a learning rate of 10^{-4} and a batch size of 64. The ASRN is trained for 100 epochs.

Since the shape parameter $\mathbf{H}_\text{B} \in \mathbb{R}^{h_\text{B}}$ varies based on the model type, we train an ASRN for each specific form of the shape parameter. In this study, we introduce two ASRNs specifically for MANO and Mixamo. For MANO, $\mathbf{H}_\text{B}$ is represented as a 10-dimensional vector, while Mixamo represents a 45-dimensional vector. The ASRNs for both MANO and Mixamo are trained using identical hyperparameters. Each network is trained on the InterHand2.6M and Mixamo datasets, but with distinct target hand models.

C ABLATION STUDY

The qualitative results of two ablated versions of our methods are illustrated in Figure 8 and Figure 9.

Figure 8 compares the results with and without the inclusion of the anatomical loss $\mathcal{L}_{\text{ana}}$. Excluding $\mathcal{L}_{\text{ana}}$ leads to a higher occurrence of unnatural finger poses, such as the abnormal splay of the interphalangeal joint of the little finger.

Figure 9 compares the results with and without using the weighting scheme described in Equation 4. The weighting scheme promotes the network's attention toward proximal joint interactions. Consequently, the whole model produces a motion where the thumb pulp contacts the index fingertip, while the ablated model fails to achieve this contact.

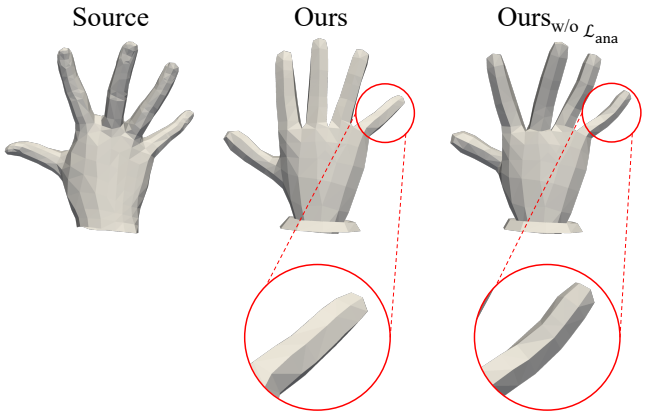

Figure 8: Comparison of results with and without the inclusion of the anatomical loss $\mathcal{L}_{\text{ana}}$.

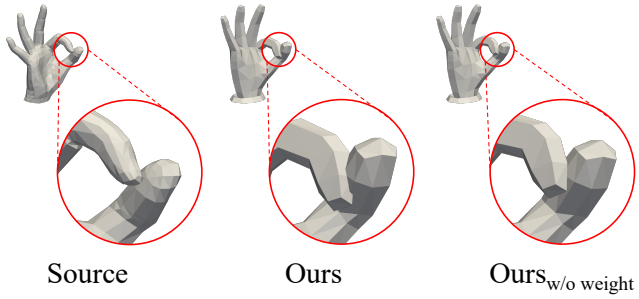

Figure 9: Comparison of results with and without the weighting scheme described in Equation 4.

D SUPPLEMENTARY QUALITATIVE RESULTS

Figure 10 presents additional qualitative results of our method. Our approach effectively preserves accurate hand motion semantics.

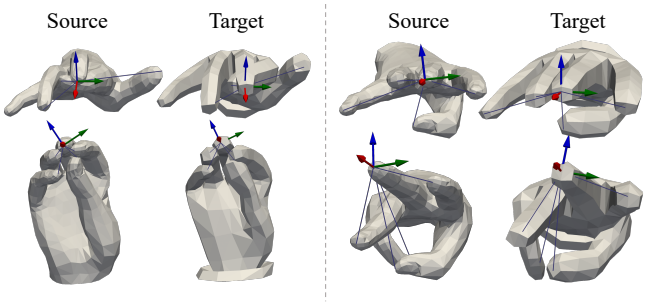

Figure 10: Our framework maintains precise spatial relationships among the fingers.