DOI: 10.1111/cgf.14645
ACM SIGGRAPH / Eurographics Symposium on Computer Animation 2022
D. L. Michels and S. Pirk
(Guest Editors)

Volume 41 (2022), Number 8

Context-based style transfer of tokenized gestures

Shigeru Kuriyama[1,3], Tomohiko Mukai[2], Takafumi Taketomi[3], and Tomoyuki Mukasa[3]

[1]Toyohashi University of Technology, Department of Computer Science and Engineering, Japan
[2]Tokyo Metropolitan University, Department of Industrial Art, Japan
[3] CyberAgent Inc., AI Lab, Japan

Abstract

Gestural animations in the amusement or entertainment field often require rich expressions; however, it is still challenging to synthesize characteristic gestures automatically. Although style transfer based on a neural network model is a potential solution, existing methods mainly focus on cyclic motions such as gaits and require re-training in adding new motion styles. Moreover, their per-pose transformation cannot consider the time-dependent features, and therefore motion styles of different periods and timings are difficult to be transferred. This limitation is fatal for the gestural motions requiring complicated time alignment due to the variety of exaggerated or intentionally performed behaviors.

This study introduces a context-based style transfer of gestural motions with neural networks to ensure stable conversion even for exaggerated, dynamically complicated gestures. We present a model based on a vision transformer for transferring gestures' content and style features by time-segmenting them to compose tokens in a latent space. We extend this model to yield the probability of swapping gestures' tokens for style-transferring. A transformer model is suited to semantically consistent matching among gesture tokens, owing to the correlation with spoken words. The compact architecture of our network model requires only a small number of parameters and computational costs, which is suitable for real-time applications with an ordinary device.

We introduce loss functions provided by the restoration error of identically and cyclically transferred gesture tokens and the similarity losses of content and style evaluated by splicing features inside the transformer. This design of losses allows unsupervised and zero-shot learning, by which the scalability for motion data is obtained.

We comparatively evaluated our style transfer method, mainly focusing on expressive gestures using our dataset captured for various scenarios and styles by introducing new error metrics tailored for gestures. Our experiment showed the superiority of our method in numerical accuracy and stability of style transfer against the existing methods.

CCS Concepts

• ***Computing methodologies*** → ***Motion processing;***

1. Introduction

Motion capture data are widely used to generate natural motions in character animation. Many methods have been proposed for constructing minute features of movements (or styles) of motion data to enhance their expressiveness. However, the existing methods mainly focus on periodic motions such as walking and punching, and capturing complicated aperiodic gestures is still challenging. Many techniques that predict human activities using recurrent neural networks [MBR17] can also generate motions, but most of them cannot treat style transformations.

Many motion style transfers modulate the means and variances of motion signals over time, in the way of adaptive instance normalizations [JPL22], similarly to image style transfers [HB17]. This success relies on the common property that the changes in phase and magnitudes can characterize the cyclic signals of gait motions, as displayed in [JPL22]. However, the shapes of motion signals are vastly changed against characteristic or exaggerated gestures; they have fewer correlations over time owing to the non-uniform change in velocities, as shown in Figure 1. The gesture style transfer should be local for gestural motions because the time ranges are irregular for more significant dynamic variations. This observation suggests that statistical transformation does not work well owing to its global property. Neural network models that convert each pose have intrinsic limitations in coping with such complicated transformations.

Moreover, the style features of motion data are physical properties, and they should preserve naturalness as human gestures; for example, joint rotations are limited to actual physical ranges. However, pure numerical approaches in image style transfer easily destroy the naturalness due to the lack of physical constraints. To overcome these limitations, we introduce an approach to swap segments of motion clips with those of the target style in a similar way to the style transfers by swapping image patches [CS16, LFY*17, SLSW18]. Our approach divides and re-arranges a style gesture to imitate the meaning (or content) of an input gesture, which is suitable for transferring styles of expressive gestures because the features of complicated styles are adaptively replicated to preserve their personality and physical constraints locally.

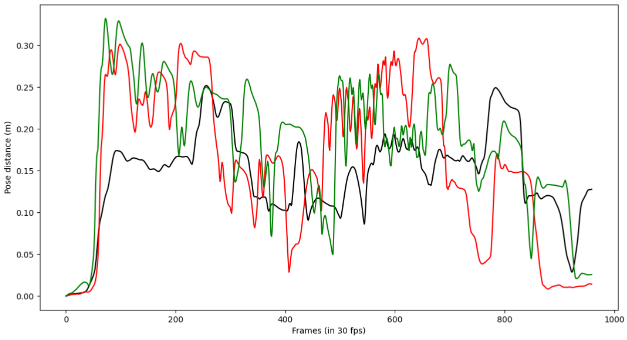


Figure 1: *Motion signals of gestures performed for the same utterance scenario. These polylines indicate time-varying pose distances, which are computed by the mean Euclidean distance of every joint's position against those of the initial pose. The gray line indicates a regular expression, and the red and green lines indicate exaggerated characteristic expressions. The green line is performed so that the motion content coincides with the regular expression. These lines reveal the various phase shifts and complicated shape deformations.*

Gestural motions can be synthesized from voice signals or texts of talks [Nef16, YYH20]. This trend implies that a sequence of structural motion units can be treated like spoken words. This assumption is theoretically supported by Kendon's continuum [Ken88], which regards each gestural unit as a component of a speech. This theory motivated us to introduce a context-based approach to motion style transfer by adopting a methodology developed in natural language fields; a transformer model [VSP*17]. In our swap-based approach, the quality of the resulting gestures is affected by the performance of pattern-matching between the motion segments of content and style gestures. Therefore, the context of gestures is a crucial factor in detecting good correspondence between the two structures. This observation introduces a neural network model suitable for extracting such contextual information.

Fortunately, the intrinsic power of transformer models has been proven in image processing fields [DBK*21,CTM*21], and we also constructed a transformer-based model, referencing existing methods proposed for image style transfers [DTD*22, TBTBD22]. Our method introduces style transfer using segmented (or tokenized) gestures by considering contextual information with a transformer model and a loss function that can be efficiently computed by splicing the transformer's variables. We also propose a gesture-specific criterion for evaluating the quality of style transfers.

2. Related works

2.1. Patch-based image style transfer

In image style transfers, many patch-based approaches [CS16, LFY*17, SLSW18] have been proposed for swapping each block region of a content image, called a patch, using the corresponding patch of a style image. In addition, the reliability of the pattern-matching process is increased by statistics-based signal regularization called whitening [SLSW18]. These piece-wise swapping approaches can be trained without the use of supervised samples.

Our method introduces such a patch-based approach by dividing motion data into segments of a fixed period, and those of a style gesture replace the segments of a contextual gesture. However, pattern-matching after the whitening cannot work well for human gestures owing to their irregular and complicated statistics features. Therefore, our method introduces context-based pattern matching to increase its accuracy. Moreover, simply swapping the segments of a motion clip, such as image-style transfers, often degrades the smoothness of human motions. Therefore, we introduce a mechanism for estimating the probability of swapping by which multiple adequate segments are blended. This soft-swapping approach can more accurately preserve gestures' contents (or outline) after transfer.

2.2. Style transfer for non-gestural motions

Existing methods identify motion-style features with a linear system [HPP05, XWCH15], or optimization in the frequency domain [YM16]. These methods require time alignment between motion clips using numerical adjustments, such as time-warping, for capturing correspondence over time. However, It is difficult to fully automate the time alignment of complicated gestural movements, which requires tedious manual preprocessing.

The recently proposed motion style transformation [HHKK17, HSK16] based on deep learning applies the drawing style transformation for images to motion data, which converts the statistical features of latent variables with a Gram matrix [GEB16]. This method revealed that the drawing-style transformation for natural images could be applied to motion data using similar mathematical tools. This approach requires no time alignment of data samples but requires re-training to add novel styles. Style transfer based on domain adaptation [MSZ*18] was introduced using a phase-functioned neural network (NN). This method successfully decomposes style components using residual adapters with compactly decomposed tensors. However, it is only applicable to periodical motions, such as gaits or punches, owing to the intrinsic properties of the phase-based model.

The generative adversarial network (GAN) model was introduced [AWL*20] for style transfer using adaptive instance normalization (AdaIN) layers trained with an unpaired dataset. This method can transfer motion styles from videos by learning a common style embedding for 2D and 3D joint positions. A spatiotemporal graph was also introduced [PJL21] to improve style translation between significantly different motions. These methods require labeled data to construct the adversarial loss of the GAN-based model, which is unsuitable for zero-shot learning. Moreover, their AdaIN-based style transfers cannot capture time-varying motion styles because they merely fit the mean and variance of content features to style features. However, this statistical modulation of AdaIN globally captures temporal features, but the features of characteristic gestures are mainly time-varying and temporally local variations. Therefore, we propose a more stable approach based on token swapping without statistical modulations.

2.3. Voice or text-driven gesture synthesis

Gestural motion synthesis from audio or text data is a vital technology for the social agent systems using virtual humans or avatars, and some methods have been developed using data-driven approaches. Levine et al. [LKTK10] proposed synthesizing gestural motions from the acoustic characteristics of utterances by optimally selecting motion segments using a probabilistic model. However, gesture style transfer using this method requires a manual setting of the control parameters for each motion unit or segment. Yang et al. [YYH20] proposed motion synthesis of conversations based on motion graphs, in which the optimal path is searched by using a stochastic greedy algorithm. This method achieves audio-motion coordination that can generate a variety of plausible motions. However, the expressiveness of gestures strongly depends on the motion clips embedded in the motion graph, and style control is not supported.

Recently, speech-driven syntheses of gestures have been proposed using NN models. As an image-based approach, Ginosar et al. [GBK*19] proposed a person-specific prediction of gestures from audio by using a video dataset, where the individual motion styles of arms and hands are learned using a temporal cross-modal translation. Alexanderson et al. [AHKB20] proposed an invertible NN model for speech-driven gesture syntheses. Although this method can control motion styles with additional annotation signals, re-training is required in adding novel styles. In addition, the training dataset must be manually labeled for supervised learning. Yoon et al. [YCL*20] proposed speech gesture synthesis from a trimodal context consisting of speech, audio, and speakers identities corresponding to motion styles. However, this method also requires training for each identity, and collecting a trimodal dataset for each style is cumbersome. Our approach can be utilized as a post-process for these speech-driven methods to enhance the expressiveness of gestures with no additional training.

2.4. Personality-based style synthesis of gestures

Chi et al. proposed a gesture synthesis method called EMOTE [CCZB00] by introducing Laban Movement Analysis and its Effort and Shape components to parameterize the qualitative features of movements. The gesture synthesis method [DKD*16] generates animations by embedding a gesture personality feature into a five-dimensional space, called OCEAN [CM92], and 39 variables were introduced to control styles whose values were determined using multivariate regressions. Smith and Neff [SN17] compressed this five-dimensional space into two dimensions by reflecting a cognitive aspect, thus enabling a more efficient classification of gestural styles. These model-based approaches can control time-dependent features using a timing control function with key poses. However, detecting good key poses from raw motion capture data is often challenging for complicated motions.

3. Architecture of style transfer

Our gesture style transfer architecture is roughly divided into an auto-encoder for embedding motion representation in a latent space and a style transformer for swapping tokens of content and style gestures, as shown in Figure 2. Our style transfer network is trained to compute the best matching between tokens of two samples to preserve the content and style similarities estimated in a latent space of the transformer, and new samples for evaluations are expected to be similarly matched by reflecting their contexts.

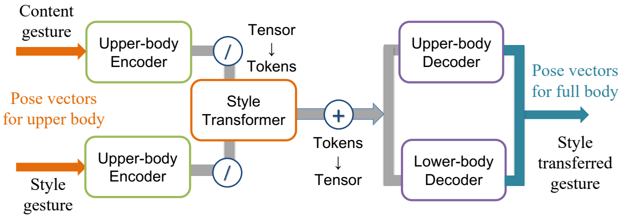


Figure 2: *Flow of motion style transfer*

Only the upper-body motion signals are fed to the auto-encoder in this architecture. The embedded variables are time-divided to create tokens that are fed to a transformer model developed for the image classification task [DBK*21]. The output tokens of the transformer were restructured to fit into the two types of decoders, and each separately generates motions for the upper- and lower bodies. Notice that our method estimates the movements of the lower body (i.e., both legs) from the transferred motions of the upper body. An ordinary gesture is denoted as a content gesture, and an expressive gesture of the target style is represented as a style gesture. These two motion clips are separately embedded to compose the tokens using the same auto-encoder.

The sizes of learnable parameters are 23.7 KB and 50.2 KB for the auto-encoder and style transformer, respectively, and the total parameter size is only 296 KB. This compact architecture enables whole processes to be run by approximately 13 microseconds per pose on average, measured on iMac Pro (2017) 3.2 GHz 8 core Intel Xeon CPU without using any GPU. This processing time corresponds to the update frequency of 77×10^3 frames per second and figure, which is small enough to allow interactive applications. The training times were measured (see Section 4.1) by the same conditions.

3.1. Gesture motion representation

Our human body model consisted of 21 joints, and finger movements were excluded. As explained in the following section, the gestural motion, given by the series of joint rotational angles, is then converted to latent variables with auto-encoders. We implemented joint rotations using a logarithmic map [Gra98] and applied no additional variables such as joint positions or constraint conditions. Therefore, the dimension of vectors representing each full-body pose therefore becomes $21 \times 3 = 63$.

For the root (or waist) joint orientations, we use the rotation angles whose components along a vertical axis are canceled and the residual rotational components to feed the auto-encoder because the global facing direction is not related to the styles of gestures. Note that the omitted rotational component is directly copied from the content gesture to the style-transfer result. Our method cannot transfer (or estimate) the positions of the root joint. We currently implement a mechanism that determines the root position at each frame to fix the lower foot position on a floor level, using forward kinematics. Since we only focus on the gestures of standing poses, this simple adjustment works well in most cases.

3.2. Feature embedding with auto-encoder

We introduce an approach to swap the fragments of representation between content and style, developed in image style transfers [CS16, SLSW18, DTD*22]. These methods first embed image (pixel) values into the latent space via convolutional deep neural networks such as VGG [SZ15]. The patches of images are then composed of blocks of pixels divided into a fixed size. In our method, the segment of motion clips corresponding to the image patch is called a *token* to stress that our model focuses on the contextual analysis using transformers.

We introduced one-dimensional convolutional neural networks (CNN) for motion data representing the time sequence of a pose consisting of skeletal joint rotations. The convolutional kernel is applied along time (or frames), and correlations in rotations among joints are implicitly considered by regarding their components as input channels. The time resolution of these embedded variables is reduced by $1/4$ with the encoder's second and third convolutional layers, whose strides are set to two. This reduced time resolution is restored with the decoder's first and second linear up-sampling layers.

This CNN-based auto-encoder takes the input signals of the 13 upper-part joints and has two types of decoders for the same upper joints and eight lower part joints consisting of two leg joints, as shown in Figure 2. This decomposition is derived from the observation that characteristic gestures mainly emerge in the upper-half body: two arms, torso, head, and root (waist) joints, and the motions in the upper half of the body drive the motions in the lower half.

These two decoders were designed with the same architecture, except for the number of channels. The detailed architecture of this CNN-based auto-encoder is explained in Appendix A.

3.3. Gesture tokens

Let the time-sequences of latent variables for a motion $z_{i=1,2,\dots,F} \in \mathbb{R}^{d_z}$ be embedded by the auto-encoder, where F denotes the full frames and d_z is the dimension of the latent space. We decompose these variables into the fixed intervals w with a stride of h to compose the n-th gesture token g_n as

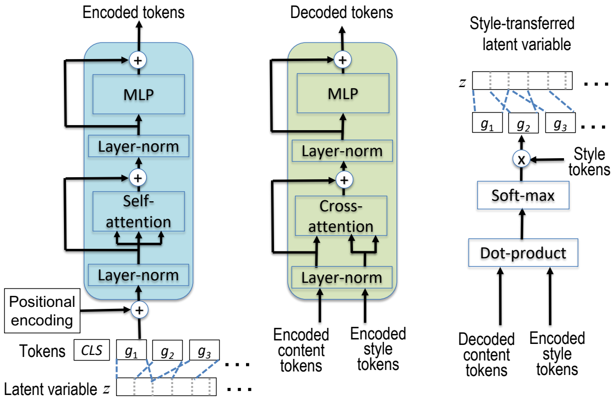
$$g_{n+1} = z_{nh+1} \oplus z_{nh+2} \oplus \cdots \oplus z_{nh+w}\ , \quad n = 0, 1, \dots, \lfloor F/h \rfloor - 1\ , \quad (1)$$

where $\oplus$ denotes the concatenation of vectors and $\lfloor\ \rfloor$ is a floor function. We call this flattened one-dimensional vectors $g_n \in \mathbb{R}^{d_g}$, $d_g = w d_z$ as a gesture token.

The stride is set to the half-size of the interval $h = w/2$; tokens are sampled while being entirely overlapped to increase the time-granularity of transfer. The transferred tokens are then converted into tensors to feed the decoder for de-embedding. The successive tokens are then linearly interpolated to merge the overlapped regions, by which the smoothness at the tokens' boundaries is obtained. We found that the size of an interval of $w = 4$ ensures an excellent balance to reflect the outline of the content gestures and the details of styles. Because the time resolution shrinks by 1/4 in the encoder, each gesture token corresponds to $4 * 4 = 16$ frames in the original time scale, corresponding to approximately a quarter second for motion data sampled at 60 frames per second.

3.4. Style transformer for gestures

The tokens of the content and style gestures are swapped in a transformer model, as shown in Figure 3. Our transformer model is designed to adequately replace every token of a content gesture with the linear blending of those of a target style gesture. This architecture can transfer styles while reflecting their contexts to obtain content similarity.



(a) Encoder (b) Decoder (c) Soft-swapper

Figure 3: *Architecture of gesture style transformer*

Encoder

Our style transfer mechanism adopts an encoder model of a visual transformer [DBK*21] for converting tokens to include contextual information, where the content and style motion clips are processed separately. This encoder model employs layer normalization as a preprocessing step and GeLU activations, where we set a single self-attention layer of a single head, as shown in Figure 3 (a). Because the linear networks for encoding queries, keys, and values have the same number of channels for input and output, the dimension of every token is preserved. The residual block with multilayer perceptron, denoted by MLP in Figure 3, has no hidden layer and is composed of two fully connected layers of the same number of channels while putting a GeLU activation layer between them.

Before feeding a token sequence of gestures to the encoder, an extra learnable classification token [DBK*21], denoted by *CLS*, is concatenated at the head of the sequence as $\boldsymbol{g} = [CLS, g_1, g_2, \ldots, g_N]$, $CLS \in \mathbb{R}^{d_g}$. In our method, the output corresponding to this classification token extracts the common features included in all gesture tokens. This value was utilized to evaluate style features, as explained in Section 3.5.

In addition, a position-encoding is adopted for every element as

$$\hat{\boldsymbol{g}} = \boldsymbol{g} + \boldsymbol{p}\,, \tag{2}$$

where $\boldsymbol{p}$ denotes a sequence of d_g dimensional positional values defined by the sine and cosine functions of different frequencies [VSP*17].

We consider that each gesture token has little effect on the distant tokens and introduce a mask operation to block the impact of distant tokens, increasing computational efficiency. Self-attention is therefore computed by masking (excluding) the tokens of $\hat{g}_k$, $k < i - \omega \cup k > i + \omega$, where we experimentally set as $\omega = 20$, which corresponds to the intervals of nearly 10 seconds centered at each token.

Decoder and soft swapper

Our transformer decoder is composed similarly to the vision transformer [DBK*21], where we set a single cross-attention layer of a single head, as shown in Figure 3 (b). The residual block with multi-layer perceptron has the same architecture as the above encoder.

The transformer decoder T_D gives the decoded content tokens $\tilde{\boldsymbol{g}}^{\mathrm{c}}$ by taking the pair of encoded tokens for content and style as:

$$\tilde{\boldsymbol{g}}^{\mathrm{c}} = T_D\left(T_E(\boldsymbol{g}^{\mathrm{c}}), T_E(\boldsymbol{g}^{\mathrm{s}})\right)\,, \tag{3}$$

where T_E is the encoder of the style transformer using the position-encoder in Equation (2).

The content tokens are then style-transferred by blending the style tokens $\boldsymbol{g}^{\mathrm{s}}$. A soft-swapper gives the weights of blending, as shown in Figure 3 (c), which computes the weights of style tokens swapped by each content token based on their similarity. Similarities between decoded content tokens and encoded style tokens are computed by their dot-products and filtered by softmax functions as follows,

$$\boldsymbol{g}^{\mathrm{s\,|\,c}} = \mathrm{softmax}(\tilde{\boldsymbol{g}}^{\mathrm{c}} \cdot T_E(\boldsymbol{g}^{\mathrm{s}}))\,\boldsymbol{g}^{\mathrm{s}}\,. \tag{4}$$

The style-transferred tokens $\boldsymbol{g}^{\mathrm{s\,|\,c}}$ are then fed to the auto-encoder to obtain the final motion representation consisting of every joint's rotations. The blending and swapping mechanism can preserve style gestures' characteristic features and physical plausibility. In addition, the final softmax operations can avoid excessive blending of many tokens, which often causes over smoothing of the resulting motions.

3.5. Loss functions

In training our CNN-based auto-encoder, the encoder's output is directly fed to the decoder while bypassing the style transfer layers. We define the loss by the mean square of the restoration error in every pose, where the angular differences between the input and output joint rotations were used as errors. Adam optimization was used for training, setting the batch size to 32 and the learning rate to 10^{-4}. Batches of training samples were randomly sampled in 256 frames with an even probability. We experimentally confirmed that 1000 epochs were sufficient to achieve convergence.

In training our style transfer mechanism, the loss function is designed using the restoration error of the latent variables between the original and synthesized motions. Identical motions were generated using the same latent tokens in the decoder of the gesture-style transformer for both the content and style gesture samples. Cyclically transferred motions are caused by transferring content tokens with style one, and the resulting tokens are back-transferred by using the same content tokens as style tokens.

In this way, we compute the identity and cyclic loss, denoted by $L_{identity}$ and L_{cyclic}, respectively, by the norm of the difference calculated as

$$\begin{aligned} L_{identity} &= \|\boldsymbol{g}^{\mathrm{c\,|\,c}} - \boldsymbol{g}^{\mathrm{c}}\|_\tau + \|\boldsymbol{g}^{\mathrm{s\,|\,s}} - \boldsymbol{g}^{\mathrm{s}}\|_\tau\,, && (5)\\ L_{cyclic} &= \|\boldsymbol{g}^{\mathrm{c\,|\,s\,|\,c}} - \boldsymbol{g}^{\mathrm{c}}\|_\tau\,, && (6)\\ \boldsymbol{g}^{\mathrm{c\,|\,s\,|\,c}} &= \mathrm{softmax}(\tilde{\boldsymbol{g}}^{\mathrm{s\,|\,c}} \cdot T_E(\boldsymbol{g}^{\mathrm{c}}))\,\boldsymbol{g}^{\mathrm{c}}\,, && (7)\\ \tilde{\boldsymbol{g}}^{\mathrm{s\,|\,c}} &= T_D\left(T_E(\boldsymbol{g}^{\mathrm{s\,|\,c}}), T_E(\boldsymbol{g}^{\mathrm{c}})\right)\,, && (8) \end{aligned}$$

where $\boldsymbol{g}^{\mathrm{c\,|\,c}}$ and $\boldsymbol{g}^{\mathrm{s\,|\,s}}$ denote the content and style tokens whose styles are transferred by themselves, $\boldsymbol{g}^{\mathrm{c\,|\,s\,|\,c}}$ corresponds to the cyclically transferred token using the style-transferred tokens $\boldsymbol{g}^{\mathrm{s\,|\,c}}$ as content and $\boldsymbol{g}^{\mathrm{c}}$ as style, and $\|\;\|_\tau$ represents the averaged L2-norm computed per each token.

These loss functions merely ensure the consistency of transformations and cannot guarantee the similarity of the content between the content gesture and the transferred gesture. Therefore, we add a loss function to ensure the consistency of meanings by introducing structural similarity proposed by the image style transfer [TBTBD22]. Let $\boldsymbol{m}(g)$ be the matrix of structural information for $\boldsymbol{g}$, which is extracted by the self-similarity of keys in the transformer encoder as

$$\boldsymbol{m}(g) = \left\{ \frac{k_i \cdot k_j}{\|k_i\|\,\|k_j\|}\,,\;\; i,j = 1,2,\ldots,n \right\}\,,\;\; k_i = \boldsymbol{m}_k\, g_i\,, \tag{9}$$

where the i-th key k_i is obtained by the linear projection of the i-th token with the linear matrix $\boldsymbol{m}_k$ for the keys. Structure loss, which is the content dis-similarity of two gestures $\boldsymbol{g}^{\mathrm{c}}$ and $\boldsymbol{g}^{\mathrm{s\,|\,c}}$, is then obtained by the Frobenious norm $\|\;\|_F$ of the two matrices as follows:

$$L_{structure} = \|\boldsymbol{m}(\boldsymbol{g}^{\mathrm{c}}) - \boldsymbol{m}(\boldsymbol{g}^{\mathrm{s\,|\,c}})\|_F\,. \tag{10}$$

Our token-wise blending automatically inherits the intrinsic features of style gestures because the resulting tokens are purely linear combinations of the original tokens for the target style gesture. However, most style gestures include tokens of normal-style motions that are likely pattern-matched to the content tokens, by which stylistic movements are missed during the transfer. Therefore, we add a loss function to enhance the tokens of characteristic motions to be more likely a match by introducing an appearance similarity [TBTBD22], which is given by the L2-norm of the class-tokens between style and transferred gestures as

$$L_{appearance} = \|CLS^{\mathrm{s}} - CLS^{\mathrm{s}\,|\,\mathrm{c}}\|_2 \,, \tag{11}$$

where CLS^{s} and $CLS^{\mathrm{s}\,|\,\mathrm{c}}$ indicate the class tokens for style and style-transferred content gestures.

Our loss function L is finally given by

$$L = L_{identity} + \lambda_c L_{cyclic} + \lambda_s L_{structure} + \lambda_a L_{appearance} \,, \tag{12}$$

where we set equal weights to $\lambda_c = \lambda_s = \lambda_a = 1.0$ for all experiments in the next section. This total loss L leads to the detailed features of the style-transferred gesture differing from those of the content gesture while preserving content similarity after style transfer and the consistency of identically and cyclically transferred results.

Our transformer model was trained by randomly sampling a pair of motion clips for content and style gestures with no behavioral synchronization or time alignment; the only requirement was that the performed motion had consistent styles within each clip. We trained the networks for style transfer with eight content gesture clips of 1280 frames per batch, and nine style gestures were utilized. In addition, we set the learning rate by 10^{-4} for 1000 epochs and adopted no dropout operations in training the attention models.

4. Experiment

4.1. Dataset and trainings

Although there are some publicity-available datasets for gestural motions, most of them do not have very expressive or exaggerated gestures, on which our method mainly focuses. We, therefore, collected motion samples for such styles of gestures by capturing them with an optical device at 60 frames per second.

A professional female actor performed ordinary gestures and expressive gestures of two types: one enhancing extraversion and the other enhancing an anime-like style by synchronizing movements with the synthetic voice of each scenario. The total length of these motion samples reached about 8 and 9 minutes, respectively, for content and style (for further details, see Appendix B). These samples are used in training our style transformer and auto-encoder. To enhance the auto-encoder training, we added samples of 230 seconds, including various types of gestures captured without using synchronization with voices (see Appendix B).

Our auto-encoder was trained by feeding the above motion samples with a batch size of 32 by randomly and evenly picking and clipping per 256 frames. The training was sufficiently converged after 10^3 epochs, which took approximately 23 minutes with the learning rate by 10^{-4} with no GPU.

The style transformer was trained by feeding the pair of content and style gestures. We randomly and evenly sampled and clipped 1280 frames among the seven content samples and nine style gestures, which produced 63 pairs. The training took approximately 6.2 minutes with a batch size of 8, using the same learning rate and the number of epochs as the auto-encoder.

For experimental evaluations, we collected samples by asking the same performer to behave in ordinary and expressive styles while synchronizing 15 scenarios of utterances spanning from 20 to 36 seconds. These scenarios were borrowed from publicly-available short voice samples for female characters. However, in this case, all characteristic gestures are performed in a freestyle according to the content of the scenario. This condition means that all evaluated gestures have different styles from the samples used in training. We collected two types of characteristic gestures: one was performed freely, and the other was conducted to have similar hand trajectories to the content sample while having the same expression as the freely performed one. The latter samples compare our style transfer method with human imitations. The total length was approximately 226 seconds for each sample type. Notice that these samples are not used in training, by which our feasibility of zero-shot learning is proved. In the evaluation phase, style gesture is augmented by computing mirror symmetry. We confirmed that the restoration accuracy of this auto-encoder is not degraded for samples used in evaluations; the restoration errors for training and testing samples have no significant differences independent of their styles.

4.2. Quantitative evaluation of style transfer

We now focus on defining the similarity measures of gestures. From the viewpoint of content preservations, we expect the output, i.e., the resulting motions after style transfer, to have the same meaningful signs or symbols shaped by the postures of arms [Mcn94]. This observation conceptually relates to the shape controls by the end-effector's key points in the EMOTE model [CCZB00]. We assume that the similarity of the arms' postures can be efficiently replaced with the similarity of both hands' trajectories because they mainly appeal to the intention or meaning of gestures. Therefore, we sum the L2-norms of both hands' positions between the content gesture and the style-transferred one as a content error after style transfer. Because we focus on the similarity of the trajectories of the hands, we measure the distance between hands of corresponding poses by allowing the shift of frames. Let $\boldsymbol{p}^{\ell} = [p^{\ell}_1, p^{\ell}_2, \ldots, p^{\ell}_F]$ be the sequence of 6D vectors representing the 3D positions in meters of both hands for the content ($\ell = \mathrm{c}$) and style-transferred ($\ell = \mathrm{x}$) gestures, and the i-th content error component $E_i^{content}$ is given as

$$\begin{aligned} E_i^{content} &= \|p^{\mathrm{c}}_{c_i} - p^{\mathrm{x}}_{x_i}\|_2 \,, && (13) \\ \boldsymbol{I} &= \mathrm{FastDTW}(\boldsymbol{p}^{\mathrm{c}}, \boldsymbol{p}^{\mathrm{x}}) \,, && (14) \end{aligned}$$

where $\boldsymbol{I} = [[c_1, x_1], [c_2, x_2], \ldots, [c_F, x_F]]$ represents the sequence of indices for corresponding pairs between $\boldsymbol{p}^{\mathrm{c}}$ and $\boldsymbol{p}^{\mathrm{x}}$ whose summation of Euclidean distances is minimized, which is efficiently computed using the linear-order dynamic warping technique denoted by FastDTW [SC07].

For evaluating the preservation of styles after a transfer, we assume that characteristic styles of gestures can be visually perceived by the quickness or slowness of meaningful arms' movements, which can also be interpreted as slow-in/out features. This observation is conceptually related to the effort control by keyframe-to-time functions in the EMOTE model [CCZB00]. However, directly computing the differences in velocities is complex and error-prone because adequately pattern-matching the fragments of velocities is difficult between the motion clips of different contents. Moreover, unlike the content features, the style features are not a local property, and their semi-global patterns emerge irregularly within some periods. From this observation, we evaluate the k-th error component of style features, denoted by E_k^{style}, using their statistical values per half-overlapped fixed interval θ as

$$E_k^{style} = \min_j \left(D_{mean}(\boldsymbol{\rho}_k^{\mathrm{x}}, \boldsymbol{\rho}_j^{\mathrm{s}}) + D_{std}(\boldsymbol{\rho}_k^{\mathrm{x}}, \boldsymbol{\rho}_j^{\mathrm{s}})\right) , \quad (15)$$

$$\boldsymbol{\rho}_{k=0,1,\ldots}^{\ell=\mathrm{x,s}} = \boldsymbol{p}^{\ell}\langle kh_\theta + 1, kh_\theta + \theta\rangle , \; h_\theta = \lfloor\theta/2\rfloor , \quad (16)$$

where $\lfloor \; \rfloor$ is a round operation, $\boldsymbol{p}^{\ell}\langle k_0, k_1\rangle$ denotes 6D hand positions segmented in the frame range from k_0 to k_1, and D_{mean} and D_{std} denote the L2-norm of the difference in means and standard deviation along frames between segments of hand positions for style-transferred $\boldsymbol{\rho}_k^{\mathrm{x}}$ and target style $\boldsymbol{\rho}_j^{\mathrm{s}}$ gestures. We experimentally set the interval $\theta = 128$ and also confirmed that the magnitude relation of the mean errors for the resulting gestures was invariant against the intervals of $\theta = 64$ and 256.

We evaluate gesture-style transfers by introducing three metrics for each error. We compute the mean of errors computed for every error component to evaluate overall performance. Moreover, we focus on the worst-case among all components because the style transfer should avoid instantaneous conspicuous deviations rather than averaged ones. Therefore, we also compute the maximum errors as a worst-case and the mean of the first 25 % most significant errors for evaluating the group of conspicuous errors among all components. We quantitatively evaluate the accuracy of style transfer by the averages of these metrics for all evaluation samples, assuming that good style transfers minimize these metrics, whereas some trade-off exists between the errors of content and style.

Here, we evaluated the reliability of our error metrics, using the samples for evaluations. Table 1 compares errors computed for content and style samples against those captured by performing supposed style-transfers as referential samples. This result shows that all content errors are smaller for the content than those for the style. On the contrary, all style errors are smaller for the style than those for the content. This relation roughly demonstrates the validity of our error metrics.

Table 1: *Error metrics of content and style samples compared with those performed as referential samples. The w25% denotes the average of worst-25% error metrics.*

Samples	Content errors $E^{content}$ ↓			Style errors E^{style} ↓		
	mean	worst	w25%	mean	worst	w25%
Content	0.139	0.51	0.3	1.14	2.65	2.14
Style	0.293	0.785	0.557	0.997	2.11	1.69

4.3. Comparison with existing methodologies

Here, we compare the performance of our method using the error metrics mentioned above. We first select a commonly used approach for motion style transfer, which considers the style features based on the mean and variances over time for each channel. This approach converts the latent variables of content gesture $\boldsymbol{z}^{\mathrm{c}} = \{z_{i=1,2,\ldots,F}^{\mathrm{c}}\}$, with the means and variances for those of style gestures $\boldsymbol{z}^{\mathrm{s}} = \{z_{i=1,2,\ldots,F}^{\mathrm{s}}\}$, in an adaptive instance normalization as,

$$z_i^{\mathrm{x}} = \sigma(\boldsymbol{z}^{\mathrm{s}})\frac{z_i^{\mathrm{c}} - \mu(\boldsymbol{z}^{\mathrm{c}})}{\sigma(\boldsymbol{z}^{\mathrm{c}})} + \mu(\boldsymbol{z}^{\mathrm{s}}) , \quad (17)$$

where $\mu(\boldsymbol{z})$ and $\sigma(\boldsymbol{z})$ denote the mean and standard deviation for $\boldsymbol{z}$, and z_i^{x} is the i-th latent variable for the style-transferred gesture. This method is regarded as the comparative study against our token-based composition of style transfer.

The second comparative method was swapping tokens using pattern-matching of content and style gestures. Let g_n^{c} and g_m^{s} be the n-th content token and m-th style token, respectively, and g_n^{c} is replaced by the most similar style token $g_{\hat{n}}^{\mathrm{s}}$ as

$$\boldsymbol{g}^{\mathrm{s}\,|\,\mathrm{c}} = [g_{i_1}^{\mathrm{s}}, g_{i_2}^{\mathrm{s}}, \ldots, g_{i_N}^{\mathrm{s}}] , \; i_n = \mathrm{argmax}_m \arccos(\bar{g}_n^{\mathrm{c}} \cdot \bar{g}_m^{\mathrm{s}}) , \quad (18)$$

where g_i^{s} is the i-th token of the style gesture, and $\bar{g}_n^{\mathrm{c}}, \bar{g}_m^{\mathrm{s}}$ are the flattened 1D vector of $g_n^{\mathrm{c}}, g_m^{\mathrm{s}}$ whose lengths are normalized to 1. We adopt regularization for the latent variables $\boldsymbol{z}^{\mathrm{c}}, \boldsymbol{z}^{\mathrm{s}}$ before and after the swapping operation using mean and standard deviation, in a similar way to image style transfer called Avatar-net [SLSW18]. This method is compared to evaluate the effectiveness of our transformer-based style transfer.

In addition, we evaluate our swapping mechanism in a hard-max manner by replacing Equation (18) with

$$i_n = \mathrm{argmax}\left(\tilde{\boldsymbol{g}}^{\mathrm{c}} \cdot T_E(\boldsymbol{g}^{\mathrm{s}})\right) . \quad (19)$$

where this approach corresponds to the ablation of our soft-swapping mechanism.

Moreover, we evaluate the similarity of the style-transferred samples against ones performed by the same actor. For each content-style pair of gesture samples used for style transfer, we asked the actor to imitate a gesture regarded as a ground truth while consciously performing the same meaning and expression as the content and style samples, respectively. We computed the error metric using these imitated gestures as a style-transferred result.

We evaluated the style transfer for two conditions: one uses content-style pairs performed in the same scenario, and another uses all style samples whose scenarios are different from content samples. In the former intra-scenario condition, the style samples have relatively similar content owing to the same scenario, whereas the actor freely performed by neglecting the correspondence to the content gesture. On the other hand, the latter inter-scenario condition is regarded as a more difficult case because the contents of style samples largely differ due to the different scenarios.

As shown in Tables 2 and 3, our method has the minimum mean errors in content and style, and this indicates superior overall performance. AdaIN has larger mean content errors and smaller worst errors, suggesting that statistics-based conversion conservatively transfers the styles due to its global property. The larger style errors in all metrics for Avatar-net and Hard swap imply the difficulty of transferring style features by one-to-one swapping without a blending technique. On the other hand, our method achieved the smallest style errors in all metrics. The magnitude correlation between intra- and inter-scenario conditions is very high because the methods of the first and secondary minimum errors mostly coincide; the exception is the tied relation in the content worst 25%

errors. As expected, all error metrics increase in the inter-scenario condition compared with intra-scenario ones. However, the degradation in style similarity is slight, considering the large degradation in content similarity.

Table 2: *Error metric comparison of **intra**-scenario style-transfer for various swapping approaches. Parenthesized number in Method denotes the index of the corresponding equation. Bold values indicate the minimum values, and italic values indicate the secondary minimum values.*

Method	Content errors $E^{content} \downarrow$			Style errors $E^{style} \downarrow$		
	mean	worst	w25%	mean	worst	w25%
AdaIN (17)	0.192	**0.421**	**0.289**	*0.763*	*1.59*	*1.34*
Avatar-net (18)	0.189	0.491	0.334	0.92	2.07	1.61
Hard swap (19)	*0.171*	0.493	0.307	0.886	2.07	1.64
Soft swap (ours)	**0.16**	*0.443*	**0.289**	**0.646**	**1.43**	**1.19**

Table 3: *Error metric comparison of **inter**-scenario style-transfer for various swapping approaches.*

Method	Content errors $E^{content} \downarrow$			Style errors $E^{style} \downarrow$		
	mean	worst	w25%	mean	worst	w25%
AdaIN (17)	0.239	**0.49**	*0.35*	*0.782*	*1.68*	*1.39*
Avatar-net (18)	0.258	0.626	0.439	1.01	2.13	1.75
Hard swap (19)	*0.212*	0.528	0.359	0.959	2.02	1.7
Soft swap (ours)	**0.206**	*0.495*	**0.348**	**0.693**	**1.49**	**1.27**

For a qualitative evaluation, we visually compare content errors with the existing methods by the poses produced at the same frame time. Figure 4 shows the snapshots of the poses whose differences against the content sample are noticeable. In all cases, the poses generated by AdaIN and Avatar-net caused significant disparities rather than those generated by our method. Although the poses by our method also differ from the content in the fourth and bottom rows, the poses have more similar shapes than those of the comparative approaches from the viewpoint of gesture's meaning.

Figure 5 shows failures generated by our methods. The results in the first and second rows imply that the target style gestures have no corresponding poses to the content poses. In contrast, the AdaIN approach works well when the shape of the content pose is relatively simple, as shown in the first column. The results in the third and fourth rows show a similar tendency for more significant differences. We found that the incorrect poses by the swap-based transfers (c), (d), and (e) have a similar shape that crossing both arms in front of a chest. On the other hand, the resulting poses in the bottom row show an opposite phenomenon; the arm-crossing pose in the content sample can not be correctly transferred in all approaches. This curious bias should be investigated. We also found that the hard swap model occasionally causes noisy vibrations, as shown in the supplementary movie. All snapshots in Figures 4 and 5 are picked up from the intra-scenario samples (see Table 7 for details).

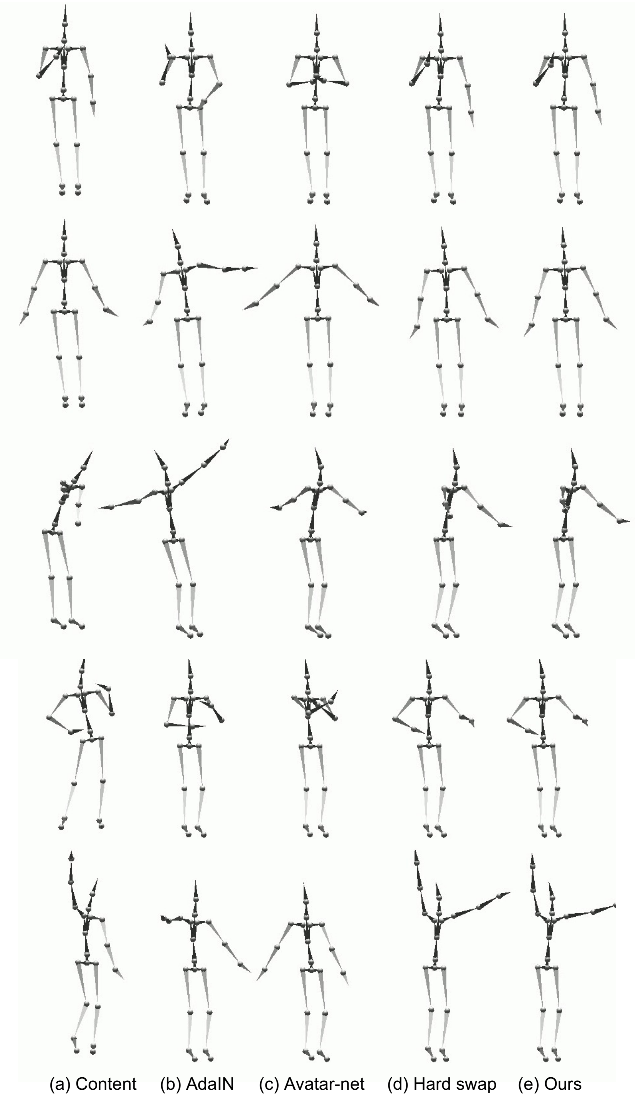
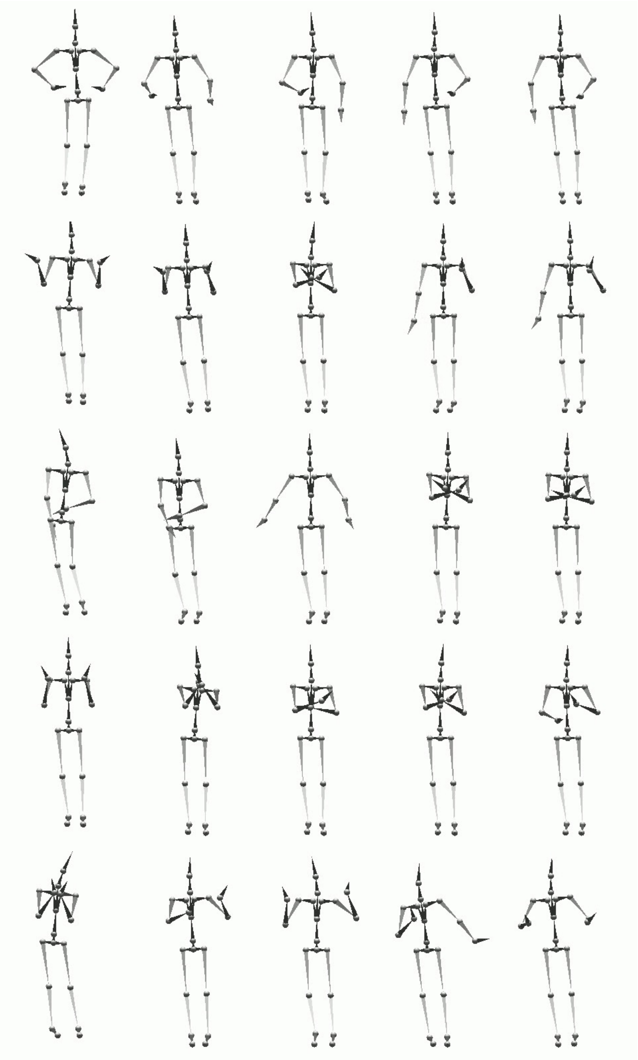


Figure 4: *Snapshot of poses generated by comparative approaches. From top to bottom, the poses are sampled from the scenarios of 10, 15, 15, 9, 13 in Table 7*

4.4. Ablation study of loss functions

Here we evaluate the effect of our loss functions. We trained the style transfer network by omitting a loss function one by one. Table 4 shows the error metrics for the combinations of intra- and inter-scenarios, which shows the degradation of accuracy. The worsen errors are marked in bold, and the minimum errors within each metric are marked in italic. This result suggests that the identity and cyclic losses, which are introduced for consistency in transfer, affect the similarity of content, and the structure and appearance losses, which are computed by splicing the transformer variables, affect the similarity of style.

Interestingly, omitting identity loss achieved the minimum style errors, and omitting structure loss achieved the minimum content



(a) Content (b) AdaIN (c) Avatar-net (d) Hard swap (e) Ours

Figure 5: *Incorrect poses caused by our method. From top to bottom, the poses are sampled from the scenarios of 8, 12, 9, 12, 5 in Table 7*

errors. These results implied some trade-offs between these losses and the potential to use two types of loss settings depending on the relative importance of the content and style. Nevertheless, the use of all losses can achieve the best performance that ensures both qualities of content and style.

5. Conclusion

We proposed a style transfer for gestural animations by incorporating a transformer model for gesture tokens. Our NN model can be trained unsupervised and requires no additional training for new style samples, similar to the relevant image-style transfer. This advantage can avoid the high cost of preparing time-aligned pairs of content and style gestural clips as supervised samples. Our token-wise transfer ensures motion plausibility because the resulting motions are composed of a linear combination of a few tokens of a style gesture in a latent space. Moreover, this mechanism robustly preserved the characteristic features of expressive gestures.

Table 4: *Error metric comparison for every ablation of a loss function, using all combinations of scenarios. The worsened results caused by each ablation are marked in bold letters, and the minimum values for every error metric are marked in italic letters.*

Loss	Content errors $E^{content}$ ↓			Style errors E^{style} ↓		
conditions	mean	worst	w25%	mean	worst	w25%
w/o $L_{identity}$	**0.216**	**0.495**	**0.36**	*0.59*	*1.26*	*1.09*
w/o L_{cyclic}	**0.208**	**0.509**	**0.356**	**0.696**	**1.5**	1.26
w/o $L_{structure}$	*0.198*	*0.488*	*0.338*	**0.709**	**1.54**	**1.3**
w/o $L_{appearance}$	0.201	0.491	0.344	**0.708**	**1.53**	**1.29**
L (all losses)	0.203	0.491	0.344	0.689	1.49	1.26

Our zero-shot learning approach can efficiently transfer style features by piecewise-reshuffling of target style gestures; its expressive power, however, still depends on the diversity or richness of the training dataset and the similarity of reshuffled gestures. The embedding space obtained by the auto-encoder restricts the expression space, and the preservation of content motion is essentially limited by the structure of the target style gesture. Although these limitations are common to the patch-based image style transfer, they are more severe and problematic for the motion data due to the lack of training dataset and the amount of information inherent in each motion clip. Although our method still needs improvement for more convincingly transferring target styles, increasing the amount and variety of samples could be a simple solution. Therefore, sophisticated data augmentation should be developed.

Recently, fully automatic gesture syntheses from voice signals have been intensively developed; most of them, however, lack flexible controls of expressive gestures. This defect becomes serious in applying entertainment fields that require rich expressions of avatars' actions and behaviors in an interactive way. The computational cost of our style transfer is very small and can be run on-the-fly, and this advantage supports the integration with the state-of-the-art gesture synthesis systems, which is the final target of this study.

Although we fixed the token size for the content and style motions, some adaptive optimizations could be developed by treating the size as learnable parameters. These extensions might increase the controllability of the styles more flexibly. Our method deals with joint rotations and requires no joint positions, which has the scalability to the change in body size. However, the performance of style transfer might be degraded when we train the auto-encoder for multiple actors at once due to its simple architecture. This possible defect motivates us to develop a more flexible auto-encoder that can adapt to various body shapes, for example, by introducing a graph-based hierarchy [YXL18, JPL22].

The current limitation of our method is the controllability of the root (or waist) joint because the bending style of the waist cannot be fully reflected in the styles. Furthermore, the actual contribution of our method is limited to upper-body movements. Our method indirectly controls the lower-body motions through the upper-body motions by assuming their correlations, but we could not evaluate their plausibility and the effects of style transfer in a meaningful way. As a by-product of this defect, kinematics constraints of the foot position are easily broken, which causes foot sliding defects. Most motion-style transfer methods suffer from kinematic problems and introduce numerical adjustments based on inverse kinematics. Although our method can also introduce such numerical solutions, the resulting gestures might lose the style feature according to the increase of adjustments. Therefore, a more intensive study is required to capture the gesture styles for a full-body scope. The motions of hand (fingers) play an important role in expressive gestures, and our future works also include the style transfer of hand motions, ideally as an integrated system.

References

[AHKB20] ALEXANDERSON S., HENTER G. E., KUCHERENKO T., BESKOW J.: Style-controllable speech-driven gesture synthesis using normalising flows. *Computer Graphics Forum* (2020). 3

[AWL*20] ABERMAN K., WENG Y., LISCHINSKI D., COHEN-OR D., CHEN B.: Unpaired motion style transfer from video to animation. *ACM Transactions on Graphics (TOG) 39*, 4 (2020), 64. 2

[CCZB00] CHI D., COSTA M., ZHAO L., BADLER N.: The emote model for effort and shape. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques* (USA, 2000), SIGGRAPH '00, pp. 173–182. 3, 6

[CM92] COSTA P. T., MCCRAE R. R.: Four ways five factors are basic. *Personality and Individual Differences 13*, 6 (1992), 653–665. 3

[CS16] CHEN T. Q., SCHMIDT M.: Fast patch-based style transfer of arbitrary style, 2016. arXiv:1612.04337. 1, 2, 4

[CTM*21] CARON M., TOUVRON H., MISRA I., JÉGOU H., MAIRAL J., BOJANOWSKI P., JOULIN A.: Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)* (2021). 2

[DBK*21] DOSOVITSKIY A., BEYER L., KOLESNIKOV A., WEISSENBORN D., ZHAI X., UNTERTHINER T., DEHGHANI M., MINDERER M., HEIGOLD G., GELLY S., USZKOREIT J., HOULSBY N.: An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations* (2021). 2, 3, 4, 5

[DKD*16] DURUPINAR F., KAPADIA M., DEUTSCH S., NEFF M., BADLER N. I.: Perform: Perceptual approach for adding ocean personality to human motion using laban movement analysis. *ACM Trans. Graph. 36*, 1 (Oct. 2016). 3

[DTD*22] DENG Y., TANG F., DONG W., MA C., PAN X., WANG L., XU C.: Stytr2: Image style transfer with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2022). 2, 4

[GBK*19] GINOSAR S., BAR A., KOHAVI G., CHAN C., OWENS A., MALIK J.: Learning individual styles of conversational gesture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019). 3

[GEB16] GATYS L. A., ECKER A. S., BETHGE M.: Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 2414–2423. 2

[Gra98] GRASSIA F. S.: Practical parameterization of rotations using the exponential map. *J. Graph. Tools 3*, 3 (mar 1998), 29–48. 3

[HB17] HUANG X., BELONGIE S.: Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV* (2017). 1

[HHKK17] HOLDEN D., HABIBIE I., KUSAJIMA I., KOMURA T.: Fast neural style transfer for motion data. *IEEE Computer Graphics and Applications 37*, 4 (2017), 42–49. 2

[HPP05] HSU E., PULLI K., POPOVIĆ J.: Style translation for human motion. *ACM Trans. Graph. 24*, 3 (July 2005), 1082–1089. 2

[HSK16] HOLDEN D., SAITO J., KOMURA T.: A deep learning framework for character motion synthesis and editing. *ACM Trans. Graph. 35*, 4 (July 2016). 2

[JPL22] JANG D.-K., PARK S., LEE S.-H.: Motion puzzle: Arbitrary motion style transfer by body part. *ACM Trans. Graph.* (jan 2022). 1, 9

[Ken88] KENDON A.: How gestures can become like words. *Crosscultural Perspectives in Nonverbal Communication* (01 1988). 2

[LFY*17] LI Y., FANG C., YANG J., WANG Z., LU X., YANG M.-H.: Universal style transfer via feature transforms. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2017), NIPS'17, Curran Associates Inc., pp. 385–395. 1, 2

[LKTK10] LEVINE S., KRÄHENBÜHL P., THRUN S., KOLTUN V.: Gesture controllers. *ACM Trans. Graph. 29*, 4 (July 2010). 3

[MBR17] MARTINEZ J., BLACK M. J., ROMERO J.: On human motion prediction using recurrent neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 4674–4683. 1

[Mcn94] MCNEILL D.: Hand and mind: What gestures reveal about thought. *Bibliovault OAI Repository, the University of Chicago Press 27* (06 1994). 6

[MSZ*18] MASON I., STARKE S., ZHANG H., BILEN H., KOMURA T.: Few-shot learning of homogeneous human locomotion styles. *Computer Graphics Forum* (2018). 2

[Nef16] NEFF M.: *Hand Gesture Synthesis for Conversational Characters.* 01 2016. doi:10.1007/978-3-319-30808-1_5-1. 2

[PJL21] PARK S., JANG D.-K., LEE S.-H.: Diverse motion stylization for multiple style domains via spatial-temporal graph-based generative model. *Proc. ACM Comput. Graph. Interact. Tech. 4*, 3 (2021). 2

[SC07] SALVADOR S., CHAN P.: Fastdtw: Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis 11*, 5 (2007), 561–580. 6

[SLSW18] SHENG L., LIN Z., SHAO J., WANG X.: Avatar-net: Multiscale zero-shot style transfer by feature decoration. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), pp. 1–9. 1, 2, 4, 7

[SN17] SMITH H. J., NEFF M.: Understanding the impact of animated gesture performance on personality perceptions. *ACM Trans. Graph. 36*, 4 (July 2017). 3

[SZ15] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015* (2015). 4

[TBTBD22] TUMANYAN N., BAR-TAL O., BAGON S., DEKEL T.: Splicing vit features for semantic appearance transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2022). 2, 5

[VSP*17] VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L., POLOSUKHIN I.: Attention is all you need. In *Advances in Neural Information Processing Systems* (2017), vol. 30, Curran Associates, Inc. 2, 5

[XWCH15] XIA S., WANG C., CHAI J., HODGINS J.: Realtime style transfer for unlabeled heterogeneous human motion. *ACM Trans. Graph. 34*, 4 (July 2015). 2

[YCL*20] YOON Y., CHA B., LEE J.-H., JANG M., LEE J., KIM J., LEE G.: Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Trans. Graph. 39*, 6 (Nov. 2020). 3

[YM16] YUMER M. E., MITRA N. J.: Spectral style transfer for human motion between independent actions. *ACM Trans. Graph. 35*, 4 (July 2016). 2

[YXL18] YAN S., XIONG Y., LIN D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the Thirty-Second AAAI* (2018), pp. 7444–7452. 9

[YYH20] YANG Y., YANG J., HODGINS J.: Statistics-based motion synthesis for social conversations. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Goslar, DEU, 2020), SCA '20, Eurographics Association. 2, 3

6. Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP19H04231 and JP21K19822

Appendix

A. Detailed architecture of feature embedding auto-encoder

The encoders are composed of three one-dimensional convolutional layers followed by a GeLU activation function, where the convolution is performed along the time (frame) axis. The first layer sets the kernel width and stride by 1, the second and third layers set kernel widths by three and strides by two, and the channels are successively converted from 39 to 64, 32, and 16, respectively. This setting was inversely replicated in the decoder.

The channels are altered as $16 \rightarrow 32 \rightarrow 64 \rightarrow 39$ and $16 \rightarrow 16 \rightarrow 32 \rightarrow 24$ for the upper-half and lower-half decoders, respectively. Instead of using multiple strides to expand the frames to the original length, we utilize an upsampling layer using a scale factor of 2 and linear interpolation before the first and second deconvolutional layers to ensure smoothness. Except for the last layer, activation with the GeLU function is similarly used after deconvolution. Note that this auto-encoder is trained independently of the style transfer networks.

B. Dataset for training

Table 5 shows the details of motion clips used for training the style transformer and auto-encoder. The parenthesized symbols in Scenario denote the capital letter of spoken languages; **E**nglish and **J**apanese, and the symbols in Style condition denote the capitals of **O**rdinary, **E**xtraversion, and **A**nime-like, respectively. The Length is the duration of the gesture in seconds for each style condition.

Table 5: *Gesture samples used for training our neural networks.*

Scenario	Style condition	Length (sec.)
Weather forecaster (E)	O, E	63
Weather forecaster (J)	O, A	54
Flight attendant (E)	O, E	62
Flight attendant (J)	O, A	75
Reading fairy tale (E)	O, E	72
Reading fairy tale (J)	O, A	82
Product introduction (J)	O, E, A	50
Encouraging talk (J)	E	46

Table 6 shows the motion clips added for training auto-encoder, where all samples were captured without using spoken voice. The Length is the duration of the gesture in seconds.

Table 6: *Gesture samples added in training auto-encoder.*

Scenario	Length (sec.)
Pointing (one hand, sideway directions)	25
Pointing (one hand, back and force directions)	37
Pointing (two hands, sideway directions)	27
Pointing (two hands, back and force directions)	47
Exaggerated pointing (various hands and directions)	40
Work operations	23
Signature poses	31

C. Dataset for evaluations

Table 7 shows the details of motion clips used for evaluations. The Content denotes the scenario of the gesture, the Style denotes the performed style, and the Length is the duration of the gesture in seconds for each style condition. Notice that every sample also includes the same duration motion clips performed as an ordinary style and a supposed transferred style.

Table 7: *Gesture samples used for evaluations.*

No.	Content	Style	Length (sec.)
1	Attack	Coldly and disappointedly	31
2	Begging	Friendly and lively	27
3	Blame	Coldly	22
4	Denying	In an panic	32
5	Excuse	In doubt	24
6	Expectation	Uneasy	35
7	Good-bye	Neatly and cleanly	35
8	Hate	Proudly	31
9	Order	High-handedly	20
10	Reminiscence	Gratefully	30
11	Talk in cafe	Actively and charmingly	36
12	Teaching	Kindly	34
13	Teenage chat	Pleasurably	34
14	Temptation	Sexy	30
15	Waking up	Hurriedly and violently	32