

# Motion Puzzle: Arbitrary Motion Style Transfer by Body Part

DEOK-KYEONG JANG, Korea Advanced Institute of Science and Technology, South Korea

SOOMIN PARK, Korea Advanced Institute of Science and Technology, South Korea

SUNG-HEE LEE, Korea Advanced Institute of Science and Technology, South Korea

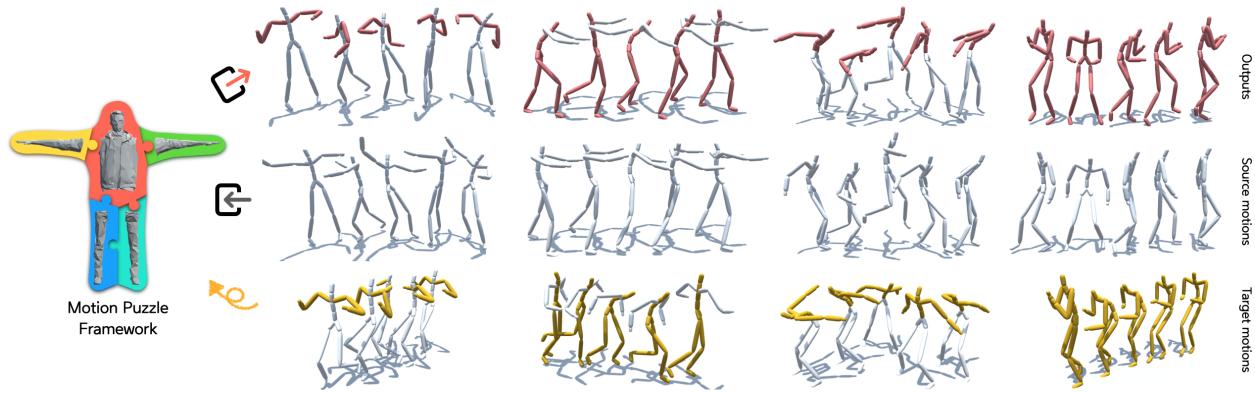


Fig. 1. Our Motion Puzzle framework translates the motion of selected body parts in a source motion to a stylized output motion. It can also compose the motion style from multiple target motions for different body parts like a puzzle to stylize the full-body output motion while preserving the content of the source motion. From the left, this figure shows examples of transferring styles in the target motions for the spine and legs, spine and arms, arms only, and whole body.

This paper presents Motion Puzzle, a novel motion style transfer network that advances the state-of-the-art in several important respects. The Motion Puzzle is the first that can control the motion style of individual body parts, allowing for local style editing and significantly increasing the range of stylized motions. Designed to keep the human's kinematic structure, our framework extracts style features from multiple style motions for different body parts and transfers them locally to the target body parts. Another major advantage is that it can transfer both global and local traits of motion style by integrating the adaptive instance normalization and attention modules while keeping the skeleton topology. Thus, it can capture styles exhibited by dynamic movements, such as flapping and staggering, significantly better than previous work. In addition, our framework allows for arbitrary motion style transfer without datasets with style labeling or motion pairing, making many publicly available motion datasets available for training. Our framework can be easily integrated with motion generation frameworks to create many applications, such as real-time motion transfer. We demonstrate the advantages of our framework with a number of examples and comparisons with previous work.

This work was supported by National Research Foundation, Korea (NRF-2020R1A2C2011541), and Korea Creative Content Agency, Korea (R2020040211).

Authors' addresses: Deok-Kyeong Jang, shadofex@kaist.ac.kr, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon, 34141, South Korea; Soomin Park, sumny@kaist.ac.kr, Korea Advanced Institute of Science and Technology, South Korea; Sung-Hee Lee, sunghlee@kaist.ac.kr, Korea Advanced Institute of Science and Technology, South Korea.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM. 0730-0301/2022/6-ART33 \$15.00 https://doi.org/10.1145/3516429

**CCS Concepts:** • Computing methodologies → Motion processing; Neural networks.

**Additional Key Words and Phrases:** Motion style transfer, Motion synthesis, Character animation, Deep learning

## ACM Reference Format:

Deok-Kyeong Jang, Soomin Park, and Sung-Hee Lee. 2022. Motion Puzzle: Arbitrary Motion Style Transfer by Body Part. *ACM Trans. Graph.* 41, 3, Article 33 (June 2022), 16 pages. <https://doi.org/10.1145/3516429>

## 1 INTRODUCTION

Style is an essential component in human movement as it reflects various aspects of people, such as personality, emotion, and intention. Thus the expressibility of style is crucial for animating human characters and avatars. However, obtaining stylized motions is a challenging task: Since motion style is somewhat subtle and subjective, professional actors or animators are often necessary to capture or edit motions, and even with them many iterations may be required because the same style can be expressed in numerous ways. Such difficulties make producing high-quality stylized motions costly and time-consuming.

An effective way of creating stylized motions is extracting a style component from example motion data (called target motion) and applying it to another motion (called source motion) containing desired content, for which researchers have achieved remarkable progress [Aberman et al. 2020b; Holden et al. 2016; Park et al. 2021; Yumer and Mitra 2016]. Yet, previous methods still suffer from several limitations.

First, they require style labeling or motion pairing in training datasets for specific loss functions or computing the difference between motions, making it difficult to use many publicly available

unlabeled motion datasets. Second, these methods only capture spatially and temporally averaged global style features of the target motion and do not capture local style features well. Thus, they are limited in extracting time-varying styles, styles exhibited by dynamic movements (e.g., fluttering or instantaneous falling). Third, they can only modify the whole-body motion while, in practical scenarios, it is often needed to change the style of only some part of the body. For example, they cannot add a collapsing style (tripping motion) only to the spine and legs of a zombie walking to create a collapsing style zombie walking. This operation is in analogy to the segment style transfer that changes styles of only the nose, eyes, or lips of the whole face in image transfer [Shen et al. 2020].

This paper presents a novel framework for motion style transfer with important improvements on these limitations. A unique advantage of our framework is that it can control the motion style of individual body parts (the arms, legs, and spine) while preserving the content of a given source motion; hence we name it *Motion Puzzle*. The ability to stylize motion by body part can dramatically increase the range of expressible motions: Rather than simply transferring a single style to a whole body, by combining various style components applied to different body parts, or by applying a new style component to only a subset of body parts, it can synthesize a wide variety of stylized motions. To the best of our knowledge, our work is the first that achieves this per-body-part motion style transfer.

*Motion Puzzle* framework consists of two encoders for extracting multi-level style features for the individual body parts from target motions and a content feature from a source motion, and one decoder to synthesize a stylized whole-body motion. For per-body-part style transfer, we develop a novel Body Part Style Network (BP-StyleNet), which takes the human body part topology into account during the style injection process, and employ a skeleton aware graph convolutional network to extract and process features while preserving the structure of body parts.

Another major advantage of our framework is that it can transfer both global and local traits of motion style via two-step transfer modules: Body Part Adaptive Instance Normalization (BP-AdaIN) and Body Part Attention Network (BP-ATN) inside BP-StyleNet. BP-AdaIN module injects global style features into the content feature by applying AdaIN [Huang and Belongie 2017] by body part. BP-ATN module transfers the locally semantic style features by constructing an attention map between content feature and style feature by body part. Especially, the time-varying motion style is captured via ATN module. For example, given a motion style of flapping with bent arms, our framework successfully extracts and transfers arm bending (global feature) and flapping (local feature).

Finally, our framework enables arbitrary (zero-shot) motion style transfer by learning to identify the style and content components of any motion. By using datasets with style labeling or motion pairing, previous methods learn style latent space labeled with general text, e.g., happy or old [Aberman et al. 2020b; Park et al. 2021]. However, motion styles are often ambiguous to describe in text, especially if they are subtle, compositive or temporally changing. This results in significant variability of text labels and a weak association between given motion and labels [Kim and Lee 2019]. Therefore, while style labeling can be effective for basic categories such as emotion, it may

hardly be extended to a broader range of styles that are challenging to label. With this awareness, we consider motion style as a specific way of movement to realize content (or task) in a given motion, and do not relate it with other styles in different motions using general text labels. Our approach allows using many publicly available motion datasets for training and transferring truly unseen arbitrary motion styles in test mode.

Figure 1 shows the idea of the Motion Puzzle framework: Given one source motion (middle row) for content and target body parts motions (yellow parts, bottom) for style, the Motion Puzzle framework translates the motion of selected body parts in a source motion to a stylized output motion (red parts, top). In addition, our framework can be easily integrated with other motion generators. We demonstrate a real-time motion style transfer by integrating with an existing motion controller.

The contributions of our work can be summarized as follows:

- We present the first motion style transfer method that can control the motion style of individual body parts.
- We propose a new two-step style adaptation network, BP-StyleNet, consisting of BP-AdaIN and BP-ATN, to transfer spatially and temporally global and local features of motion style, which presents a distinctive advantage: time-varying style features can be captured and translated.
- Our architecture that considers the human skeleton structure and body parts makes arbitrary (zero-shot) style transfer possible without style labeling or motion pairing of training data.

## 2 RELATED WORK

In this section, we introduce previous research on stylizing human motion and some image style transfer methods that contributed much to motion stylization.

### 2.1 Arbitrary Image Style Transfer

The methods for transferring arbitrary image styles aim at zero-shot learning to synthesize a content image adopting an arbitrary unlabeled style of another image. Gatys et al. [2016] showed that the style of an image can be represented by the Gram matrices, which compute correlations between the feature maps of pre-trained convolutional networks. Their method includes a computationally heavy optimization step, which can be replaced with faster feed-forward networks [Johnson et al. 2016]. Huang and Belongie [2017] introduced the adaptive instance normalization (AdaIN), which learns to adapt the affine parameters to the style input to enable transferring arbitrary styles. Since the AdaIN simply modifies the mean and variance of the content image, it cannot sufficiently transfer the semantic features of style images. Li et al. [2017] transformed content features into a style feature space by applying whitening and coloring transformation (WCT) with aligning the covariance. WCT is computationally heavy to deal with high dimensional features and it does not perform well when content and style images are semantically different. Sheng et al. [2018] introduced Avater-Net, a patch-based style decorator that can transfer style features to the semantically nearest content features. However, it cannot capture

both global and local style patterns since the scale of patterns in the style images depends on the patch size. Park et al. [2019] proposed a style-attention network (SANet) to integrate the local style features according to the semantically nearest content features for achieving good results with evident style patterns.

In our work, we develop a method that applies the AdaIN in a way that preserves the structure of human body to be able to transfer the global style feature of individual body parts. In addition, we leverage the attention network to transfer local style features to semantically matched content features.

## 2.2 Motion Style Transfer

Previous work of motion style transfer using early machine learning methods infers motion style with manually defined features. The work of [Hsu et al. 2005] models the style difference of two motions with a linear time-invariant (LTI) system, which can transform the style of a new motion that is similar in content. Ma et al. [2010] modeled style variation of individual body parts with latent parameters, which are controlled by user-defined parameters through a Bayesian network. Xia et al. [2015] proposed a method to construct the mixtures of autoregressive (MAR) models online to represent style variations locally at the current pose and apply linear transformations to control the style. Another approach is to represent the motion style in the frequency domain using the Fourier Transform [Unuma et al. 1995; Yumer and Mitra 2016], which allows extracting style features from a small dataset without a need to conduct spatial matching between data. These studies on motion style transfer are usually adequate only for a limited range of motions and may require special processing, such as time-warping and alignment, of the example motions or searching motion database.

Deep learning-based approaches have greatly improved the quality and the possible range of motion stylization. Holden et al. [2017a; 2016] showed that the motion style can be transferred based on the Gram matrices [Gatys et al. 2016] through motion editing in the latent space. Du et al. [2019] presented a conditional Variational Autoencoder (CVAE) with the Gram matrices to construct the style-conditioned distribution for human motion. These approaches require much computing time to extract style features through optimization and have a limitation in capturing complex or subtle motion features, making style transfer between motions with significantly different contents ineffective. Mason et al. [2018] applied few-shot learning to synthesize stylized motions with limited motion data. However, similarly to [Du et al. 2019], this method is confined to specific types of motion, such as locomotion.

Recently, Aberman et al. [2020b] applied the generative adversarial networks (GAN) based architecture with the AdaIN from FUNIT [Liu et al. 2019a] model used in image style transfer. They alleviated the restrictions on training data by allowing for training the networks with an unpaired dataset with style labels while preserving motion quality and efficiency. It, also notably, can transfer style from videos to 3D animations by learning a shared style embedding for both 3D and 2D joint positions. Park et al. [2021] constructed a spatio-temporal graph to model a motion sequence, which improves style translation between significantly different actions. Their framework generates diverse stylization results for a target

style by introducing a network that maps random noise to various style codes for the target style. However, the above approaches have two limitations. First, they require labeled data for content or style to construct adversarial loss of GAN-based model, making zero-shot (arbitrary) motion style transfer impossible. They cannot use many publicly available, unlabeled motion datasets, while ease of obtaining a wide range of data is critical for a deep learning method to generalize to the data in the wild. Second, style transfer using only the AdaIN cannot capture time-varying motion style. Since the AdaIN simply modifies the mean and variance of the content features to translate the source motion, it captures temporally global features well but loses temporally local features. Wen et al. [2021] proposed a flow-based motion stylization method. Its probabilistic structure allows to generate various motions with a specific set of style, context and control, but it also suffers from capturing time-varying motion styles.

In contrast, our framework achieves high-quality motion style transfer by using a novel model structure that does not require data labeling for adversarial loss. Therefore, our framework can transfer truly unseen arbitrary motion styles. In addition, we solve the problem of capturing time-varying motion style by developing BP-StyleNet, which combines AdaIN-based module for the global style feature and the attention module for the local style feature.

## 2.3 Skeleton-based Graph Networks

Since the human skeleton has specific connectivity, researchers have proposed network structures and associated operations that respect the skeleton topology to model correlation between joints. Zhou et al. [2014] constructed a partition-based structure of human skeleton and used a graphical representation of the conditional dependency between joints to model variations of human motion. Yan et al. [2018] proposed a spatio-temporal graph convolution network (STGCN) to deal with skeleton motion data. STGCN consists of spatial edges that connect adjacent joints in the human body and temporal edges that connect the same joint in consecutive frames. The skeleton-aware graph structures have been successfully adopted in many applications [Huang et al. 2020; Li et al. 2019; Shi et al. 2019a,b; Si et al. 2018]. Aberman et al. [2020a] proposed skeleton-aware operators for convolution, pooling and unpooling for motion retargeting between different homeomorphic skeletons. Park et al. [2021] used STGCN for motion style transfer.

Most previous work on deep learning-based motion style transfer [Aberman et al. 2020b; Du et al. 2019; Holden et al. 2017a, 2016; Smith et al. 2019] did not consider the hierarchical spatial and temporal structure of the human skeleton. Park et al. [2021] used STGCN to consider the skeletal joint structure during the convolution but did not take the human body part topology into account for the style injection. In this work, we also use STGCN to represent spatial-temporal change of human motion but go a significant step further by proposing novel style adaptation networks, BP-StyleNet that consider human body parts during style injection process. Thanks to this skeleton-aware transfer network, our framework can control style transfer per body part, which previous studies have not made possible.

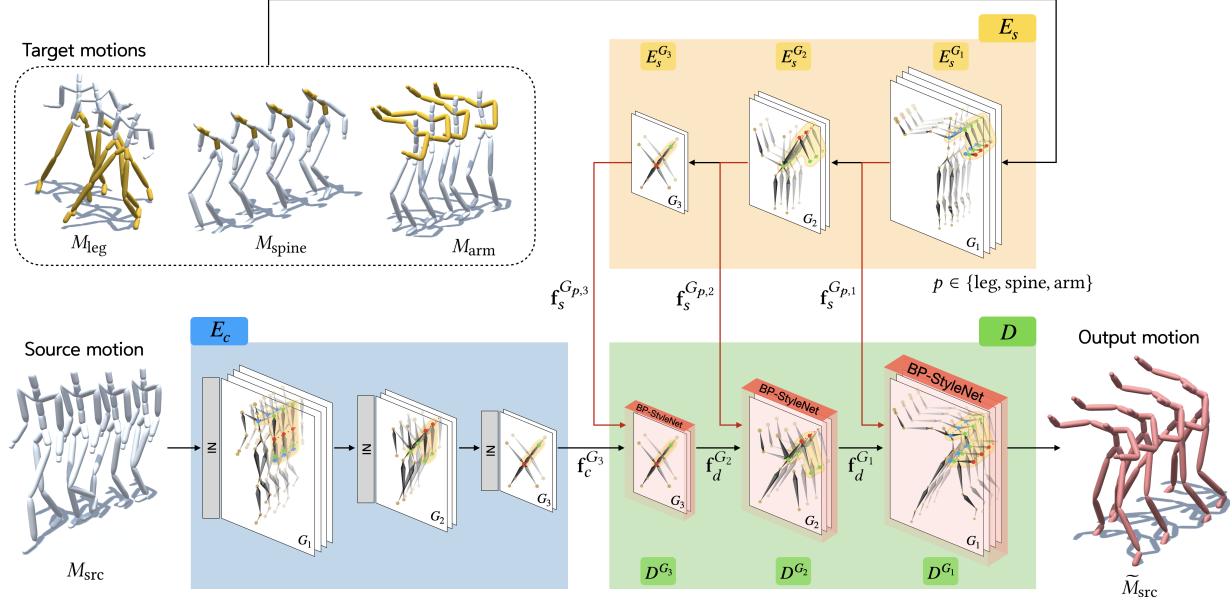


Fig. 2. The overall network architecture of our Motion Puzzle framework. Our framework can transfer the styles of some body parts in the target motion to the corresponding body parts in the source motion. This figure illustrates a case where three target motions, each for the leg  $M_{\text{leg}}$ , spine  $M_{\text{spine}}$ , and arm  $M_{\text{arm}}$ , are used to stylize the source motion  $M_{\text{src}}$  to generate the stylized output motion  $\tilde{M}_{\text{src}}$ . Target motions are encoded into multi-level style features  $f_s^{G_{p,i}}$  for the legs, arms and spine, respectively, through the style encoder  $E_s$ , and a source motion passes through the content encoder  $E_c$  to extract a content feature  $f_c^{G_3}$ . The decoder  $D$  of the framework receives the content feature and per-body-part multi-level style features, and progressively synthesizes a stylized whole-body intermediate features  $f_d^{G_{i-1}}$  through a novel per-body-part style adaption network  $\text{BP-StyleNet}$ .

### 3 MOTION DATA REPRESENTATION AND PROCESSING

Before discussing our framework in detail, we first explain the representation of motion data and the construction of the motion dataset.

We denote a human motion set by  $\mathcal{M}$  and its subset of  $T$  total number of frames by  $\mathcal{M}_T \subset \mathcal{M}$ , with  $\mathcal{M}_T$  being a random variable of  $\mathcal{M}_T$ . We represent a motion of length  $T$  as  $\mathcal{M}_T = [\mathbf{M}_1, \dots, \mathbf{M}_t, \dots, \mathbf{M}_T]$ , where  $\mathbf{M}_t$  denotes a pose feature matrix at frame  $t$ . Specifically, we define  $\mathbf{M}_t = [\mathbf{m}_j^p, \mathbf{m}_j^r, \dot{\mathbf{m}}_j^p, \mathbf{r}^x, \mathbf{r}^z, \mathbf{r}^a]_{j=1}^{n_{\text{joint}}} = [\mathbf{M}_{t,j}]_{j=1}^{n_{\text{joint}}}$ , where  $\mathbf{m}_j^p \in \mathbb{R}^3$ ,  $\mathbf{m}_j^r \in \mathbb{R}^6$  and  $\dot{\mathbf{m}}_j^p \in \mathbb{R}^3$  are the local position, rotation and velocity of joint  $j$ , expressed with respect to the character forward-facing direction in the same way as [Holden et al. 2016]. A joint rotation is represented by the 2-axis (forward and upward vectors) rotation matrix as in [Zhang et al. 2018]. Under the assumption that the root motion plays a critical role in motion style, each joint information is accompanied by the root velocity information:  $\mathbf{r}^x \in \mathbb{R}$  and  $\mathbf{r}^z \in \mathbb{R}$  are the root translational velocities in  $X$  and  $Z$  directions relative to the previous frame  $t - 1$ , represented with respect to the current frame  $t$ , and  $\mathbf{r}^a \in \mathbb{R}$  is the root angular velocity about the vertical ( $Y$ ) axis.

Therefore, the dimension of our human motion feature with  $T$  frames is  $\mathcal{M}_T \in \mathbb{R}^{T \times n_{\text{joint}} \times d_{\text{joint}}}$  and that of our pose feature matrix is  $\mathbf{M}_t \in \mathbb{R}^{n_{\text{joint}} \times d_{\text{joint}}}$ , where  $d_{\text{joint}} = 15$  is the degrees of freedom of a joint. The data of our motion feature is arranged as a tensor of  $T \times n_{\text{joint}} \times d_{\text{joint}}$  dimensions for our network architecture, whereas

[Holden et al. 2016] used  $T \times d_{\text{body}}$  matrices, where  $d_{\text{body}}$  is the degrees of freedom of a pose.

#### 3.1 Dataset Construction

We construct a motion dataset using the CMU motion data [CMU ], which provides various motions with unlabeled styles. All motion data are 60 fps and retargeted to a single 21-joint skeleton with the same skeleton topology as the CMU motion data. An additional dataset of [Xia et al. 2015] is used as unseen test data for quantitative evaluation in Sec. 7.2. While motion data can have variable frame lengths at runtime, we train our networks with motion clips of 120 frames for convenience. For this, a long motion sequence is divided into 120-frame motion clips overlapping 60 frames with adjacent clips.

*Data augmentation.* To enrich the training dataset, we augment data in several ways. First, we mirror the skeleton to double the number of original motion clips. Second, we obtain additional data by changing the velocity of motion. For this, we randomly sample a sub-motion clip  $M_{\Delta t}$  of length  $\Delta t \in [\frac{T}{2}, T]$  ( $T = 120$  in our experiment), and scale  $M_{\Delta t}$  by a random factor  $\gamma$ , where  $\gamma \in [1, 2]$  if  $\Delta t < \frac{3}{4}T$  or  $\gamma \in [0.5, 1]$  otherwise. The former effectively decreases the velocity of the motion, and the latter increases it. The scaled motion is either cut or padded to make 120 frames. We call this data augmentation method *temporal random cropping*. We perform the temporal random cropping on input motion clip data with the rate

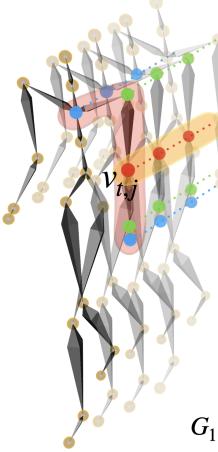


Fig. 3. Illustration of the spatial-temporal graph. Vertices are denoted as yellow circles. Red, green and blue vertices denote 0, 1 and 2 edge distances from  $v_{t,j}$ . Black bones constitute the spatial edges, and dotted lines make the temporal edges. Note that this figure is only to visualize the connections of vertices from the perspective of the skeleton structure. The features of each vertex in the graph are not directly related to vertex positions.

of 0.2 during training. Finally, we build a training dataset of about 120K motion clips of 120 frames in 60fps.

#### 4 MOTION PUZZLE FRAMEWORK

Figure 2 shows the overall architecture of the Motion Puzzle framework. Inputs to our framework are one source motion for the content and multiple target motions, each for stylizing one or more body parts, and the output is a stylized whole-body motion.

To encode and synthesize complex human motions in which numerous joints move over time in spatially and temporally correlated manners, we use a spatial-temporal graph convolutional network as the basis of our framework and employ a graph pooling-unpooling method to keep the graph in accordance with human’s body part structure (Sec. 4.1). Maximum five target motions are encoded into multi-level style features for the legs, arms and spine, respectively, through the style encoder (Sec. 4.2), and a source motion passes through the content encoder to extract a content feature (Sec. 4.3). The decoder of the framework receives the content feature and per-part multi-level style features, and synthesizes a stylized whole-body motion. In this process, each level of style feature is adapted progressively into the spatial-temporal feature for the corresponding body part through a novel per-body-part style transfer network named *BP-StyleNet* (Sec. 4.4).

##### 4.1 Skeleton based Graph Convolutional Networks

*Skeletal graph construction.* A human motion  $M_T$  has a temporal relationship in the sequence of pose features  $M_1, \dots, M_T$ , each of which shows hierarchical spatial relationship among the skeletal joints. We use a spatial-temporal graph to model these characteristics. Specifically, we follow the graph structure proposed by [Yan et al. 2018], summarized below.

Figure 3 shows an example of a spatial-temporal graph  $G_1 = (V_1, E_1)$  featuring both the intra-body and inter-frame connections for a skeleton sequence with  $n_{\text{joint}}$  joints and  $T$  frames. The vertex set  $V_1 = \{v_{t,j} | t = 1, \dots, T \text{ and } j = 1, \dots, n_{\text{joint}}\}$  corresponds to all the joints in the skeleton sequence. The edge set  $E_1$  consists of two kinds: Spatial edges  $E_B = \{v_{t,j}v_{t,i} | (j, i) \in S\}$ , where  $S$  is a set of connected joint pairs, connect adjacent joints at each frame  $t$ , and

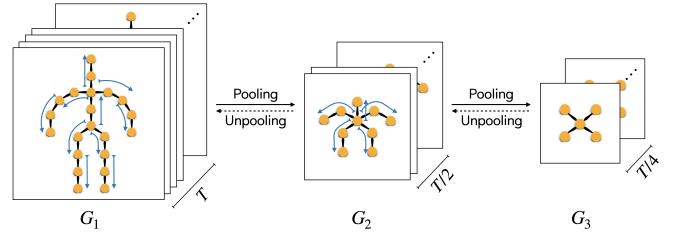


Fig. 4. Body part preserving pooling and unpooling of the spatial-temporal graph. Blue arrows indicate the vertices used by each pooling kernel along the kinematics chain directions.

the temporal edges  $E_F = \{v_{t,i}v_{t-1,i}\}$  connect the vertices for the same joint in adjacent time frames.

*Graph convolution.* We describe the spatial-temporal graph convolution using the case of vertex  $v_{t,j}$  in graph  $G_1$  shown in Fig. 3. The spatial graph convolution at frame  $t$  is formulated as follows [Yan et al. 2018]:

$$f_{\text{out}}(v_{t,j}) = \sum_{v_{t,i} \in \mathcal{B}_{t,j}} \frac{1}{Z_{t,j}} f_{\text{in}}(v_{t,i}) \cdot \mathbf{w}(d(v_{t,i}, v_{t,j})), \quad (1)$$

where  $f$  denotes the feature map, e.g.,  $f(v_{t,j}) = M_{t,j}$  for graph  $G_1$  if no operation has been performed yet, and  $\mathcal{B}_{t,j}$  is the neighboring vertices for convolving  $v_{t,j}$ , defined as the  $K$  edge-distance neighbor vertices  $v_i$  from the target vertex  $v_j$ :

$$\mathcal{B}_{t,j} = \{v_{t,i} | d(v_{t,i}, v_{t,j}) \leq K\}, \quad (2)$$

where  $K = 2$  for  $G_1$ .

The weight function  $\mathbf{w}$  provides a weight vector for the convolution, and we choose to model it as a function of edge distance  $d(v_{t,i}, v_{t,j})$ . That is, the vertices of the same color in Fig. 3 are given the same weight vector  $\mathbf{w}$  to convolve  $v_{t,j}$ . Note that while the number of convolution weight vectors is fixed, the number of vertices in  $\mathcal{B}_{t,j}$  varies, of which effect is normalized by dividing by the number  $Z_{t,j}$  of vertices of the same edge distance. The convolution in the temporal dimension is conducted straightforwardly as regular convolution because the structure of the temporal edge is fixed. Graph convolutions on  $G_2$  and  $G_3$  are performed in the same way with  $K = 1$ .

*Graph Pooling and Unpooling.* Recent image transfer methods [Karras et al. 2019; Liu et al. 2019a] show remarkable performance in translating images by gradually increasing resolution and adding details while unpooling (upsampling) high-level features and injecting style features. Following this idea, we incorporate the pooling and unpooling method into our spatial-temporal graph to handle and extract features from the low (local) to high (global) level.

We use a standard average pooling method for the temporal dimension of the graph as it is uniformly structured with vertices sequentially connected along time. However, as its spatial dimension is non-uniformly structured, standard pooling methods are not applicable. For this, several methods [Aberman et al. 2020a; Yan et al. 2019] exist to preserve the skeletal structure for motion retargeting and synthesis. We devise a similar graph pooling method (Fig. 4),

with a difference being that our pooling operator (blue arrow) averages consecutive vertices along the kinematic chain only within the same body part to retain the individuality of the body parts. As a result of pooling, our framework uses graph structures  $G_i = (V_i, E_i)$  in varying resolutions. Specifically,  $|V_1| = 21 \times T$ ,  $|V_2| = 10 \times \frac{T}{2}$  and  $|V_3| = 5 \times \frac{T}{4}$ . Note that each vertex  $v_{t,j}$  in  $G_3$  corresponds to each of the 5 body parts. The edges  $E_2$  and  $E_3$  are constructed similarly as  $E_1$ .

Unpooling is conducted in the opposite direction of the pooling. We use the nearest interpolation for upscaling the temporal dimension. For the spatial dimension, we unpool the joint dimension by mapping the pooled vertices to the previous skeleton structure.

#### 4.2 Style Encoder

We develop our style encoder to extract the style features in multi-level for gradual translation. The style encoder  $E_s$  is a concatenation of multi-level encoding blocks  $E_s^{G_i}$ ,  $i \in \{1, 2, 3\}$  corresponding to graph  $G_i$ , each of which consists of a STGCN layer followed by a graph pooling layer. Each encoding block progressively extracts the intermediate style feature  $\mathbf{f}_s^{G_i} = E_s^{G_i}(\mathbf{f}_s^{G_{i-1}})$ . As a result, given a target motion  $M_p$ , the style encoder  $E_s$  extracts multi-level style features.

$$[\mathbf{f}_s^{G_{p,i}}]_{i=1}^3 = E_s(M_p), \quad (3)$$

where  $G_{p,i}$  is the graph from  $M_p$ ,  $\mathbf{f}_s^{G_{p,1}} \in \mathbb{R}^{21 \times T_p \times C}$ ,  $\mathbf{f}_s^{G_{p,2}} \in \mathbb{R}^{10 \times \frac{T_p}{2} \times 2C}$  and  $\mathbf{f}_s^{G_{p,3}} \in \mathbb{R}^{5 \times \frac{T_p}{4} \times 4C}$  with  $T_p$  being the frame number of  $M_p$  and  $C$  being the feature dimension. These style features are used to transfer styles progressively during the decoding process.

Each level of feature  $\mathbf{f}_s^{G_{p,i}}$  can be divided into five part-features that correspond to the five body parts thanks to STGCN and the part-preserving pooling scheme.

$$\mathbf{f}_s^{G_{p,i}} = [\mathbf{f}_s^{LL_{p,i}}, \mathbf{f}_s^{RL_{p,i}}, \mathbf{f}_s^{SP_{p,i}}, \mathbf{f}_s^{LA_{p,i}}, \mathbf{f}_s^{RA_{p,i}}], \quad (4)$$

where  $LL$ ,  $RL$ ,  $SP$ ,  $LA$ , and  $RA$  denote subset vertices of graph  $G$  corresponding to the left leg, right leg, spine, left arm, and right arm, respectively. This structure-aware style transfer is reasonable for the graph convolution-based approach and leads to a better stylization performance, as will be shown in Sec. 7.2.

Leveraging this division, from one target motion, we can select only a subset of features corresponding to the target body parts to apply the encoded style, and for other body parts, we can use a subset of encoded style features of other target motion. Thus, the whole set of style features can be made by combining these part-features from up to 5 different motions  $p, \dots, t$ .

$$\mathbf{f}_s^{G_i} = [\mathbf{f}_s^{LL_{p,i}}, \mathbf{f}_s^{RL_{q,i}}, \mathbf{f}_s^{SP_{r,i}}, \mathbf{f}_s^{LA_{s,i}}, \mathbf{f}_s^{RA_{t,i}}]. \quad (5)$$

Note that each part-feature is allowed to have different temporal lengths. This approach gives much freedom for controlling styles. We can use only one target motion for the whole body parts ( $p = \dots = t$ ) or three motions for the arm, spine, and leg ( $p = q$  and  $s = t$ ), both of which can be general use cases of our method. For those parts that we want to preserve the original style of the source motion, the style feature of the source motion can be used. Hereafter, we will omit motion IDs  $p, \dots, t$  but assume that  $\mathbf{f}_s^{G_i}$  is made by one or more motions. Figure 2 shows a case where three target motions

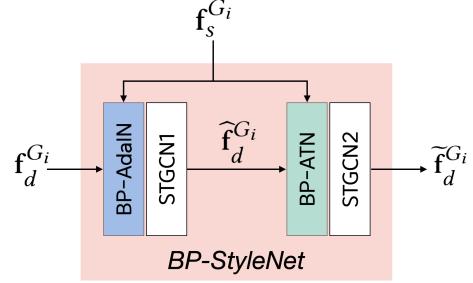


Fig. 5. BP-StyleNet on  $G_i$  level. It transfers the global and local characteristics of a style feature  $\mathbf{f}_s^{G_i}$  to  $\mathbf{f}_d^{G_i}$  via two-step transfer modules, BP-AdaIN and BP-ATN.

$M_{\text{leg}}$ ,  $M_{\text{spine}}$ , and  $M_{\text{arm}}$  are used for transferring styles to the legs, spine, and arms.

#### 4.3 Content Encoder

The content encoder  $E_c$  has a similar structure as the style encoder. The differences are that every STGCN layer is preceded by Instance Normalization (IN) to normalize feature statistics and remove style variations, and only the final output  $\mathbf{f}_c^{G_3}$  is used for the content feature. As a result, the content encoder extracts the style-invariant latent representation for motion. Given a source motion  $M_{\text{src}}$ , the encoding process (blue part in Fig. 2) is written as:

$$\mathbf{f}_c^{G_3} = E_c(M_{\text{src}}), \quad (6)$$

which maps the low-level joint feature  $M_{\text{src}} \in \mathbb{R}^{21 \times T_c \times d_{\text{joint}}}$  to a high-level body part content feature  $\mathbf{f}_c^{G_3} \in \mathbb{R}^{5 \times \frac{T_c}{4} \times 4C}$ .

Same as the style feature, the content feature has the graph structure in accordance with human's five body parts, which plays a key role not only in better reconstructing human motion but also in performing per-body-part style transfer during the decoding step.

#### 4.4 Decoder

The decoder  $D$  (green part in Fig. 2) transforms the content feature  $\mathbf{f}_c^{G_3}$  into an output motion  $\tilde{M}_{\text{src}}$  using the multi-level target style features  $[\mathbf{f}_s^{G_i}]_{i=1}^3$ .

$$\tilde{M}_{\text{src}} = D(\mathbf{f}_c^{G_3}, [\mathbf{f}_s^{G_i}]_{i=1}^3) \quad (7)$$

The decoder network takes an inverse form of the encoder networks. Similarly to the multi-level style adaptation module for image style transfer [Sheng et al. 2018], we progressively generate the intermediate features  $\mathbf{f}_d^{G_{i-1}} = D^{G_i}(\mathbf{f}_d^{G_i})$  starting from  $\mathbf{f}_d^{G_3} (= \mathbf{f}_c^{G_3})$ , with  $\dim(\mathbf{f}_d^{G_i}) = \dim(\mathbf{f}_c^{G_i})$ . Each level of decoding block  $D^{G_i}$  is composed of our novel style adaption network *BP-StyleNet* and an unpooling layer.

**BP-StyleNet.** In order to transfer motion styles, we propose *BP-StyleNet*. As shown in Fig. 5, it translates a decoded content feature  $\mathbf{f}_d^{G_i}$  into feature  $\tilde{\mathbf{f}}_d^{G_i}$  with style feature  $\mathbf{f}_s^{G_i}$  via two-step transfer modules, namely BP-AdaIN and BP-ATN. In the first step, BP-AdaIN (Adaptive Instance Normalization) transfers the global statistics of style feature to generate  $\tilde{\mathbf{f}}_d^{G_i}$ . Next, BP-ATN (attention network)

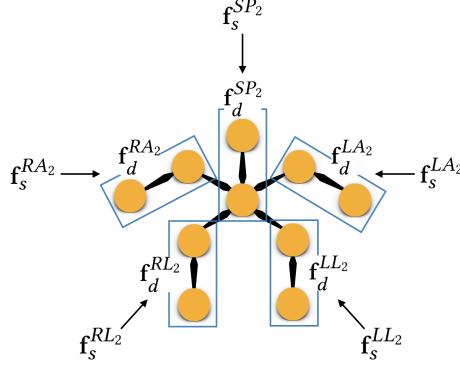


Fig. 6. Example of applying BP-AdaIN (blue boxes) on  $G_2$  level. The BP-AdaIN module injects style feature  $f_s^{G_2}$  into  $f_d^{G_2}$  by body part.

transfers the local statistics of style feature to create  $\tilde{f}_d^{G_i}$ , which now reflects both the global and local traits of style feature.

The BP-AdaIN module takes  $f_d^{G_i}$  as input and injects style feature  $f_s^{G_i}$  by applying AdaIN by body part:

$$\text{BP-AdaIN}(f_d^{G_i}, f_s^{G_i}) = [\text{AdaIN}(f_d^{LL_i}, f_s^{LL_i}), \dots, \text{AdaIN}(f_d^{RA_i}, f_s^{RA_i})]. \quad (8)$$

Figure 6 shows an example of how BP-AdaIN injects style features into each body part in  $G_2$ . The style injection process of AdaIN is conducted as:

$$\text{AdaIN}(f_d^{P_i}, f_s^{P_i}) = \gamma^{P_i}(f_s^{P_i}) \left( \frac{f_d^{P_i} - \mu(f_d^{P_i})}{\sigma(f_d^{P_i})} \right) + \beta^{P_i}(f_s^{P_i}), \quad (9)$$

where  $P \in \{LL, RL, SP, LA, RA\}$ ,  $\mu$  and  $\sigma$  are the channel-wise mean and variance, respectively. AdaIN scales the normalized  $f_d^{P_i}$  with a learned affine transformation with scales  $\gamma^{P_i}$  and biases  $\beta^{P_i}$  generated by  $f_s^{P_i}$ . Finally, we feed output feature into STGCN1 layer and generate the first stage of the stylized feature.

$$\tilde{f}_d^{G_i} = \text{STGCN1}(\text{BP-AdaIN}(f_d^{G_i})). \quad (10)$$

In the second step, the BP-ATN module transfers the locally semantic style features of  $f_s^{G_i}$  via constructing an attention map with  $\tilde{f}_d^{G_i}$  by body part. The BP-ATN module was inspired by [Fu et al. 2019; Park and Lee 2019; Wang et al. 2018], which use attention block for action recognition, image segmentation, and image style transfer. We feed the globally-stylized decoded feature  $\tilde{f}_d^{G_i}$  and style feature  $f_s^{G_i}$  to the BP-ATN module that maps the correspondences between the part-features of the same body part.

$$\text{BP-ATN}(\tilde{f}_d^{G_i}, f_s^{G_i}) = [\text{ATN}(\tilde{f}_d^{LL_i}, f_s^{LL_i}), \dots, \text{ATN}(\tilde{f}_d^{RA_i}, f_s^{RA_i})]. \quad (11)$$

Figure 7 illustrates the ATN module. It channel-wise normalizes the given style feature  $f_s^{P_i}$  and the decoded content feature  $\tilde{f}_d^{P_i}$  to make  $\overline{f_s^{P_i}}$  and  $\overline{\tilde{f}_d^{P_i}}$ , and maps them into new feature spaces  $m$  and  $n$  by convolution layers. Then we reshape the mapped features to

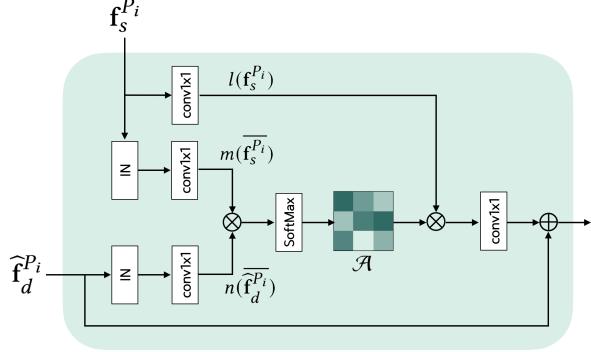


Fig. 7. Applying BP-ATN module by body part  $P_i$  of  $G_i$  level. The BP-ATN module transfers the locally semantic style features of  $f_s^{P_i}$  via constructing an attention map with  $\tilde{f}_d^{P_i}$ .

$m(\overline{f_s^{P_i}}) \in \mathbb{R}^{C_i \times |V_i|_s}$  and  $n(\overline{\tilde{f}_d^{P_i}}) \in \mathbb{R}^{|V_i|_d \times C_i}$ , where  $|V_i|$  is vertex cardinality of part-features and  $C_i$  is the feature dimension of  $G_i$ . After that, we perform a matrix multiplication between  $m$  and  $n$ , and apply a softmax layer to construct an attention map  $\mathcal{A} \in \mathbb{R}^{|V_i|_s \times |V_i|_d}$ :

$$\mathcal{A}_{\beta,\alpha} = \frac{\exp(m(\overline{f_s^{P_i}})_\alpha \cdot n(\overline{\tilde{f}_d^{P_i}})_\beta)}{\sum_{\alpha=1}^{|V_i|_s} \exp(m(\overline{f_s^{P_i}})_\alpha \cdot n(\overline{\tilde{f}_d^{P_i}})_\beta)}, \quad (12)$$

where  $\mathcal{A}_{\beta,\alpha}$  measures the similarity between the  $\alpha$ -th vertex of the style feature and the  $\beta$ -th vertex of the decoded feature. The more similar feature representation of the two vertices signifies the greater spatial-temporal correlation between them. Meanwhile, we feed  $f_s^{P_i}$  to another convolution layer and reshape it to generate a new feature,  $l(f_s^{P_i}) \in \mathbb{R}^{|V_i|_s \times C_i}$ . Then, we perform a matrix multiplication between  $\mathcal{A}$  and  $l(f_s^{P_i})$  to adjust the attention map to the style feature. After that, we apply a convolution layer to the result and add to  $\tilde{f}_d^{P_i}$  element-wise:

$$\text{ATN}(\tilde{f}_d^{P_i}, f_s^{P_i}) = \mathcal{A}^T \otimes l(f_s^{P_i}) \oplus \tilde{f}_d^{P_i}, \quad (13)$$

where  $\otimes$  denotes matrix multiplication and  $\oplus$  element-wise addition. Finally, we feed the output feature into STGCN2 layer and generate the second step of the stylized feature  $\tilde{f}_d^{G_i}$ :

$$\tilde{f}_d^{G_i} = \text{STGCN2}(\text{BP-ATN}(\tilde{f}_d^{G_i}, f_s^{G_i})). \quad (14)$$

Implementation details for the whole architecture are provided in Appendix A.1.

## 5 TRAINING THE MOTION PUZZLE FRAMEWORK

Given a source motion  $M_{\text{src}} \in \mathcal{M}$  and a target motion  $M_{\text{tar}} \in \mathcal{M}$ , we train the entire networks end-to-end by using the following techniques and loss terms.

**Mixing style code by body part.** A naive way of preparing the training data for our per-body-part style transfer would be to construct a batch of motions by combining one source and many target motions, but it requires too much memory. To avoid this, we devise a per-body-part style mixing scheme following the idea of [Karras et al. 2019] to achieve a similar effect with only one source and one

**ALGORITHM 1:** Style mixing by body part.

---

**Input:** Style feature set  $\mathbf{f}_s^{G_i}$  from source motion and  $\mathbf{f}_s^{G_i'}$  from target motion

**Output:** Mixed style feature set  $\mathbf{f}_{\text{mix}}^{G_i}$

/\*  $\gamma$ : predefined probability of style mixing \*/

if  $\text{Rand}([0, 1]) < \gamma$  then

    /\* Number of parts to apply  $\mathbf{f}_s^{G_i}$  \*/

$n_{\text{switch}} = \text{RandInt}([1, 5])$

    /\* Randomly selected  $n_{\text{switch}}$  number of body parts \*/

$\text{idx}_{\text{switch}} = \text{Sample}(\text{idx}_{\text{all}}, n_{\text{switch}})$

    /\* Mix style codes \*/

$\mathbf{f}_{\text{mix}}^{G_i}[\text{idx}_{\text{switch}}] = \mathbf{f}_s^{G_i}$

$\mathbf{f}_{\text{mix}}^{G_i}[\text{idx}_{\text{all}} - \text{idx}_{\text{switch}}] = \mathbf{f}_s^{G_i'}$

else

    /\* Only use style code  $\mathbf{f}_s^{G_i'}$  \*/

$\mathbf{f}_{\text{mix}}^{G_i}[\text{idx}_{\text{all}}] = \mathbf{f}_s^{G_i'}$

end

---

target motion. The idea is, instead of explicitly preparing target style features, a style feature  $\mathbf{f}_s^{G_i'}$  from a target motion and a style feature  $\mathbf{f}_s^{G_i}$  from a source motion are mixed with a predetermined probability to apply the target and source style codes to some randomly selected body parts, as implemented in Algorithm 1.

**Motion reconstruction.** By using the same motion for both the source and target motions, the reconstruction loss enforces the translator to generate output motions identical to the input motions.

$$\begin{aligned} \mathcal{L}_{\text{rec}} &= \mathbb{E}_{M_{\text{src}}} [\|\tilde{M}_{\text{src}}^{\text{rec}} - M_{\text{src}}\|_1] + \mathbb{E}_{M_{\text{tar}}} [\|\tilde{M}_{\text{tar}}^{\text{rec}} - M_{\text{tar}}\|_1], \\ \tilde{M}_{\text{src}}^{\text{rec}} &= D(\mathbf{f}_c^{G_3}, [\mathbf{f}_s^{G_i}]) \text{ (reconstructed source motion)}, \\ \tilde{M}_{\text{tar}}^{\text{rec}} &= D(\mathbf{f}_c^{G_3'}, [\mathbf{f}_s^{G_i'}]) \text{ (reconstructed target motion)}, \end{aligned} \quad (15)$$

where  $\mathbf{f}_c^{G_3} = E_c(M_{\text{src}})$ ,  $[\mathbf{f}_s^{G_i}] = E_s(M_{\text{src}})$  while  $\mathbf{f}_c^{G_3'}$  and  $[\mathbf{f}_s^{G_i'}]$  are from the target motion ( $[\cdot]_{i=1}^3$  is written as  $[\cdot]$  for brevity). This loss term helps our encoder-decoder framework achieve the identity map.

**Preserving content and style features.** To guarantee that the translated motion preserves the style-invariant characteristics of its input source motion  $M_{\text{src}}$  and also preserves the content-invariant characteristics of its input target motion  $M_{\text{tar}}$ , we employ the cycle consistency loss [Choi et al. 2020; Lee et al. 2018; Zhu et al. 2017].

$$\begin{aligned} \mathcal{L}_{\text{cyc}} &= \mathbb{E}_{M_{\text{src}}, M_{\text{tar}}} [\|\tilde{M}_{\text{src}}^{\text{cyc}} - M_{\text{src}}\|_1 + \|\tilde{M}_{\text{tar}}^{\text{cyc}} - M_{\text{tar}}\|_1], \\ \tilde{M}_{\text{src}}^{\text{cyc}} &= D(E_c(D(\mathbf{f}_c^{G_3}, [\mathbf{f}_{\text{mix}}^{G_i'}])), [\mathbf{f}_s^{G_i'}]), \\ \tilde{M}_{\text{tar}}^{\text{cyc}} &= D(\mathbf{f}_c^{G_3'}, E_s(D(\mathbf{f}_c^{G_3}, [\mathbf{f}_s^{G_i'}]))), \end{aligned} \quad (16)$$

where  $[\mathbf{f}_{\text{mix}}^{G_i}]$  is a mixed style feature set from  $[\mathbf{f}_s^{G_i}]$  and  $[\mathbf{f}_s^{G_i'}]$  (Alg. 1).

**Preserving root motion.** We encourage the output motion to preserve the linear and angular velocities of the root of the source motion  $M_{\text{src}}$  because the root motion plays a critical role in motion content. The preservation of the root motion gives an additional benefit of reducing foot sliding and improving temporal coherence

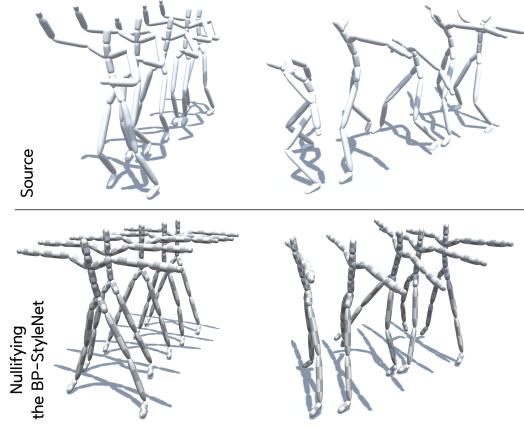


Fig. 8. Results of output motions (bottom) only with content features from source motion (top), i.e., without applying style features.

of output motion. To this end, we employ a root loss as follows:

$$\mathcal{L}_{\text{root}} = \mathbb{E}_{M_{\text{src}}, M_{\text{tar}}} [\|RV(D(\mathbf{f}_c^{G_3}, [\mathbf{f}_{\text{mix}}^{G_i'}])) - RV(M_{\text{src}})\|_1], \quad (17)$$

where  $RV(M)$  extracts the root velocity terms  $[\dot{\mathbf{r}}^x, \dot{\mathbf{r}}^z, \dot{\mathbf{r}}^a]$  from a motion.

**Smoothness.** To ensure that the output pose changes smoothly in time, we add a smoothness term between temporally adjacent pose features.

$$\mathcal{L}_{\text{sm}}(\tilde{M}, M) = \mathbb{E}_M [\|V(\tilde{M}) - V(M)\|_1] \quad (18)$$

$$\begin{aligned} \mathcal{L}_{\text{sm}} &= \mathcal{L}_{\text{sm}}(\tilde{M}_{\text{src}}^{\text{rec}}, M_{\text{src}}) + \mathcal{L}_{\text{sm}}(\tilde{M}_{\text{tar}}^{\text{rec}}, M_{\text{tar}}) \\ &\quad + \mathcal{L}_{\text{sm}}(\tilde{M}_{\text{src}}^{\text{cyc}}, M_{\text{src}}) + \mathcal{L}_{\text{sm}}(\tilde{M}_{\text{tar}}^{\text{cyc}}, M_{\text{tar}}) \end{aligned} \quad (19)$$

where  $V(M)$  is the sequence of pose feature differences between adjacent frames.

The total objective function of the Motion Puzzle framework is thus:

$$\min_{E_c, E_s, D} \mathcal{L}_{\text{rec}} + \lambda_{\text{cyc}} \mathcal{L}_{\text{cyc}} + \lambda_{\text{root}} \mathcal{L}_{\text{root}} + \lambda_{\text{sm}} \mathcal{L}_{\text{sm}}, \quad (20)$$

where  $\lambda_{\text{cyc}}$ ,  $\lambda_{\text{root}}$  and  $\lambda_{\text{sm}}$  are hyperparameters for each loss term. Please see training details in Appendix A.2.

## 6 DISCUSSION ON CONTENT FEATURE AND ATTENTION MAP

This section examines what information is stored by the content feature and how the attention map in BP-ATN represents the correlation between the style feature and the decoded content feature by visualizing them.

Figure 8 visualizes the content feature generated from our model. Figure 8 shows the output motions generated only by the content features obtained from a dancing motion (bottom) and a Pterosauria motion (top). These motions were generated by nullifying the BP-StyleNet by setting the scale and bias of BP-AdaIN to 1 and 0, and the attention map of BP-ATN to 0. One can see that the content feature retains the overall walking phase of the source motion while filtering out its style variation of all body parts. This shows that our content encoder mainly extracts the phase information of each body

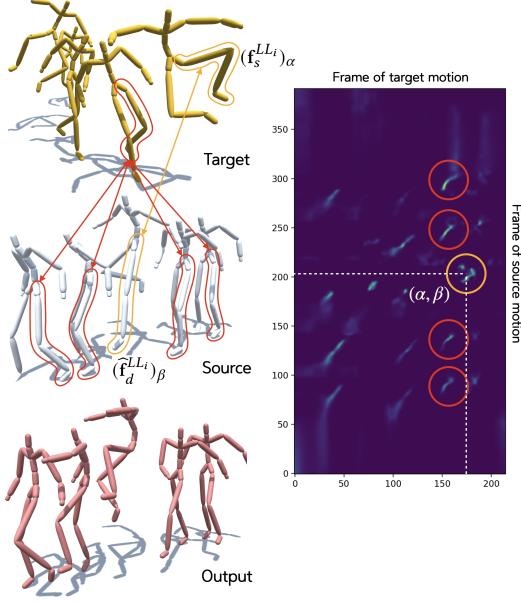


Fig. 9. Visualization of the left leg attention map between the decoded content feature from the source motion and the style feature from the target motion (right), and corresponding poses to the marked high attention regions (left).

part from a source motion while other characteristic information of the motion is stored in the style feature.

Next, we visualize attention map  $\mathcal{A}$  (Eq. 12) that represents the correlation between the style feature  $f_s^{P_i}$  and the decoded content feature  $\hat{f}_d^{P_i}$ . Figure 9 (right) shows the visualization of the attention map of the left leg between a source (jumping while walking) and target motions (repeated swing jumping). To examine the temporal correlation between features in the attention map, the spatial dimension of each feature is averaged out, i.e.,  $\mathcal{A} \in \mathbb{R}^{T_s \times T_d}$  where  $X$  and  $Y$  axes denote  $T_s$  and  $T_d$ . In the attention map visualization, the higher the attention or correlation, the brighter the color. We identify corresponding poses in the source and target motions to some regions with a high attention value. The yellow circle corresponds to the left leg pose in the middle of jumping in both the source and target motions (yellow borders), which shows that the jumping poses of the source and target motions are found to have a high correlation, promoting the transfer of jumping style of the target to the jumping motion of the source. In addition, the four red circles correspond to a landing pose in the target motion in a high correlation with four landing poses in the source motion (red borders). This indicates that periodic poses in a motion can be matched to a single pose in the other motion, and in this case, the landing motion style of the target will be transferred to all landing motions of the source.

The output motion shows that the swing jumping style and the landing motion style from the target are well transferred to the jumping motion and the periodic landing motions in the source. This result suggests that the attention map creates correspondences between similar actions in the source and target motions, driving style transfer between these matched actions. For instance, if a

target motion contains walking, jumping, and kicking motions and a source motion contains only a jumping motion. The style transfer would occur from only the jumping motion of the target to the source motion. By finding the correlation between the content and style features, it significantly contributes to successful motion style transfer, especially capturing time-varying motion style well.

## 7 EXPERIMENTS

We conduct various experiments to present interesting features and advantages of our method and evaluate its performance. First, we examine a unique capability of our method, transferring style by body part, with respect to content preservation and per-part stylization. Second, we perform a comparison with other previous methods [Aberman et al. 2020b; Holden et al. 2016] and variations of our model with BP-ATN or BP-StyleNet ablated. Lastly, we showcase a real-time motion style transfer combined with an existing character controller.

Motion clips in the test dataset and the new unseen dataset used in the experiment have variable frame lengths, as allowed by our convolutional network-based architecture. All results are presented after post-processing by the same method in [Aberman et al. 2020b] including foot-sliding removal. In the figures, the source, target, and resulting motion are shown in white, yellow, and red skeleton, respectively. We used AI4Animation framework [Starke 2021] for visualizing results. The supplemental result video shows the resulting motions from the experiments.

### 7.1 Motion Style Transfer by Body Part

First, we show several examples in Fig. 1 to demonstrate the utility of the per-part style transfer approach. Suppose that a zombie walking motion is given, and we want to modify it to make a staggering zombie walking motion. One way would be to add a staggering motion to the trunk and leg while keeping the style of waddling with stretched arm in the source motion. Our method realizes this by transferring the style of a staggering motion to only the leg and spine while keeping the style of the given zombie motion for all other body parts as shown in Fig. 1 (the 2nd column). In contrast, previous whole-body style transfer methods may replace the zombie style with a staggering style over the whole body. Figure 1 (the 1st column) shows a case that adds a flapping style to a dance motion by transferring the style of the arm motions of the target to the corresponding parts in the source motion. Figure 1 (the 3rd and 4th columns) shows the cases of changing the style of parts of a body and the whole body, respectively.

**Separate transfer to three body parts.** Figure 10 shows the results of transferring arbitrary unseen styles from 3 target motions, each for a different body part, to a single source motion. For convenience, each sub-figure is referred to as  $(i, j)$ , meaning the  $j$ -th image in the  $i$ -th row.

First, comparing the source motions and the output motions shows that the contents of the source motions are well preserved in the output motions. Although all target motions stylizing the leg, arm and spine are clearly different from the source motion, the output motions preserve the phases of walking (1, 5), (3, 4), and

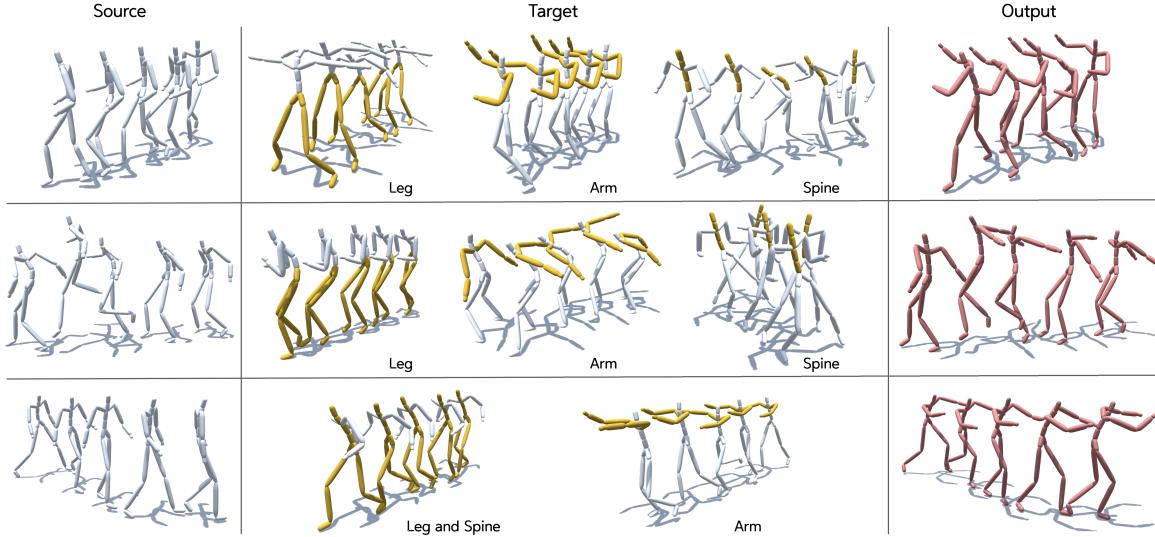


Fig. 10. Results of our per-body-part motion style transfer. For each row, two or three separate target motions (middle) for the leg, arm, and spine styles are applied to the source motion (left, gray) to make the output motion (right, red). Note that the style code for each target motion is extracted from the target parts marked in yellow.

jumping (2, 5) motions. This result suggests that our architecture disentangles content and style well.

In the output motions, each body part shows a realistic motion reflecting the style of its corresponding target motion despite the differences in style and content between target motions. For example, the output motion of (2, 5) takes the bent leg style of (2, 2), the stretched arm style of (2, 3) like Pterosaur, and the spine style (2, 4) of proud dancing while maintaining the jumping content of the source (2, 1) motion. Also, the output motion of (3, 4) takes the old spine and leg style of (3, 1) and the childlike arm style of (3, 3) while maintaining the walking content of source (3, 1) motion. This suggests that our per-body-part style transfer is localized well. The locality is analyzed further next.

**Locality.** An essential requirement of our motion style transfer using *BP-StyleNet* is that its effect is local to each body part while maintaining the content of the source motion. To validate this, we conduct the following experiment. We transfer a velociraptor style from a target motion to only one part among the legs, arms and spine of the source motion while keeping the original style for the remaining body parts (Fig. 11). If our style transfer is well localized, when the target style is transferred only to one part, other body parts should keep their original style with only the affected part reflecting the velociraptor style. Figure 11 shows that our Motion Puzzle framework achieves this.

To quantitatively evaluate the locality, we measure the mean squared displacement (MSD) for each joint between the source motion and four output motions, in which three have only one part stylized (arm, leg and spine) and one has all parts stylized (all) for the experiments in Fig. 11. The more an output motion deviates from the source motion, the larger MSD is. Table 1 shows that only the style transferred joints have large MSDs, quantitatively confirming the locality of our method. The table shows that the shoulder joints

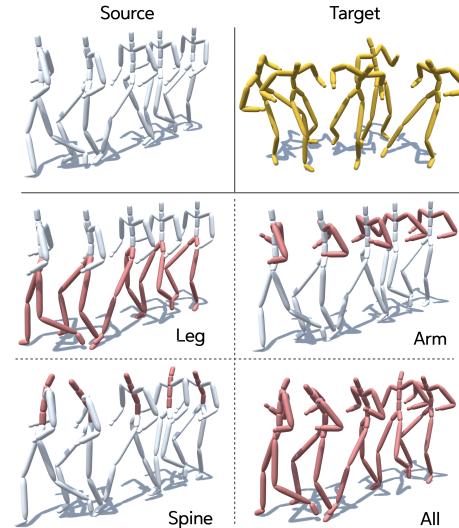


Fig. 11. Locality test. The target style (yellow) is applied only to one body part (red) while keeping the original style with the remaining body parts. The bottom right figure shows the result when the target style is applied to all body parts.

(13 and 17) are mildly affected ( $0.1 \geq MSD > 0.05$ ) by the spine style transfer, a natural result obtained by the graph convolution across parts.

**Per-part style interpolation.** This experiment tests whether our style transfer can be interpolated intuitively between two different style motions. Specifically, two style features  $[f_s^{G_i}]_{i=1}^3$  and  $[f_s^{G_i'}]_{i=1}^3$  from two different target motions are interpolated in the

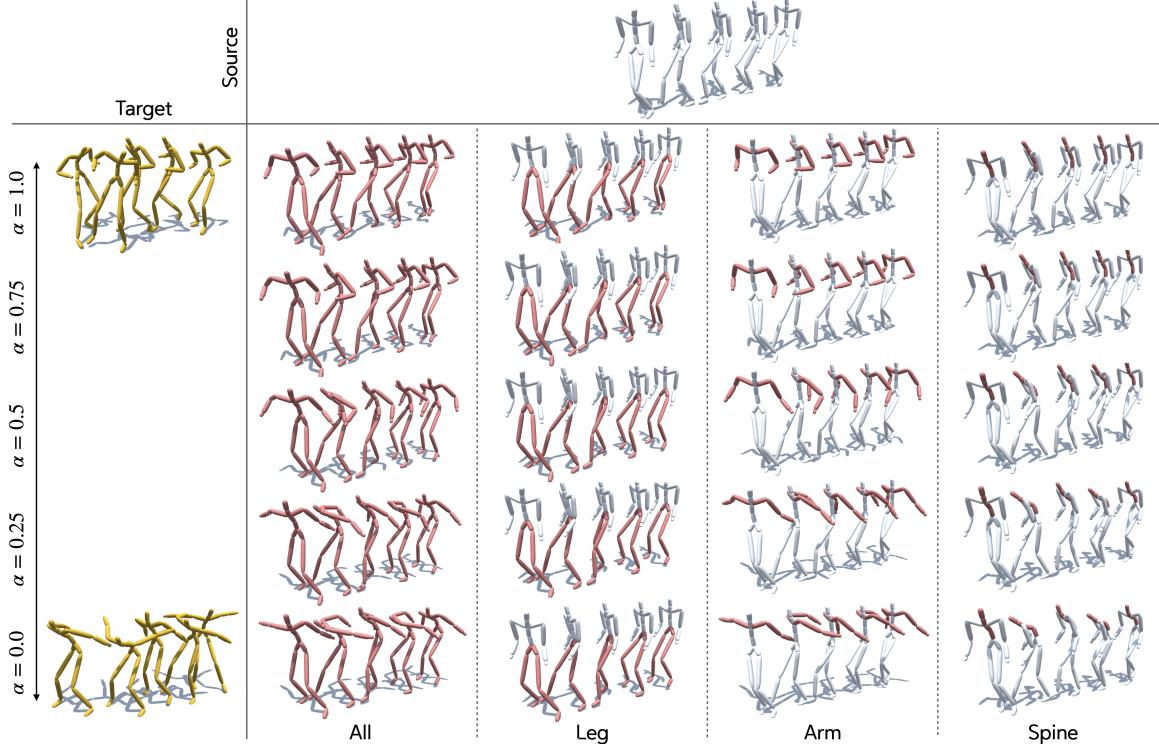


Fig. 12. Results of style interpolation by part. Interpolation parameter  $\alpha$  of the latent style space varies from 0 (Pteranodon style) to 1 (Velociraptor style).

MSD	Leg	Arm	Spine
>0.1	3, 4, 7, 8	15, 16, 19, 20	12
>0.05	2, 3, 4, 6, 7, 8	14, 15, 16, 18, 19, 20	11, 12, 13, 17

Table 1. The indices of joints with mean squared distance (MSD) of 0.1 or more and 0.05 or more when only one part (leg, arm, or spine) is style transferred for the experiment in Fig. 11. The displacement is measured with respect to the global positions of each joint between source and output motions. Length is normalized by dividing by the skeleton height. Joint indices are the leg (1-8), arm (13-20), and spine (0, 9-12).

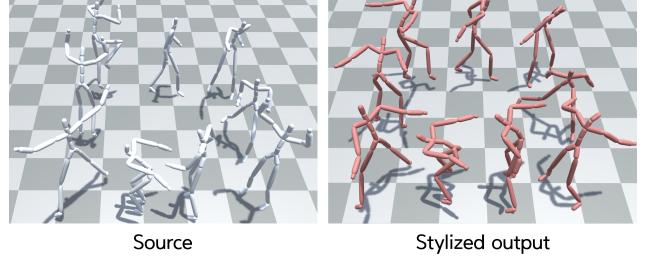


Fig. 13. Transferring a Pteranodon style to a long-term heterogeneous source motion containing various behaviors.

continuous latent style feature space and then applied to the target body part. Figure 12 shows the result of transferring interpolated style codes  $(1 - \alpha)[\mathbf{f}_s^{G_i}]_{i=1}^3 + \alpha[\mathbf{f}_s^{G_i'}]_{i=1}^3$  from  $\alpha = 0$  to  $\alpha = 1$  to each body part. Since our style latent space is not normalized, linear interpolation was used rather than spherical linear interpolation. The experiment shows that the interpolation works well while maintaining locality.

**Long-term heterogeneous motion.** We test the effectiveness of our framework for long-term heterogeneous motions involving a variety of behaviors. Figure 13 shows the whole body style transfer results of a Pteranodon style to a neutral source motion over 3600 frames (1 min), including walking, jumping, transition, kicking and punching. We found that our method can be applied robustly to

long motions with multiple behaviors. Additional results of long-term heterogeneous motion with various styles are provided in the supplementary video.

## 7.2 Ablation Study and Comparison with Prior Work

**7.2.1 Qualitative evaluation.** We conduct an ablation study to examine the effect of BP-StyleNet and BP-ATN modules. In addition, we compare our framework with those in previous work [Aberman et al. 2020b; Holden et al. 2016]. Since the original framework of [Aberman et al. 2020b] is effective for style-labeled motion data, we instead construct a comparable framework (dubbed Conv1D+AdaIN) that takes 1D convolution and AdaIN components from [Aberman

et al. 2020b], and train it using the same losses as ours to enable arbitrary style transfer.

Figure 17 shows the comparison of motion style transfer results with our method (a) and its two variations (b and c), Conv1D+AdaIN (d), and a 1D convolution model with Gram matrix as proposed by [Holden et al. 2016] (e). For an ablation study, “w/o BP-ATN” (b) is made by removing BP-ATN module from our framework, only using BP-AdaIN to examine the effect of BP-ATN module, and STGCN+AdaIN (c) is made by removing BP-StyleNet and performing AdaIN over the whole body, not by body part, to compare BP-AdaIN and AdaIN. Please refer to the supplemental result video to clearly view the differences between the resulting motions.

*Effect of BP-ATN.* The results of “w/o BP-ATN” (b) in Fig. 17 preserve the content of the source motion well but reflect the style of the target motion less than our results. In the second and fourth rows, through the attention map in BP-ATN, jumping parts in the target and source motions are matched to transfer style between corresponding motions. The output motions (a) in the first and third rows also show the advantage of BP-ATN. In the first row, the target motion includes the style of a flapping bent arm, but the output of (b) shows only bent arms that do not flap. Similarly, in the third row, the target motion includes tripping while walking, but the output motion of (b) does not exhibit the tripping style at all and only preserves the content of stair climbing. Since AdaIN simply modifies the mean and variance of the content features to translate the source motion, it captures temporally global features (e.g., bending arms) well but loses temporally local features (e.g., flapping motion). This comparison demonstrates the importance of BP-ATN in transferring local features.

*BP-AdaIN vs. AdaIN.* We found that the style of the target motion is not transferred except for the spine part if AdaIN is used (c) instead of BP-AdaIN for STGCN-based networks. This is because, in the process of transferring style, individual motion styles of each body part are not separated but instead averaged out over the whole body. The overall extent of style transfer is even poorer than Conv1D+AdaIN, a topology-agnostic approach. This result shows that the BP-AdaIN should be used instead of AdaIN for high-quality style transfer for graph-based networks.

As Conv1D+AdaIN (d) uses only AdaIN for style transfer, it reflects the global style features well but may not express temporally local features. In addition, the output motions sometimes jiggle and are distorted, which we attribute to the topology-agnostic 1D convolution scheme. In contrast, the STGCN-based models, (b) and (c), consider human topology, which helps generate plausible motions without motion distortion.

Results of (e) show that the degree of content preservation and the plausibility of motion is lower than the compared models. Since the content and style features are extracted from the same deep features, leading to a dependency between content and style, this approach has a limitation in transferring style between motions with different contents.

**7.2.2 Quantitative evaluation.** We quantitatively measure the degree of generation quality, content preservation, and style reflection of three generative models: [Holden et al. 2016], Conv1D+AdaIN

Table 2. Quantitative evaluation on Xia dataset. We calculate FMD, CRA, and SRA on 2500 stylized samples generated by each method on each trial. The table reports the mean ( $\pm$  standard deviation) values for each metric over 10 trials.

Methods	FMD $\downarrow$	CRA $\uparrow$ (%)	SRA $\uparrow$ (%)
Real motions ( $\mathcal{M}_{\text{gen}}$ )		96.04	90.24
Ours	<b>9.38 <math>\pm</math> 1.10</b>	29.83 $\pm$ 1.35	<b>54.94 <math>\pm</math> 2.09</b>
Conv1D+AdaIN	24.22 $\pm$ 2.29	29.09 $\pm$ 1.29	41.97 $\pm$ 2.01
[Holden et al. 2016]	29.40 $\pm$ 2.34	<b>38.93 <math>\pm</math> 2.09</b>	41.92 $\pm$ 1.77

from [Aberman et al. 2020b], and ours. Specifically, we use three metrics; Fréchet Motion Distance (FMD), content recognition accuracy (CRA), and style recognition accuracy (SRA). We compute FMD, CRA, and SRA on the stylized set with all possible combinations of source (content) and target (style) motions generated by each model.

The FMD follows the approach of the widely used Fréchet Inception Distance (FID) [Heusel et al. 2017], which is the distance between feature vectors calculated from real and generated images, extracted from Inception v3 network [Szegedy et al. 2016] trained on ImageNet [Russakovsky et al. 2015]. It is used to evaluate the quality and diversity of image generative models. The FMD measures distance between feature vectors for motion. To this end, we train a content classifier using the method of [Yan et al. 2018] and use the feature vector obtained from the final pooling layer to measure the FMD between the real and generated motions. A lower FMD suggests higher generation quality.

We use the same content classifier to measure the CRA on generated motions since a model with a higher degree of content preservation would generate more correctly classified motions. The higher CRA indicates better performance with respect to content preservation. Similarly, we measure the SRA on stylized sets using another classifier trained to predict the style label of motion. A higher SRA means better performance for style reflection.

We test generative models using the Xia dataset  $\mathcal{M}_{\text{test}}$  [Xia et al. 2015], which is unseen to the models during training. We divide  $\mathcal{M}_{\text{test}}$  into two halves,  $\mathcal{M}_{\text{cls}}$  for classification and  $\mathcal{M}_{\text{gen}}$  for generation. We pre-train the action classifiers with  $\mathcal{M}_{\text{cls}}$  on six content types (*walking*, *running*, *jumping*, *kicking*, *punching*, *transitions*) and on eight style types (*angry*, *childlike*, *depressed*, *neutral*, *old*, *proud*, *sexy*, *strutting*) for content and style classification, respectively. We generate a stylized set using  $\mathcal{M}_{\text{gen}}$  with each generative model.  $\mathcal{M}_{\text{gen}}$  is additionally used as a validation set for the action classifiers and a reference distribution for FMD calculation. The pre-trained content and style classifier show 96.04% and 90.24% accuracies on  $\mathcal{M}_{\text{gen}}$ , respectively, and each stylized set is evaluated by these classifiers.

Table 2 shows the result. Reasonably, there is a trade-off between the content and style classification. Although the method of [Holden et al. 2016] shows the highest CRA, the style classification is the lowest. Conv1D+AdaIN gives a slightly improved SRA with a lower CRA. Our method increases the SRA performance by a large margin while maintaining a higher CRA than Conv1D+AdaIN. Further, our

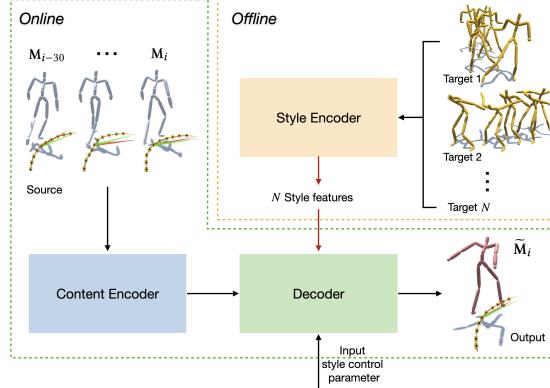


Fig. 14. Overview of a real-time motion stylizer.

method outperforms the others in terms of FMD. This result suggests that our method produces higher-quality stylized motions with a greater degree of style reflection while not sacrificing the content preservation performance too much. The reason why our method shows strong SRA compared with CRA is presumably due to the nature of the content feature. As the content feature mainly includes the phase information while other characteristic aspects are stored in the style feature (Sec. 6), output motions are strongly affected by the style feature, which may lead to high SRA in comparison with CRA.

**7.2.3 Effect of post-processing.** In the case of using the content feature and style feature from the same motion, reconstructed motion is sufficiently plausible without post-processing. However, combining features from different motions may create foot sliding in the stylized output motion. To remedy this, we compute foot contact labels from the source motion and then use them to correct the output foot positions to maintain the contact phase with inverse kinematics. The target foot position is set as the average foot position in the contact phase. The supplementary video shows the comparison between output motions with and without post-processing.

### 7.3 Real-time Motion Style Transfer

In this experiment, we integrate our Motion Puzzle framework with a state-of-the-art motion controller, phase-functioned neural networks (PFNN) [Holden et al. 2017b], to demonstrate real-time motion style transfer. Our real-time system provides style control parameters to select the target body part, target style, and the degree of style reflection. To achieve real-time performance, we reduced the required memory size by simplifying the networks while keeping the core structure. For the detailed structure, please refer to Appendix A.1. Figure 14 shows the overview of our real-time system to generate the  $i$ -th frame of the stylized output pose. In the offline process, our style encoder extracts  $N$  ( $= 4$  in our experiment) style features from  $N$  target motions and store them. In the online process, the content encoder extracts content features from the source motion of the duration  $[i - 30, i]$  ( $= 0.5\text{sec}$ ) generated by PFNN. When the user selects a target style and target body part, the system

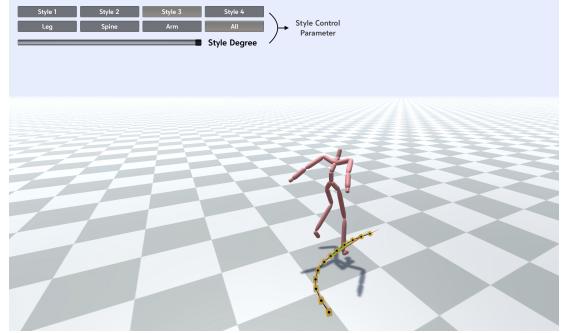


Fig. 15. Snapshot of a real-time motion stylizer combined with a PFNN motion controller.

applies the selected style feature to the content feature to generate the  $i$ -th frame pose with the translated style. Post-processing is not conducted for this experiment as the contact information of the source motion is not available. This real-time motion-style controller runs in 30 FPS. Figure 15 shows a snapshot of the system.

## 8 LIMITATIONS AND FUTURE WORK

Our framework has limitations that need to be overcome by future research. First, our framework does not consider contact or interaction between the human body and the environment, so the motion content related to contact is not well preserved. For example, Fig. 16 (top) shows a failure case that the content of crawling in the source motion is not preserved when a zombie walking motion is given as a style target. As a future step, we will improve the architecture to include the contact information as an important content feature.

Second, our method is trained to preserve the root motion of the source to the output, which may degrade style transfer quality if the root motion of the target contains important style characteristics, such as dancing. One way to improve this would be to develop an additional module for creating the root velocity to reflect the style of the target motion.

Another limitation is that our framework may not deal with rapidly changing, dynamic motion because it is challenging to identify correspondences between abruptly changing motions within long sequences of the content and target motions. Figure 16 (bottom) shows such a case that the output motion fails to preserve the content of rapid rotation in the source motion. A possible solution would be to segment the long motions into sub-sequences and apply the style transfer between the sub-sequences with similar content.

Since the style of each body part is controlled independently, the resulting whole-body motion may not look coordinated or physically plausible if very different styles are applied over the body parts. An important future work to address this problem would be to add a final step of adjusting the part-stylized motion to improve the naturalness of the motion from holistic viewpoint.

Our framework assumes a single motion as content. We observed that sometimes strong preservation of motion content degrades the effect of stylizing some part, which may be improved if the content component can also be edited. In general, the applicability of a motion editing method will be enhanced if both content and

style can be concurrently edited. Thus, an important future research direction would be to control both components by part. A recent technique [Starke et al. 2021] that allows per-part motion editing can be a good reference for this direction.

In this paper, we dealt with the motion style of a single person. Solving the problem of stylizing the motion of a group of people, such as collaboration or competition, is another interesting direction for future research.

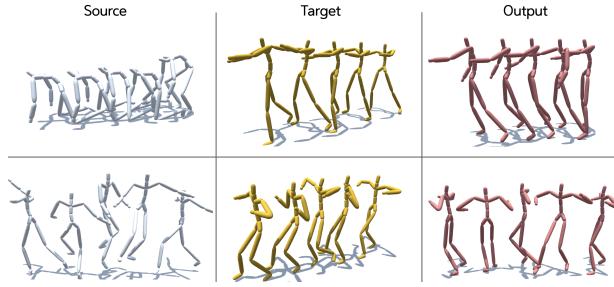


Fig. 16. Failure cases. Top: Transferring a zombie style to a crawling source motion breaks the contact between the hand and the ground. Bottom: Rapid rotation in a source motion is collapsed.

## 9 CONCLUSION

In this paper, we presented the Motion Puzzle framework for motion style transfer. It was carefully designed to respect the kinematic structure of the human skeleton, and thus it can control the motion style of individual body parts while preserving the content of the source motion. At the core of the Motion Puzzle framework is our style transfer network, BP-StyleNet, which applies both the global and local traits of target motion style to the source motion. As a result, our framework can deal with temporally varying motion styles, which was impossible in previous studies. In addition, our model can be easily integrated with motion generation frameworks, allowing a wide variety of applications, including real-time motion transfer.

## REFERENCES

- Kfir Aberman, Peizh Uo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. 2020a. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 62–1.
- Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. 2020b. Unpaired motion style transfer from video to animation. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 64–1.
- Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. 2020. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8188–8197.
- CMU. "Carnegie-Mellon Mocap Database". <http://mocap.cs.cmu.edu>
- Han Du, Erik Herrmann, Janis Sprenger, Klaus Fischer, and Philipp Slusallek. 2019. Stylistic locomotion modeling and synthesis using variational generative models. In *Motion, Interaction and Games*. 1–10.
- Jun Fu, Jing Liu, Haijin Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. 2019. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3146–3154.
- 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*. 2414–2423.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).
- Daniel Holden, Ikhlasul Habibie, Ikuo Kusajima, and Taku Komura. 2017a. Fast neural style transfer for motion data. *IEEE computer graphics and applications* 37, 4 (2017), 42–49.
- Daniel Holden, Taku Komura, and Jun Saito. 2017b. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–13.
- Daniel Holden, Jun Saito, and Taku Komura. 2016. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–11.
- Eugene Hsu, Kari Pulli, and Jovan Popović. 2005. Style translation for human motion. In *ACM SIGGRAPH 2005 Papers*. 1082–1089.
- Linjiang Huang, Yan Huang, Wanli Ouyang, and Liang Wang. 2020. Part-level graph convolutional network for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11045–11052.
- Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*. 1501–1510.
- Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*. Springer, 694–711.
- Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations*.
- Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4401–4410.
- Hyue Ji Kim and Sung-Hee Lee. 2019. Perceptual Characteristics by Motion Style Category. In *Eurographics 2019 - Short Papers*. Paolo Cignoni and Eder Miguel (Eds.). The Eurographics Association. <https://doi.org/10.2312/egs.20191000>
- Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. 2018. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*. 35–51.
- Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. 2019. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3595–3603.
- Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017. Universal style transfer via feature transforms. *arXiv preprint arXiv:1705.08086* (2017).
- Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2019b. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265* (2019).
- Ming-Yu Liu, Xun Huang, Arun Mallay, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. 2019a. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*. 10551–10560.
- Wanli Ma, Shiheng Xia, Jessica K Hodges, 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*. 21–30.
- Ian Mason, Sebastian Starke, He Zhang, Hakan Bilen, and Taku Komura. 2018. Few-shot Learning of Homogeneous Human Locomotion Styles. In *Computer Graphics Forum*, Vol. 37. Wiley Online Library, 143–153.
- Dae Young Park and Kwang Hee Lee. 2019. Arbitrary style transfer with style-attentional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5880–5888.
- Soomin Park, Deok-kyeong Jang, and Sung-Hee Lee. 2021. Diverse motion stylization for multiple style domains via spatial-temporal graph-based generative model. *Proceedings of the ACM on Computer Graphics and Interactive Techniques (PACMCIGT)* (2021).
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115, 3 (2015), 211–252.
- Yujun Shen, Jinjin Gu, Xiaozou Tang, and Bolei Zhou. 2020. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9243–9252.
- Lu Sheng, Ziyi Lin, Jing Shao, and Xiaogang Wang. 2018. Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8242–8250.
- Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019a. Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7912–7921.
- Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019b. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12026–12035.
- Chenyang Si, Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. 2018. Skeleton-based action recognition with spatial reasoning and temporal stack learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 103–118.

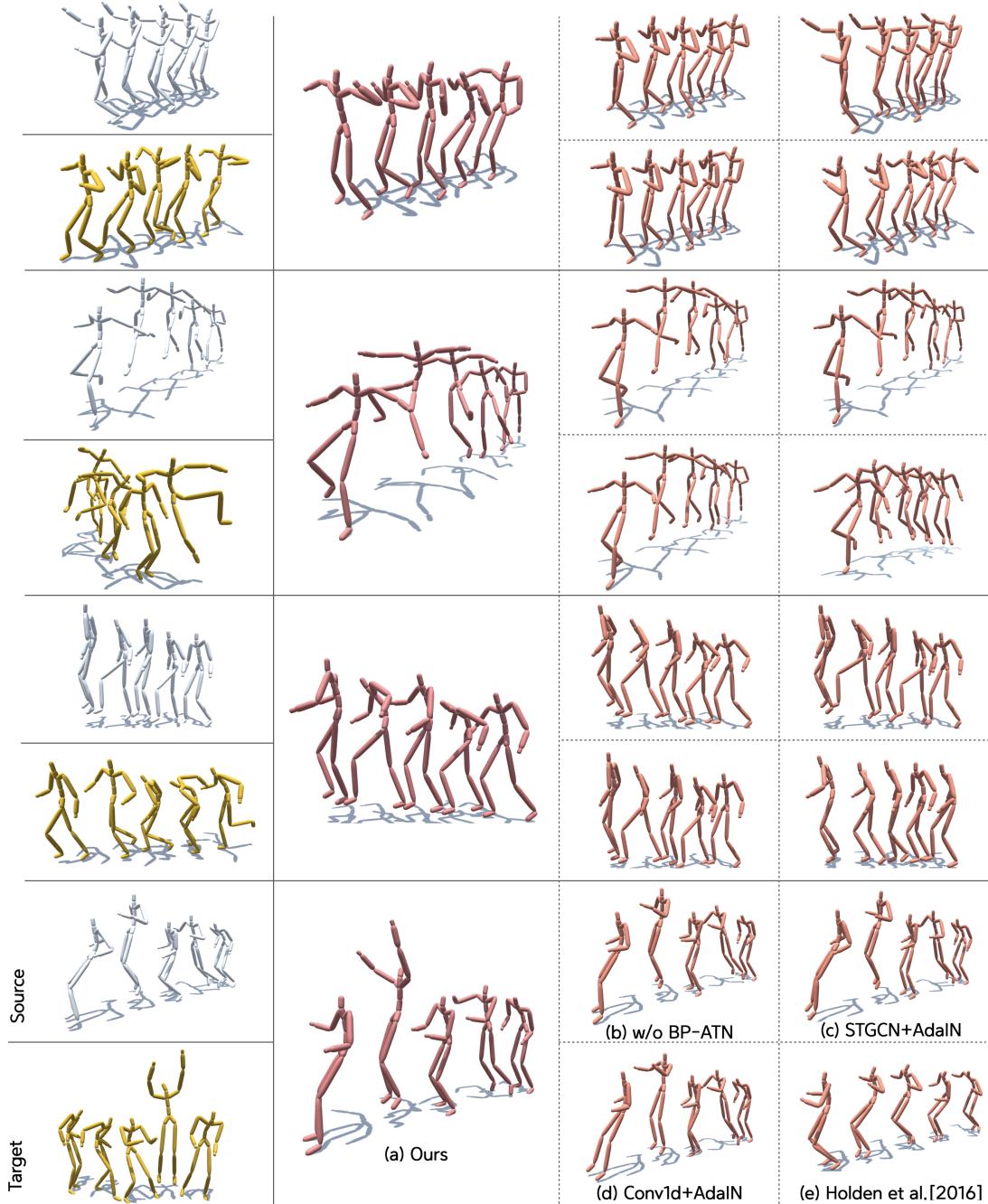


Fig. 17. Comparison results of motion style transfer methods. The style of target motions (yellow) are applied to the source motions (grey) to make output motions (our results in red, compared results in orange).

Harrison Jesse Smith, Chen Cao, Michael Neff, and Yingying Wang. 2019. Efficient neural networks for real-time motion style transfer. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 2, 2 (2019), 1–17.  
 Sebastian Starke. 2021. *AI4Animation framework*. <https://github.com/sebastianstarke/AI4Animation>  
 Sebastian Starke, Yiwei Zhao, Fabio Zinno, and Taku Komura. 2021. Neural Animation Layering for Synthesizing Martial Arts Movements. *ACM Trans. Graph.* 40, 4, Article

92 (jul 2021), 16 pages. <https://doi.org/10.1145/3450626.3459881>  
 Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.  
 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*. 91–96.

- Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7794–7803.
- Yu-Hui Wen, Zhipeng Yang, Hongbo Fu, Lin Gao, Yanan Sun, and Yong-Jin Liu. 2021. Autoregressive Stylized Motion Synthesis With Generative Flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13612–13621.
- Shihong Xia, Congyi Wang, Jinxiang Chai, and Jessica Hodgins. 2015. Realtime style transfer for unlabeled heterogeneous human motion. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 1–10.
- Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. 2019. Convolutional sequence generation for skeleton-based action synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4394–4402.
- Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
- M Ersin Yumer and Niloy J Mitra. 2016. Spectral style transfer for human motion between independent actions. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–8.
- He Zhang, Sebastian Starke, Taku Komura, and Jun Saito. 2018. Mode-adaptive neural networks for quadruped motion control. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–11.
- Liuyang Zhou, Lifeng Shang, Hubert PH Shum, and Howard Leung. 2014. Human motion variation synthesis with multivariate Gaussian processes. *Computer Animation and Virtual Worlds* 25, 3-4 (2014), 301–309.
- Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.

## A APPENDIX: IMPLEMENTATION

### A.1 Network architecture

Table 3 shows the details of the network architecture of Motion Puzzle framework. The size of spatial-temporal convolutional block (STConv) filter is denoted as  $k_s$ (# spatial kernel size) $k_t$ (# temporal kernel size)s(# stride). The neighbor distances for STConv are  $K = (3, 2, 2)$  for  $G_1, G_2$  and  $G_3$ , respectively (Eq. 2). All graph pooling (Gr-Pool) and graph unpooling (Gr-UnPool) are performed before each layers.

The style encoder  $E_s$  consist of 3 STConv with 2 Gr-Pool, followed by a STConv residual block (ResBlk).  $E_s$  extracts style features for each level graph  $G_i$  from target motion  $M_{tar}$  and produces  $[\mathbf{f}_s^{G_i}]_{i=1}^3$ . The output dimension of style features  $\mathbf{f}_s^{G_i}$  are  $T_1 \times J_1 \times C_1 = T \times 21 \times 128$ ,  $T_2 \times J_2 \times C_2 = T/2 \times 10 \times 256$  and  $T_3 \times J_3 \times C_3 = T/4 \times 5 \times 512$ . Here, residual block follows the form used in [Liu et al. 2019a].

The content encoder  $E_c$  receives a source motion as input, passes it through 3 STConv and 2 Gr-Pool layers, followed by a STConv residual block (ResBLK) to generate content feature  $\mathbf{f}_c^{G_3}$ . Every STConv and ResBlk layer has Instance Normalization (IN) for removing style variation. Note that we only extract content feature for  $G_3$  level as mentioned in Sec. 4.3.

The decoder  $D$  has roughly a reverse structure of the encoder. For each level, BP-StyleNet applies a corresponding style feature to the decoded content feature. Details of BP-StyleNet are provided in Table 3(c).

The BP-StyleNet on  $G_i$  level consists of BP-AdaIN, BP-ATN and 2 STConv between them. The BP-StyleNet receives a decoded content feature  $\mathbf{f}_d^{G_i}$  and a style feature  $\mathbf{f}_s^{G_i}$  to generate a translated feature  $\tilde{\mathbf{f}}_d^{G_i}$ . Output shape  $T_i \times J_i \times C_i$  and  $k_s[i]k_t[i]s1$  are noted in each level of decoder part.

For the real-time motion style transfer in Sec. 7.3, we simplified the networks to reduce the memory size. We removed the residual blocks from the encoders and decoder and applied BP-StyleNet only

Layer	Filter	Act.	Norm.	Resample	Output shape		
$M_{tar}$					$T \times 21 \times 15$		
Conv1×1	$k_s 1k_t 1s1$	-	-	-	$T \times 21 \times 64$		
STConv	$k_s 3k_t 7s1$	LReLU	-	-	$T \times 21 \times 128$		
STConv	$k_s 2k_t 5s1$	LReLU	-	Gr-Pool	$T/2 \times 10 \times 256$		
STConv	$k_s 2k_t 5s1$	LReLU	-	Gr-Pool	$T/4 \times 5 \times 512$		
ResBlk	$k_s 2k_t 3s1$	LReLU	-	-	$T/4 \times 5 \times 512$		
$[\mathbf{f}_s^{G_i}]_{i=1}^3$					$[T_i \times J_i \times C_i]_{i=1}^3$		
(a) Style Encoder $E_s$ .							
Layer	Filter	Act.	Norm.	Resample	Output shape		
$M_{src}$	-	-	-	-	$T \times 21 \times 15$		
Conv1×1	$k_s 1k_t 1s1$	-	-	-	$T \times 21 \times 64$		
STConv	$k_s 3k_t 7s1$	LReLU	IN	-	$T \times 21 \times 128$		
STConv	$k_s 2k_t 5s1$	LReLU	IN	Gr-Pool	$T/2 \times 10 \times 256$		
STConv	$k_s 2k_t 5s1$	LReLU	IN	Gr-Pool	$T/4 \times 5 \times 512$		
ResBlk	$k_s 2k_t 3s1$	LReLU	IN	-	$T/4 \times 5 \times 512$		
$\mathbf{f}_c^{G_3}$					$T/4 \times 5 \times 512$		
(b) Content Encoder $E_c$ .							
Layer	Filter	Resample	Output shape				
$\mathbf{f}_c^{G_3}$	-	-	$T/4 \times 5 \times 512$				
BP-ResBlk	$k_s 2k_t 3s1$	-	$T/4 \times 5 \times 512$				
BP-StyleNet	$k_s 2k_t 5s1$	-	$T/2 \times 10 \times 256$				
BP-StyleNet	$k_s 2k_t 5s1$	Gr-UnPool	$T \times 21 \times 128$				
BP-StyleNet	$k_s 3k_t 7s1$	Gr-UnPool	$T \times 21 \times 64$				
Conv1×1	$k_s 1k_t 1s1$	-	$T \times 21 \times 15$				
$\tilde{M}_{src}$			$T \times 21 \times 15$				
(c) Decoder $D$ .							
Layer	Filter	Output shape					
$\mathbf{f}_d^{G_i}, \mathbf{f}_s^{G_i}$		$T_i \times J_i \times C_i$					
BP-AdaIN	-	-					
LReLU	-	-					
STConv	$k_s[i]k_t[i]s1$	$T_i \times J_i \times C_i$					
BP-ATN	-	-					
STConv	$k_s[i]k_t[i]s1$	$T_i \times J_i \times 2C_i$					
$\tilde{\mathbf{f}}_d^{G_i}$		$T_i \times J_i \times C_i/2$					
(d) BP-StyleNet on $G_i$ level.							

Table 3. Architecture of Motion Puzzle framework.

in  $G_3$  and  $G_2$  levels. And i-frame translated pose is obtained by slicing in [i-30, i] translated motion. As a result, we reduced the memory size by more than 80% while maintaining a suitable quality in the translated output motion.

### A.2 Training Details

Our architecture is trained for 10 epochs with a batch size of 32. Learning rates for all networks are set to  $10^{-4}$ . Training time is about 5 hours with two NVIDIA GTX 2080ti. All networks are implemented with PyTorch. We set  $\lambda_{cyc} = 1$ ,  $\lambda_{root} = 1$  and  $\lambda_{sm} = 1$  in Eq. (20). We use the RAdam optimizer [Liu et al. 2019b] with  $\beta_1 = 0$  and  $\beta_2 = 0.99$ . For better results, we adopt the exponential moving averages [Karras et al. 2018] over all parameters of every network.