Implicit Compositional Generative Network for Length-Variable Co-Speech Gesture Synthesis

Chenghao Xu, Jiexi Yan, Yanhua Yang, and Cheng Deng, Senior Member, IEEE

Abstract—Co-speech gesture synthesis is a practical yet challenging task that aims to generate body motion sequences in line with speech audio. Most of the existing methods can only generate the gesture sequence with a fixed number of frames, which does not satisfy the high-quality requirement of the virtual speech video in real-world applications. In this paper, we propose a novel Implicit Compositional Generative Network (ICGN) for length-variable co-speech gesture synthesis. In ICGN, the implicit neural representation is captured and optimized for a whole gesture sequence of arbitrary length with temporal embeddings. Moreover, to enforce the synthesized gestures more realistic and consistent, we compositionally generate the gesture sequence through a well-designed asymmetric two-stream network that effectively captures and utilizes the rich correlations between speech audio and human body motions. In this way, the coarse and fine-grained gestures are synthesized, respectively, according to the corresponding content-aware and emotion-aware audio components. Extensive experiments on four widely-used benchmarks demonstrate that the proposed method renders realistic human gestures and achieves the superior performance against several state-of-the-art methods.

Index Terms—Co-speech gesture synthesis, implicit neural representations, compositional generation.

I. INTRODUCTION

In the real world, we often make spontaneous body motions to help convey our thoughts when talking with others. Such non-verbal behaviors enhance interaction experiences and improve comprehension between talkers and listeners with implicit semantic information [1], [2], [3], [4]. Therefore, automatically generating vivid co-speech gestures for given audio clips is of great importance in the field of realistic digital human synthesis [5], [6], [7]. To this end, many studies have been proposed to explore the task of co-speech gesture synthesis [8], [9], [10], [11], [12], [13], [14] that aims to generate harmonious body motions synchronized with the corresponding speech audio. It gives us a new opportunity for realistic human-human interaction in virtual platforms.

Recently, audio-driven gesture synthesis methods leverage deep learning techniques to learn the compact representation of audio and map it to holistic human pose sequence through exploring the correlations between speech and body language [11], [12]. Due to the lack of cross-modal semantic information, such a straightforward approach can only capture the averaged motion that neglects the diversity of generated body motions. To tackle this issue, some latest studies take the cross-modal semantic relationships between audio and the corresponding gesture sequence into consideration [11], [15]. Leveraging the concrete text in audio, each audio can be decoupled into different semantic components that are adopted to generate the corresponding poses.

Existing methods are constrained to generating body motion sequences with a fixed number of frames, mirroring those seen in the training data. However, this approach proves inadequate when producing high-quality virtual avatar speech videos necessitating a higher frames-per-second (FPS) output. As depicted in Fig. 1, if a gesture sequence with a heightened frame rate is desired, we must resort to video frame interpolation based on the original gesture sequence generated by conventional methods. Unfortunately, due to the inherent characteristics of frame interpolation, the motion speed in the interpolated frames remains constant, resulting in a lack of smoothness and realism.
Extensive experiments on 3D and 2D datasets demonstrate that our model generates natural and smooth motions well synced with the speech, outperforming other SOTA methods in many aspects.

II. RELATED WORK

A. Co-Speech Gesture Generation

Synthesizing co-speech gestures has been of important interest in the field of computer vision [16], graphics [9], and robotics [17], [18], [19]. Generally, methods for this problem take the speech audio as input and generate corresponding body motions to simulate the real speaker. The existing methods can be roughly divided into two categories, i.e., rule-based and data-driven methods.

Data-driven Method: To reduce the manual effort required for rule generation, data-driven methods try to discover gesture synthesis rules in data utilizing machine learning techniques. Probabilistic model [14], [20], [21] and neural classification model [22] are proposed for speech-gesture mapping. With recent advancements in deep learning, an end-to-end approach directly using raw gesture data is possible, which frees the human efforts of creating the mapping rules and gesture templates. Specifically, gesture generation is formulated as a regression problem that does not require crafting unit gestures and rules. In this way, co-speech gestures can be synthesized using deterministic models, including multi-layer perception (MLP), convolution neural networks (CNNs) [8], recurrent neural networks (RNNs) [10], [11], [12], [23], and transformers [24]. Moreover, generative models such as VAEs [8] and normalizing flow [9], can also be adopted to learn co-speech human body motion synthesis.

B. Implicit Neural Representation

In recent years, implicit neural representation (INR) has been extensively employed in various generation tasks, such as inverse graphics [25], [26], image synthesis [27], [28], and scene generation [29], [30], due to its continuous, efficient, and differentiable nature [31], [32]. The usual implementation of INRs includes a fully-connected neural network that maps coordinates to the corresponding values of data, essentially encoding data in terms of continuous functional relationships between signals. In this paper, at the temporal level by a transformer-based architecture.

III. IMPLICIT COMPOSITIONAL GENERATIVE NETWORK

A. Problem Formulation

Co-speech gesture synthesis aims to generate the corresponding body motion sequence for a given speech data (i.e., audio and text). Though the text information of the speech is useful for learning the map from speech to body motion, it is not accessible in the real-world scenario. Therefore, in this paper, we consider audio-driven gesture generation without text information. Our
goal is to generate length-variable gesture sequences with natural poses and synchronized motions that match well with any given speech and arbitrary frames.

Given the training speech audio set $D = \{X_1, X_2, \ldots, X_N\}_{i=1}^{N}$ with the corresponding body motion sequences $G = \{Y_1, Y_2, \ldots, Y_N\}_{i=1}^{N}$ represented by the human pose landmarks [33, 34, 35], we need to establish the mapping $\theta: D \rightarrow G$, i.e., $Y_i = f(X_i; \theta)$ parameterized by the model parameter $\theta$. Each $X_i \in \mathbb{R}^{T \times D_X}$ or $Y_i \in \mathbb{R}^{T \times D_Y}$ corresponds to $T$ consecutive frames in the speech videos, where $D_X$ and $D_Y$ denote the feature dimensions of audio and gesture, respectively. For example, when processing 2D pose, $D_Y = 2 \times L$, i.e., the ground truth $Y_i$ contains the 2-D coordinates of $L$ pose landmarks of the $i$-th identity. Moreover, following [36], the text transcripts $\{T_1, T_2, \ldots, T_N\}$ are available for training, which is also useful to gesture synthesis.

B. Overall Framework

The overall flowchart of the proposed ICGN model is illustrated in Fig. 3. Our proposed method consists of two main modules, i.e., the compositional audio encoding module and the implicit gesture generation module.

Different from straightforwardly using MFCC [37] feature as input, we conduct a compositional audio encoding (CAE) module to decouple the original audio signals. We derive a series of encoded audio components, i.e., Mel-frequency cepstral coefficient, volume, and rhythm. We will utilize different compositions of these components as input to employ the specific gesture generation blocks in the following.

In the implicit gesture generation (IGG) module, we first learn the implicit feature representation with variable length $T'$ with respect to the content-aware and emotion-aware audio components. And then, we propose an asymmetric two-stream network to synthesize the coarse and fine-grained gestures, respectively. Specifically, the content-aware audio component is adapted to generate the coarse gesture through a transformer-based network, while the emotion-aware component is input to synthesize the fine-grained gesture through a convolution-based network. To this end, we combine the coarse and fine-grained gestures and achieve the final co-speech gesture with variable length.

C. Implicit Gesture Generation

In order to satisfy the demand for high-quality co-speech gestures, we propose the implicit gesture generation module to adaptively synthesize smooth and consistent gesture sequences with any frame requirement.

Implicit Feature Representation: To meet the different variable frame requirements, we extensively employ implicit neural representation (INR) in our co-speech gesture synthesis. INR has recently attracted great attention in many computer vision tasks due to its continuous, efficient, and differentiable nature. Combined with INR, we can encode the original audio feature with fixed frame $T$ into implicit feature representation with any frame $T'$.

Following the existing INR methods [38, 39], we regard the feature representation as a continuous function $g : t \rightarrow v$, where $t$ is the temporal coordinate on the normalized speech audio, and $v$ is the corresponding implicit feature vector. Specifically, we construct a normalized temporal grid and project it into high-dimensional vectors with the Fourier encoding [25, 40]. The Fourier encoded temporal coordinate is calculated as

$$Q = F(z) \in \mathbb{R}^{r T \times D_X},$$

where $z = [-1, \eta - 1, 2\eta - 1, \ldots, 1]$ refer to the temporal coordinate with gaps computed as $\eta = \frac{T - 1}{2^r}$. Since $\tau$ can be any continuous value, the granularity of the temporal coordinate can
be varying. And then, we map the Fourier encoded temporal coordinate into a sequence of implicit vectors with the input concatenated feature sequence \( \Lambda_i \in \mathbb{R}^{T \times D_c} \) as conditional factor through a multi-head attention block as follows:

\[
V_i = \text{MHA}(W_Q \cdot Q, W_K \cdot \Lambda_i, W_V \cdot \Lambda_i) \in \mathbb{R}^{T \times D_v},
\]

where \( Q \) is regarded as the query with a learnable weight \( W_Q \), and \( W_K, W_V \) are the learnable weights that project the input audio features to key and value, respectively. \( T = rT \) is the desired number of frames. Note that, we concatenate the audio feature \( X_i \), text feature \( T_i \), and ID feature \( I_i \) to obtain the concatenated feature sequence \( \Lambda_i \). In this way, we obtain the implicit feature sequence \( V \) with the desired number of frames \( T' \) and utilize it as input to generate the final gesture sequence with the corresponding length \( T' \).

**Compositional Gesture Synthesis**: With the frame-variable implicit feature sequence, we can straightforwardly generate the corresponding frame-variable gesture sequence through a simple conv1d block following the existing methods [10, 23]. However, such a straightforward approach fails to capture the structural information that is helpful for realistic and consistent synthesis. In order to effectively mine and exploit the structural information, we give an intuitive assumption that human body motions are structural and influenced by the speech content and the emotion. Inspired by this assumption, we consider generating the co-speech gesture in a compositional way.

In this compositional scheme, we distinguish between content-related broad motions and emotion-related subtle movements, denoting them as “coarse” and “fine-grained” gestures, respectively. Subsequently, we introduce an asymmetric two-stream network designed to simultaneously synthesize these coarse and fine-grained gestures. To ensure that the generated gestures contain distinct semantic information, we employ two separate networks for generating coarse and fine-grained gestures. Acknowledging the inherent characteristics where transformer-based architectures attenuate high-frequency signals while convolutional-based architectures amplify high-frequency components [41, 42], we adopt a transformer decoder and a convolution-based network for generating coarse and fine-grained gestures, respectively. The combination of these two blocks enables us to capture diverse structural information and generate the final gestures in a complementary fashion.

Formally, given the encoded implicit feature sequence \( V_i \), we can produce the coarse gesture

\[
G_i^C = \text{Decoder}(V_i),
\]

and the fine-grained gesture

\[
G_i^F = \text{Conv1d}(V_i),
\]

where \( \text{Decoder}(\cdot) \) refer to the transformer decoder. Aggregating the above gestures results in the final output

\[
G_i = G_i^C + G_i^F.
\]

**Loss Function**: The coarse and fine-grained gestures are two independent components of the overall gesture synthesis. To better guide the optimization of these independent branches during training, we separate the ground truth \( Y_i \) into two different components that focus on different structural information. In the coarse aspect, we conduct a mean filter with the window width \( W \) to extract the average motion outline as the coarse component:

\[
Y_{i,t}^C = \frac{1}{W} \sum_{a}^{b} Y_{i,t},
\]

where \( Y_{i,t} \) refers to the landmark vector of the \( t \)-th frame in the \( Y_i \), and \( a = t - \frac{W-1}{2}, b = t + \frac{W-1}{2} \) denotes the lower and upper bounds of the summation, respectively. Correspondingly, the subtle dynamics for the fine-grained branch can be derived as follows:

\[
Y_{i,t}^F = Y_{i,t}^C - Y_{i,t}.
\]

To this end, the corresponding loss function is denoted as:

\[
L_c = \sum_{i=1}^{N} \sum_{t=1}^{T} \text{HL}(G_{i,t}^C, Y_{i,t}^C),
\]

\[
L_f = \sum_{i=1}^{N} \sum_{t=1}^{T} \text{HL}(G_{i,t}^F, Y_{i,t}^F),
\]

where \( \text{HL}(\cdot) \) denotes the Huber loss [43] and \( G_{i,t}^C, G_{i,t}^F \) refer to the landmark vectors of the \( t \)-th frame in the \( G_i^C \) and \( G_i^F \), respectively. Moreover, we apply the regularization loss on the final result as follow:

\[
L_{\text{reg}} = \sum_{i=1}^{N} \sum_{t=1}^{T} \text{HL}(G_{i,t}, Y_{i,t}).
\]

Overall, we summarize the reconstruction loss function as:

\[
L_{\text{REC}} = L_c + L_f + L_{\text{reg}}.
\]

Moreover, following the standard protocol [11, 36], we also employ a style-aware diverse loss to avoid the posterior collapse on speaker identify \( I_i \). Specifically, we constrain the any two speaker identities as follows:

\[
L_{\text{SD}} = \sum_{i \neq j} \min \left( \frac{\text{HL}(G_{i,j}, G_{j,i})}{\| I_i - I_j \|_1}, \eta \right),
\]

where \( \eta \) is the numerical clipping parameter and \( \| \cdot \|_1 \) denotes the \( \ell_1 \) norm. Meanwhile, we regard our model as a generator \( G \) and leverage an adversarial loss for preserving realism as follows:

\[
L_{\text{ADV}} = \min_{G} \max_{D} \mathbb{E}[p][\log D(p)],
\]

\[+ \mathbb{E}[\log(1 - D(G(\Lambda))]].
\]

To sum up, the overall learning objective for our whole framework is as follows:

\[
L = \lambda_1 L_{\text{REC}} + \lambda_2 L_{\text{SD}} + \lambda_3 L_{\text{ADV}},
\]

where \( \lambda_1, \lambda_2, \lambda_3 \) are trade-off parameters.
D. Compositional Audio Encoding

As illustrated in [44], [45], audio can be decoupled into different components such as content, volume, and rhythm. Gestures are associated with different levels of audio information. To ensure the harmony between speech and gestures, we utilize different audio components to generate the corresponding gestures rather than only using the MFCC feature of speech audio for gesture synthesis.

Specifically, we decouple the original audio signals \( \{ x_i \in \mathbb{R}^T \}_{i=1}^N \) into three components, i.e. MFCC factor \( \hat{M}(x_i) \), volume factor \( \hat{V}(x_i) \) and rhythm factor \( \hat{R}(x_i) \). The MFCC factor can be easily calculated following [37]. We exploit the amplitudes and onsets of audio to stand for the volume and rhythm factors, respectively. It is natural that large amplitude values of the audio signal imply large volumes. Based on this, we define the volume factor as:

\[
\hat{V}_i(x_i) = \begin{cases} 
0 & F_i(x_i) < \sum_{t=1}^{T} \frac{F_i(x_i)}{T} \\
1 & F_i(x_i) \geq \sum_{t=1}^{T} \frac{F_i(x_i)}{T}, 
\end{cases}
\]

(15)

where \( F(x_i) \in \mathbb{R}^T \) denotes the amplitude of the audio signal \( x_i \). Meanwhile, the rhythm factor is represented with the onset strength envelope [46], [47], [48], where onset [47], [48] states the start points of the sound while the strength envelope implies the probabilities of the onset detected in the original audio signal.

We adopt the MFCC factor \( \hat{M}(x_i) \) to encode the content-aware audio features \( \{ x_i \}_{i=1}^N \) with the specific encoder \( E_c \) and combine the volume and rhythm information \( \hat{V}(x_i), \hat{R}(x_i) \) to jointly encode the emotion-aware audio features \( \{ x_i \}_{i=1}^N \) with the specific encoder \( E_c \). To this end, we can well build the association between the speech audio and gestures in a compositional manner.

We map the content-aware audio component \( \{ x_i \}_{i=1}^N \) into the content-related coarse gesture as follow:

\[
\begin{align*}
\hat{V}_i^C &= \text{MHA}(W_Q \cdot Q, W_K \cdot \hat{X}_i, W_V \cdot \hat{X}_i), \\
\hat{G}_i^C &= \text{Decoder}(\hat{V}_i).
\end{align*}
\]

(16)

And the emotion-aware audio component \( \hat{X}_i \) is projected into the emotion-related fine-grained gestures:

\[
\begin{align*}
\hat{V}_i^F &= \text{MHA}(W_Q \cdot Q, W_K \cdot \hat{X}_i, W_V \cdot \hat{X}_i), \\
\hat{G}_i^F &= \text{Conv1d}(\hat{V}_i).
\end{align*}
\]

(17)

The updated \( \hat{G}_i^C, \hat{G}_i^F \) will be adopted to calculate the loss function summarized in (11) for optimizing the overall implicit compositional generative network.

IV. EXPERIMENTS

In this section, we provide experimental analysis and comparative evaluation to demonstrate the effectiveness of our proposed method ICGN.

A. Experimental Settings

Datasets: Following the existing methods [8], [11], [12], [36], we evaluate our method on several widely-used benchmark datasets as follows:

1) Speech2Gesture: In the 2D Co-Speech Gesture Synthesis task, we conduct all the experiments on the Speech2Gesture dataset, which contains speaker-specific videos of television anchors. Due to the Video Copyright Issues, we utilize the speech videos of 8 speakers in Speech2Gesture dataset.

2) Speech2Gesture Extended: In order to more comprehensively evaluate the performance of our method, we follow [8] and use the speech videos of Oliver and Kubinec in original Speech2Gesture, and two different identities, Xing and Luo, for experiments. We denote this mixed dataset as Speech2Gesture Extended. In particular, these four people have upper body keypoints and face keypoints as groundtruth, while the original Speech2Gesture dataset only has upper body keypoints.

3) TED Gesture: TED Gesture dataset [19], [36] is a large-scale English-language dataset for speech-driven motion synthesis, which contains 1766 TED videos of different narrators covering various topics. The extracted 3D human skeletons, aligned English transcripts and speech audio are all available. Following [11], [36], we resample human poses with 15 FPS and sample the consecutive 34 frames with the stride of 10 frames as input segments. We finally get 252109 segments with a length of 106 h. In this dataset, the human pose is represented by direction vectors of 10 upper body joints.

4) TED-Expressive: Following the [11], we test our method with a more delicate dataset. The TED-Expressive dataset has more expressive finger and body movements than the TED Gesture dataset by using the state-of-art 3D pose estimator ExPose [49] to capture the pose information in videos fully. TED Expressive annotates the 3D coordinates of 43 keypoints, including 13 upper body joints and 30 finger joints.

Evaluation Metrics: To comprehensively verify the effectiveness of our method, we adopt some widely-used metrics as follows:

1) \( L_2 \): To evaluate the authenticity, we calculate the distance between each landmark of the prediction and the ground truth using \( L_2 \) norm.

2) \( \mathcal{E}_{\text{lip}} \): The normalized lip-sync error (\( \mathcal{E}_{\text{lip}} \)) [8] is a proxy metric for gesture synchronization. We can calculate it as follow:

\[
\mathcal{E}_{\text{lip}} = \frac{1}{T} \sum_{t=1}^{T} \frac{\| \hat{d}_t^g - d_t^y \|_2}{\max_{1 \leq n \leq T, 1 \leq t \leq T} \| d_t^y \|_2},
\]

(18)

where \( \hat{d}_t^g \) is the distance between the center keypoints of upper and lower lip in the \( t \)-th frame of the generated gesture sequence \( \hat{G}_t \), and \( d_t^y \) is the corresponding distance for ground-truth gesture sequence \( Y_t \).

3) LVD: Landmark velocity difference (LVD) [50], [51] calculates the average Euclidean distance between reference landmark velocities and predicted ones, which indicates how well the model captures the relation of speech and motion, especially since we do not expect to generate motions the same as the ground truth.
TABLE I
QUANTITATIVE COMPARISON WITH THE STATE-OF-THE-ART METHODS ON SPEECH2GESTURE EXTENDED DATASET USING L₂, ε₁IP, AND LVD METRICS (LOWER IS BETTER)

<table>
<thead>
<tr>
<th></th>
<th>Oliver</th>
<th>Kubince</th>
<th>Xing</th>
<th>Luo</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L₂</td>
<td>ε₁IP</td>
<td>LVD</td>
<td>L₂</td>
</tr>
<tr>
<td>Audio to Body [10]</td>
<td>52.9</td>
<td>0.19</td>
<td>6.68</td>
<td>71.3</td>
</tr>
<tr>
<td>S2G [23]</td>
<td>55.7</td>
<td>0.24</td>
<td>5.41</td>
<td>61.0</td>
</tr>
<tr>
<td>MoGlow [9]</td>
<td>51.3</td>
<td>0.20</td>
<td>5.26</td>
<td>68.1</td>
</tr>
<tr>
<td>TriCon [36]</td>
<td>50.6</td>
<td>0.20</td>
<td>5.11</td>
<td>78.1</td>
</tr>
<tr>
<td>Tmpt [8]</td>
<td>62.4</td>
<td>0.17</td>
<td>4.86</td>
<td>100.7</td>
</tr>
<tr>
<td>Ours</td>
<td>45.2</td>
<td>0.17</td>
<td>4.53</td>
<td>53.7</td>
</tr>
</tbody>
</table>

The best results are indicated as bold.

TABLE II
QUANTITATIVE COMPARISON WITH THE STATE-OF-THE-ART METHODS ON SPEECH2GESTURE DATASET USING LVD METRIC (LOWER IS BETTER)

<table>
<thead>
<tr>
<th></th>
<th>Almaram</th>
<th>Angelica</th>
<th>Chemistry</th>
<th>Conan</th>
<th>Oliver</th>
<th>Seth</th>
<th>Shelly</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio to Body [10]</td>
<td>13.5</td>
<td>16.3</td>
<td>8.5</td>
<td>12.6</td>
<td>10.6</td>
<td>13.1</td>
<td>10.1</td>
<td>17.6</td>
</tr>
<tr>
<td>S2G [23]</td>
<td>11.0</td>
<td>9.1</td>
<td>14.9</td>
<td>8.4</td>
<td>8.4</td>
<td>5.2</td>
<td>4.0</td>
<td>11.7</td>
</tr>
<tr>
<td>MoGlow [9]</td>
<td>10.6</td>
<td>8.7</td>
<td>8.7</td>
<td>7.2</td>
<td>9.1</td>
<td>5.6</td>
<td>4.4</td>
<td>12.1</td>
</tr>
<tr>
<td>TriCon [36]</td>
<td>8.3</td>
<td>9.3</td>
<td>5.6</td>
<td>6.7</td>
<td>7.5</td>
<td>5.7</td>
<td>4.1</td>
<td>13.8</td>
</tr>
<tr>
<td>Tmpt [8]</td>
<td>9.1</td>
<td>6.3</td>
<td>5.6</td>
<td>7.7</td>
<td>7.3</td>
<td>3.3</td>
<td>4.2</td>
<td>11.8</td>
</tr>
<tr>
<td>Ours</td>
<td>7.3</td>
<td>5.7</td>
<td>4.4</td>
<td>5.8</td>
<td>6.9</td>
<td>5.2</td>
<td>4.2</td>
<td>10.9</td>
</tr>
</tbody>
</table>

The best results are indicated as bold, and the second ones are indicated as underline.

TABLE III
QUANTITATIVE COMPARISON WITH THE STATE-OF-THE-ART METHODS ON SPEECH2GESTURE AND SPEECH2GESTURE EXT. DATASETS USING DIVERSITY (MORE IS BETTER)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Speech2Gesture Ext.</th>
<th>Speech2Gesture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio2Body</td>
<td>21.65</td>
<td>23.53</td>
</tr>
<tr>
<td>S2G [23]</td>
<td>31.44</td>
<td>35.73</td>
</tr>
<tr>
<td>MoGlow [9]</td>
<td>32.48</td>
<td>37.34</td>
</tr>
<tr>
<td>TriCon [36]</td>
<td>35.05</td>
<td>42.31</td>
</tr>
<tr>
<td>Tmpt [8]</td>
<td>39.29</td>
<td>45.86</td>
</tr>
<tr>
<td>Ours</td>
<td>42.51</td>
<td>50.03</td>
</tr>
</tbody>
</table>

The best results are indicated as bold.

4) **Diversity**: Diversity [52] measures how many different motions have been generated within a long motion. If a method makes some static motions for a long time, we should give it low scores. We follow the testing process in former works [11], where we randomly sample 60 speech audios and compute the average feature distance we generated between 500 random pairs.

5) **FGD**: Fréchet Gesture Distance directly measuring the distance between a generated gesture sequence and the ground truth discourages variety, which is similar to Fréchet Gesture Inception Distance. We use FGD to measure the distribution distance between the synthesized and the real gesture sequences among a group of samples rather than a single sample.

6) **BC**: Beat Consistency Score (BC) is wildly used in motion-audio tasks [11], [24], [53], [54], which computes the average distance between every music beat and its nearest kinematic beat. We follow previous work to define kinematic beats as the local minima of the kinetic velocity.

**Implementation Details**: We implement our method in PyTorch [55]. For audio feature, we use torchaudio1 and librosa2 to obtain MFCC and strength envelope. During training, we take input spectrograms corresponding to about 4.2 seconds of audio and predict 64 pose vectors on both datasets, corresponding to about 15 FPS (Frames Per Second). We set $D_x$ to 256, the window width $W$ to 5, and the number of heads in multi-head attention to 1. We set $λ_1$, $λ_2$, $λ_3$ as 50, 0.01 and 1.

---


Fig. 5. Examples of generated gestures with our implicit synthesis model. The top line states the ground-truth speech video, and the bottom line shows the corresponding gesture sequence synthesized by our method. It is easy to find that our results can correspond well with the natural speech video.

Fig. 6. Visualized comparison of the generated gestures on Speech2Gesture Extended dataset [8]. The frequency and amplitude of the body motions we generate are consistent with the ground truth, while the result of Tmpt [8] does not match well with the ground truth.

TABLE IV  

<table>
<thead>
<tr>
<th>Methods</th>
<th>TED Gesture</th>
<th>TED Expressive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FGD↓</td>
<td>BC♂</td>
</tr>
<tr>
<td>Ground Truth</td>
<td>0</td>
<td>0.698</td>
</tr>
<tr>
<td>Attention Seq2Seq [19]</td>
<td>18.154</td>
<td>0.196</td>
</tr>
<tr>
<td>S2G [23]</td>
<td>19.254</td>
<td>0.668</td>
</tr>
<tr>
<td>Joint Embedding [56]</td>
<td>22.083</td>
<td>0.200</td>
</tr>
<tr>
<td>TriCon [36]</td>
<td>3.729</td>
<td>0.677</td>
</tr>
<tr>
<td>HA2G [11]</td>
<td>3.072</td>
<td>0.672</td>
</tr>
<tr>
<td>Ours</td>
<td>2.415</td>
<td>0.681</td>
</tr>
</tbody>
</table>

The best results are indicated in bold.

B. Evaluation on 2D Datasets

To evaluate the performance of our methods, we conduct authenticity and diversity evaluation with state-of-the-art methods, including Audio2Body [10], S2G [23], MoGlow [9], TriCon [36] and Tmpt [8], on two 2D datasets, i.e., Speech2Gesture and Speech2Gesture Extended. To perform a fair comparison, we use Adam [57] optimizer with the learning rate of 0.0001 on one GTX 2080Ti GPU and set the batch size as 32, which is kept consistent to these compared methods. The experimental results are shown in following.

Quantitative Evaluation: We compare our method with the state-of-the-art methods on both datasets, and the corresponding results are reported in Tables I and II. We only use the LVD metric for the Speech2Gesture dataset because the face keypoint is lacking in this dataset. It can be easily observed that our
method achieves the best performance on both datasets, which demonstrates that our proposed method can effectively generate realistic and consistent human gesture sequences, resulting in better performance.

To measure the diversity of the synthesized gestures, we compare our method with the state-of-the-art on Speech2Gesture and Speech2Gesture Extended datasets, and the Diversity metric is adopted for evaluation. The experimental results are reported in Table III. Our method obviously achieves the best performance on both datasets, indicating that our method can synthesize diverse speech gesture sequences. The diversity score of Audio2Body [10] is the lowest because the output of the RNN-based method quickly gets stuck into some poses, resulting in long static motion afterward.

Subjective Evaluation by User Study: Our goal is to generate realistic motions based on the audio information. The evaluation above can only evaluate the quality in a single frame or the diversity of motions sequence. To evaluate the performance from the perspective of humans, we further conduct user studies, where Tmpt [8] is used as the compared method. We generate 10 test video clips for each method, which consists of the results synthesized by our methods, Tmpt, and the ground truth in random order. Notably, all the clips are synthesized using the models trained on the Speech2Gesture Extended dataset, and the duration of each clip is around 15 s. We not only test the models at 15 FPS but also test them at 60 FPS. We recruit 30 volunteer participants to participate in our study. When participants watch each video clip, they will be asked to answer three questions and rate the video from 1 to 5, with 1 being the worst, 3 being average, and 5 being the best. The three questions are: (1) degree of naturalness, (2) degree of fluency, and (3) speech-to-gesture beat matching.

We report the subjective evaluation results in Fig. 4. It is easy to find that our method outperforms the state-of-the-art regarding the extent of naturalness fluency and matching. More specifically, our model achieves the best results on higher FPS. The existing methods get a lower naturalness score on higher FPS due to the non-smoothness of the frame interpolation. The interpolation operation calculates the average of two adjacent frames as the interpolated frame ignoring the body dynamics. Our implicit generative method can adaptively synthesize the gesture sequence with higher FPS, which helps us achieve a more realistic and consistent human body motion sequence with higher FPS.

Moreover, when a frame result makes a mistake, the interpolation operation spreads the mistake to all the nearby results generated by interpolation. As for ours, the predictions of the different results are relatively independent and do not influence each other.

Visualization: To further illustrate the superiority of our method, we show the generation results compared with the ground truth and other state-of-the-art methods in Figs. 5 and 6. As shown in Fig. 5, the generative examples of our method can correspond well with the ground-truth speech video, which indicates that our ICGN can effectively synthesize more realistic and consistent co-speech gesture sequences. These visualization results in Fig. 6 demonstrate that our method can synthesize more stable gestures compared with the other methods. It is because our method focuses more on sounding regions and covers the sounding regions more accurately and comprehensively, which is beneficial for accurate temporal localization prediction. A common problem for co-speech motion generation is permanent drastic changes, which can be observed in the results of Tmpt [8]. Tmpt directly makes the generated results diverse, leading to continually changing results, while our model has a much more stable generation quality, which is similar to tangible results.

C. Evaluation on 3D Datasets

To better verify the effectiveness of our proposed ICGN, we make a comprehensive comparison with state-of-the-art methods, including Attention Seq2Seq [19], S2G [23], Joint Embedding [56], TriCon [36], and HAG2G [11], on two 3D datasets, i.e., TED Gesture and TED-expressive. To perform a fair comparison, We use Adam [57] optimizer with the learning rate of 0.001 on 1 GTX 2080Ti GPU and set the batch size as 128, which are kept the same as the compared methods [11], [12].

Quantitative Evaluation: Following the existing methods, we compare our proposed ICGN with all the baselines with three evaluation metrics, i.e., FGD, BC, and Diversity, on two datasets, i.e., TED Gesture and TED-expressive. We report the experimental results in Table IV. We can find that our ICGN outperforms
all these compared methods by a large margin in terms of all three evaluation metrics. The significant improvements indicate the effectiveness and superiority of our implicit compositional generation.

Visualization: To comprehensively evaluate our method, we conduct a qualitative analysis of our method and other compared approaches. The keyframes comparison of our method against ground truth and SOTA baselines is shown in Fig. 8. Since TED-expressive requires a higher ability of generative models, we select one case of TED-expressive [11] for comparison. Compared with the baseline, our method can generate better diverse human-like poses consistent with the ground truth, which further demonstrates the effectiveness of our method.

D. Ablation Study

Effect of Different Components: To gain more insights into the two different components of the ground truth $Y_i$, we test the performance in different window widths $W$. We examined how important an asymmetric two-stream network needs to solve the co-speech gesture synthesis. When $W$ set to 1, $Y$ is decomposed into $Y^C$ and $Y^F$, and we just use the $\mathcal{L}_{reg}$ to optimize the network. When the value of $W$ is bigger than 1, $Y$ is decomposed, and the total of the three losses optimizes the network. The results in Table V demonstrate that the decomposition option is necessary, and increasing the size of $W$ deteriorates the performance gradually. We summarize this as a mean filter that shortens the length of both ends. The more enormous value $W$ is, the more uncertainty appears to optimize on both ends.

Animation Study: To further demonstrate the superiority of our implicit generation, we analyze the learned coordinate values of our method and traditional co-speech gesture synthesis with the video frame interpolation technique, such as linear interpolation. We upsample the frame per second to four times with linear interpolation and INR. In particular, our method only requires changing the dimensions of the template during the inference phase and no other modifications are required. According to the experimental results shown in Fig. 7, we can see that, unlike the linear interpolation, our method exhibits non-linear deformation trajectories, which are more in line with the actual situation.

V. CONCLUSION

In this paper, we propose a novel Implicit Compositional Generative Network (ICGN) for length-variable co-speech gesture synthesis. To generate the gesture sequence with an arbitrary number of frames, we first introduce implicit neural representation and efficiently combine it with our gesture sequence generation through the temporal embeddings. In addition, an asymmetric two-stream network is employed to mine and take full advantage of the correlations between speech audio and human body motion. Specifically, the independent content-aware and emotion-aware components are extracted from each speech audio, respectively. And then, these components are input to different networks to generate the coarse and fine-grained gestures. Our model generates highly smooth and visually realistic body motions with arbitrary length in a compositional manner.

REFERENCES


Jiexi Yan received the B.E. and Ph.D. degrees from Xidian University, Xi’an, China, in 2017 and 2022, respectively. He is currently a Lecturer with the School of Computer Science and Technology, Xidian University. His research interests include machine learning and computer vision.

Yanhua Yang received the B.E., M.S., and Ph.D. degrees in signal and information processing from Xidian University, Xi’an, China, in 2004, 2007, and 2017, respectively. She is currently an Associate Professor with the School of Computer Science and Technology, Xidian University. Her main research interests include complex action recognition and event detection.

Cheng Deng (Senior Member, IEEE) received the B.E., M.S., and Ph.D. degrees in signal and information processing from Xidian University, Xi’an, China, in 2004, 2007, and 2017, respectively. He is currently a Full Professor with the School of Electronic Engineering, Xidian University. He is the author and the coauthor of more than 100 scientific articles at top venues, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CYBERNETICS, NeurIPS, ICML, CVPR, ICCV, AAAI, IJCAI, and KDD. His research interests include computer vision, pattern recognition, and information hiding.