Adult2child: Motion Style Transfer using CycleGANs

Yuzhu Dong
University of Florida
yuzhudo1g@ufl.edu

Andreas Aristidou
University of Cyprus
RISE Research Centre
a.aristidou@ieee.org

Ariel Shamir
The Interdisciplinary Center Herzliya
arik@idc.ac.il

Moshe Mahler
Carnegie Mellon University
mmahler@andrew.cmu.edu

Eakta Jain
University of Florida
ejain@cise.ufl.edu

ABSTRACT

Child characters are commonly seen in leading roles in top-selling video games. Previous studies have shown that child motions are perceptually and stylistically different from those of adults. Creating motion for these characters by motion capturing children is uniquely challenging because of confusion, lack of patience and regulations. Retargeting adult motion, which is much easier to record, onto child skeletons, does not capture the stylistic differences. In this paper, we propose that style translation is an effective way to transform adult motion capture data to the style of child motion. Our method is based on CycleGAN, which allows training on a relatively small number of sequences of child and adult motions that do not even need to be temporally aligned. Our *adult2child* network converts short sequences of motions called motion words from one domain to the other. The network was trained using a motion capture database collected by our team containing 23 locomotion and exercise motions. We conducted a perception study to evaluate the success of style translation algorithms, including our algorithm and recently presented style translation neural networks. Results show that the translated adult motions are recognized as child motions significantly more often than adult motions.

CCS CONCEPTS

• **Computing methodologies → Motion capture**; **Motion processing**; **Machine learning**; **Animation**.

KEYWORDS

Style transfer, CycleGAN, Unpaired data, Motion Analysis

ACM Reference Format:
Yuzhu Dong, Andreas Aristidou, Ariel Shamir, Moshe Mahler, and Eakta Jain. 2020. Adult2child: Motion Style Transfer using CycleGANs. In *Motion, Interaction and Games (MIG '20), October 16–18, 2020, Virtual Event, SC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3424636.3426909

1 INTRODUCTION

Children in the age group 8 to 11 years old have been found to spend as much as 8 hours weekly on video games [Johnson 2018]. Such trends make children important markets for the video game and electronic entertainment industry. Games such as Just Dance, by Ubisoft, and Ring Fit Adventure, by Nintendo, are designed to motivate children to exercise. As a result, there is a need to identify methods for synthesizing child motions.

Keyframing requires hours of manual effort from trained animators to create realistic and compelling motion. Motion capture (mocap), the leading technology for creating animated characters from actual human motion data, has the advantage of maintaining realism, capturing subtle secondary movements, and following real world physics [Menache 2000]. However, motion capturing children is full of difficulties. Children get confused with the instructions, lack patience, and are hard to collaborate with [Piaget 2015], especially at very young ages. These difficulties are the reason there are few online motion repositories. The most well-known mocap repositories, such as the CMU [2020] and OSU [2020] databases, consist only of adult motions. Currently, the Kinder-gator [Aloba et al. 2018] and the Human Motion Database [Guerra-Filho and Biswas 2012] are the only publicly accessible repositories that contain child motion. For games in particular, an abundance of action types, repetitions and variations allows for realism in real time play.

One way to overcome this scarcity of child motion data is to retarget easily available motion from adults to a child sized skeleton. However, retargeting mostly involves changes in the dimensions of limbs, so mapping adult motion directly on child characters fails to transfer the *style* and nuances of the children motion such as speed and variability. Style translation, that is, learning a mapping between two labeled motion capture sequences, has been extensively studied, starting with approaches by Brand and Hertzmann [2000] and Gleicher [1998] to recent advances made by deep neural networks [Aberman et al. 2020; Du et al. 2019a; Holden et al. 2017, 2016, 2015; Mason et al. 2018; Smith et al. 2019].

In this paper we devise an adult-to-child motion translation algorithm based on the CycleGAN [Zhu et al. 2017] architecture. CycleGAN has been successfully used in the past for transforming image styles *without* paired training data. This characteristic is critical for adult-to-child translation due to the very limited availability of child data. Generative Adversarial Networks (GANs) have rarely been used in character animation because of the difficulty to train a mapping that exhibits temporal dynamic behavior and generates temporally coherent and realistic movements. We show

MIG '20, October 16–18, 2020, Virtual Event, SC, USA

ACM ISBN 978-1-4503-8171-0/20/10…$15.00
https://doi.org/10.1145/3424636.3426909

that GANs have the capacity to learn the mapping between two distinct distributions, thus enabling style translation from the adult domain to a child domain in the absence of motion alignment.

We first collected 23 motions from both adults and children, which we plan to make available for future research. Motion capturing children is uniquely challenging and our motion dataset creates an authentic corpus for future research on style translation and motion generation. To train our *adult2child* network, we sliced motion sequences into shorter temporal windows called *motion words*, in a similar manner as Aristidou et al. [2018a]. The use of motion words helps the network to learn both the spatial and temporal information about that motion. The synthesis of motion words is conditioned using several losses, including an adversarial, cycle consistency, and temporal coherence. All these conditions contribute to a realistic and smooth motion synthesis that enhances the stylistic variability in childish motion.

We evaluated our *adult2child* translation in terms of naturalness and child-like-ness via a perception study. User responses indicate that our method produces motion that can be distinguished from adult motion in terms of child-like-ness, similar to state-of-the-art methods by Holden et al. [2015] and Aberman et al. [2020], but without needing motion alignment, and with significantly lesser pre-processing and training data.

The main scientific contributions of the paper are: (1) **Architecture:** We are the first to adapt a cycleGAN architecture for motion style transfer in such a way that the neural network is able to alter the timing of the motion. We redesigned the generators and the discriminators to extract meaningful features from motion inputs. Our demonstration of this adapted architecture opens the path forward for style transfer networks that do not need temporally aligned data. We further demonstrate the advantage of temporal coherence loss terms within the cycleGAN framework to create natural and smooth output motions. (2) **Representation:** We espouse joint angles as an animation-centric representation scheme for this architecture and task. This representation sets us apart from previous machine learning-centric work that has used joint positions to make it easier to train the network. An animation-centric approach, in contrast, looks ahead to how the output of the neural network will be bound to a skeleton and skinned. We further add to evidence in favor of motion words as motion representation schema that can encode both temporal and spatial changes. (3) **Dataset:** We release a high quality dataset of children's movements on a publicly accessible repository. This dataset is the first of its kind as it captures via an optical motion capture system the natural behavior of preteen children in response to verbal prompts.

2 RELATED WORK

In this section, we discuss briefly the literature that has studied typically developing children in contrast to typically developing adults, methods for motion retargeting and style transfer, and examine how the assumptions and design constraints underlying state of the art algorithms impact the *adult2child* style translation problem.

Domain Knowledge about Child Motion: A child body is not a mini-version of an adult body. There are several differences: for instance, the ratio of the size of the head to body height continues to decrease as children grow; the center of gravity for children is located higher than adults, etc. [Huelke 1998]. Several studies have been conducted that study the differences between the child and adults behavior and motion, in terms of mass body, motor control, skillful, coordination, and energy [Aloba 2019; Hraski et al. 2015; Huelke 1998; Nader et al. 2008]. More particularly, Jain et al. [2016] found that naive viewers can identify if a motion was performed by a child or an adult, even when they are completing the same actions. It follows that child motions are stylistically different from adult motions.

Motion Retargeting and Dynamic Scaling: Previous graphics research has extensively studied an approach to adapt motion from one skeleton to a differently sized skeleton, namely, motion retargeting [Choi and Ko 1999; Gleicher 1998; Hecker et al. 2008]. One could say that easily available adult motion capture data can be retargeted to a child sized skeleton. However, while skeletal retargeting captures changes in motion needed to accommodate changes in limb lengths, it does not account for biomechanical differences. Dynamic scaling accounts for biomechanics by scaling both length and time in such a way that gravity is preserved [Hodgins and Pollard 1997; Raibert and Hodgins 1991]. However, as shown by Dong et al. [2017], scaling adult motion directly on child characters fails to transfer the childish, playful, and carefree style of the children motion. Dong et al. [2017] found also that the dynamically scaled motions look more childlike to naive viewers, but not as childlike as motion captured from actual children. We have adapted their perceptual evaluation framework so as to rank various approaches by their effectiveness at the *adult2child* problem.

Style Transfer: Distinct from motion retargeting, a number of methods have been developed in the literature to transform motion so that it has a different style even if it is on the same skeleton, for example, a neutral walk to a happy walk [Amaya et al. 1996; Unuma et al. 1995; Witkin and Popović 1995]. Because style is a subjective characteristic, difficult to express with mathematics, methods were developed to infer style features given exemplars in the form of large databases [Aristidou et al. 2017; Brand and Hertzmann 2000; Ikemoto et al. 2009; Ma et al. 2010; Shapiro et al. 2006; Taylor and Hinton 2009; Wang et al. 2007; Yumer and Mitra 2016]. In this class of methods, Dong et al. [2018] applied the method of Hsu et al. [2005] to the *adult2child* problem. They trained a linear time-invariant model to extract the stylistic aspect of motion using a database of matched pairs of adult and child motion. Methods that rely on paired training data have an inherent burden associated with them: collecting child motion capture data in such a way that they respond to the same prompts as adults. In contrast, a method that does not require matched pairs makes it much easier to collect exemplar child motions.

Deep Neural Networks: In recent years, deep learning methods have shown promising results in areas such as image processing and style transfer, e.g., [Gatys et al. 2016]. For instance, Holden et al. [2015] stylized motion by minimizing an optimization function that preserves naturalness and smoothness as well as matches a desired style that was extracted via autoencoders. There were subsequent approaches that sought to reduce the amount of paired training data needed for style transfer, requiring only a limited number of style examples or a short exemplar motion clip for the desired style supplemented with a large database of neutral motions [Du et al. 2019a,b; Mason et al. 2018]. These methods, also only work for locomotion and cyclic motions. We note that even a database of neutral motions requires children to be motion captured, which is a bottleneck in itself. Thus motivated, our approach investigates the use of CycleGANs to learn style transfer with a relatively small number of exemplars.

Recently, Smith et al. [2019] introduced a computationally efficient method using three multi-layer neural networks that motion to be adjusted in a latent style space, thus achieving real-time style transfer with low memory requirements. The main limitation in their approach is that they have a separate network to learn the timing, and this timing adjustment is applied as a post-process. Aberman et al. [2020] encode motion into two latent codes, one for motion content and one for style. Their approach injected a new style by altering the deep features of motion, but did not change the timing. In contrast to these recent approaches, we learn timing along with pose within the stylization network.

Generative Adversarial Networks (GANs) work well with a small dataset compared to other deep learning architectures. They also do not require coupled training data. However, despite the success of GANs in image and video processing Isola et al. [2017], they are not popular in character animation because of the difficulty in modeling the temporal dynamics of movements. The few methods that use adversarial learning for motion synthesis and stylization [Barsoum et al. 2018; Wang et al. 2020] only deal with pose changes and not with timing. CycleGAN is an architectural variant of GANs that allows for unpaired datasets to be leveraged for style transfer [Zhu et al. 2017]. Our contribution is to adapt this framework for the temporal dimension. Similar to patch-based image style translation, we divide motion sequences into short temporal windows named motion words. Motion words allow the network to learn both temporal and spatial information about the motion [Aristidou et al. 2018b]. We further extend the original CycleGAN network by introducing two new loss terms in the network's architecture, one to consider the temporal evolution of motion, and the other for smooth blending between the motion words. Our ablation study evaluated the impact of each introduced loss term separately.

3 DATA ACQUISITION

Existing motion capture datasets, such as the CMU [2020] motion capture database, are publicly available but do not provide examples of children. Other datasets that include motion data of children, such as the Aloba et al. [2018] use a Microsoft Kinect v1.0 device for data collection. The use of a Kinect v1.0 is limiting in several ways: it is a low-resolution RGB-depth camera with a 320x240 16-bit depth sensor and a 640x480 32-bit color sensor, at a capturing rate of 30Hz. This low spatial and temporal resolution favours interactive gaming experiences over accurate pose reconstruction, resulting in the loss of key information for faster motions. A kinect v1.0 device also limits the capture space to a small region of about two square meters. As a result, the Aloba et al. [2018] dataset features noisy information, with unstable foot contacts, and limits the subject to perform motions such as walking and running in-place.

In this work, we acquired adult and child similar motions using a 10-camera Vicon optical motion capture system (1080p resolution). This commercial system is capable of capturing retro-reflective 3d markers, with sub-millimeter accuracy, at a frame rate of 120Hz. We use a total 53 markers per subject. Each marker is carefully placed on the subject, denoting pivot points and joint segmentation, creating an accurate reconstruction of a fully articulated subject. We instruct the subjects to perform their actions in a capture volume of 10×8 meters, a space large enough to capture full locomotion cycles and other dynamic and expressive motions.

We invited nine adults (older than 18 years) and eight children (all participants were from 5-10 years old) and recorded a variety of dynamic motions. We collected a variety of takes that can be categorized into three types of motions: (a) discrete actions, (b) cyclic locomotion, and (c) dynamic combinations. In our observations, children's actions are playful, less predictable than adults, and appear uncoordinated at times; this is precisely what makes a child's motion authentic. We captured the following *discrete action* examples: "Throw a ball with left arm", "Throw a ball with right arm", "Punch", "Kick", "Jump with one leg","Jump with the other leg", "Idle", "Broad Jump Forward", "Jump as high as you can in place" "Jump", "5 Jumping jacks"; *cyclic locomotion* examples:"Walk", "Walk as fast as you can", "Hop Scotch", "Sneaky Walk", "Happy Walk", "Jog", "Run as fast as you can", "Skip"; and *dynamic combination* examples: "Run and Jump", "Walk, step over obstacle". We captured 2-4 repetitions for each action type for each subject.

Raw motion capture data were converted to .bvh file format (using the default Vicon application); data acquisition problems, such as missing data due to marker occlusions, were fixed with using default features in Shogun post software. The motion data was categorized and labeled with the motion type, subject type (adult or child), and subject ID. We released this data on a publicly accessible repository for future studies (in fbx, bvh, and csv formats). (URL: https://jainlab.cise.ufl.edu/publications.html#Adult2ChildCycleGAN)

More details about data acquisition and the peculiarities of capturing and dealing with young children can be found in Appendix A.

4 METHODOLOGY

The main premise of this work is that motion sequences can be broken down to smaller movements and then we can apply image translation algorithms in a similar manner to image patches; those short temporal windows can carry the temporal and spatial properties of motions, including the temporal evolution, and capture their essences and stylistic behavior.

4.1 Data Representation

Motion data are represented using joint angles wherein each articulated skeleton consists of 25 joints. The choice of joint angles as the motion representation sets us apart from most existing work where joint positions are used as input to neural networks. This design decision reflects an animation-centric approach to the style transfer problem, rather than a machine learning-centric approach. With this approach we elevate the desirability of a skinned final result, which makes it necessary to preserve the rotation along the bone axes. Another key consideration is that limb lengths are very different between children and adults, and across children themselves. Joint angles capture that the elbow is bent about the same for a child and an adult but the knee is not (for example) without needing an intermediate retargeting step to a canonical skeleton.

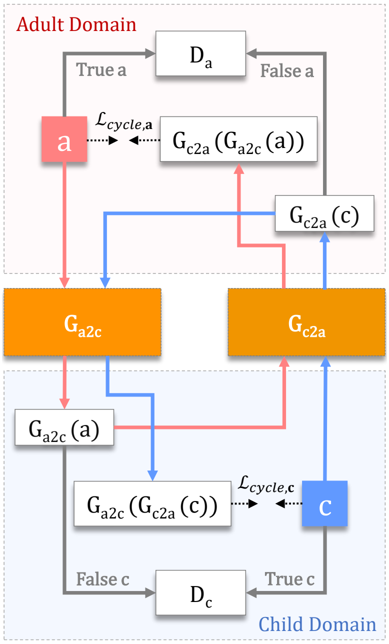


Figure 1: Our *adult2child* framework, based on the CycleGAN architecture.

All joint angles are converted into unit quaternion format; the quaternion representation is free of gimbal lock, and it supports interpolation. Since the subjects may face different directions during the experiment, we translated the motions into local root space and align their direction. The original motions were captured in 120fps. We downsampled the motions to 60fps. Our motion sequences are divided into *motion words* with a temporal window of 60 frames, e.g., [Aristidou et al. 2018a]. and an 20 frames stride; these overlapping frames are later used for blending purposes. We normalized the data inputs to the range of [-1,1] using the overall max and min values, and then, at a post-processing step, we de-normalized the values back to their original range.

4.2 Network architecture

We adopted the CycleGAN architecture to learn the mapping between adult motions to childlike motions. The network was trained on one motion word pair at a time. Usually CycleGAN networks are trained using unpaired data from two specific domains, e.g., horse and zebra images. Similarly, in our case, we train the network with motion words of the same motion type (e.g. adult jump with child jump, adult kick with child kick, etc.).

Our network architecture consists of two GANs: one for *adult2child* translation and the other one for *child2adult* translation (see Figure 1). According to Zhu et al. [2017], having the two GANs forming a cycle enables training without paired data and prevents modes collapse. We denote adult motion as **a** and child motions as **c** . We have two generators:$\mathbf{G_{a2c}}$ maps adult motions to fake child motions while $\mathbf{G_{c2a}}$ maps the child motions back to adult motions. We also have two discriminators: $\mathbf{D_a}$ that distinguish original adult motions **a** from the fake adult motions $\mathbf{G_{c2a}(c)}$, and $\mathbf{D_c}$ that differentiates the original child motions **c** from the fake child motions $\mathbf{G_{a2c}(a)}$.

Figure 2 describes the network generator; we used a stride-1 depth-wise convolution layer with filter size 7, followed by two stride-2 depth-wise convolution layers with filter size 3. The joints are arranged along the depth dimension of a layer. The kernel only convolves along the quaternion dimension and the temporal dimension, leaving individual joints separate. The output of the third convolution layer feeds into nine residual blocks. Each residual block consists of two convolution layers with filter size 3 with a skip layer connection. The decoder of the generator (Figure 3) consists of two stride-1/2 deconvolution layers and one stride-1 convolution layer of stride-1 with filter size 7. Similarly to the original CycleGAN natwork, the discriminator is a four-layer convolutional network with 64, 128, 256, 512 filters in each layer. All the layers have a filter size 4. The first three layers have a stride length of 2, while the fourth and fifth layers have a stride length of 1.

4.3 Loss function

The goal of the GANs is to generate a mapping $\mathbf{M} : \mathbf{i} \mapsto \mathbf{j}$, by creating instances similar to the target distribution conditional to the input. The original CycleGAN consists of two losses: an *adversarial loss*, that makes sure that the distribution of the generated motion $\mathbf{M(i)}$ is indistinguishable from the distribution of the real motion $\mathbf{j}$, and a *cycle consistency loss* that couples it with an inverse mapping $\mathbf{I} : \mathbf{j} \mapsto \mathbf{i}$, to enforce motion to go back to its original domain. We introduced two additional loss terms: the *temporal coherence loss* and the *transition loss*. The temporal coherence loss ensures that motions are smooth, preventing sudden jerks, and the transition loss penalizes the differences between the overlapping frames in the adjacent motion words. However, we observed that the transition loss introduced some artifacts, such as unusual poses for overlapping frames, and thus we left the transition loss out in the final loss function. We show those artifacts in the supplementary video.

4.3.1 *Adversarial loss.*

Equation 1 shows the loss for mapping the adult domain to the child domain, while Equation 2 shows the loss for the opposite mapping.

In our adversarial framework, the generator aims to create motion words that will be recognized as being a child's motion, while the discriminator aims to catch generated instances as being translated motion rather than real motion capture.

The discriminator D_c learns to assign 1 to motions that were captured from child actors and 0 to motions to style translated motion. The discriminator D_a correspondingly learns to assign 1 to real adult motions and 0 to style translated motion. We adopted the least square loss to avoid the vanishing gradient problem [Mao et al. 2017].

$$\mathcal{L}_{\mathrm{D_c}} = 0.5 * \mathbb{E}_{\mathrm{c}\sim p(\mathrm{c})}[\mathrm{D_c(c)} - 1] + 0.5 * \mathbb{E}_{\mathrm{a}\sim p(\mathrm{a})}[\mathrm{D_c(G_{a2c}(a))}] \tag{1}$$

$$\mathcal{L}_{\mathrm{D_a}} = 0.5 * \mathbb{E}_{\mathrm{a}\sim p(\mathrm{a})}[\mathrm{D_a(a)} - 1] + 0.5 * \mathbb{E}_{\mathrm{c}\sim p(\mathrm{c})}[\mathrm{D_a(G_{c2a}(c))}] \tag{2}$$

$$\mathcal{L}_{\mathrm{G_{c2a}}} = 0.5 * \mathbb{E}_{\mathrm{c}\sim p(\mathrm{c})}[\mathrm{D_a(G_{c2a}(c))} - 1] \tag{3}$$

$$\mathcal{L}_{\mathrm{G_{a2c}}} = 0.5 * \mathbb{E}_{\mathrm{a}\sim p(\mathrm{a})}[\mathrm{D_c(G_{a2c}(a))} - 1] \tag{4}$$

4.3.2 *Cycle consistency loss.*

We adopted the cycle consistency loss from [Zhu et al. 2017] to increase the visual quality of the output motion. Because our dataset contains different motion types, having

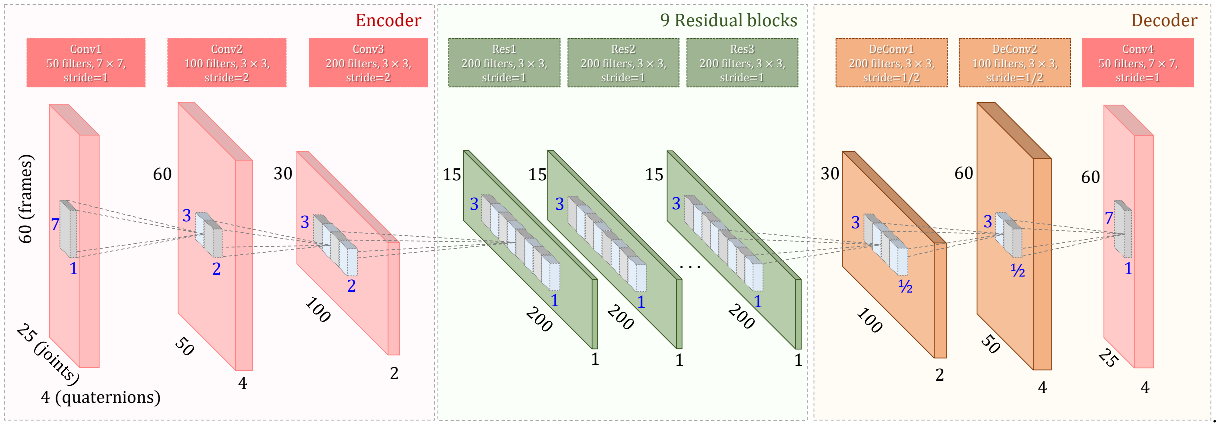

Encoder
9 Residual blocks
Decoder
Conv1 50 filters, 7 × 7, stride=1
Conv2 100 filters, 3 × 3, stride=2
Conv3 200 filters, 3 × 3, stride=2
Res1 200 filters, 3 × 3, stride=1
Res2 200 filters, 3 × 3, stride=1
Res3 200 filters, 3 × 3, stride=1
DeConv1 200 filters, 3 × 3, stride=1/2
DeConv2 100 filters, 3 × 3, stride=1/2
Conv4 50 filters, 7 × 7, stride=1
60 (frames)
25 (joints)
4 (quaternions)
7
1
60
3
2
50
4
30
3
2
100
2
15
3
1
200
1
15
3
1
200
1
...
15
3
1
200
1
30
3
½
100
2
60
3
½
50
4
60
7
1
25
4

Figure 2: Architecture of the generator: it consists of a encoder with three convolutional layers, 9 residual blocks, and a decoder with three deconvolutional layers. Number of kernels, kernel size, and stride length of each layer are also denoted.

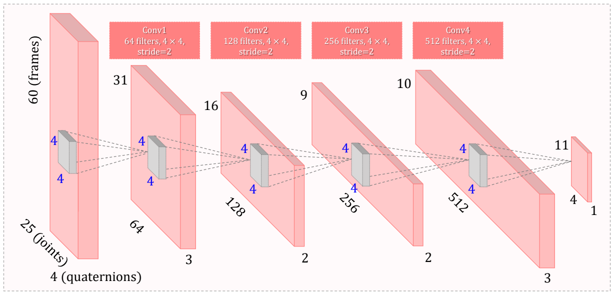

Conv1 64 filters, 4 × 4, stride=2
Conv2 128 filters, 4 × 4, stride=2
Conv3 256 filters, 4 × 4, stride=2
Conv4 512 filters, 4 × 4, stride=2
60 (frames)
25 (joints)
4 (quaternions)
4
4
31
4
4
64
3
16
4
4
128
2
9
4
4
256
2
10
4
4
512
3
11
4
1

Figure 3: Architecture of the discriminator

cycle consistency loss helps to preserve the content of the input motion. It is defined as the differences between $\mathbf{G_{a2c}}(\mathbf{G_{c2a}}(\mathbf{c}))$ and $\mathbf{c}$ for child domain (see Equation 5) and the differences between $\mathbf{G_{c2a}}(\mathbf{G_{a2c}}(\mathbf{a}))$ and $\mathbf{a}$ for adult domain (see Equation 6).

$$\mathcal{L}_{cycle,\mathrm{c}} = \mathbf{G_{a2c}}(\mathbf{G_{c2a}}(\mathbf{c})) - \mathbf{c} \tag{5}$$

$$\mathcal{L}_{cycle,\mathrm{a}} = \mathbf{G_{c2a}}(\mathbf{G_{a2c}}(\mathbf{a})) - \mathbf{a} \tag{6}$$

4.3.3 Temporal coherence loss. The temporal coherence loss (Equation 7) is introduced to increase the smoothness and stability of the output motion. We compute the first derivative of the output motion as the delta between two consecutive poses for all the degrees of freedom across all the frames in one motion word. It limits the angle differences in the adjacent frames to prevent sudden changes in the motions.

$$\mathcal{L}_{coherence,\mathrm{a}} = \sum_{t} \sum_{DOF} ||\mathbf{G_{a2c}}(\mathbf{a})(t) - \mathbf{G_{a2c}}(\mathbf{a})(t-1)|| \tag{7}$$

$$\mathcal{L}_{coherence,\mathrm{c}} = \sum_{t} \sum_{DOF} ||\mathbf{G_{c2a}}(\mathbf{c})(t) - \mathbf{G_{c2a}}(\mathbf{c})(t-1)|| \tag{8}$$

4.3.4 Transition loss. Our aim in adding the transition loss (Equation 10) is to create a smooth transition between motion words. The transition loss penalizes the differences in the overlapping frames of adjacent motion words. We use the average of the overlapping frames in the adjacent motion words for blending. We denote overlapping frame number as $t_{overlap}$, and the motion word index as i.

$$y = \mathbf{G_{c2a}}(\mathbf{c}) \tag{9}$$

$$\mathcal{L}_{transition,\mathrm{c}} = \sum_{t} \sum_{DOF} ||y_i(t_{overlap:end}) - y_{i+1}(0 : t_{overlap})|| \tag{10}$$

4.3.5 Overall loss function. The overall loss function (Equation 11) can be computed as the weighted sum of all the terms described above. $\lambda_{.}$ denotes the weights of each loss term, that were decided experimentally via ablation studies, as shown in Section 5.

$$\begin{aligned}\mathcal{L}_{\mathrm{G_{a2c},G_{c2a},D_a,D_c}} = {} & \mathcal{L}_{\mathrm{G_{a2c}}} + \mathcal{L}_{\mathrm{G_{c2a}}} + \mathcal{L}_{\mathrm{D_a}} + \mathcal{L}_{\mathrm{D_c}} \\ & + \lambda_{cycle} \times (\mathcal{L}_{cycle,\mathrm{c}} + \mathcal{L}_{cycle,\mathrm{a}}) \\ & + \lambda_{coherence} \times (\mathcal{L}_{coherence,\mathrm{a}} + \mathcal{L}_{coherence,\mathrm{c}}) \\ & + \lambda_{transition} \times L_{transition,\mathrm{c}}\end{aligned} \tag{11}$$

We trained two generators to minimize the generator loss and trained two discriminators to minimize the discriminator loss.

$$L = \arg \min_{\mathrm{G_{a2c},G_{c2a}}} \min_{\mathrm{D_a,D_c}} \mathcal{L}_{\mathrm{D_a,D_c,G_{a2c},G_{c2a}}} \tag{12}$$

The contribution of each of these components is evaluated in the ablation study in Section 5.3.

4.4 Post-processing

In the post-processing step, we scale the values back to the original range of values using the saved min and max values for the childlike motions. Then, we stitch the output (translated) motion words of the childlike motions $\mathbf{G_{a2c}}(\mathbf{a})$ back to the original length in order. We blend adjacent motion words by taking the average of the overlapping frames. We then applied peak removal filter and Butterworth filter with cutoff frequency of 7Hz to further smooth out the motions.

5 RESULTS AND DISCUSSION

In this section, we provide several experiments to evaluate the performance of our *adult2child* motion translation in terms of realism and style expressiveness of the generated motion, and compare it with two alternative baseline methods: [Aberman et al. 2020] and [Holden et al. 2016]. Adult-to-child motion translation is a relatively special case of style transfer in character animation, and this poses several challenges in the evaluation and comparison of our method, mainly due to the lack of available motion datasets; thus, we have evaluated our method with others using only the dataset we have collected. We also conducted ablation studies to quantitatively examine the performance and contribution of each integrated losses. Last but not the least, we conducted a perceptual study to evaluate the quality of our results.

5.1 Implementation Details

We implemented and trained the model on Google Colab Pro with NVIDIA P100 graphics card. The model was written in Python using TensorFlow. More specifically, we trained the generators and discriminators in the following order: $\mathbf{G_{a2c}}, \mathbf{D_c}, \mathbf{G_{c2a}}, \mathbf{D_a}$, and we set $\mathcal{L}_{coherence} = 30$ and $\mathcal{L}_{cycle} = 15$. We used the Adam solver for gradient descent, while the batch size is 1 to incorporate the transition loss. The network was trained 180 epochs with a learning rate of 0.0002, and the training takes approximately 7 hours. After 100 epochs, the learning rate is decreased at a rate of 1% per epoch. The size of the trained model is 8.67MB. The training data includes 90% of motions from three mocap actors. Our test data includes the remaining 10% of data from the training set plus motions from two different actors. The purpose is to test the generalizability of the model, and how it performs on motions that are similar to the training data as well as motions that are different from the training data. We additionally trained two baseline methods for comparison purposes. For Holden et al. [2016]'s architecture, we adopted the pretrained network and fed our adult motion and child motion pairs from the same motion types. For Aberman et al. [2020], we found the pretrained network performed poorly with our adult motion, therefore we trained the network for another 1000 epochs with our dataset.

We applied the *adult2child* architecture on our motion dataset, using two different architectures, as shown in Figure 4 joint-wise convolution and overall convolution. For joint-wise convolution, we computed separate kernels for each joint. The kernels convolves the motion words along the time axis and the quaternion axis. For overall convolution. the kernels traversed each motion word along the time axis and the joint axis. We have observed that the overall convolution networks failed to create natural and childlike motions. This happens since the same 2D kernels were applied to all the joints, thus the network failed to extract the style from the entire skeleton. On the other hand, the joint-wise convolution networks operates on joint basis. In this case, different kernels were applied on each joint separately, with different set of weights; allowing the network to learn the transfer function on per joint makes style modeling much easier.

5.2 Experimental Results

In this work, we introduce the *adult2child* framework to translate adult motions to child-like motions without assuming the availability of temporally aligned pairs of motion sequences in our training database. Anecdotally, we have noticed that our output results exhibit some childlike characteristics: faster pace and swinging arms. Figures 5 and 6 demonstrate "jump-as-high-as-you-can"and "walk-as-fast-as-you-can" motions that have been translated from an adult's motion. We show frame snapshots taken every 5 frames. The top row shows the input adult motion used to generate the fake child motions shown below, the second and third rows show a generated childlike translated motion using the Holden et al. [2016] and Aberman et al. [2020] methods, respectively, the forth row shows results using our *adult2child* method, whereas the fifth row shows a reference child motion. It can be observed that both the timing and the poses for our output childlike motions are more similar to the movements of a child than adults. In particular, the spine of the fake childlike motion generated using our method is more exaggerated, while the movement of the hands rise-up appears to be carefree and playful, compared to the tight and structured movement of the adult; indeed, the methods of Holden et al. [2016] and Aberman et al. [2020] modified the adults motion to look more like a child, but the poses of the generated motion do not look as close to the actual child motion as ours. Finally, in contrast to the method of Aberman et al. [2020], our method takes into consideration the timing; the output childlike motion starts at the same pose as the input adult motion, but towards the end, the output motion is about 10 frames (one snapshot) faster than the input adult motion. Readers are also encouraged to watch more examples of the results in the accompanying video.

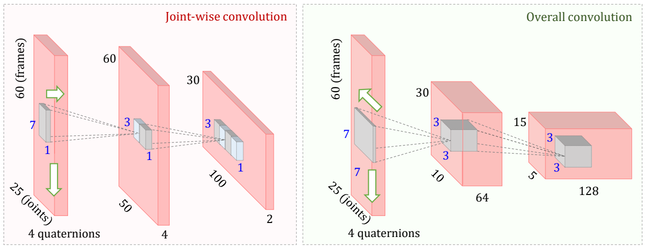
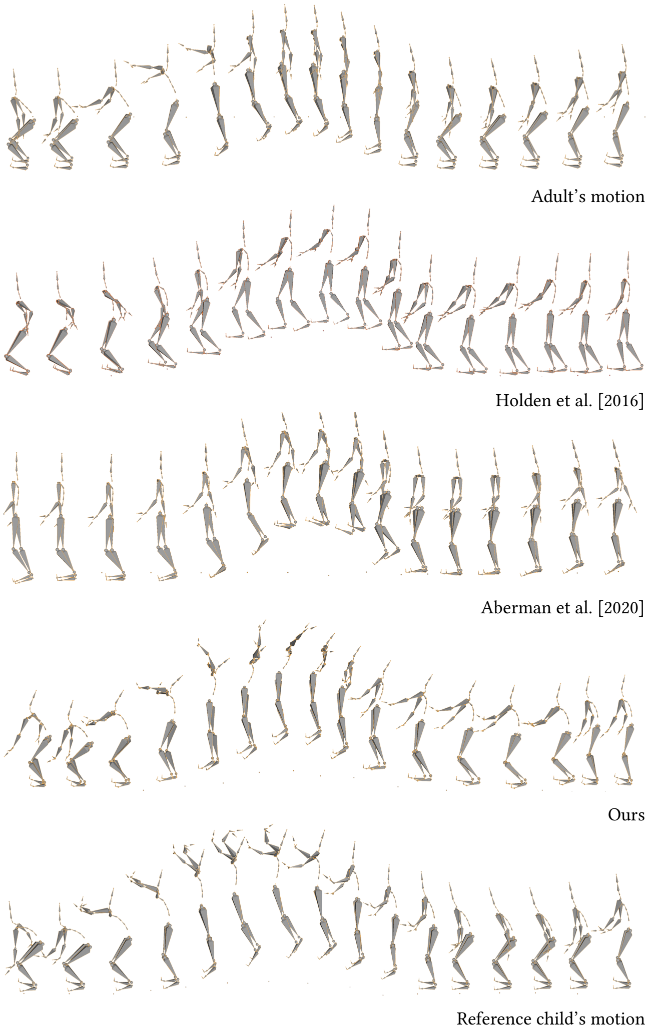


Figure 4: The 1D and 2D convolutional networks tested in our experiments. The joint-wise convolution convolves along the temporal axis (height) and the quaternion axis (depth). The overall convolution convolves along the temporal axis and the joint axis (width).

5.3 Ablation study

We conducted an ablation study to evaluate the effectiveness of each component of our loss function $\mathcal{L}_{\mathbf{G_{a2c}},\mathbf{G_{c2a}},\mathbf{D_a},\mathbf{D_c}}$, as described in Eq. 11. We re-trained the entire network for 180 epochs, but each time, we removed one component term in the loss function. Then, we generated childlike motions using the same input motion for each re-trained network.

Figure 7 illustrates the results of our ablation study in keyframe poses; in this example, frame snapshots are taken every 3 frames. The top row shows the output childlike motion when translated from an adult's one, with all losses included, while the second row shows the same *adult2child* translation, but this time without employing the temporal coherency loss. It can be observed that the generated motion, when the temporal coherence loss is excluded, is not smooth, neither natural, while the movement of some key



Adult's motion

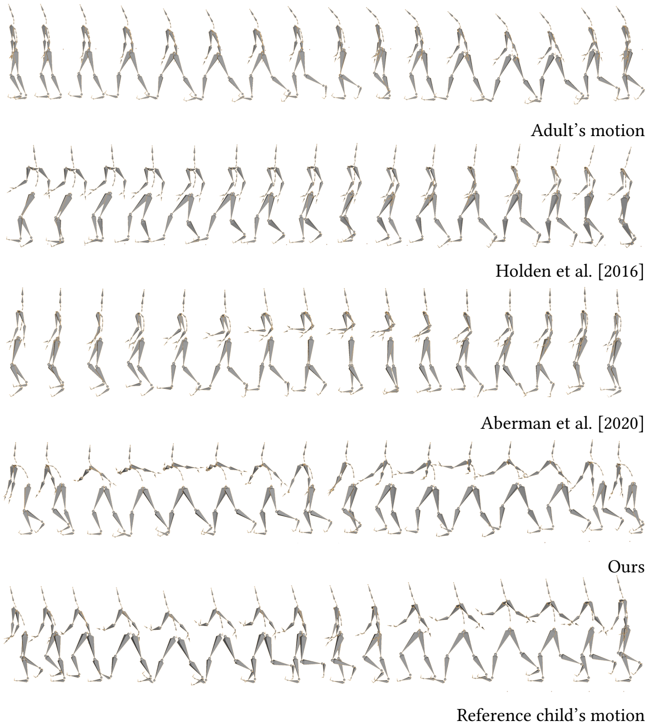
Holden et al. [2016]

Aberman et al. [2020]

Ours

Reference child's motion

Figure 5: Results on "jump-as-high-as-you-can". It can be seen that the spine of the translated childlike motion using our method is more exaggerated than those of Holden et al. [2016] and Aberman et al. [2020], to look more similar to the reference child motion, while the hands are risen up in a more playful way.



Adult's motion

Holden et al. [2016]

Aberman et al. [2020]

Ours

Reference child's motion

Figure 6: Results on "walk-as-fast-as-you-can". It can be observed that the output childlike motion using our method has the arms move in a carefree, uncoordinated, and playful mode during the walking cycle to mimic the child walking, whereas this is not seen in the original adult motion.

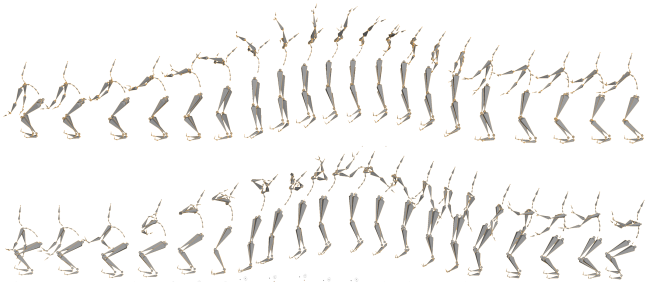


Figure 7: Ablation study: The top row shows the output childlike motion with all losses, while the bottom row illustrates the same motion translation without the temporal coherence loss. It can be clearly observed that the translated motion in the latter case is not as smooth as the motion when all losses are used.

joints are jittery with absurd changes in direction. In contrast, we can see that the translated motion, when all loss terms are included, is natural, temporally smooth, and consistent. Note that, we only demonstrate the results visually by showing selected key frames since, as pointed out by Barsoum et al. [2018], the loss function does not reflect the quality of the motion, and thus is not a reasonable metric for evaluating the performance of GANs.

We initially included a transition loss term to minimize the differences between the overlapping frames of the adjacent motion words. In theory, the transition loss is responsible to smoothly connect two motion words, ensuring that the output motion word will not have incomplete cycles, e.g., a character that start a new step without finishing its previous one. However, we noticed that, in some cases, this loss term caused the overlapping frames to be similar, but drastically different from the rest of the sequence. The network therefore creates weird results to minimize the loss function but in an abnormal way, resulting in jerky and unstable motions. Hence it was excluded from the final loss function. The effect of the loss in the ablation study is better seen in the supplementary video.

5.4 Perceptual study

To qualitatively evaluate the effectiveness and performance of our *adult2child* algorithm, we conducted a perceptual study. Our study consists of five conditions: adult, child, our method, [Holden et al. 2016], that was used as baseline, and Aberman et al. [2020], that is a very recent unpaired motion style transfer paper. To avoid the study being too long, we included one cyclic and one non-cyclic motion for each condition: "jump-as-high-as-you-can" and "walk-as-fast-as-you-can". We selected those two movements for our experiment since in these examples the child actors performed with more exaggerated poses, and hence style appears to be more distinctive between the two domains. Each stimuli is 10 seconds long; for motions clips that are shorter than 10 seconds, we looped the motions to extend its length (a transition slide was added between each clip saying "Repeat").

In total, 41 participants (14 females, 26 males, 1 other) were presented with 50 animations, in random order, and asked to evaluate the childlike style, and the naturalness of the presented motion. More specifically, following each animation, participants were asked three questions: "Does this motion belong to a child or an adult?" with two options to select from: "Adult", "Child". In the second question, we asked the viewer to indicate the naturalness of the motion on a 7-point Likert scale. As a sanity check to make sure viewers were attentive to the video, we asked the viewer "What is the action being performed?". The statistics of our perceptual study are listed in Figure 8. We excluded five participants whose overall correctness score are equal or below 50%. We calculated the percentage child responses as the number of motions a participant labeled as child motions divided by the total number of motions in that condition. We showed the box plot of percentage child responses and naturalness score in Figure 8. The bold bar in the middle shows the median value. The upper edge and the lower edge of the box mark the first quartile and the third quartile of the distribution, respectively. It can be observed that the participants distinguished the results of our method from those of adult's motion with similar statistics to state-of-the-art. More specifically, our algorithm achieves similar percent scores as Aberman et al. [2020], in terms of childlike motion perception and the naturalness of the translated motion, while our method rated higher than Holden et al. [2016] in the latter comparison. We found a significant main effect by actor type through a one-way ANOVA test ($F_{4,140} = 23.66$, $p < 0.0001$). The results from Tukey HSD post-hoc test shows that our results received significantly more child responses than the adult motions ($p < 0.001$). These observations, combined with the ability of our method to modify the timing in motion using less training data, and without the requirement to be paired or temporally aligned, indicate that CycleGAN is a reliable architecture to enable style transfer between motions of different domains.

5.5 Adult-to-Child-to-Adult

To further evaluate the performance of our method, we compared the double translated motions $\mathbf{G_{c2a}}(\mathbf{G_{a2c}}(\mathbf{a}))$ with the input motion $\mathbf{a}$, and $\mathbf{G_{a2c}}(\mathbf{G_{c2a}}(\mathbf{c}))$ with the input motion $\mathbf{c}$. We observed that the double translated motions appears a bit jittery and disconnected, but the timing and poses are aligned; we believe that during the double transformation, any noise that has been introduced into the motion through the network is magnified. Please refer to our supplementary video for animated results.

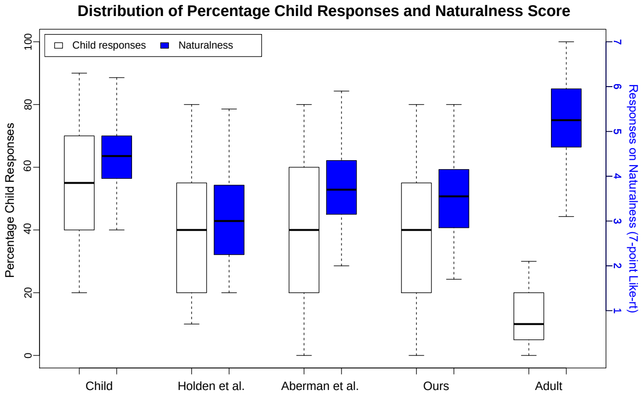


Figure 8: Distributions of the percentage of times participants selected this category of videos as belonging to a child are shown in white. The distributions of the naturalness score are shown in blue for each condition.

5.6 Discussion

In this work, we demonstrate that our *adult2child* framework, based on the well-known CycleGAN architecture, can be used to extract the style component and transfer it from one motion to another: the output motions appear more childlike compared to the original adult motions. It is also capable of learning different motion types within one network. Our work shows that CycleGAN's can be adapted to transfer both pose style and timing in a single network, in contrast to other recent works which do not handle temporal differences in style [Aberman et al. 2020; Smith et al. 2019]. Another advantage compared to the traditional style translation methods that use time-warping [Hsu et al. 2005; Smith et al. 2019], is that our algorithm can work on unaligned data and learn the temporal mapping from adult motions to child motions. Dynamic time warping is a time-consuming process, and importantly, it relies on having a beginning and and end frame that match. This match is hard to pinpoint and skipping this process can save both effort and time.

Moreover, our network can translate style from one motion to a different motion type. For instance, in the child running motion, the actors have their arms swinging in the air. This behavior was not observed in the original child walking and fast walking motions in training data. However, our output walking motion picked up this behavior in the running action and lifted the arms in the air for the fast walking motion. Indeed, users may perceive motions to be more childlike even if they do not exist in our limited children database. The truth is that there are many ways to be a child, and most networks will only learn those ways of being a child that are in the training data. Our method is capable to learn the stylistic characteristics from the entire dataset, and translate the learnt style from one action to another, even though the specific stylistic behavior do not exist in the original motion.

For example, in acyclic motions such as jumping high or punching, the output childlike motion learns from several input child motions, and can perform multiple repetitions of movements (such as punches or hand flinging) even if the input adult motion contains only one. This is different from the traditional style transfer algorithms, where the stylized motions follow exactly the movement and are different from only in terms of postures. Our method can alter both the temporal and the frequency/repetition of motions, which, we believe, are important aspects of childlike style. Nevertheless, a possible future direction could see the enrichment of the database, by sampling many different children to avoid learned bias: different ethnicities, developmental stages, personality types, etc.

The adult-to-child translated motions do not always look like original children motion. Indeed, there are cases where children perform an action in a way that adults simply do not do, e.g., a dab while walking. In such cases, our *adult2child* framework cannot encode this specific action as a stylistic feature to transfer it on the generated child motion. This happens since this unpredictable childish movement (e.g., the dab) is instant, not common, neither continuous, so as to be characterized as style.

6 CONCLUSIONS

We have presented a method that allows *adult2child* style transfer, converting the movement of adults so that it is perceived as a child's movement. We have introduced two additional losses to condition the network, one to learn the temporal coherency of motion, and another to minimize the transition cost. The use of motion words helps the network to learn both the spatial and temporal information about that motion, enabling the stylization network to learn timing along with pose. We also dealt with the current lack of children motion data availability by acquiring high quality children movements that will be released on a publicly accessible repository. Results demonstrate that our method can transfer a childlike style to adult movements, while a perceptual study confirms that our results are natural, and are perceived as child motions significantly more often than adult.

Limitations and Future Work: We have noticed that our method works well for cyclic motions but there is an artifact with non-cyclic motions. This happens because the network changes the timing of the input motion, but the sequence length remains the same. The output motions which are acyclic sometimes start from a later frame and fail to complete the motions before the sequence ends. Future work should investigate mechanisms to change the sequence length and allow smooth blending. We also noticed that our translated motions have foot-contact artifacts. As future work, the network could be trained to learn and label the foot contact information of each foot in the original training data, similar to [Shi et al. 2020], and then use inverse kinematics to enforce foot contact constraints in the output motions. In addition, the translated motion can be jittery; the same artifact appears also in the original CycleGAN work, where the translated image looked noisy. We suspect that this is intrinsic to the training process. The temporal loss term helped to mitigate this artifact and made the results look much smoother, but it did not solve entirely the problem. Nevertheless, this artifact can be further addressed by applying a peak removal filter and a smoothing filter (e.g., the Butterworth filter)

ACKNOWLEDGMENTS

This work was supported in part by Gartner Group Graduate fellowship; the Israel Science Foundation (grant No. 1390/19); the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 739578 and the Government of the Republic of Cyprus through the Directorate General for European Programmes, Coordination and Development.

REFERENCES

Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. 2020. Unpaired Motion Style Transfer from Video to Animation. *ACM Trans. Graph.* 39, 4 (July 2020).

Aishat Aloba. 2019. Tailoring Motion Recognition Systems to Children's Motions. In *2019 International Conference on Multimodal Interaction (ICMI '19)*. ACM, NY, USA, 457–462.

Aishat Aloba, Gianne Flores, Julia Woodward, Alex Shaw, Amanda Castonguay, Isabella Cuba, Yuzhu Dong, Eakta Jain, and Lisa Anthony. 2018. Kinder-Gator: The UF Kinect Database of Child and Adult Motion. In *Proceedings of the 39th Annual European Association for Computer Graphics Conference: Short Papers (EG)*. Eurographics Association, 13–16.

Kenji Amaya, Armin Bruderlin, and Tom Calvert. 1996. Emotion from Motion. In *Proceedings of the Conference on Graphics Interface '96 (GI '96)*. Canadian Info. Proc. Society, CAN, 222–229.

Andreas Aristidou, Daniel Cohen-Or, Jessica K. Hodgins, Yiorgos Chrysanthou, and Ariel Shamir. 2018b. Deep Motifs and Motion Signatures. *ACM Trans. Graph.* 37, 6, Article 187 (Nov. 2018), 13 pages.

Andreas Aristidou, Daniel Cohen-Or, Jessica K. Hodgins, and Ariel Shamir. 2018a. Self-similarity Analysis for Motion Capture Cleaning. *Comput. Graph. Forum* 37, 2 (May 2018), 297–309.

Andreas Aristidou, Qiong Zeng, Efstathios Stavrakis, Kangkang Yin, Daniel Cohen-Or, Yiorgos Chrysanthou, and Baoquan Chen. 2017. Emotion Control of Unstructured Dance Movements. In *Proceedings of the ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA '17)*. ACM, NY, USA, Article 9, 10 pages.

Emad Barsoum, John Kender, and Zicheng Liu. 2018. HP-GAN: Probabilistic 3D Human Motion Prediction via GAN. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE Computer Society, 1418–1427.

Matthew Brand and Aaron Hertzmann. 2000. Style Machines. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '00)*. ACM Press/Addison-Wesley Publishing Co., USA, 183–192.

Kwang-Jin Choi and Hyeong-Seok Ko. 1999. On-Line Motion Retargetting. In *Proceedings of the 7th Pacific Conference on Computer Graphics and Applications (PG '99)*. IEEE Computer Society, USA, 32.

CMU. 2020. Carnegie Mellon University MoCap Database: http://mocap.cs.cmu.edu/. [Online; Retrieved July, 2020].

Yuzhu Dong, Aishat Aloba, Lisa Anthony, and Eakta Jain. 2018. Style Translation to Create Child-like Motion. In *Proceedings of the 39th Annual European Association for Computer Graphics Conference: Posters (EG '18)*. Eurographics Association, Goslar, DEU, 31–32.

Yuzhu Dong, Aishat Aloba, Sachin Paryani, Lisa Anthony, Neha Rana, and Eakta Jain. 2017. Adult2Child: Dynamic Scaling Laws to Create Child-like Motion. In *Proceedings of the Tenth International Conference on Motion in Games (MIG '17)*. ACM, NY, USA, Article 13.

Han Du, Erik Herrmann, Janis Sprenger, Noshaba Cheema, Somayeh Hosseini, Klaus Fischer, and Philipp Slusallek. 2019a. Stylistic Locomotion Modeling with Conditional Variational Autoencoder. In *40th Annual Conference of the European Association for Computer Graphics, Eurographics 2019*, Paolo Cignoni and Eder Miguel (Eds.). The Eurographics Association, 9–12.

Han Du, Erik Herrmann, Janis Sprenger, Klaus Fischer, and Philipp Slusallek. 2019b. Stylistic Locomotion Modeling and Synthesis Using Variational Generative Models. In *Motion, Interaction and Games (MIG '19)*. ACM, NY, USA, Article 32, 10 pages.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image Style Transfer Using Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. 2414–2423.

Michael Gleicher. 1998. Retargetting Motion to New Characters. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '98)*. ACM, NY, USA, 33–42.

Gutemberg Guerra-Filho and Arnab Biswas. 2012. The Human Motion Database: A Cognitive and Parametric Sampling of Human Motion. *Image Vision Comput.* 30, 3 (March 2012), 251–261.

Chris Hecker, Bernd Raabe, Ryan W. Enslow, John DeWeese, Jordan Maynard, and Kees van Prooijen. 2008. Real-Time Motion Retargeting to Highly Varied User-Created Morphologies. *ACM Trans. Graph.* 27, 3 (Aug. 2008), 1–11.

Jessica K. Hodgins and Nancy S. Pollard. 1997. Adapting Simulated Behaviors for New Characters. In *Proceedings of the 24th Annual Conference on Computer Graphics and*

Interactive Techniques (SIGGRAPH '97). ACM Press/Addison-Wesley Publishing Co., USA, 153–162.

Daniel Holden, Taku Komura, and Jun Saito. 2017. Phase-Functioned Neural Networks for Character Control. *ACM Trans. Graph.* 36, 4, Article 42 (July 2017).

Daniel Holden, Jun Saito, and Taku Komura. 2016. A Deep Learning Framework for Character Motion Synthesis and Editing. *ACM Trans. Graph.* 35, 4, Article 138 (July 2016).

Daniel Holden, Jun Saito, Taku Komura, and Thomas Joyce. 2015. Learning Motion Manifolds with Convolutional Autoencoders. In *SIGGRAPH Asia 2015 Technical Briefs (SA '15)*. ACM, NY, USA, Article 18.

Marijana Hraski, Željko Hraski, and Ivan Prskalo. 2015. Comparison of standing long jump technique performed by subjects from different age groups. *Baltic Journal of Sport and Health Sciences* 98, 3 (2015), 2.

Eugene Hsu, Kari Pulli, and Jovan Popović. 2005. Style Translation for Human Motion. *ACM Trans. Graph.* 24, 3 (July 2005), 1082–1089.

Donald F. Huelke. 1998. An overview of anatomical considerations of infants and children in the adult world of automobile safety design. *Annual Proceedings / Association for the Advancement of Automotive Medicine* 42 (1998), 93–113.

Leslie Ikemoto, Okan Arikan, and David Forsyth. 2009. Generalizing Motion Edits with Gaussian Processes. *ACM Trans. Graph.* 28, 1, Article 1 (Feb. 2009).

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 1125–1134.

Eakta Jain, Lisa Anthony, Aishat Aloba, Amanda Castonguay, Isabella Cuba, Alex Shaw, and Julia Woodward. 2016. Is the Motion of a Child Perceivably Different from the Motion of an Adult? *ACM Trans. Appl. Percept.* 13, 4, Article 22 (July 2016).

Joseph Johnson. 2018. Hours children spend gaming weekly in the UK from 2013 to 2017, by age group. https://www.statista.com/statistics/274434/time-spent-gaming-weekly-among-children-in-the-uk-by-age [Online; Retrieved July, 2020].

Wanli Ma, Shihong Xia, Jessica K. Hodgins, Xiao Yang, Chunpeng Li, and Zhaoqi Wang. 2010. Modeling Style and Variation in Human Motion. In *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '10)*. Eurographics Association, 21–30.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*. 2794–2802.

Ian Mason, Sebastian Starke, He Zhang, Hakan Bilen, and Taku Komura. 2018. Few-shot Learning of Homogeneous Human Locomotion Styles. *Comput. Graph. Forum* 37, 7 (2018), 143–153.

Alberto Menache. 2000. *Understanding motion capture for computer animation and video games.* Morgan kaufmann.

Philip R. Nader, Robert H. Bradley, Renate M. Houts, Susan L. McRitchie, and Marion O'Brien. 2008. Moderate-to-vigorous physical activity from ages 9 to 15 years. *JAMA* 300, 3 (2008), 295–305.

OSU. 2020. Ohio State University MoCap Database https://accad.osu.edu/research/motion-lab/mocap-system-and-data. [Online; Retrieved July, 2020].

Jean Piaget. 2015. *The Grasp of Consciousness (Psychology Revivals): Action and Concept in the Young Child.* Psychology Press.

Marc H. Raibert and Jessica K. Hodgins. 1991. Animation of Dynamic Legged Locomotion. In *Proceedings of the 18th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '91)*. ACM, NY, USA, 349–358.

Ari Shapiro, Yong Cao, and Petros Faloutsos. 2006. Style Components. In *Proceedings of Graphics Interface 2006 (GI '06)*. Canadian Info. Proc. Society, CAN, 33–39.

Mingyi Shi, Kfir Aberman, Andreas Aristidou, Taku Komura, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. 2020. MotioNet: 3D Human Motion Reconstruction from Monocular Video with Skeleton Consistency. *ACM Trans. Graph.* (June 2020).

Harrison Jesse Smith, Chen Cao, Michael Neff, and Yingying Wang. 2019. Efficient Neural Networks for Real-Time Motion Style Transfer. *Proc. ACM Comput. Graph. Interact. Tech.* 2, 2, Article 13 (July 2019).

Graham W. Taylor and Geoffrey E. Hinton. 2009. Factored Conditional Restricted Boltzmann Machines for Modeling Motion Style. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09)*. ACM, NY, USA, 1025–1032.

Munetoshi Unuma, Ken Anjyo, and Ryozo Takeuchi. 1995. Fourier Principles for Emotion-Based Human Figure Animation. In *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '95)*. ACM, NY, USA, 91–96.

Jack M. Wang, David J. Fleet, and Aaron Hertzmann. 2007. Multifactor Gaussian Process Models for Style-content Separation. In *Proceedings of the 24th International Conference on Machine Learning (ICML '07)*. ACM, NY, USA, 975–982.

Qi Wang, Thierry Artières, Mickael Chen, and Ludovic Denoyer. 2020. Adversarial learning for modeling human motion. *Vis. Comp.* 36 (2020), 141–160.

Andrew Witkin and Zoran Popović. 1995. Motion Warping. In *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '95)*. ACM, NY, USA, 105–108.

Shihong Xia, Congyi Wang, Jinxiang Chai, and Jessica Hodgins. 2015. Realtime Style Transfer for Unlabeled Heterogeneous Human Motion. *ACM Trans. Graph.* 34, 4, Article 119 (July 2015).

M. Ersin Yumer and Niloy J. Mitra. 2016. Spectral Style Transfer for Human Motion between Independent Actions. *ACM Trans. Graph.* 35, 4, Article 137 (July 2016).

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 (ICCV'17)*. 2242–2251.

A APPENDIX

A.1 Child motion capture

In this section, we provide methodological innovations specific to young child motion capture.

Motion Selection: Before capturing a dataset, motion selection is an important design choice. Younger children may have a difficult time interpreting the actions. Actions that require much interpretation is not suitable for children to perform. We chose actions that children already incorporate in their daily life, such as hopscotch, skipping. Actions that are less likely to encounter in day to day life, such as "Fly like a bird" are not included in the dataset. In addition, to capture motions as training data to our system, we also captured combo motions as a testing dataset, such as run then jump.

Consent: The motion capture acquisition and study were approved by our university review board (IRB). Each subject signed an informed consent form before participating in the study. For child participants, we had an age-appropriate version of the consent form while their parents received the full consent form prior to performing the motion capture.

Space Setup: The main challenge with motion capturing children is that they are easily distracted. Because of the constraints with the lab space, the computer running Vicon Shogun software is co-located with the motion capture space. During capturing, some child participants turned around and focused on the monitor to see their embodied virtual avatar inside the Shogun software. To avoid this distraction, we turned the monitor around or moved it away so the participants could not see it. Another approach is to have research staff standing in front of the participants and remind them always to look ahead.

Markers and Suits: When performing the same action, children usually have a larger range of motions than adults. Their motion tends to be more dramatic than their adult counterpart. The markers are easily moved around. A research staff was responsible for paying attention to the marker positions to make sure they all stay in place during the study. If a marker fell off, we would pause the study and place the marker back to the original spot. If too many markers are disturbed or deviate from their original position the subject would need to be re-calibrated.

Vicon provides motion capture suits tailored for children including velcro straps to put on the children's shoes. We found that these straps often fell off during the motion capture session. As a result, we made customized motion capture shoes using duct tape to fix the markers to child shoes of different sizes. Child participants were asked to try out different sizes to find out the best fit.

Study Duration: The study duration is about 1.5 hours for each child participant. It is important to keep the children motivated for this entire time. At the beginning of the study, we related what the child was doing with the motion capture technology to popular animated movies or video games. We told them their motions would be used to animate their favorite character. During the capture session, every time they completed an action, we provided positive reinforcement by telling them how well they did. Even when they made some mistakes or deviated from the instruction, we always stayed positive and patiently corrected them to the extent feasible.

Ask rather than Assume: Children sometimes are less willing to communicate their needs to our experiment staff. During the study, we asked the child participants every 15-20 minutes if they need water, a break, or if there is anything make them uncomfortable. We noticed that most children would not bring their needs up unless the study staff asked them. If we noticed the participants appeared tired or if they were sweating, we paused the study and suggested they take a break, and resume when they are ready.

A.2 The *adult2child* dataset

Our novel child dataset is very different from the one released by Xia et al. [2015]. There are four broad categories of differences that directly impact the training of the neural network parameters:

Improvisation/Not following instruction exactly: Xia et al.'s dataset captured adults acting like a children, whereas our dataset has actual children, who interpret instructions quite differently, and are prone to improvisation when they are having fun. For example, we noticed in the child subject *child004*, the subject did multiple dabs while performing walking motion. In contrast, adult participants follow instructions as is; they do not perform any extra motions or improvise; that was the case in Xia et al's dataset, where all motions match the motion types with no redundant motions.

Extra movements before and after the prompted action: All the motions from our dataset started with a T-pose. We cut off the t-pose at the beginning using an algorithmic approach. The angles of the shoulder joints were used to determine the frame when the T-pose ends. After the motion is completed, some child participants swung their arms around or crossed their legs. These valuable extra motions were captured in our dataset, and can be used in different or further studies.

Age/biomechanics/skeleton of motion sequences: In the training set, we have 3 adults (adult001, adult002, adult003) and 3 child participants (child001, child004, child005), where each one has a skeleton with different proportions. In contrast, all motions in Xia et al's database share the same skeleton size and template. Thus, either their data were retargeted to a uniform template, or they captured only one adult. Our dataset, on the other hand, consists of motions taken from actual child participants. For instance, the skeletal proportions of the five-year-olds are quite different (appreciably smaller, different head-to-body ratio) from those of the ten-year-olds. They also have different cognitive abilities: we found that the ten-year-olds were more likely to be closer to what we expected when we gave them a certain prompt.

Capture setup In the dataset of Xia et al. [2015], subjects only move in one direction and there is no turning. In our dataset, we asked our subjects to turn when they reached the end of the capture space and do a loop in order to record long continuous sequences. As a result, another difference between the two datasets is that for locomotion, i.e, walk, fast walk, and jog, our dataset included 180-degree turns.