

Skinned Motion Retargeting with Residual Perception of Motion Semantics & Geometry

Jiaxu Zhang^{1*} Junwu Weng² Di Kang² Fang Zhao² Shaoli Huang²
Xuefei Zhe² Linchao Bao² Ying Shan² Jue Wang² Zhigang Tu^{1†}
¹Wuhan University ²Tencent AI Lab {zjiaxu, tuzhigang}@whu.edu.cn

Abstract

A good motion retargeting cannot be reached without reasonable consideration of source-target differences on both the skeleton and shape geometry levels. In this work, we propose a novel **Residual RETargeting network (R²ET)** structure, which relies on two neural modification modules, to adjust the source motions to fit the target skeletons and shapes progressively. In particular, a skeleton-aware module is introduced to preserve the source motion semantics. A shape-aware module is designed to perceive the geometries of target characters to reduce interpenetration and contact-missing. Driven by our explored distance-based losses that explicitly model the motion semantics and geometry, these two modules can learn residual motion modifications on the source motion to generate plausible retargeted motion in a single inference without post-processing. To balance these two modifications, we further present a balancing gate to conduct linear interpolation between them. Extensive experiments on the public dataset Mixamo demonstrate that our R²ET achieves the state-of-the-art performance, and provides a good balance between the preservation of motion semantics as well as the attenuation of interpenetration and contact-missing. Code is available at <https://github.com/Kebi/R2ET>.

1. Introduction

As a process of mapping the motion of a source character to a target character without losing plausibility, *motion retargeting* is a long-standing problem in the community of computer vision and computer graphics. It has a wide spectrum of applications in game and animation industry, and is a cornerstone of the digital avatar and metaverse technologies [25]. In recent years, learning-based retargeting methods started sparkling in the community. Among them, the neural motion retargeting [1, 16, 24, 25], which has advantages in intelligent perception and stable inference, be-

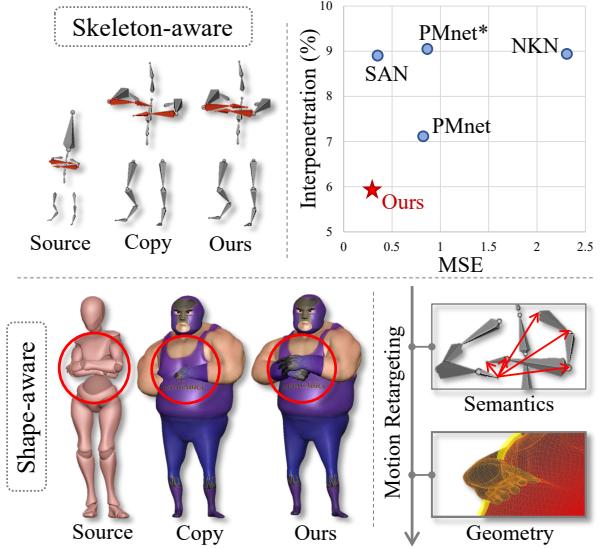


Figure 1. Our R²ET fully considers the source-target differences on both the skeleton and shape geometry levels. The retargeted motion of R²ET preserves motion semantics, eliminates interpenetration, and keeps self-contact without post-optimizations.

comes a new research trend. The previous learning-based methods utilize a *full-motion mapping* structure, which decodes joint rotations of the target skeleton as outputs, with joint positions [16, 24, 25] or joint rotations [1] as inputs. However, due to the gap between the Cartesian coordinate space and the rotation space, the full joint position encoding unavoidably introduces motion distortion. Meanwhile, the full joint rotation encoding always leads to discontinuity in the rotation space [1, 26].

In animation, we observe that artists usually copy the motion of the source character, and then manually modify it to preserve motion semantics and avoid translation artifacts, e.g., interpenetration, during motion reuse in new characters. Inspired by this observation, we design a **Residual RETargeting network (R²ET)** with a *residual* structure for motion retargeting. This structure takes the source motion as initialization and involves neural networks to imitate the modifications from animators, as illustrated in Figure 1.

*Most of this work was done during Jiaxu's internship at Tencent.

†Corresponding author: tuzhigang@whu.edu.cn

With this design, the coherence of the source motion is well maintained, and the search space for retargeting solutions during training is effectively reduced.

The key to achieving physically plausible single-body motion retargeting is to understand two main differences between the source and target characters, 1) the differences in bone length ratio; 2) the differences in body shape geometry. To reach this goal, we explore two modification modules, i.e., the skeleton-aware module and the shape-aware module, to perceive the two differences.

On the skeleton level, the *skeleton-aware* module takes the skeleton configurations as input to assist the transfer of the source motion semantics, such as arm folding and hand clapping, to the target character. To overcome the lack of paired and semantics-correct ground truth, we directly take the supervision signal from the input source motion. The motion semantics is explicitly modeled as a normalized Distance Matrix (DM) of skeleton joints. Accordingly, the semantics preservation is achieved by aligning the DM between the source and target motions (Figure 1-Semantics).

On the shape geometry level, the *shape-aware* module senses the compatibility between the target character mesh and the skeleton adjusted after motion semantics preservation to avoid interpenetration and contact-missing. To train the module end-to-end, we introduce two voxelized Distance Fields, i.e., the Repulsive Distance Field (RDF) and the Attractive Distance Field (ADF) (Figure 1-Geometry), as the measurement tools for interpenetration and contact. We sample the distance of the query vertices on the target character mesh to the body surface in these two fields to estimate the degree of interpenetration and contact. With this manner, the whole process is differentiable during training.

In practice, we find there always exists a contradiction between the preservation of motion semantics and the avoidance of interpenetration. We, therefore, propose a *balancing gate* to make a trade-off between the skeleton-level and geometry-level modifications by learning an adjusting weight. By leaving the weight to the user, our R²ET also accepts interactive fine control from users.

With the above main designs, our R²ET preserves the motion semantics of the source character and avoids interpenetration and contact-missing issues in a single-pass without post-processing. We evaluate our method on various complex motion sequences and a wide range of character geometries from skinny to bulky. The qualitative and quantitative results show that our R²ET outperforms the existing learning-based methods by a large margin. The contributions of this work are summarized in three-fold:

- A novel residual network structure is proposed for neural motion retargeting, which involves a skeleton-aware modification module, a shape-aware modification module, and a balancing gate.
- A normalized joint Distance Matrix is presented to

guide the training of the skeleton-aware module for explicit motion semantics modeling, and two Distance Fields are introduced to achieve differentiable pose adjustment learning.

- Extensive experiments on the Mixamo [2] dataset demonstrate that our R²ET achieves the state-of-the-art performance qualitatively and quantitatively.

2. Related Work

Motion Retargeting. Motion retargeting is pioneered by [7], which identifies features of the source motions as kinematic constraints and solves the space-time optimization problem. Following it, many optimization-based motion retargeting methods were proposed successively by introducing specific constraints, e.g., dynamics constraints [22], inverse kinematics [14], joint angle constraints [5], Euclidean distance [4], and trajectory constraints [6]. Recently, there has been a surge of interest in studying deep-learning-based motion retargeting. Jang *et al.* [10] used a deep autoencoder combining the DC-IGN [13] and the U-Net [21] to generate human motions. Villegas *et al.* [25] trained a Neural Kinematic Network for unsupervised motion retargeting. Aberman *et al.* [1] proposed a Skeleton-Aware Network for retargeting motion between skeletons with different topologies. Lim *et al.* [16] developed a novel architecture which separately learns frame-by-frame poses and overall movement. Li *et al.* [15] introduced an iterative method to yield retargeted motions based on a trained motion autoencoder. All these methods were performed on articulated skeletons while ignoring the shape geometry of the characters. Thus, they are prone to produce unrealistic results on skinned motions. Villegas *et al.* [24] presented a latent-space optimization method for skinned motions to preserve self-contacts and prevent interpenetration, but this post-processing method is cumbersome and is unsuitable for a stable real-time system. In contrast to the above methods, we propose a new network R²ET for skeletal and skinned motion retargeting with respect to motion semantics and character shapes, which can be trained in an end-to-end fashion.

Geometry-aware Motion Modeling. There is tremendous work on learning geometry-aware deep motion representations [20]. Gomes *et al.* [8] leveraged the human body shape in the retargeting process while considering the physical constraints of the motion in the 2D image domain. Peng *et al.* [19] introduced the neural blend weight fields to reconstruct an animatable human model from a multi-view video. These works extract human motion from RGB-based videos or images, while our work looks at the motion retargeting of humanoid characters in 3D space. Jin *et al.* [11] designed a volumetric mesh that surrounds a character’s skin to preserve the spatial relationships of humans. Basset *et al.* [3] proposed an optimization-based method to de-

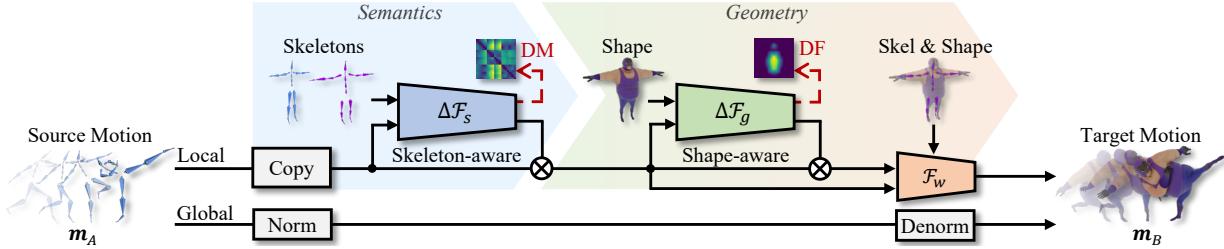


Figure 2. Overview of the proposed network R^2ET , which has three decoupled modules, i.e., the skeleton-aware module $\Delta\mathcal{F}_s$, the shape-aware module $\Delta\mathcal{F}_g$, and the balancing gate \mathcal{F}_w . The Distance Matrix (DM) and the Distance Field (DF) are two types of distance measurements that guide the network to learn the information of semantics and geometry.

form the source shape in the desired pose using three energy functions. In this work, we designed two neural residual modules with distance-based losses to learn the motion semantics and the character geometries for motion retargeting. Our method mainly copes with the animated characters with articulated skeletons, but can be also extended to retarget the motion from Skinned Multi-Person Linear Model (SMPL) [17] estimated from RGB videos.

3. Method

Given the motion sequence \mathbf{m} of the input character A and the skeleton and mesh of a target character B under rest-pose, two branches are involved to retarget the global root motion, i.e., velocities and rotations $\{\mathbf{v}^t\}_{t=1}^T, \mathbf{v} \in \mathbb{R}^4$, and the local joint rotation quaternions $\{\mathbf{q}^t\}_{t=1}^T, \mathbf{q} \in \mathbb{R}^{N \times 4}$, respectively, as Figure 2 shows. N and T indicate the number of joints and the sequence length. The time-index t is ignored in the following for simplification. The global root movement is simply processed by normalizing and de-normalizing with respect to the heights of the source and target character. The motion is translated framewise. Inspired by the creation process in animation, we design a residual learning structure model R^2ET to achieve skeleton-aware and shape-aware local motion retargeting automatically. In particular, our R^2ET takes the source motion \mathbf{q}_A as an initialization. A skeleton-aware modification module $\Delta\mathbf{q}_s = \Delta\mathcal{F}_s(\cdot)$ is introduced to maintain motion semantics, and a shape-aware modification module $\Delta\mathbf{q}_g = \Delta\mathcal{F}_g(\cdot)$ is involved to tackle the interpenetration and contact-missing issues. A balancing gate \mathcal{F}_w is located at the end of this pipeline to balance the two motion modifications. The whole framework is then formulated as:

$$\mathbf{q}_B = \mathcal{F}_w(\mathbf{q}_A, \Delta\mathbf{q}_s, \Delta\mathbf{q}_g), \quad (1)$$

In contrast to [16, 25] which take joint positions as input, we focus on retargeting the motions in the rotation space as [1]. The details of $\Delta\mathcal{F}_s$, $\Delta\mathcal{F}_g$ and \mathcal{F}_w are introduced in Section 3.2, Section 3.3 and Section 3.4, respectively.

3.1. Base Losses

One of the challenges in neural motion retargeting is that there is always no paired ground truth as target motion supervision. Following [16], we utilize the self-reconstruction principle and adversarial learning as basic training rules to

train $\Delta\mathcal{F}_s$, $\Delta\mathcal{F}_g$ and \mathcal{F}_w in an unsupervised way.

During training, the self-reconstruction regularization is followed in two ways: 1) reconstructing the exact source motion in source character (see Section 3.2), 2) minimizing the pose modifications in pose adjustment (see Section 3.3). To avoid the quaternion ambiguity and reduce the position error accumulation along the kinematic chain, the joint rotations and positions are reconstructed simultaneously. Accordingly, the reconstruction loss is defined as:

$$\mathcal{L}_{rec}(\mathbf{q}, \hat{\mathbf{q}}) = \|\mathbf{q} - \hat{\mathbf{q}}\|_2^2 + \|f_K(\mathbf{q}, \gamma) - f_K(\hat{\mathbf{q}}, \gamma)\|_2^2, \quad (2)$$

where \mathbf{q} is the input rotation and $\hat{\mathbf{q}}$ is the estimated one. f_K is a Forward Kinematics (FK) layer [25] that maps the local joint rotations to the global joint positions by referring the rest-pose configuration $\gamma \in \mathbb{R}^{N \times 3}$.

To achieve realistic motion retargeting, a discriminator $\delta(\cdot)$ is introduced to differentiate the translated motion sequences from the genuine ones. A motion discrimination loss is designed based on the adversarial training [9]:

$$\mathcal{L}_{adv}(\hat{\mathbf{q}}) = \mathbb{E}_{\mathbf{m} \sim p_{real}} [\log \delta(\mathbf{m})] + \mathbb{E}_{\mathbf{m} \sim p(\hat{\mathbf{q}})} [\log (1 - \delta(\mathbf{m}))], \quad (3)$$

in which $p(\cdot)$ represents the distribution of the *real* motions or *fake* retargeted motions controlled by $\hat{\mathbf{q}}$.

Besides, a rotation constraint loss [25] is introduced to limit the y -axis Euler angles within a range, which avoids excessive joint twisting:

$$\mathcal{L}_{rot}(\hat{\mathbf{q}}) = \left\| \max(0, |\epsilon_y(\hat{\mathbf{q}})| - \alpha) \right\|_2^2. \quad (4)$$

The function $\epsilon_y(\cdot)$ converts the input quaternion to the Euler angle of y -axis, and α is the angle limitation bound. $\max(\cdot)$ is an element-wise function that returns the maximum number between the two inputs.

With these definitions, the base loss is thereafter defined by the weighted summation of the above losses as follows:

$$\mathcal{L}_{base}(\mathbf{q}, \hat{\mathbf{q}}) = \mathcal{L}_{rec} + \lambda \mathcal{L}_{adv} + \mu \mathcal{L}_{rot}, \quad (5)$$

where λ and μ are the loss balancing factors.

3.2. Skeleton-aware Module

The motion residual design can help maintain the motion coherence from the source character and meanwhile provides a good initialization for motion translation. However,

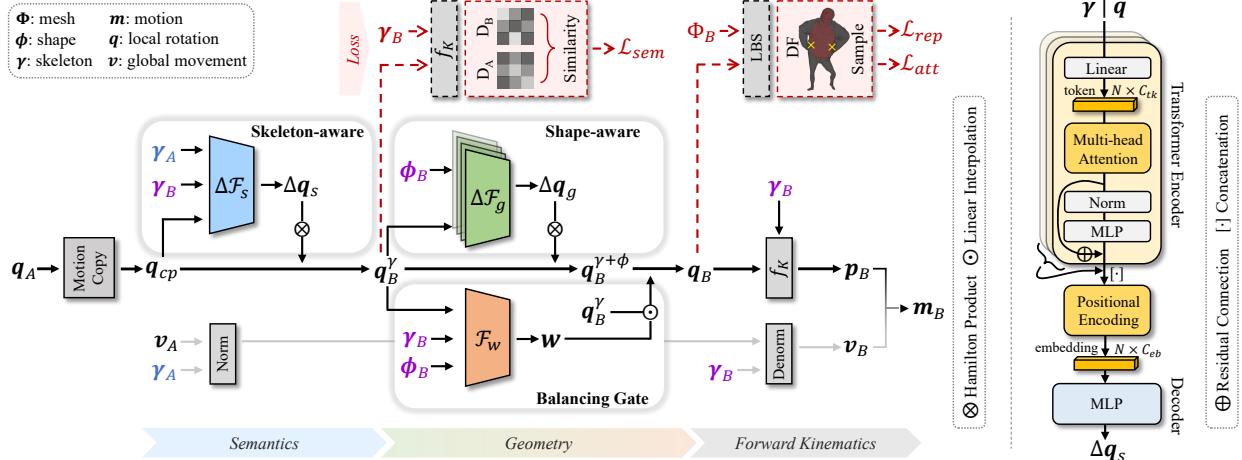


Figure 3. The detailed architectures of the proposed network R²ET and the transformer-based skeleton-aware module.

owing to the differences of bone length and skeleton proportion between source and target characters, motion copy $q_{cp} = q_A$ may ignore the source motion semantics. In this section, we will introduce the skeleton-aware module $\Delta\mathcal{F}_s$ for motion semantics preservation.

The skeleton-aware module takes the rest-pose skeletons γ_A and γ_B as well as the copied motion q_{cp} as input, and outputs the semantics-oriented quaternion modification $\Delta\mathbf{q}_s \in \mathbb{R}^{N \times 4}$. With this estimated modification, a Hamilton Product is applied to modify the motion copy q_{cp} and thereafter to preserve the motion semantics in the output q_B^γ , namely:

$$\begin{aligned} q_B^\gamma &= \Delta\mathbf{q}_s \otimes q_{cp} \\ &= \Delta\mathcal{F}_s(\gamma_A, \gamma_B, q_{cp}; \theta_s) \otimes q_{cp}, \end{aligned} \quad (6)$$

where θ_s indicates the parameter of module $\Delta\mathcal{F}_s$.

Semantics Preservation. There is no paired ground truth as strong semantics supervision. By utilizing the property of motion retargeting, we here take the supervision signals from the source motion. We model the motion semantics preservation as the maintaining of the source normalized pair-wise joint distances in the target character pose, frame by frame. We introduce a normalized Distance Matrix (DM) $D \in \mathbb{R}^{N \times N}$ to represent the motion semantics, i.e., pair-wise joint distances as shown in Figure 4. The columns of the matrix indicate the query joints, and the rows represent the reference joints. The element $d_{i,j}$ of D represents the Euclidean distance from query joint i to reference joint j . We extract the pose DM from the source character and regard it as a supervision signal to guide the learning of target pose DM. With this design, a Semantics Similarity loss is then defined as:

$$\mathcal{L}_{sem} = \left\| \eta\left(\frac{D_A}{h_A}\right) - \eta\left(\frac{D_B}{h_B}\right) \right\|_2^2, \quad (7)$$

where h is the height of the skeleton. $\eta(\cdot)$ is an L1 normalization performed on each row of the distance matrix. This normalization operation eliminates the difference of bone

lengths and heights between the source and target skeletons.

To support the pair-wise joint relationship learning, we introduce a Transformer structure [23], whose attention mechanism is suitable for pair-wise learning, to build the skeleton-aware module. As the right of Figure 3 shows, our Transformer-based structure consists of two Transformer encoders and one MLP decoder. The Transformer encoders process γ and q independently. In this process, N joint features are treated as N tokens with C_{tk} channels, and they are encoded by an Multi-head Attention and a Layer Normalization operations. Then, the feature of γ and q are concatenated and position-encoded to obtain an embedding with C_{eb} channels. In the end, an MLP is shared within N joints to decode the rotation modifications $\Delta\mathbf{q}_s$ for these joints. With the loss and model structure introduced above, the skeleton-aware module can be trained by:

$$\min_{\theta_s} \mathcal{L}_{base}(q_{cp}, q_B^\gamma) + \nu \mathcal{L}_{sem}, \quad (8)$$

where ν is the loss balancing factor. The reconstruction loss is applied when the source and the target characters are the same. We sample the target character as the source one with a probability of 0.5.

3.3. Shape-aware Module

We introduce a shape-aware module $\Delta\mathcal{F}_g$ in this section to ensure the retargeted skinned motion is interpenetration-free and contact-preserved, as illustrated in the middle part of Figure 3. The shape-aware module takes the shape information ϕ of each body part in the target character as well as the $\Delta\mathcal{F}_s$ -modified local joint rotation q_B^γ as input, and outputs the geometry-oriented quaternion modification $\Delta\mathbf{q}_g \in \mathbb{R}^{N \times 4}$. $\phi \in \mathbb{R}^{N \times 3}$ is represented by the edge-lengths of the body part bounding box corresponding to each joint in the rest-pose. As we observed, most of the interpenetration and contact-missing issues occur between the limbs and the main body. We, therefore, choose to only adjust the rotations of four target limbs and introduce four MLPs to estimate rotation modifications for them independently. With the estimated rotation modifications, the ad-

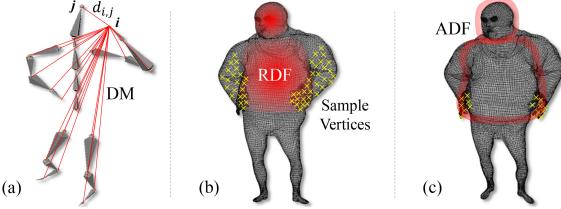


Figure 4. Illustration of the distance measurements. DM is the normalized Distance Matrix of the skeleton joints. RDF and ADF are the Distance Fields inside and outside the main body.

justed joint rotation $\mathbf{q}_B^{\gamma+\phi}$ is then defined as:

$$\begin{aligned} \mathbf{q}_B^{\gamma+\phi} &= \Delta\mathbf{q}_g \otimes \mathbf{q}_B^\gamma \\ &= \Delta\mathcal{F}_g(\phi_B, \mathbf{q}_B^\gamma; \theta_g) \otimes \mathbf{q}_B^\gamma, \end{aligned} \quad (9)$$

where θ_g represents the parameter of module $\Delta\mathcal{F}_g$.

Penetration-free & Contact-preserving. To achieve differentiable pose adjustment learning in respect of mesh geometry, we introduce two truncated distance fields, the Repulsive Distance Field (RDF) ψ_R inside the main body and the Attractive Distance Field (ADF) ψ_A around the body surface. These two fields are illustrated in Figure 4 (b,c). RDF assists the network to force the penetrating vertices to be apart from the interpenetration area, and the ADF attracts the near-contact vertices adjusted after motion semantics preservation to be close to the body surface. With the adjusted rotation $\mathbf{q}_B^{\gamma+\phi}$, we first deform the mesh vertex set of the target character Φ_B by applying Linear Blend Skinning (LBS). Hereafter, ψ_R and ψ_A are estimated by voxelizing the deformed target mesh. Each node on the voxel grid records its distance to the body surface from inside or outside, and we can thus measure the body-surface deviation of each query vertex e on the deformed target mesh by interpolating four node distances on its surrounding tight voxel grid. With this mechanism, our model can be trained to handle the interpenetration and contact-missing problem in an end-to-end manner, and translate a source motion to a plausible target motion during inference without post-processing. These two fields are embedded into two losses, the Repulsive and Attractive loss, respectively to achieve end-to-end training, and they are defined as:

$$\mathcal{L}_{rep} = \frac{1}{N_l} \sum_{e \in E_l} \psi_R(e), \quad \mathcal{L}_{att} = \frac{1}{N_h} \sum_{e \in E_h} \psi_A(e), \quad (10)$$

where $E_l = \{e_i\}_{i=1}^{N_l}$ and $E_h = \{e_i\}_{i=1}^{N_h}$ are the vertices set of the deformed target mesh's limbs and hands, respectively. N_l and N_h are the corresponding numbers of vertices. $\psi(e)$ samples the ψ value for each vertex e in a differentiable way. $\Delta\mathcal{F}_g$ consists of four independent networks corresponding to the four limbs of the character. Accordingly, these four networks are optimized by four related $\mathcal{L}_{rep}^{E_l}$ as:

$$\min_{\theta_g} \mathcal{L}_{base}(\mathbf{q}_B^\gamma, \mathbf{q}_B^{\gamma+\phi}; \theta_g) + \kappa \sum_{i=1}^4 \mathcal{L}_{rep}^{E_l^i}(\cdot; \theta_g^i), \quad (11)$$

where κ is the balancing hyper-parameter, $\theta_g = [\theta_g^i]_{i=1}^4$. As repulsing and attracting the mesh vertices simultaneously

would cause unstable training convergence, we do not involve \mathcal{L}_{att} here but leave it in the next Balancing module.

3.4. Balancing Gate

In practice, it is challenging to learn the motion semantics preservation at the skeleton level and meanwhile train the network to tackle the interpenetration as well as contact-missing issues at the shape-geometry level. Let's take the bottom of Figure 1 as an instance. When retargeting motion from a thin character to an obese character, if only the relative positions of the joints are maintained, it will inevitably lead to interpenetration. On the other side, if only the target shape is considered, the retargeted motion may lose motion semantics. To overcome this obstacle, we introduce an additional MLP module \mathcal{F}_w to balance the influence between the two modifications $\Delta\mathbf{q}_s$ and $\Delta\mathbf{q}_g$ by a learned balancing factor $\mathbf{w} \in \mathbb{R}^N$. This balancing process is achieved by a linear interpolation between \mathbf{q}_B^γ and $\mathbf{q}_B^{\gamma+\phi}$:

$$\mathbf{q}_B = (\mathbf{1} - \mathbf{w}) \cdot \mathbf{q}_B^\gamma + \mathbf{w} \cdot \mathbf{q}_B^{\gamma+\phi}. \quad (12)$$

in which $\mathbf{w} = \mathcal{F}_w(\gamma_B, \phi_B, \mathbf{q}_B^\gamma; \theta_w)$. θ_w indicates the parameters of \mathcal{F}_w . Each element of w is ranged from 0 to 1. The symbol “.” indicates scaling each row of q by each element of w . By leaving the vector w to the user, we can also manually adjust its value at each joint to finely control the retargeted results. To reach an optimized balancing, \mathcal{F}_w can be trained by:

$$\min_{\theta_w} \mathcal{L}_{base}(\mathbf{q}_B^{\gamma+\phi}, \mathbf{q}_B) + \kappa \mathcal{L}_{rep} + \iota \mathcal{L}_{att} + \tau \mathcal{L}_{reg}, \quad (13)$$

where κ , ι , and τ are hyper-parameters. \mathcal{L}_{reg} is a L_2 regularization loss for w .

4. Experiments

Datasets. We evaluate our R²ET on the Mixamo dataset [2], which is an animation repository performed by multiple 3D virtual characters with different skeletons and shapes. For training, we collect 1952 non-overlapping motion sequences of seven characters and randomly sample 60 frames from each sequence. For testing, we collect 800 motion sequences of 11 characters and each sequence has 120 frames. We have unseen character (UC), unseen motion (UM), seen character (SC), and seen motion (SM) so that four splits UC+UM, UC+SM, SC+UM, SC+SM are considered in the experiment. Around 3/4 of the test samples are unseen. The Mixamo dataset does not provide clean Ground Truth (GT): many of the motions may have interpenetration or contact-missing issues which makes geometry learning challenging. For a fair comparison, we follow the spirit of [25] to implement experiments.

Implementation details. We select $N = 22$ joints for each character in our experiments. The hyper-parameters λ , μ , ν , κ , ι , and τ in loss functions are set as 2.0, 10.0, 100.0, 0.5, 0.5, and 0.005, respectively. The margin factor α in the Rotation Constraint loss is defined as 100. We implement

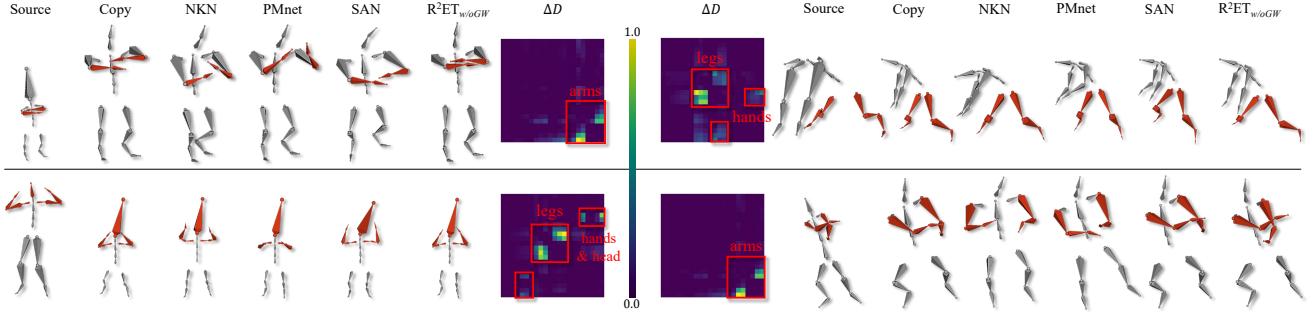


Figure 5. Qualitative results of skeletal motion retargeting. ΔD indicates the DM difference between the motion copy and our result.

our model based on the PyTorch framework [18] and apply the Adam optimizer [12] to train the network. The training process is divided into two stages: firstly train the skeleton-aware module independently, then freeze its parameters and train the shape-aware module and the balancing gate. Please see Sup. Mat. for more details.

Evaluation metrics. We evaluate the results from two aspects: skeleton and geometry levels. For skeletal motions, the metric Mean Square Error (MSE) and local MSE are to measure how close the retargeted joint positions are to the GT provided in the Mixamo. The local MSE is calculated when the root of the retargeted motion is aligned with that of the GT. The joint position error is normalized by the character height [25]. The geometric evaluation includes the measurement of interpenetration and self-contact. Interpenetration is scored by the ratio of the number of penetrated limb vertices to the total number of limb vertices in each frame. Self-contact is measured by the average distance from the vertex of the hands to the surface of the body. Additionally, we also implement qualitative analysis and user study to verify the advance of our method. The methods included for comparison are recent deep-learning-based methods NKN [25], PMnet [16], and SAN [1].

4.1. Qualitative Results

Semantics. Figure 5 visualizes the effect of the skeleton-aware module of R²ET on motion semantics preservation in skeletal motion retargeting. R²ET_{w/oGW} means our method with only the Skeleton-aware Module equipped. We retarget the motion of the bones among small, medium, and large skeletons with different bone length ratios. The skeleton-aware module can well preserve the motion semantics according to the characteristics of the target character’s skeleton. For example, in the top-left row of Figure 5, the “arm folding” pose is retargeted from a small character to a large one, but the existing methods cannot accurately preserve the semantics, and the result of motion copy is more like “hand clapping”. Our R²ET can perceive the key differences between these two skeletons, i.e., the bone length ratio of the arms to the forearms, and adaptively adjust the copied rotations to generate a reasonable retargeted motion. The following three cases further demonstrate that our method, driven by the Semantics Similarity Loss, can

well perceive the skeleton motion semantics and translate it between skeletons with large differences.

Geometry. Figure 6 shows the results of skinned motion retargeting among characters with different shapes. The target characters of the last two rows are unseen characters. The existing methods barely consider the shape geometry of target characters and their results suffer from severe interpenetration and contact-missing issues. Our results, which are based on the semantics-preserved motion and adjusted by the shape-aware module as well as the balancing gate, can well reduce these implausible issues while maintaining motion semantics as much as possible without the post-processing. Figure 6 also visualizes the w of each joint predicted by the balancing gate. The predicted w have higher responses on the joints whose succeeding body parts are interpenetrated, and have lower responses when the motion semantics of the corresponding parts need to be preserved.

Figure 7 (a) shows the change of the geometry-oriented modification Δq_g with time of a motion sequence in the Mixamo dataset. The target 1,2,3 are unseen characters. The vector Δq_g is converted to the average z-axis Euler angle value of two arms for simple illustration. Our shape-aware module can accurately perceive the poses with interpenetration problems as the change of time and apply reasonable adjustments to them. For the poses that are not suffering from interpenetration, our R²ET hardly adjusts them, so as to keep the original motion semantics as much as possible. At the same time, the modification changes smoothly as time goes on, which ensures the coherence and naturalness of the retargeted motion. Figure 7 (b) shows our results of retargeting one source motion to multiple targets. For characters with different body shapes, our R²ET can sensitively perceive their geometries and make precise adjustments to the source motion. Overall, Figure 7 demonstrates that our R²ET is robust to a variety of poses and characters.

An automatic algorithm can provide visual results that follow the pre-defined learning constraints designed by engineers, such as the avoidance of interpenetration and contact-missing, but cannot always satisfy the aesthetic needs of the animators. Thanks to the flexible balancing gate, our R²ET overcomes this drawback as Figure 8 shows. By gradually scaling w , we can get results that vary

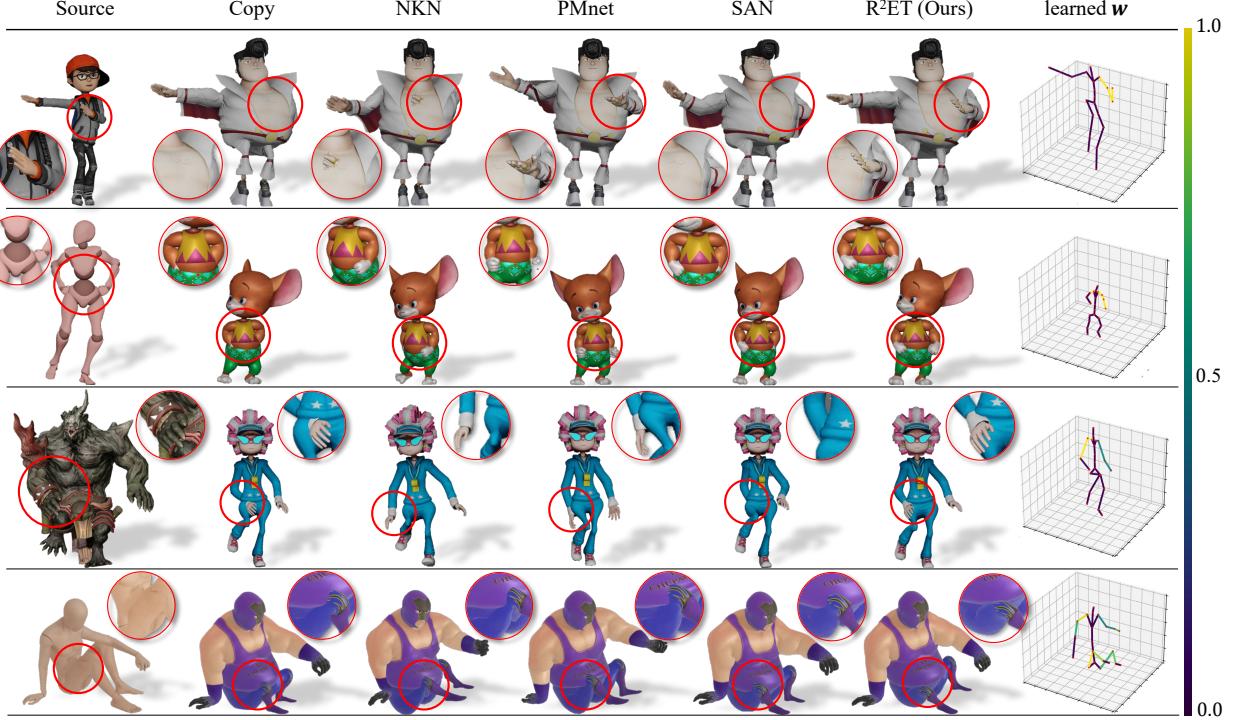


Figure 6. Qualitative results of skinned motion retargeting. w is the adjusting weight that learned by the balancing gate.

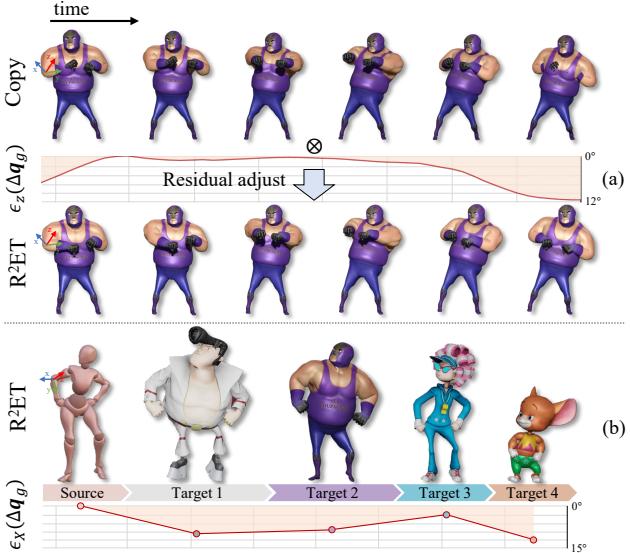


Figure 7. (a) The change of the geometry-oriented modification Δq_g of a motion sequence on time domain. (b) Our results of retargeting one source to multi-targets.

smoothly from the arm pose that preserves the motion semantics to the one that avoids interpenetration, thereby interactively selecting the visually best results.

4.2. Quantitative Results

Comparison with the state-of-the-arts. Table 1 shows the comparison between our method and the state-of-the-arts. Considering that the Mixamo dataset may create a new character with an archived motion by using motion copy, the “Copy” has the lowest MSE and local MSE. However,

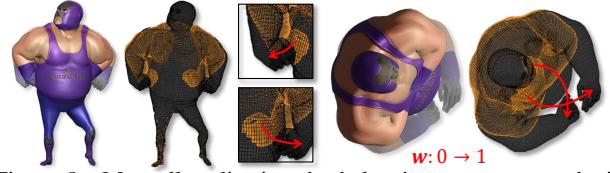


Figure 8. Manually adjusting the balancing gate w can obtain smooth motion adjustment.

this does not mean that the motion copy is the best choice (See Figure 5 and Figure 6). We here just treat MSE as an auxiliary reference metric for comparison. Compared with NKN, PMnet, and SAN which focus on skeletal motion retargeting, our $R^2ET_{w/oGW}$ reduces the MSE by 87% (0.297 vs 2.298), 63% (0.297 vs 0.806), and 7% (0.297 vs 0.321), respectively. The MSE of the PMnet with rotation-input (PMnet*) is lower than the PMnet with position-input but is also worse than ours. The above results show that our method can well reconstruct the source motion while adjusting the local motion according to the skeletal configurations to make it more in line with the motion semantics.

As Table 1 shows, the GT of the Mixamo dataset bears the interpenetration and contact-missing issues. Our $R^2ET_{w/oW}$, with the help of the shape-aware module, can perceive the geometry of characters and reduce the interpenetration effectively. Compared with the GT, our $R^2ET_{w/oW}$ reduces the penetration rate by more than 48% (4.68 vs 9.02). Without the balancing gate equipped, the contact can not be well maintained while reducing interpenetration. Our full model, i.e., R^2ET , reaches a good balance among these three quantitative metrics and also obtains the best qualitative visualization results (see Figure 6). More results about

Table 1. Comparison with the state-of-the-arts. MSE^{lc} is the local MSE. $R^2ET_{w/oGW}$ is the model with the skeleton-aware module only. $R^2ET_{w/oW}$ is the model without the balancing gate. Copy \dagger is the motion copy without the global motion normalization.

Methods	Inp.	MSE_{\downarrow}	MSE^{lc}_{\downarrow}	$Pen.\%_{\downarrow}$	$Con.^{cm}_{\downarrow}$
GT	-	-	-	9.02	4.92
NKN [25]	Pos.	2.298	0.575	8.96	4.42
PMnet [16]		0.806	0.281	7.11	14.7
Copy	Rot.	<u>0.267</u>	<u>0.060</u>	9.23	4.95
Copy \dagger		3.087	0.060	9.23	4.95
SAN [1]		0.321	0.118	8.91	4.86
PMnet*		0.374	0.120	9.03	5.24
$R^2ET_{w/oGW}$	Rot.	0.297	0.094	9.09	4.93
$R^2ET_{w/oW}$		0.378	0.178	4.68	5.31
R^2ET (Ours)		0.318	0.116	5.94	3.57

Table 2. Ranking results of the user study. We invite 100 users to compare our retargeting results to that of the recent methods from three aspects, i.e., overall quality (Q), semantics preservation (S), and motion details (D).

Methods	Skeletal Motion			Skinned Motion		
	Q \downarrow	S \downarrow	D \downarrow	Q \downarrow	S \downarrow	D \downarrow
Copy	1.88	1.83	1.84	1.84	1.84	1.93
NKN [25]	3.37	3.45	3.40	3.44	3.44	3.42
PMnet [16]	3.06	3.06	3.06	3.10	3.07	3.00
R^2ET (Ours)	1.69	1.67	1.70	1.63	1.65	1.64

the seen/unseen splits can be seen in Sup. Mat.

Figure 9 shows the comparisons of the penetration rates of the character’s four limbs between the motion copy and our models. The interpenetration issues mainly occur between the arms and the body of the character, and our models can significantly reduce their penetration rates.

The Contact-aware Model in [24] focuses on skinned motion retargeting and can also effectively reduce interpenetration and preserve self-contacts. However, unlike our R^2ET , the Contact-aware Model adopts a post-processing method to optimize the latent space of motion feature, which may not generate plausible results in a single inference pass, and it may result in unstable real-time inference. Therefore, our method is more efficient and easier to use.

Figure 10 shows the change of the end-effector’s height of a retargeted motion on time domain. Compared to the NKN and SAN that are based on the full-motion mapping structure, our R^2ET with the residual structure can obtain smooth and stable retargeted motion in time series.

4.3. User Study

We conduct a user study to evaluate the performance of our R^2ET against the relevant methods NKN, PMnet, and motion copy. We invited 100 users and gave them six skeletal action videos and seven shape action videos in total to evaluate. Each video includes one source motion and four anonymous results. We ask users to rank the four results

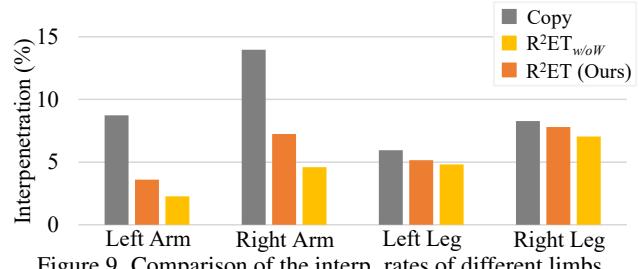


Figure 9. Comparison of the interp. rates of different limbs.

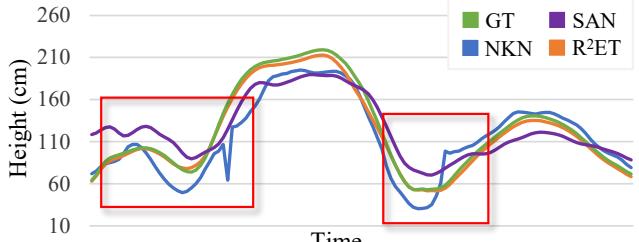


Figure 10. The change of the left-hand end-effector’s height of a retargeted motion on time domain.

in three aspects: overall quality (Q), semantics preservation (S), and motion details (D). After excluding abnormal questionnaires, we collect 3120 ranking comparisons statistics in total and the average rank of the methods is summarized in Table 2. For skeletal motion, our method ranks 1.68 on average. For skinned motion, our method ranks 1.64 on average. In general, more than 71.2% of users prefer the retargeting results of our method.

5. Conclusions

A novel network R^2ET with residual structure is proposed for neural motion retargeting. In R^2ET , two motion modification modules are explored to assist in generating plausible target motion. The skeleton-aware module adjusts the input motion to preserve source motion semantics in the target character. The shape-aware module senses the compatibility between the target shape and semantics-preserved pose to avoid interpenetration and contact-missing. Besides, a balancing gate is designed to make a trade-off between the skeleton-level and geometry-level modifications by learning an adjusting weight. With the help of two distance-based measurements, R^2ET is trained end-to-end. Extensive experiments on the Mixamo dataset demonstrate that our R^2ET achieves the state-of-the-art motion retargeting performance. It provides a good balance between the preservation of motion semantics and the reduction of interpenetration and contact-missing without post-processing.

Limitations. One potential drawback lies in the noisy motion data of the adopted Mixamo dataset. We endeavor to reduce noise interference for future work. Foot contact is not our focus but can be tackled by a method in [1].

Acknowledgements. This work was supported by the fund of Tencent AI Lab RBFR2022012, and the Natural Science Fund for Distinguished Young Scholars of Hubei Province under Grant 2022CFA075.

References

- [1] Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics*, 39(4):62–1, 2020. [1](#), [2](#), [3](#), [6](#), [8](#)
- [2] Adobe. Mixamo. <https://www.mixamo.com/>. [2](#), [5](#)
- [3] Jean Basset, Stefanie Wuhrer, Edmond Boyer, and Franck Multon. Contact preserving shape transfer: Retargeting motion from one shape to another. *Computers & Graphics*, 89:11–23, 2020. [2](#)
- [4] Antonin Bernardin, Ludovic Hoyet, Antonio Mucherino, Douglas Gonçalves, and Franck Multon. Normalized euclidean distance matrices for human motion retargeting. In *Proceedings of the Tenth International Conference on Motion in Games*, pages 1–6, 2017. [2](#)
- [5] Kwang-Jin Choi and Hyeong-Seok Ko. Online motion retargetting. *The Journal of Visualization and Computer Animation*, 11(5):223–235, 2000. [2](#)
- [6] Andrew Feng, Yazhou Huang, Yuyu Xu, and Ari Shapiro. Automating the transfer of a generic set of behaviors onto a virtual character. In *International Conference on Motion in Games*, pages 134–145. Springer, 2012. [2](#)
- [7] Michael Gleicher. Retargetting motion to new characters. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 33–42, 1998. [2](#)
- [8] Thiago L Gomes, Renato Martins, João Ferreira, Rafael Azevedo, Guilherme Torres, and Erickson R Nascimento. A shape-aware retargeting approach to transfer human motion and appearance in monocular videos. *International Journal of Computer Vision*, 129(7):2057–2075, 2021. [2](#)
- [9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. [3](#)
- [10] Hanyoung Jang, Byungjun Kwon, Moonwon Yu, Seong Uk Kim, and Jongmin Kim. A variational u-net for motion retargetting. In *SIGGRAPH Asia 2018 Posters*, pages 1–2. 2018. [2](#)
- [11] Taeil Jin, Meekyoung Kim, and Sung-Hee Lee. Aura mesh: Motion retargeting to preserve the spatial relationships between skinned characters. In *Computer Graphics Forum*, volume 37, pages 311–320. Wiley Online Library, 2018. [2](#)
- [12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *The International Conference on Learning Representations*, 2015. [6](#)
- [13] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. *Advances in neural information processing systems*, 28, 2015. [2](#)
- [14] Jehee Lee and Sung Yong Shin. A hierarchical approach to interactive motion editing for human-like figures. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 39–48, 1999. [2](#)
- [15] Shujie Li, Lei Wang, Wei Jia, Yang Zhao, and Liping Zheng. An iterative solution for improving the generalization ability of unsupervised skeleton motion retargeting. *Computers & Graphics*, 104:129–139, 2022. [2](#)
- [16] Jongin Lim, Hyung Jin Chang, and Jin Young Choi. Pmnet: learning of disentangled pose and movement for unsupervised motion retargeting. In *30th British Machine Vision Conference*. British Machine Vision Association, BMVA, 2019. [1](#), [2](#), [3](#), [6](#), [8](#)
- [17] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics*, 34(6):1–16, 2015. [3](#)
- [18] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. [6](#)
- [19] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021. [2](#)
- [20] Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *European conference on computer vision*, pages 71–87. Springer, 2020. [2](#)
- [21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. [2](#)
- [22] Seyoon Tak and Hyeong-Seok Ko. A physically-based motion retargeting filter. *ACM Transactions on Graphics*, 24(1):98–117, 2005. [2](#)
- [23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. [4](#)
- [24] Ruben Villegas, Duygu Ceylan, Aaron Hertzmann, Jimei Yang, and Jun Saito. Contact-aware retargeting of skinned motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9720–9729, 2021. [1](#), [2](#), [8](#)
- [25] Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. Neural kinematic networks for unsupervised motion retargetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8639–8648, 2018. [1](#), [2](#), [3](#), [5](#), [6](#), [8](#)
- [26] Yi Zhou, Connally Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. [1](#)