# ImitationNet: Unsupervised Human-to-Robot Motion Retargeting via Shared Latent Space



Yashuai Yan<sup>\*1</sup>, Esteve Valls Mascaro<sup>\*1</sup>, and Dongheui Lee<sup>1,2</sup> evm7.github.io/UnsH2R

Fig. 1: Our human-to-robot motion retargeting connects robot control with diverse source modalities, such as a text description, an RGB video, or key poses. Our approach can encode human skeletons into a shared latent space between humans and robots, and subsequently decode these latent variables into the robot's joint space, enabling direct robot control. Additionally, our approach facilitates the generation of smooth robot motions between human key poses (represented as green and blue dots) through interpolation within the latent space (indicated by the orange dots).

Abstract-This paper introduces a novel deep-learning approach for human-to-robot motion retargeting, enabling robots to mimic human poses accurately. Contrary to prior deeplearning-based works, our method does not require paired human-to-robot data, which facilitates its translation to new robots. First, we construct a shared latent space between humans and robots via adaptive contrastive learning that takes advantage of a proposed cross-domain similarity metric between the human and robot poses. Additionally, we propose a consistency term to build a common latent space that captures the similarity of the poses with precision while allowing direct robot motion control from the latent space. For instance, we can generate in-between motion through simple linear interpolation between two projected human poses. We conduct a comprehensive evaluation of robot control from diverse modalities (i.e., texts, RGB videos, and key poses), which facilitates robot control for non-expert users. Our model outperforms existing works regarding human-to-robot retargeting in terms of efficiency and precision. Finally, we implemented our method in a real robot with self-collision avoidance through a wholebody controller to showcase the effectiveness of our approach.

\*Contributed equally to the work.

## I. INTRODUCTION

In recent years, human-robot interaction (HRI) has gained significant attention as it plays a leading role in deploying robots into our daily lives. For a natural HRI, the robot needs not only to capture the human movements but also to understand the human motion intentions behind them. To enhance HRI, it is also crucial to intuitively retarget these human motions onto robots while preserving their similarity and improving robot autonomy. This paper addresses the challenge of enabling robots to mimic human motions while preserving the likeness of the original movement.

However, retargeting human motions to robots is a complex task due to the fundamental differences between human and robot anatomies, kinematics, and motion dynamics. Unlike humans, robots possess rigid bodies, different form factors, and distinct physical limitations. Consequently, directly mapping human motion to robot actuators often leads to unnatural and suboptimal robot behavior, undermining the objective of achieving human-like movements. For example, when retargeting the motion of a human touching his head with his right hand, it is crucial that the retargeted robot poses also reproduce this touching behavior in their motion. Solely replicating the specific arm movements could lead to the robot hand not being close to the head due to the different robot kinematics. Encoding such motions in the retargeting

<sup>&</sup>lt;sup>1</sup>Yashuai Yan, Esteve Valls Mascaro, and Dongheui Lee are with Autonomous Systems Lab, Technische Universität Wien (TU Wien), Vienna, Austria (e-mail: {yashuai.yan, esteve.valls.mascaro, dongheui.lee}@tuwien.ac.at).

<sup>&</sup>lt;sup>2</sup>Dongheui Lee is also with the Institute of Robotics and Mechatronics (DLR), German Aerospace Center, Wessling, Germany.

task is essential to ensure that robots have more natural and intuitive behaviors, leading to better and easier HRI.

While motion retargeting is a long-standing challenge in the robotic and animation community, most recent research has been focused on the exploration of large human motion capture datasets [1], [2] to learn and synthesize human motions from different modality inputs: text [3], 3D scene [4], audio [5] or conditioned by key poses [6]. Our primary goal in this research is to develop a novel method that eliminates the reliance on data annotation, thereby accomplishing the learning of a shared representation space in which human and robot poses are mutually and integrally represented. A good representation space ensures that similar poses from both domains are positioned close to each other while dissimilar poses are far apart. While previous research [7], [8] requires manually annotating human and robot pairs performing the same pose to learn this retargeting process, we consider an unsupervised training technique that does not require pairing data. Consequently, we can reduce the implementation costs for retargeting human poses to new robots.

To this end, we propose an encoder-decoder architecture to construct a latent space that preserves the spatial relationships between human joints as well as the likeness of the original human motion. We achieve this process through the synergy of multiple losses. First, we adopt adaptive contrastive learning to autonomously construct the common latent space based on a proposed similarity metric. Then, we incorporate a reconstruction loss on robot data to ensure the regeneration of the same motion from the latent space. that the robot faithfully follows the movement of the human. Finally, we enforce a consistency term to constrain that the robot faithfully follows the movement of the human. As a consequence, the constructed latent space remains tractable via simple operations. For instance, we are able to generate smooth robot motions between key poses by simply using linear interpolation in the latent space. This intuitive behavior facilitates motion control and also showcases the robustness of our learned latent space. Finally, our decoder can translate the latent representations to robot motion control commands. Contrary to prior methods that adopt soft safety measures in learned approach [9], we implement our method in a real robot with a whole-body controller that ensures self-collision avoidance in the retargeted motion.

Our pipeline allows for the seamless and real-time translation of human skeleton data into robot motion control. Additionally, our model can be easily integrated into the aforementioned deep learning architectures [3]–[5] to accommodate robot motion control from various modalities, enabling flexible and intuitive control over robot behavior. By addressing this challenge, we anticipate significant advancements in HRI. Our research has broad applications, including robot-assisted therapy, entertainment, teleoperation, and industrial robotics. Enabling robots to replicate human motion and intention opens up new possibilities for intuitive and natural HRI, enhancing user experience and fostering acceptance of integrating robots into our daily lives. Our work leads to the following contributions.

- 1) Unsupervised deep learning approach to learn humanto-robot retargeting without any paired human and robot motion data.
- Robust and tractable latent space to generate smooth robot motion control through simple linear interpolation.
- 3) Direct mapping from human skeletons to robot control commands via an encoder-decoder neural network.
- Evaluating control of a real robot from various modalities: text, video, or conditioned by key poses, which ensures user-friendly robot control, particularly for non-experts.

# **II. RELATED WORK**

Existing literature on human-to-robot motion retargeting techniques is reviewed next, highlighting limitations and the need for advancements in translating human motion's overall expressivity and naturality.

## A. Motion retargeting in animation

Human motion retargeting onto animated characters has been a long-standing challenge in the computer graphics community. By bridging the gap between human motion and animation, motion retargeting enhances the quality and naturality of character animation, opening up possibilities for various applications in fields such as film, gaming, and virtual reality.

Classical motion retargeting approaches [10]-[13] involved manually defining kinematic constraints and simplifying assumptions to map human motion onto animated characters. These methods were limited in their ability to handle complex motions and could not accurately capture human movement's nuances. However, with the increased availability of motion capture data [1], [2], data-driven approaches emerged as a more attractive alternative. These approaches offer the potential to overcome the limitations of classical methods and achieve more natural and nuanced motion transfer. [7], [8] learned a shared latent representation to translate motions between different kinematic agents. However, they required paired training data, which is costly and specific for each robot. To cope with the cost of pairing data, [14] used a recurrent neural network to learn motion retargeting without those pairs using adversarial training and cycle consistency. [15] showed that disentangling pose from movement in the retargeting process leads to more natural outcomes. However, these data-driven approaches required the same source and target kinematics. Inspired by the intuition that different kinematics can be reduced to a common primal skeleton, [16] proposed explicitly encoding the different skeleton topologies and projecting those into a shared latent space without pairing data. [16] adopted a latent consistency loss to ensure that the retargeted poses remain faithful to the source. Our work is inspired by their consistency idea, but we construct a more robust shared latent space through a contrastive loss which improves the retargeting outcome. Recently, [17], [18] focused on the motion retargeting but considering the mesh constraints of the animated characters, and thus adjusting motions to reduce interpenetration and feasibility of the motions. Contrary to the aforementioned works that consider self-collision avoidance as an additional feature for more realistic animation, our work ensures the feasibility of the retarget motion by implementing self-collision in the whole-body control of a real robot while preserving the source motion likeness. Finally, [18] proposed an Euclidean distance matrix to account for the motion retargeting, which is relevant for skeletons with similar proportions but underperforms when the targets have different trunk-to-arms ratios, as in our case. On the contrary, we propose to formulate this similarity through global rotations, which precisely capture the likeness in the retargeting task.

## B. Motion retargeting in robotics

Despite the great success of motion retargeting for character animation, their community has only been considering the feasibility of the movements in terms of physical constraints [12], [17]–[19]. Besides ensuring motion's feasibility, robotics research also requires adequate control of the appropriate robot based on the source motion. [19], [20] considered constrained optimization algorithms to retarget a human motion in a simulated robot but required learning a given trajectory and can not quickly overcome new variations. [21] proposed Bayesian optimization and inverse kinematics (IK) to tackle natural retargeting, but their approach required manually selecting joints of interest and was constrained to a few specific motions. Likewise, [19], [22], [23] considered whole-body retargeting by mapping human link orientation to robots and solving IK. [22] introduced a dynamic filter to enforce robot stability, which also over-smoothed the robot poses, thus failing to capture the motion nuances in the retargeting. Moreover, [22] method did not generalize to new kinematics. To cope with that issue, [23] proposed to solve the IK over the robot model, which facilitated the generalization to new robots. For that, [23] orients the robot links closer to the corresponding human links to better capture the likeness in the retargeting. We adopt a similar approach by considering the global rotation of body links as the similarity measurement between humans and the retargeted robot pose. However, all these previous works failed to overcome the manual morphing problem [24]: the challenge of mapping in the joint space from human to robot, which requires similar joint orders among the human and robot. On the contrary, our work does not focus on the task of retargeting the poses while keeping the robot balanced, [22], [23], but on the generalization of a unique method for human-robot retargeting with accuracy and capturing the nuances. Closer to our work, [25] proposed a learned-based footstep planner and a whole-body controller to retarget the human locomotion to a robot while being coherent with the generated footsteps. However, [25] only considered locomotion retargeting and assumed that the robot had at least one known contact with the environment at any time. Therefore, [25] was inappropriate for contact-free motions such as jumping or running.

Deep learning has become a solution to ensure the retargeting process generalizes in terms of kinematics and diversity in the motions while being efficient. First, [26] proposed to construct a shared latent space to retarget human motion to humanoid robots, and the shared latent space is constructed with annotated human-to-robot pair data. Gathering a sufficient quantity of paired data for constructing the latent space is a laborious and time-intensive process and hardens the generalization to new configurations. [9] extended this approach by creating an automated paired data generation process. However, both works have to use nonparametric optimization in the latent space to retrieve similar robot poses to control the robot, which is inefficient if the dataset to retrieve is large. Contrastingly, our method learns a direct mapping from human poses to robot control commands. Therefore, our approach can control a robot at a high rate without being constrained by the quantity of training data.

# III. METHODOLOGY

In this section, we present an overview of our proposed framework for unsupervised human-to-robot motion retargeting via a shared latent space. First, we formulate the humanto-robot retargeting task. Then, we describe our encoderdecoder deep learning architecture, illustrated in Figure 2.

## A. Problem Formulation

Let  $\mathbf{x}_h = [x_{h,1}, \dots, x_{h,J_h}] \in \mathbb{R}^{J_h \times n}$  be a human pose composed by  $J_h$  joints. Similarly,  $\mathbf{x}_r = [x_{r,1}, \dots, x_{r,J_r}] \in \mathbb{R}^{J_r \times s}$  represents a robot pose. Then, the task of human motion retargeting can be formulated as finding a function f that maps a  $\mathbf{x}_h$  to  $\mathbf{x}_r$  ( $f : \mathbf{x}_h \mapsto \mathbf{x}_r$ ) so that  $\mathbf{x}_r$  preserves the human-like naturality of the pose  $\mathbf{x}_h$ . However, the joints for humans and robots usually have different configurations: a human joint (e.g., wrist joint) can have more than 1DoF, while one robot joint usually has only 1DoF. To cope with such differences, we describe each human joint  $x_{h,j}$  as its quaternion representation referring to its parent (n = 4), while each robot joint  $x_{r,j}$  (i.e., revolute joint) is described as its joint angle (s = 1).

In our particular case, and contrary to all works focusing on character animation, we are interested in the direct control of a robot. Robots can be controlled via their joint angles. As joint angles for robots and humans have different configurations, it makes little sense to compare joint angles to measure their similarity. Inspired by [23], we propose to use the global rotation of body links to compare the similarity between human and robot poses, which better captures their likeness and allows for better generalization to different kinematics. The similarity metric is defined in Section III-B.

Previous works [9], [26] rely on the acquisition of a dataset of mapped motions between the human and the robot to retarget, which we describe as a  $\{\mathbf{x}_h, \mathbf{x}_r\}$  pair. These works learn the retargeting function f in a supervised manner. On the contrary, we consider the retargeting task without collecting the correct  $\{\mathbf{x}_h, \mathbf{x}_r\}$  pair and learn without supervision how to approximate f better. To this end, our model first learns to project human  $\mathbf{x}_h$  and robot  $\mathbf{x}_r$  poses to



Fig. 2: **Model overview.** Two human poses  $(\mathbf{x}_h^i, \mathbf{x}_h^j)$  are encoded into latent variables  $(z^i, z^j)$  within the shared space using the function  $Q_h$ . Similarly, a robot data  $\mathbf{x}_r^k$  is mapped into  $z^k$  by  $Q_r$ . Given three samples  $(z^i, z^j, z^k)$ ,  $z^i$  is randomly chosen as an anchor  $z_o^i$ , and  $z^j, z^k$  are estimated as a negative  $z_-^j$  and positive  $z_+^k$  sample through similarity metric in Equation 1. The triplet loss  $\mathcal{L}_{triplet}$  constrains the construction of the latent space by bringing  $z_o^i$  and  $z_+^k$  closer and pushing  $z_o^i$  and  $z_-^j$ apart. The decoder  $D_r$  decodes latent variable  $z^k$  into  $\mathbf{\hat{x}}_r^k$  that should be consistent with the robot data  $\mathbf{x}_r^k$  regarding  $\mathcal{L}_{rec}$ . The latent variable  $z^j$  from the human data  $\mathbf{x}_h^j$  is mapped into a robot data  $\mathbf{\hat{x}}_r^j$ . To ensure that  $\mathbf{\hat{x}}_r^j$  is from the same distribution as  $\mathbf{x}_r^k$ ,  $Q_r$  encodes  $\mathbf{\hat{x}}_r^j$  back to latent variable  $\hat{z}^j$ , and  $\mathcal{L}_{ltc}$  minimizes the distance between  $\hat{z}^j$  and  $z^j$ . During the inference phase,  $\mathbf{\hat{x}}_r^j$  is used to control the robot directly to mimic human pose  $\mathbf{x}_h^j$ .

the same representation space. Then, we decode the learned representation to robot joint angles, which allows us to control the robot directly.

# B. Cross-domain similarity metric

To create a shared latent space in an unsupervised way, we initially define a similarity metric that captures the likeness of the poses between humans and robots. Contrary to prior works that use the local quaternions [16] or the relative XYZ position of the end effector [18], we consider the global rotation of body limbs as the similarity metric that better preserves the skeleton visual appearance. By using global rotation, our model captures the complete 3D orientation and remains invariant to coordinate systems and articulation variations. Let  $q_{h,j}$  and  $q_{r,j}$  represent the global quaternions of the same limbs (e.g., shoulder-to-elbow, elbow-to-wrist, etc.) of a human pose  $\mathbf{x}_h$  and a robot pose  $\mathbf{x}_r$ . As a human pose is represented as limb quaternions, it is straightforward to obtain  $q_{h,i}$  from  $\mathbf{x}_h$ . To get limb quaternions of a robot, we utilize forward kinematics to map robot joints  $\mathbf{x}_r$  to its limb quaternions  $q_{r,j}$ . Then, the distance between the two poses can be computed as shown in Equation 1, where <,>denotes the dot product between two vectors.

$$S_{GR}(\mathbf{x}_h, \mathbf{x}_r) = \sum_j (1 - \langle q_{h,j}, q_{r,j} \rangle^2)$$
(1)

 $S_{GR}$  is employed to measure the similarity between two poses used for contrastive learning in Section III-C.

## C. Human-to-Robot shared representation

We formulate the task of motion retargeting as the translation between two domains. We adopt two multi-layer perceptron (MLP) encoders  $(Q_h, Q_r)$  to project the human and robot poses to a shared representation space, respectively. This way,  $Q_h$  projects  $\mathbf{x}_h \in \mathbb{R}^{J_h \times n}$  to  $z \in \mathbb{R}^d$  while  $Q_r$ translates  $\mathbf{x}_r \in \mathbb{R}^{J_r \times s}$  to  $z \in \mathbb{R}^d$ . Given a human pose  $\mathbf{x}_h$ , our shared latent space is used as a bridge to generate  $\mathbf{x}_r$  while conserving its similarity defined in Section III-B.

We propose to learn the retargeting function  $f: \mathbf{x}_h \mapsto \mathbf{x}_r$ without any paired human and robot motion data. Inspired by the recent success of contrastive learning methods (e.g., CLIP [27]), we propose to construct a shared latent space between two domains (here human and robot poses) in an unsupervised manner. Contrastive learning is a training technique that aims to learn from unlabeled data by comparing and contrasting different instances according to given similarity metrics. To do that, a neural network is optimized to maximize the agreement between positive pairs (similar instances) and minimize the agreement between negative pairs (dissimilar instances).

Let us assume a large set of data that contains feasible human poses  $\mathbf{x}_h$  and robot poses  $\mathbf{x}_r$ . Our method randomly selects triplets of projections from these data instances. As shown in Figure 2,  $\mathbf{x}_h^i, \mathbf{x}_h^j$  and  $\mathbf{x}_r^k$  are a triplet. Then, we first encode them to the shared latent space through  $Q_h$  and  $Q_r$ , respectively. For the encoded triplet  $(z^i, z^j, z^k), z^i$  is randomly selected as an anchor  $z_o^i$ , which serves as the reference. We compute the global rotation distance  $S_{GR}$  detailed in Equation 1 to obtain the similarity between our anchor pose  $z_o^i$  and the two other poses  $(z^j, z^k)$ . The dissimilar  $z^j$  is a negative sample  $z_-^j$  while  $z^k$  is a positive sample  $z_+^k$ .

Then, we adopt the Triplet Loss [28] that pulls similar samples (anchor  $z_o^i$  and positive  $z_+^k$ ) close while simultane-

ously pushing dissimilar samples (anchor  $z_o^i$  and negative  $z_o^I$ ) away in the latent space. This allows a representation space where similar instances are clustered together and dissimilar instances are pushed apart. Equation 2 shows the Triplet Loss  $\mathcal{L}_{triplet}$  used in our scenario, where  $\alpha = 0.05$ .

$$\mathscr{L}_{triplet} = max(||z_o^i - z_+^k||_2 - ||z_o^i - z_-^j||_2 + \alpha, 0)$$
(2)

# D. Shared representation to robot control

Our proposed encoders allow us to project human poses and robot poses into a shared representation space. Therefore, the next step is to learn how to decode latent variables **z** sampled from the shared space into robot joint space that can be directly used to control the robot. As shown in Figure 2, the decoder  $D_r$  decodes the latent variables  $z^j$  and  $z^k$  to robot data  $\hat{\mathbf{x}}_r^j$  and  $\hat{\mathbf{x}}_r^k$ , respectively. As  $z^k$  is encoded from the robot data  $\mathbf{x}_r^k$ , we employ a standard reconstruction loss over  $\hat{\mathbf{x}}_r^k$  and  $\mathbf{x}_r^k$ , as shown in Equation 3. Additionally, to ensure that the predicted robot data  $\hat{\mathbf{x}}_r^j$  from human data  $\mathbf{x}_h^j$ is from the same distribution as the real robot data, we adopt the latent consistent loss shown in Equation 4 to encourage direct mapping in the retargeting process, similar to [16].

$$\mathscr{L}_{rec} = ||\mathbf{x}_{\mathbf{r}} - D_r(Q_r(\mathbf{x}_{\mathbf{r}}))||_1$$
(3)

$$\mathscr{L}_{ltc} = ||Q_h(\mathbf{x_h}) - Q_r(D_r(Q_h(\mathbf{x_h}))))||_1$$
(4)

Our approach employs an end-to-end training strategy, enabling the encoder to learn a shared representation space for both human and robot poses unsupervised while ensuring that this representation space is reconstructible to robot control through our decoder. The total loss employed during training is a weighing sum as described in Equation 5, where  $\lambda_{triplet} = 10, \lambda_{rec} = 5$ .

$$\mathscr{L} = \lambda_{triplet} \mathscr{L}_{triplet} + \lambda_{rec} \mathscr{L}_{rec} + \mathscr{L}_{ltc}$$
(5)

## IV. EXPERIMENTS

The experimental setup and datasets used to evaluate the performance of our model are presented, along with the metrics and benchmarks employed to assess the accuracy and fidelity of the retargeted robot motions.

#### A. Experiment Settings

The hyperparameter configurations used in our framework are listed in this subsection. The network consisting of two encoders and one decoder is trained end-to-end with a learning rate of 0.001 and batch size of 256. The encoder and decoder are Multi-Layer Perceptrons with the same structure; they have 6 hidden layers, each with 128 units. The shared latent space is of 8 dimensions. Adam [29], a momentum-based method, is utilized to optimize the loss function during training. We trained our model for 2.5 hours until the losses reached convergence. We did not experiment with the hyperparameters but chose default values to simplify the training. We acknowledge that further finetuning of those parameters could result in improvements in our results. We use a Ubuntu 22.04 and RTX A4000 Graphic card for our experiment.

Additionally, we employ a bi-manual TiaGo++ robot that integrates two 7-DoF arms. In this paper, we focus on the motion of the upper and lower arm parts. We ignore the motion of the two hands because the HumanML3D human motion dataset [2] used does not contain hand motions. Therefore, the similarity metric  $S_{RD}$  in Equation 1 is defined on four limbs: left shoulder-to-elbow, left elbow-to-wrist, right shoulder-to-elbow, and right elbow-to-wrist.

To control the robot in the real world, we send joint commands to the whole-body-controller [30] integrated in Tiago++ robot. The whole-body controller handles joint angle limits, joint velocity limits, and self-collision avoidance.

## B. Data collection

We present a robot pose generation procedure that requires only the robot's kinematic information. First, we sample the robot joint angles from its configuration space. The robot pose can be computed by following its forward kinematics. In such a way, we collect around 15M poses from the TiaGo++ robot by randomly sampling angles per joint. For human motions, we use the HumanML3D dataset [2] that consists of 14616 motions with a total length of 28.59 hours, summing up to around 20M poses. HumanMl3D covers human daily activities (e.g., 'walking', 'jumping'), sports (e.g., 'playing golf'), acrobatics (e.g., 'cartwheel'), and artistry (e.g., 'dancing'). In HumanML3D, a human pose is represented by its skeletons. As robot poses are sampled randomly from the configuration space, they are not matched to human poses in HumanML3D.

#### C. Baseline

We implement  $S^{3}LE$  [9] as our baseline. To train  $S^{3}LE$ , we use a similar method as mentioned in [9] to generate paired data. We generate the same amount of paired data as in [9], 200K, by selecting the pairs with minimal rotation distance measured by Equation 1. The paired data is only used to train our baseline method.

## D. Quantitative evaluation

To evaluate the performance of each retargeting method, we annotated 11 distinct motions that were not observed while training. The annotated motions serve as the ground truth for our evaluation. We employ the Mean Square Error (MSE) of joint angles between ground truth and predicted results to quantify our proposed method. Furthermore, our method endeavors to address motion retargeting in realtime scenarios. We thoroughly evaluated the computational efficiency and speed at which our model operates.

Table I compares our method with the baseline in Section IV-C. Our method outperforms the baseline in terms of MSE of joint angles. Furthermore, our novel approach demonstrates a notable increase in operational efficiency, surpassing the baseline by more than a factor of three. With a speed of 1.5kHz, our method readily fulfills the requirements of most advanced robot control systems.

TABLE I: Performance of our proposed method and the baseline. The Mean Square Error (MSE) of joint angles between ground truth and predicted results are compared here. Bold fonts indicate better results.



Fig. 3: Human Retargeting comparison for different key poses. Various human skeleton key poses are retargeted to the Thiago robot. Our model captures the initial pose's visual similarity and is closely related to the manually annotated ground-truth poses.

## E. Qualitative evaluation

Visually compelling examples and comparisons between the original human and retargeted robot motions are showcased in Figure 3. For the selected human motions, we annotated their ground truth shown in the second row. Our method accurately retargets the motion when the input skeleton lifts hands above his head, lifts hands to his chest, or performs T-pose, whereas the baseline fails.

## F. Ablation Study

An ablation study is conducted to systematically analyze the impact of individual loss components in our proposed model. We utilize three loss components in our approach to optimize retargeted motions. When analyzing the results in Table II, it becomes apparent that the removal of the latent consistency loss  $\mathcal{L}_{ltc}$  results in a slight reduction in the performance of our method. On the contrary, the Triplet loss  $\mathcal{L}_{triplet}$  is indispensable for the optimization process. As supported by the experimental results, eliminating  $\mathcal{L}_{triplet}$ significantly increases the loss value, rising from 0.21 to 0.57. This underscores the crucial role played by  $\mathcal{L}_{triplet}$  in achieving improved optimization outcomes, contrary to all previous works that do not explore our contrastive training.

## G. From RGB videos to robot motions

The proposed method can generate natural and visually similar motions from RGB videos. We adopt [31] to obtain TABLE II: Ablation study of proposed loss components. Mean Square Error (MSE) of joint angles between ground truth and predicted results. Bold fonts indicate better results.

$\mathscr{L}_{triplet}$	$\mathcal{L}_{rec}$	$\mathscr{L}_{ltc}$	MSE
1	1	X	0.24
X	1	1	0.57
1	1	1	0.21

human 3D skeletons from RGB images in real-time. We extended [31] with the state-of-the-art YOLOv8 [32] for human detection and tracking to optimize the speed. Since there is no ground truth, we only show snapshots of reference images and corresponding TiaGo poses in Figure 4 for qualitative evaluation. We implement the whole pipeline that runs in real-time to control the robot's motions based on the human video.

#### H. From texts to robot motions

Text is an essential modality for human motions. Using a pre-trained motion synthesis model, Text-to-Motion Retrieval [33], our method can generate robot motions with texts. To this end, we first retrieve human motions from texts with Text-to-Motion Retrieval and then retarget human motions to the TiaGo++ robot. Figure 5 shows two examples of retargeting motion from texts. More examples can be found on our webpage.

## I. From key poses to robot motions

Our training strategy allows us to build a shared latent space that covers diverse motions. The contrastive loss  $\mathscr{L}_{triplet}$  makes similar poses close and dissimilar poses far away in latent space. In such a way, our proposed method learns a smooth latent space, which enables us to interpolate motions between key poses. In Figure 6, we show three key poses: A, B, and C, and the interpolated in-between motions. For interpolation, two key (e.g., A and B) poses are mapped into two points in latent space, and intermediate steps can be linearly interpolated in between them. The in-between motions are decoded from these interpolated steps.

# J. Future work

Our work proposed to construct a likeness-aware latent space that unifies human and robot representations seamlessly and allows for real-time robot control. While our model exhibits high precision in the retargeting process, we still observe room for improvement. Better exploring the similarity metrics between the different domains as well as connecting the shared space to higher-level representations (textual descriptions of the poses), will be considered in the future to enhance human-to-robot retargeting.

## V. CONCLUSIONS

In this paper, we presented an unsupervised motion retargeting method that ensures a shared latent space for motion generation. To this end, we use contrastive learning combined with deep latent space modeling to incorporate human and robot motion data. To construct a shared representation of



Fig. 4: Video-to-Motion. We leverage the state-of-the-art off-the-shelf 3D human pose estimator [31] to translate RGB images into human skeletons. Then we employ our proposed method to achieve direct motion control from human skeletons.



Fig. 5: **Text-to-Motion.** Our model can connect as a pipeline to pre-trained motion synthesis models. In this case, we first use Text-to-Motion Retrieval [33] to get human motion in skeleton representation. Then, we utilize our proposed method to translate the motion into robot control commands (i.e., joint angles) to mimic it.



Fig. 6: **Key poses-to-Motion.** The proposed method enables motion generation with key poses. Given distinct key poses, natural in-between motions can be generated by linearly interpolating key poses in our learned latent space. Our results provide the potential for direct motion control in latent space

human and robot motion, we define a cross-domain similarity metric based on the global rotation of different body links. Similar motions are clustered together, and dissimilar motions are pushed apart while constructing the latent space. Furthermore, our decoder maps the shared representation to robot joint angles to control a robot directly without any additional optimization process. Additionally, we connect our model with existing pre-trained models to achieve motion retargeting from different modalities, such as controlling the robot with given texts or retargeting from RGB videos. Moreover, our learned latent space remains tractable and allows for the generation of smooth motion inbetweening between two distinct key poses through linear interpolation in the projected latent space. We showcase all results and the robustness of our model through various experiments, both quantitatively and qualitatively.

# ACKNOWLEDGMENT

This work is funded by Marie Sklodowska-Curie Action Horizon 2020 (Grant agreement No. 955778) for project 'Personalized Robotics as Service Oriented Applications' (PERSEO).

#### REFERENCES

- Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.
- [2] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision* and Pattern Recognition (CVPR), pages 5152–5161, June 2022.

- [3] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023.
- [4] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20460–20469, 2022.
- [5] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. ACM Trans. Graph., 42(4):1–20, 2023.
- [6] Esteve Valls Mascaro, Shuo Ma, Hyemin Ahn, and Dongheui Lee. Robust human motion forcasting using transformer-based model. In 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 10674–10680, 2022.
- [7] Brian Delhaisse, Domingo Esteban, Leonel Rozo, and Darwin Caldwell. Transfer learning of shared latent spaces between robots with similar kinematic structure. 02 2017.
- [8] Hanyoung Jang, Byungjun Kwon, Moonwon Yu, Seong Kim, and Jongmin Kim. A variational u-net for motion retargeting. pages 1–2, 12 2018.
- [9] Sungjoon Choi, Min Jae Song, Hyemin Ahn, and Joohyung Kim. Selfsupervised motion retargeting with safety guarantee. In 2021 IEEE International Conference on Robotics and Automation (ICRA), pages 8097–8103. IEEE, 2021.
- [10] Michael Gleicher. Retargetting motion to new characters. *Proceedings* of the 25th annual conference on Computer graphics and interactive techniques, 1998.
- [11] 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*, SIGGRAPH '99, page 39–48, USA, 1999. ACM Press/Addison-Wesley Publishing Co.
- [12] Christian Ott, Dongheui Lee, and Yoshihiko Nakamura. Motion capture based human motion recognition and imitation by direct marker control. In *Humanoids 2008 - 8th IEEE-RAS International Conference on Humanoid Robots*, pages 399–405, 2008.
- [13] Dongheui Lee, Christian Ott, and Yoshihiko Nakamura. Mimetic communication model with compliant physical contact in human—humanoid interaction. *The International Journal of Robotics Research*, 29(13):1684–1704, 2010.
- [14] 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.
- [15] Jongin Lim, Hyung Jin Chang, and Jin Young Choi. Pmnet: learning of disentangled pose and movement for unsupervised motion retargeting. In *Proceedings of the 30th British Machine Vision Conference (BMVC 2019)*. British Machine Vision Association, BMVA, September 2019. 30th British Machine Vision Conference (BMVC 2019) ; Conference date: 09-09-2019 Through 12-09-2019.
- [16] 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 (TOG), 39(4):62–1, 2020.
- [17] Ruben Villegas, Duygu Ceylan, Aaron Hertzmann, Jimei Yang, and Jun Saito. Contact-aware retargeting of skinned motion, 2021.
- [18] Jiaxu Zhang, Junwu Weng, Di Kang, Fang Zhao, Shaoli Huang, Xuefei Zhe, Linchao Bao, Ying Shan, Jue Wang, and Zhigang Tu. Skinned motion retargeting with residual perception of motion semantics & geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13864–13872, 2023.
- [19] Kai Hu, Christian Ott, and Dongheui Lee. Online human walking imitation in task and joint space based on quadratic programming. In 2014 IEEE International Conference on Robotics and Automation (ICRA), pages 3458–3464, 2014.
- [20] Waldez Gomes, Vishnu Radhakrishnan, Luigi Penco, Valerio Modugno, Jean-Baptiste Mouret, and Serena Ivaldi. Humanoid wholebody movement optimization from retargeted human motions. In 2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids), pages 178–185, 2019.
- [21] Sungjoon Choi and Joohyung Kim. Towards a natural motion generator: a pipeline to control a humanoid based on motion data. In 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 4373–4380, 2019.

- [22] Luigi Penco, B. Clement, V. Moduano, Enrico Mingo Hoffman, Gabriele Nava, Daniele Pucci, Nikos Tsagarakis, J. Mourert, and Serena Ivaldi. Robust real-time whole-body motion retargeting from human to humanoid. pages 425–432, 11 2018.
- [23] Kourosh Darvish, Yeshasvi Tirupachuri, Giulio Romualdi, Lorenzo Rapetti, Diego Ferigo, Francisco Javier Andrade Chavez, and Daniele Pucci. Whole-body geometric retargeting for humanoid robots. In 2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids), pages 679–686. IEEE, 2019.
- [24] Kourosh Darvish, Luigi Penco, Joao Ramos, Rafael Cisneros, Jerry Pratt, Eiichi Yoshida, Serena Ivaldi, and Daniele Pucci. Teleoperation of humanoid robots: A survey. *IEEE Transactions on Robotics*, 2023.
- [25] Paolo Viceconte, Raffaello Camoriano, Giulio Romualdi, Diego Ferigo, Stefano Dafarra, Silvio Traversaro, Giuseppe Oriolo, Lorenzo Rosasco, and Daniele Pucci. Adherent: Learning human-like trajectory generators for whole-body control of humanoid robots. *IEEE Robotics* and Automation Letters, PP:1–1, 01 2022.
- [26] Sungjoon Choi, Matthew Pan, and Joohyung Kim. Nonparametric motion retargeting for humanoid robots on shared latent space. 07 2020.
- [27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.
- [28] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network, 2018.
- [29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
- [30] Nicolas Mansard, Olivier Stasse, Paul Evrard, and Abderrahmane Kheddar. A versatile generalized inverted kinematics implementation for collaborative working humanoid robots: The stack of tasks. In 2009 International Conference on Advanced Robotics, pages 1–6, 2009.
- [31] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021.
- [32] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, January 2023.
- [33] Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-tomotion retrieval using contrastive 3D human motion synthesis. arXiv preprint, 2023.