DisCo: Disentangled Implicit Content and Rhythm Learning for Diverse Co-Speech Gestures Synthesis

Haiyang Liu*
The University of Tokyo

Naoya Iwamoto
Huawei Technology Japan K.K.

Zihao Zhu
Keio University

Zhengqing Li
Huawei Technology Japan K.K.

You Zhou
Huawei Technology Japan K.K.

Elif Bouzert
Huawei Turkey R&D Center.

Bo Zheng
Huawei Technology Japan K.K.

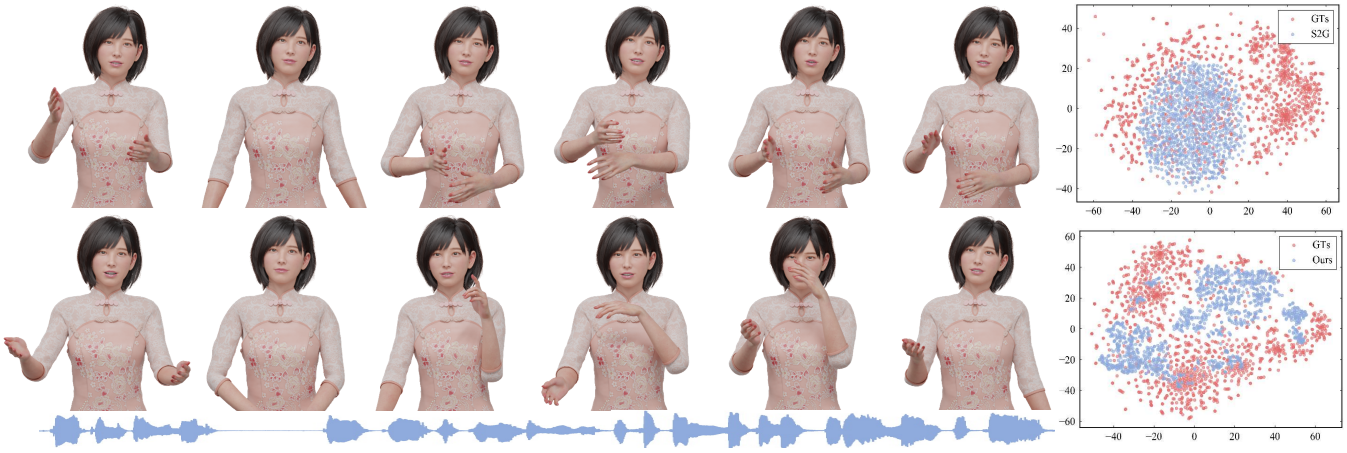


Figure 1: DisCo is a framework for co-speech gestures synthesis in a disentangled manner, which generates more diverse gestures without sacrificing speech-motion synchrony. *Left*: the input speech (bottom), generated body and hands motion by S2G [28] (top) and Ours (middle). *Right*: T-SNE results for motion segments on Trinity [24] test data. The demo adopts avatar by artists with recorded facial expressions.

ABSTRACT

Current co-speech gestures synthesis methods struggle with generating diverse motions and typically collapse to single or few frequent motion sequences, which are trained on **original data distribution** with customized models and strategies. We tackle this problem by temporally clustering motion sequences into content and rhythm segments and then training on **content-balanced data distribution**. In particular, by clustering motion sequences, we have observed for each rhythm pattern, some motions appear frequently, while others appear less. This imbalance results in the difficulty of generating low frequent occurrence motions and it cannot be easily solved by resampling, due to the inherent many-to-many mapping between content and rhythm. Therefore, we present DisCo, which disentangles motion into implicit content and rhythm features by contrastive loss for adopting different data balance strategies. Besides, to model the inherent mapping between content and rhythm features, we design a diversity-and-inclusion network (DIN), which firstly generates content features candidates and then selects one candidate by learned voting. Experiments on two public datasets, Trinity and S2G-Ellen, justify that DisCo generates more realistic and diverse motions than state-of-the-art methods. Code and data are available at https://pantomatrix.github.io/DisCo/

CCS CONCEPTS

• **Computing methodologies** → **Animation**; **Machine learning**; *Natural language processing*; Computer vision.

KEYWORDS

Co-Speech gestures synthesis, Cross-Modal mapping

ACM Reference Format:
Haiyang Liu, Naoya Iwamoto, Zihao Zhu, Zhengqing Li, You Zhou, Elif Bouzert, and Bo Zheng. 2022. DisCo: Disentangled Implicit Content and Rhythm Learning for Diverse Co-Speech Gestures Synthesis. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3548400

*liuhaiyang@kmj.iis.u-tokyo.ac.jp

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548400

1 INTRODUCTION

Co-Speech gestures synthesis aims to synthesize body motions corresponding to concurrent speech, which has the late advancements in game development and social media content creation domains using digital characters. Some early works related to gesture synthesis including [4, 24, 28, 31, 64], could synthesize plain, repetitive motion with speech-motion rhythm synchrony. However, plain motion, as shown in Figure 1 (top), is only a subset of real co-speech gestures, which should also includes motions (vivid, intense, *etc*). Besides, humans are also not satisfied with monotonous performance and sensitive to the proportion of vivid motions. Thus, solutions, which can synthesize gestures that are more diverse and closer to real distribution are expected to overcome current plain, monotonous performance.

While in real co-speech gestures, the plain motion appears more often than vivid motion, the drawback of previous methods is generating frequent motions in the dataset more compared to less-frequent motions, which yields to plain motion sequences. In this paper, we consider the problem of synthesizing diverse gestures as less-frequent motions synthesis. To achieve the same goal of diverse gestures synthesis, recent state-of-the-art data-driven methods [47, 59] have improved the model's ability to fit the original speech-motion data distribution. However, for data-driven approaches, the final performance of the model is determined by a combination of model capability and the training data distribution (As shown in Figure 2). Changing the distribution of training data, such as re-sampling [18] and data augmentation [70], brings a non-negligible performance improvement in several research fields such as large-scale image recognition, speech recognition [40, 63], and motion estimation [60]. Unfortunately, it is difficult to apply these methods directly in the context of co-speech gesture synthesis for two reasons: *lack of category-level annotation* and *inherent many-to-many mapping between motion content and rhythm.*

Lack of category-level annotation. Most of the publicly available audio-visual co-speech gesture datasets [24, 28, 29, 52, 69] only provide low-level motion information such as joints position/rotation and frame-rate, which lack more interpretable information as the category-level annotations/statistics, *e.g.,* motion categories. As a result, the main benchmark for evaluating realism and diversity of proposed data-driven methods is directly using average joints position/rotation Euclidean distance [24, 28, 47], velocity distance [42] or feature-level distance, *e.g.,* FGD/FTD [9, 59, 69]. These metrics only provide an average estimation of diversity but cannot objectively mark where the model underperforms when generating the motion sequence. Based on these observations, resampling strategies are difficult to reproduce/employ in our case. In this framework, we propose to use category-level statistics by clustering motion content and rhythm in Sec. 3.

Inherent many-to-many mapping. Resampling methods may weaken the inherent synchrony between motion content and rhythm. Ideally, the rhythm and content of the motion should be uncorrelated, *i.e.,* when different sampling strategies are adopted for motion content, motion rhythm should remain uniformly distributed, where resampling can transform both distributions to uniform distribution and facilitate the learning for models. However, we demonstrate through the correlation matrix that there is a many-to-many correlation between rhythm and content. Thus, resampling the motion content will impair the uniformity of rhythm, which reduces the synchrony of the results.

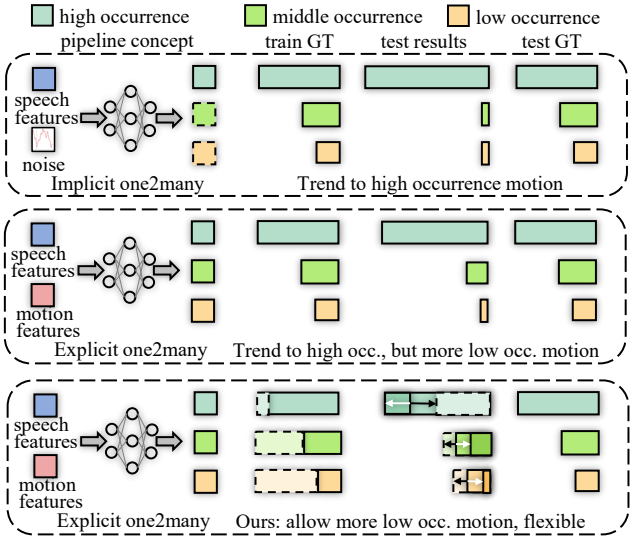
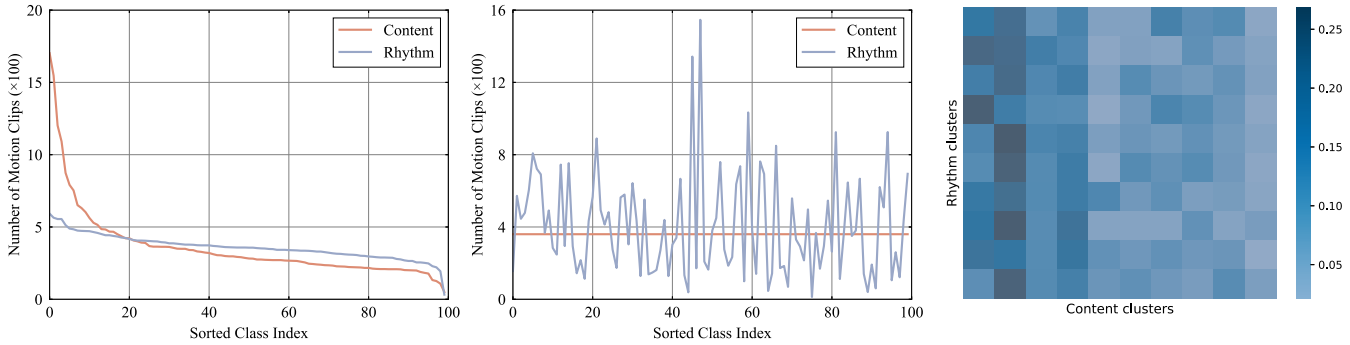


Figure 2: Concept Comparison. Gestures synthesis is a one-to-many task. *Top*: Early baselines [4, 24, 28, 31, 64] used noise to realize one-to-many implicitly. *Middle*: Recent works [47, 59] designed explicit one-to-many models to increase diversity, the data imbalance during training cause the failure of synthesis low occurence motions. *Bottom*: We balance the distribution of training data and achieve the distribution controllable motions.

To solve these problems, we propose DisCo in Sec. 4, which disentangles motion into implicit features of content and rhythm by contrastive loss. Content and rhythm features are extracted with different sampling strategies to alleviate the distribution difference between rhythm and content, which ensures diversity is improved while preserving the synchrony between the two modalities. To model the inherent mapping between content and rhythm, we design the Diversity-and-Inclusion Network (DIN), which consists of multiple sub-networks (diversity networks) and a learned selector (inclusion network) to generate content candidates first and select one candidate by learned voting. Our contributions can be summarized as follows:

- We propose to improve motion diversity by balancing the training data distribution and conduct a **category-level analysis** on motion data that resolves the shortcomings of the traditional resampling methods.
- We propose **DisCo** to disentangle motion into implicit content and rhythm features in the context of co-speech gestures synthesis by contrastive loss and adopt different sampling strategies on disentangled features to increase motion diversity without sacrificing synchrony.
- We present **DIN** to model the inherent mapping between motion content and rhythm, which improves the motion diversity in the training phase and realizes rhythm-feature or user controlled motion synthesis during inference.



(a) Original Distribution (b) Resampled Distribution (c) Correlation Matrix

Figure 3: A Study of Motion Content and Rhythm on Trinity [24] by Clustering. a) Different from rhythm, content exhibits a long-tailed, unbalanced distribution in the number of each category, resulting in generated motions being biased towards the head category. b) Resampling based on content will destroy the uniformity of rhythm, resulting in decreasing speech-motion synchrony. c) Visualization of the inherent correlation between content and rhythm, distributions of content for each rhythm class are not always the same.

2 RELATED WORK

2.1 Co-Speech Gestures Synthesis.

The prior art of co-speech gestures synthesis can be grouped into rule-based [10–12, 15, 16, 37, 57] and traditional data-driven methods [19, 20, 45, 46, 55], *e.g.,* Hidden Markov Model. Recently, deep learning based gesture synthesis methods overperform the previous state-of-the-art HMM based methods by modelling discriminatory features of input data. Lately, several audio-visual datasets have been introduced for researching human co-speech gesture synthesis. Trinity [24] presented a mocap dataset and used autoencoder to establish a baseline. Subsequently, Speech2Gestures [28] used OpenPose [14] to extract 2D poses from large-scale YouTube videos as training data, called S2G Dataset, and used a fully convolution-based GAN model to generate pose by speech. Similarly, MultiContext [69] used VideoPose3D [58] to build on the TED dataset. Subsequent studies present methods trained on these datasets or their extentions [29]. However, these datasets only provide low-level representations of motion, such as joint rotations or positions.

Additionally, several frameworks [3, 12, 25, 26, 42, 54, 66, 67] try to improve the performance of the baseline model by input/output representation selection, adversarial training and various types of generative modeling techniques. As an example, StyleGestures [4] uses Flow-based model [33]. More recently, [47, 59] have begun to shift the focus from realism to the diversity of generated motions. Audio2Gestures [47] improved the realism and diversity of the generated motions by dividing the feature space into shared code and motion-specific code representations. SpeechDrivenTemplate [59] proposed to improve motion diversity by sampling implicit motion features from datasets and combining them with audio features to generate motions. Different from previous works, we focus on improving the performance by balancing the training data.

2.2 Class Imbalance

Resampling methods have been commonly adopted to alleviate the adverse effect of the class imbalance problem in machine learning [21, 35, 38, 72]. The basic idea of resampling-based methods is to over-sample the minority categories [18, 30] or to under-sample the frequent categories during the training process [13, 23]. In particular, Class-aware sampling [61] proposes to choose samples of each category with equal probabilities, which is widely used in computer vision related tasks [27, 53]. However, unlike visual classification, gestures generation does not have annotated class labels and has multiple evaluation metrics. Therefore, it is challenging to apply resampling methods directly for co-speech gestures synthesis.

2.3 Disentangled Representation Learning

Similar to [48], in this framework, we use the term motion rhythm to refer to the duration (beats) of each basic action, *i.e.,* one local minimum of motion velocity. As for the definition of content for motion [6], we refer to the duration-agnostic part of motion. Several works are proposed to disentangle representation in image field [17, 22, 43, 44, 51, 62, 65, 71], however, there are few methods [2, 50] to disentangle motion features. Recently, [1] and [68] disentangle the motion into content and style by adaptive instance normalization and adversarial training, respectively. The methods mentioned above are proposed for motion retargeting, where content and style disentanglement is important. Unlike these methods, we disentangle the motion into implicit content and rhythm features to enable different resampling strategies.

3 A STUDY OF MOTION DIVERSITY

In this section, we discuss the quantitative study of content and rhythm distribution for co-speech gestures through clustering (See Sec. 2 for the definitions of motion content and rhythm).

3.1 Clustering of Motion Content and Rhythm

3.1.1 Motion Words. We refer to [6] for clustering motion content and rhythm, where we use multiple, overlapped, same length motion clips as descriptors of a motion sequence. We create motion clips $\mathbf{M} = \{\mathbf{m}_i \mid \mathbf{m}_i \in \mathbb{R}^{47\times3\times w}, i = 1, \ldots, n\}$, which are called *motion words* as in [5, 6], by downsampling the motion data into 15 fps and using $w = 30$ (*c.f.,* [8]) frames per motion word with stride $s = 5$.

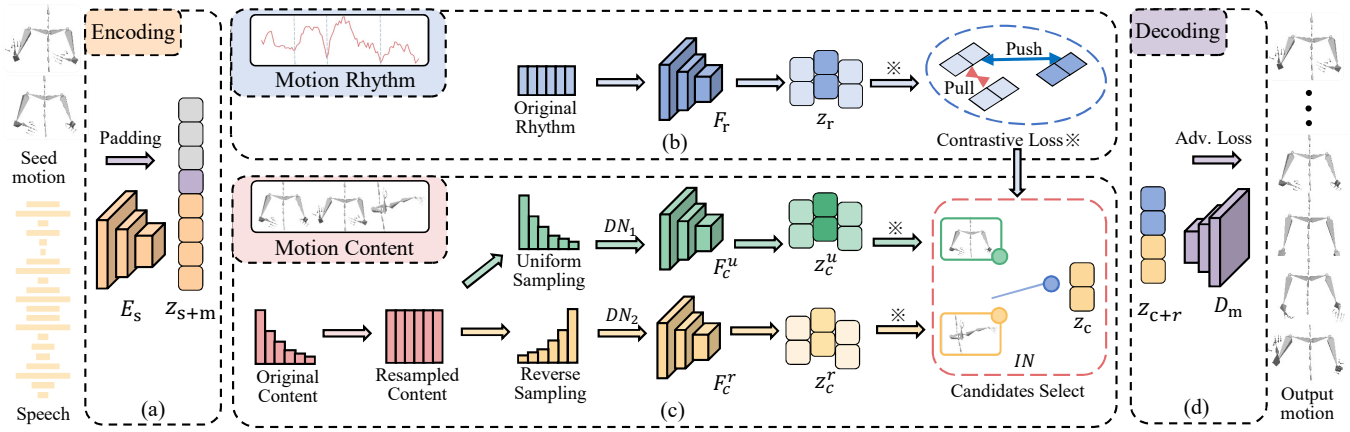


Figure 4: Overview. We disentangles motion into implicit content and rhythm features by contrastive loss on both branches. Afterwards, features are learned through different branches with different sampling strategies. a). Input speech is encoded and concatenated with the seed motion. b). Rhythm features learning with original distribution. Contrastive loss disentangles features by pushing and pulling the distance between negative and positive samples. c). Content branch first generates feature candidates through diversity networks, which consist of two sub-network and adopt uniform and reverse sampling. Then selects the candidates are based on rhythm feature by the inclusion network. d). Motion decoder takes rhythm and selected content features jointly with adversarial training to synthesize the final motion.

3.1.2 Content Clustering. Similar to [6], we use $\mathbf{M}$ as content motion words and Dynamic Time Warping (DTW) distance to measure similarity between two content motion words, *i.e.,* $d^{\mathrm{c}}_{i,j} = DTW(\mathbf{m}_i, \mathbf{m}_j)$ where $j \neq i, j = 1, \ldots, n$. We run the KMeans by DTW and experimentally set the number of classes as 100.

3.1.3 Rhythm Clustering. Similar to [48], we calculate that the rhythmic similarity of the motion depends on the underlying motion beat. Specifically, we calculate the local minimum of the motion velocity $\mathbf{v} = \left\{\mathbf{v}_i \mid \mathbf{v}_i \in \mathbb{R}^{47\times(w-1)}\right\}$ as motion rhythm. Thus, binary motion rhythm is $\mathbf{R} = \left\{\mathbf{r}_i \mid \mathbf{r}_i \in \mathbb{R}^{w-1}\right\}$, *e.g.,*$\mathbf{r} = [0_0, 0_1, 1_2, 0_3, \ldots, 0_{w-1}]$. The Euclidean distance is used to measure the rhythm similarity, *i.e.,* $d^{\mathrm{r}}_{i,j} = \left|\mathbf{r}_i, \mathbf{r}_j\right|$ for clustering them into 100 classes by KMeans.

3.2 Distribution of Content and Rhythm

As shown in Figure 3(a), We find that the distribution of the motion content satisfies the imbalanced, long-tail distribution. Specifically, on the Trinity dataset, the sample account of the tail motion class and the head motion class is close to 1:10. However, since each type of tail motion does not approach zero, 40.72% of the data is gathered in the 60 tail classes. Through visualization, we find that the tail class covers more detailed motion in speech, so increasing the recall rate on the tail class will significantly improve the subjective realism and diversity of the generated motion. We observe similar distributions for the Trinity, TED, and S2G datasets and further show that distribution only exists in the context of conversational gestures synthesis (not in audio2dance, audio2sign-language, etc., recommend reading supplementary materials). It trends to a general problem for this task, so it is difficult to balance various motion content when recording the dataset artificially. On the other hand, the motion rhythm differs from the motion content in terms of data distribution. The results in Figure 3(a) show that the distribution of motion rhythm trends to uniform distribution (only the number of the last category is reduced to 67), which will not lead to the prediction boundary shift [36].

3.3 Discovery Bottleneck by Resampling

Since the existing data faces the long-tail problem in the content distribution, we utilized a standard solution to re-balance the data is class-level resampling. However, the resampled distribution, shown in Figure 3(b), presents that resampling destroys the uniformity of the rhythm distribution. We show the reason is the inherent many-to-many mapping between content and rhythm in Figure 3(c), for different rhythm conditioned content distribution is different, *e.g.,* rhythm from rows 4 and 10 have higher relevance with content 1 and 2, respectively. Therefore, it demonstrates that a simple resampling cannot solve the content class imbalance problem since it may affect the distribution of rhythm, leading to a decrease in the rhythm synchrony.

4 METHOD

Our method aims to better model the long-tail distribution of motion content without sacrificing speech-motion synchrony. As shown in Figure 4, our method divides co-speech gestures synthesis into two branches: rhythm and content branch to synthesize gestures. We first describe the concept-level pipeline design in (Sec. 4.1), then introduce the detailed motion content and rhythm features generation in three phases: disentanglement (Sec. 4.2), resampling (Sec. 4.3), and refinement (Sec. 4.4).

4.1 Overview

We describe the input, output and middle stages of our method in this section but omit how we generate these middle stages, which will be introduced in the later sections. Our model inputs the speech $\mathbf{S} = \left\{\mathbf{s}_i \mid \mathbf{s}_i \in \mathbb{R}^{16000\times w/f}\right\}$, where f is the frame rate, and padded seed motion $\mathbf{m}^{\mathrm{seed}}_i = [\mathbf{m}_{i,1}, \mathbf{m}_{i,2}, \ldots, 0_{i,w-1}, 0_{i,w}]$.

For the given speech audio, we extract their latent features $\mathbf{z}_\mathrm{s}$ through $\mathbf{z}_\mathrm{s} = E_\mathrm{s}(\mathbf{s})$, where E_s represents the speech encoder. The $\mathbf{m}^{\mathrm{seed}}$ is considered as motion features $\mathbf{z}_\mathrm{m}$.

The concatenated features $\mathbf{z}_{\mathrm{s}\oplus\mathrm{m}}$ by speech and seed motion are adopted to generate latent space of motion, which is divided into two sub-spaces, $\mathbf{z}_\mathrm{r}$, and $\mathbf{z}_\mathrm{c}$. The $\mathbf{z}_\mathrm{r}$ and $\mathbf{z}_\mathrm{c}$ are associated with the attribute of rhythm and content, and generated by rhythm and content translation network, F_r and F_c, respectively. The contrastive loss ℓ_{cont} is adopted for features disentanglement (Sec. 4.2). The motion decoder D_m, uses the disentangled motion content and rhythm features to reconstruct motion, as in Eq. 1.

$$\begin{aligned} \hat{\mathbf{m}} &= D_\mathrm{m}(\mathbf{z}_\mathrm{r}, \mathbf{z}_\mathrm{c}) \\ \ell_{\mathrm{rec}} &= \mathbb{E}\left[\|\mathbf{m} - \hat{\mathbf{m}}\|_1\right]. \end{aligned} \tag{1}$$

We also add adversarial loss for the motion decoder D_m, as

$$\ell_{\mathrm{adv}} = -\mathbb{E}[\log(Dis(\hat{\mathbf{m}}))]. \tag{2}$$

Overall, the final loss function of our framework is

$$\ell_{\mathrm{all}} = \lambda_1 \ell_{\mathrm{cont}} + \lambda_2 \ell_{\mathrm{rec}} + \lambda_3 \ell_{\mathrm{adv}}, \tag{3}$$

λ_1, λ_2, and λ_3 are hyperparameters and obtained by grid search.

4.2 Content and Rhythm Disentanglement

The first phase of motion content and rhythm features generation is disentanglement. we describe how to enforce disentanglement during training by contrastive loss. Let $\hat{\mathbf{m}} = [\hat{\mathbf{m}}_1, \hat{\mathbf{m}}_2, ..., \hat{\mathbf{m}}_b]$ denote synthesized motion from one-batch, where b is the number of batch size and $\hat{\mathbf{m}}_b$ is synthesized from $\mathbf{z}_{\mathrm{r},b}$ and $\mathbf{z}_{\mathrm{c},b}$. Besides, we adopt 100-class cluster for motion content and rhythm referring to Sec. 3 as the label of content and rhythm categories, $\mathbf{k}_\mathrm{c}$ and $\mathbf{k}_\mathrm{r}$, respectively. Then, for one specific feature pair, *e.g.*, $\mathbf{z}_\mathrm{r}^i$ and $\mathbf{z}_\mathrm{r}^j$, we could define the positive and negative samples based on $\mathbf{k}_\mathrm{r}^i$ and $\mathbf{k}_\mathrm{r}^j$. Our contrastive loss is defined as,

$$\ell_{\mathrm{cont}}\left(\mathbf{z}_\mathrm{r}^i, \mathbf{z}_\mathrm{r}^j\right) = \begin{cases} \frac{1}{C_\mathrm{r}^+} \max\left(D_\mathrm{r}\left(\hat{\mathbf{m}}_i, \hat{\mathbf{m}}_j\right) - \tau_\mathrm{r}^+, 0\right), & \mathbf{k}_\mathrm{r}^i = \mathbf{k}_\mathrm{r}^j \\ \frac{1}{C_\mathrm{r}^-} \max\left(\tau_\mathrm{r}^- - D_\mathrm{r}\left(\hat{\mathbf{m}}_i, \hat{\mathbf{m}}_j\right), 0\right), & \text{otherwise} \end{cases} \tag{4}$$

here i, j denotes the i-th and j-th features in one-batch, *i.e.*, $n = b$, $\mathbf{k}_\mathrm{r}^i = \mathbf{k}_\mathrm{r}^j$ indicates the same cluster category. τ_r are the thresholds of the positive and negative samples. C_r are constants that normalize the loss according to the number of positive and negative loss components, *i.e.*, $C_\mathrm{r}^+ = \sum_{i,j} \mathbb{1}\left\{\mathbf{k}_\mathrm{r}^i = \mathbf{k}_\mathrm{r}^j\right\}$ and $C_\mathrm{r}^- = \sum_{i,j} \mathbb{1}\left\{\mathbf{k}_\mathrm{r}^i \neq \mathbf{k}_\mathrm{r}^j\right\}$. The function D_r is a distance metric, which could be $L1$, $L2$, and cosine distance, *etc.* We calculate the D_r as,

$$D_\mathrm{r}\left(\hat{\mathbf{m}}_i, \hat{\mathbf{m}}_j\right) = \frac{H_\mathrm{r}(\hat{\mathbf{m}}_i) \cdot H_\mathrm{r}(\hat{\mathbf{m}}_j)}{\|H_\mathrm{r}(\hat{\mathbf{m}}_i)\| \|H_\mathrm{r}(\hat{\mathbf{m}}_j)\|}, \tag{5}$$

where function H_r is extracting the latent features from a pretrained motion rhythm classifier. Similiar as [62], we calculate the distance between embedding vectors using cosine-distance. The contrastive loss for content features $\mathbf{z}_\mathrm{c}$ is constructed in the same manner to better disentangle both content and rhythm features.

4.3 Resampling on Different Branches

The second phase of motion content and rhythm features generation is resampling training data. In this section, we discuss different resampling strategies that different branches should use. Based on the experiments in Sec. 3, we observe that the original distribution of rhythm and content is uniform and imbalanced. Therefore, the rhythm should be trained on original distribution, and content should be resampled to uniform distribution (in Figure 4).

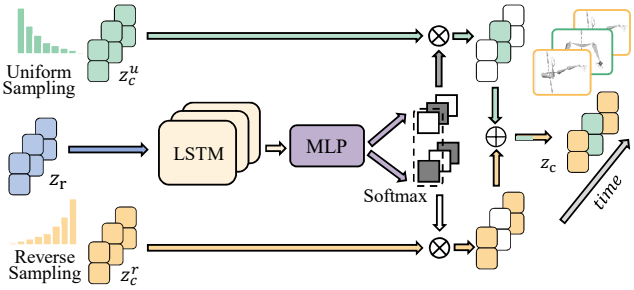


Figure 5: Inclusion Network Details. The purpose of the inclusion network is to linearly weight the motion content features generated by the diversity network in a time series and select the motion content features that are most suitable for the current rhyrhm.

Next, we describe the implementation of above concept. During training, paired speech-motion clips will be resampled based on the content distribution. The resample ratio η for each category is, $\eta^i = \max_k(n^1, n^2, ..., n^k)/n^i$.

For the rhythm translation network, F_r, the data distribution leading to its gradient update should be the original data distribution. Thus, we ignore the weight updating of each sample instance by the possibility p, where $p^i = 1 - 1/\eta^i$.

Finally, we discuss the weight updating of speech encoder E_s, decoder D_m and discriminator $Dim.$, It has been demonstrated in [36] that the original data distribution performs better for representation learning, *e.g.*, the CNN based backbone for image recognition, and balanced distribution performs better for decoding, *e.g.*, the MLP based classifier. We transfer this theory into the context of co-speech generation. In particular, we update weights of E_s based on original distribution as it is proposed to extract speech representation, in contrast, the D_m and $Dis.$ are updated based on balanced distribution.

4.4 Diversity-and-Inclusion Network

The final phase is to model the inherent mapping between content and rhythm. As observed in Sec. 3, the motion content is related to the rhythm, which motivates us to explicitly refine the content features based on rhythm features after generating them separately. Although the basic design choice for refinement is generating the refined content features based on the concatenated content and rhythm features, it cannot be observed a clear difference between refined and original content features based on our experiments, which demonstrates the difficulty of translation based refinement. Thus, instead of translation, we propose a selection based refinement method, the Diversity-and-Inclusion Network, which firstly generates content features candidates and then selects one candidate by learned voting.

4.4.1 Diveresity Networks. DIN is built on the content branch and it takes the $\mathbf{z}_{\mathrm{s}\oplus\mathrm{m}}$ and $\mathbf{z}_\mathrm{r}$ as inputs to generate selected motion content features. In particular, the DNs predicts the content of motion that is most likely to be generated under different distributions. We

adopt uniform sampling and reverse sampling, *i.e.*, in case of the class sample ratio is η, the sampling rate will be $1/\eta$. Besides, we set the diversity network consisting of two sub-networks, F_c^u and F_c^r, corresponding to the head type (plain motions) and the tail type (vivid motions), to force the generated content candidates bias toward different motion types. Diversity networks will generate corresponding motion content features $\mathbf{z}_c^u$ and $\mathbf{z}_c^r$ as Eq. 6,

$$\begin{aligned} \mathbf{z}_c^u &= F_c^u\left(\mathbf{z}_{s\oplus m}^u\right) \\ \mathbf{z}_c^r &= F_c^r\left(\mathbf{z}_{s\oplus m}^r\right), \end{aligned} \tag{6}$$

where different motion content features will also be calculated contrastive loss for rhythm and content disentanglement.

Due to different distributions of input content features, the networks F_c^u and F_c^r have different decision boundaries of the motion content for the same speech. Specifically, F_c^u will tend to generate motions of low occurrence frequency on the original dataset.

4.4.2 The Inclusion Network. IN selects one content candidate for final motion decoding, which is a classification network containing selective kernels attention [49]. As shown in Figure 5, it generates selected motion content by weighted sum for each frame, as

$$\begin{aligned} \mathbf{z}_c &= \alpha \mathbf{z}_c^u + \beta \mathbf{z}_c^r, \\ \alpha, \beta &= IN\left(\mathbf{z}_r\right), \text{ where } \alpha + \beta = 1, \end{aligned} \tag{7}$$

where IN consists of LSTM, MLP and Softmax activation layers. In addition, since $\mathbf{z}_{sc}$ and $\mathbf{z}_{sr}$ consist of long-tailed $\mathbf{z}_{sc}^u$ and its reverse sampling representation $\mathbf{z}_{sc}^r$, the inclusion network is trained in a class-balanced fashion, so the problem of decision boundary shift is avoided [36]. We adopt a supervised manner to train IN, *i.e.*, for each motion clips we will calculated its occurrence probability on different distribution as the groundtruth for α and β. In fact, we found that for a completely unsupervised sampler, IN tends to select the same sub-network's results for long-duration speech, *e.g.*, all head classes. One possible reason is the limitation of LSTM's ability for long-term data. Therefore, we adopt additional supervision to train the IN.

The selection of content features based on rhythm features by the IN forces the model to learn the inherent mapping between features. Besides, IN as a selective module enables the result generated by the final model is not overfitting which tends to be the head class or the tail class as a whole, instead, the correct class should be generated at the correct time location, and the final data distribution is as close as possible to the long-tailed distribution of the real data.

5 EXPERIMENTS

5.1 Setup

5.1.1 Datasets. We report results on commonly-used datasets, **Trinity** [24] and **S2G-Ellen** [28]. Trinity contains 244 minutes of paired speech-motion data in 23 sequences, around ten minutes each for 56 joints. We select 19, 2, 2 sequences for training, validation, and test. Following [47], we also experiment on the S2G-Ellen dataset, which is a subset of the S2G dataset, contains positions of 49 2D body joints estimated from 504 videos. At the training stage, we reduce the motion data to 15 fps, and downsample audio by to the sample rate of 16K using Librosa [56]. We crop $w = 34$ motion frames segments for training. The first four frames are fed as the seed pose to enable multiple results for the same speech audio.

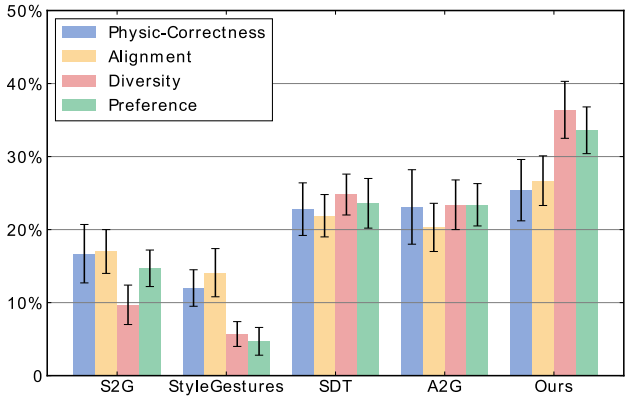


Figure 6: User Study Results. We calculate the win rate for evaluation (in percentage). "S2G", "SDT", "A2G" are abbreviations for [28, 47, 59]. Our methods show higher participants' overall preference, and also outperform state-of-the-art methods in the other aspects, especially the diversity. The error bar is calculated by standard deviation.

5.1.2 Network. We use the TCN [7] with skip-connection [32] as audio encoder. The final content and rhythm features are 128 and 8-dimensional, respectively. For each translation and decoder network, we use the structure of LSTM [34] and MLP [41].

5.1.3 Training. We train our model into three stages: Firstly, train rhythm and content classifier using the Adam [39] optimizer with learning rate (LR) 3e-4 for 200K iterations. Secondly, we train the other part exclude IN and set the LR to 5e-4 for 120k iterations, we directly feed all content features candidates from diversity network to the decoder. Finally, we train the Inclusion network with LR 3e-5 and adjust the rest to 3e-6 for 30K iterations to get the final training results. Besides, the gradient normalization is set to 150 and 50 to avoid the gradient exploding of the LSTM layers on Trinity and S2G-Ellen datasets, respectively. The final $\lambda_1, \lambda_2, and \lambda_3$ and λ_4 are set to 20, 500, 1 on Trinity and 20, 100, 1 on S2G-Ellen.

5.2 Metrics

In this section, we discuss evaluation metrics.

1) Distribution Distance: We use **Frechet Gesture Distance (FGD)** [69] to evaluate the similarity between synthesized motion and ground truth. Similar to perceptual loss in image generation tasks, FGD computes the distance between latent features extracted by a pretrained network, as Eq. 8,

$$\mathrm{FGD}(\mathbf{m}, \hat{\mathbf{m}}) = \left\|\mu_r - \mu_g\right\|^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r \Sigma_g\right)^{1/2}\right), \tag{8}$$

where μ_r and Σ_r are the first and second moments of the latent features distribution z_r of real human gestures $\mathbf{m}$, and μ_g and Σ_g are the first and second moment of the latent features distribution z_g of generated gestures $\hat{\mathbf{m}}$. We use an LSTM-based autoencoder as the pretrained network.

2) Diversity: Following [47], we measures **Diversity** by average L1 distance from different N motion clips, as

$$L1div. = \frac{1}{2N(N-1)} \sum_{t=1}^{N} \sum_{j=1}^{N} \left\|p_t^i - \hat{p}_t^j\right\|_1, \tag{9}$$

	Body			Hands		
	FGD ↓	Align. ↑	Diversity ↑	FGD ↓	Align. ↑	Diversity ↑
Speech2Gesture [28]	21.47	0.146	5.99	42.17	0.053	0.048
StyleGestures [4]	22.67	0.072	3.79	56.76	0.036	0.041
SpeechDrivenTemplate [59]	17.13	0.153	7.02	44.35	0.056	0.053
Audio2Gestures [47]	16.97	0.147	6.52	44.96	0.051	0.045
DisCo (Ours)	**14.08**	**0.158**	**7.39**	**38.63**	**0.059**	**0.061**

Table 1: Evaluation on Trinity Dataset. DisCo outperforms previous state-of-the-art algorithms on the FGD, diversity, and alignment metrics, demonstrating that DisCo generates more diverse gestures without sacrificing speech-motion synchrony. Due to the uncertainty in the training results of the generative models, we train the given models five times and report their average scores.

	FGD ↓	Align. ↑	Diversity ↑
Speech2Gesture [28]	4.79	0.059	0.89
StyleGestures [4]	6.03	0.038	0.67
SpeechDrivenTemplate [59]	3.67	0.061	0.83
Audio2Gestures [47]	3.57	0.054	0.81
DisCo (Ours)	**3.03**	**0.065**	**0.91**

Table 2: Evaluation on S2G-Ellen dataset. We report the average score for body and hands for comparison. We keep the same evaluation dataset as [47], the results show our method outperforms previous methods in 2D gesture generation dataset.

where p_t is the position of joints in frame t, we split the generated motions into equal-lengthed non-overlapping motion clips by 60 frames and calculate diversity with $N = 50$. The resulting diversity of motions is effective when achieving similar FGD distances.

3) Alignment: There is no objective audio beat in the speech audio, unlike the music-conditioned dance generation. In this paper, similar as [48] we take the RMS onset of speech as the potential speech audio beat and calculate the speech-motion synchrony by **Align.**, as Eq 10,

$$\text{Align.} = \frac{1}{M}\sum_{b_M \in M} \exp\left(-\frac{\min_{b_S \in S}\|b_M - b_S\|^2}{2\sigma^2}\right), \qquad (10)$$

where M, S is the set of gesture beat and audio beat, respectively. We also require the generated motion to have a low FGD distance as an additional constraint.

5.3 Comparison with state-of-the-art methods

We compare our method with four recent representative state-of-the-art methods, including Speech2Gesture [28], StyleGestures [4], SpeechDrivenTemplate [59] and Audio2Gestures [47]. The final output layer of Speech2Gesture has been adjusted to predict 3D joint rotations. We trained two models for all methods with only body joints and body + hands joints for Trinity dataset. For body parts, we keep the best score reported Audio2Gestures [47]. Only for the body + hands version. All methods are trained with this filtered data. Subjectively, as Figure 1 and Figure 7, DisCo generates diverse motions without sacrificing speech-motion synchrony, *e.g.*, key-frame two.

5.3.1 Quantitative Results. As listed in Table 1 and Table 2 results, our method outperforms previous state-of-the-art algorithms on the FGD, diversity and alignment metrics, demonstrating that distribution balanced training can improve the diversity of generated motion while maintaining synchrony.

	Re	DM	DN	IN	FGD ↓	Align. ↑	Diversity ↑
Baseline					21.79	0.122	4.61
	✓				22.19	0.093	6.63
		✓			20.53	0.157	5.01
	✓	✓			17.61	0.154	7.12
	✓	✓	✓		15.13	0.155	**7.41**
DisCo	✓	✓	✓	✓	**14.08**	**0.158**	7.39

Table 3: Ablation Study on Trinity Dataset. Re, DM, MS, IN are abbreviations of Resamping, Disentangled Motion, Diversity Network and Inclusion Network, respectively. Results show the effectiveness of each proposal. Resampling will increase the motion diversity significantly but decrease the synchrony, motion disentanglement makes it possible to integrate the advantages of resampling.

5.3.2 User Studies. For user study, we provide each user with 30 generated video clips with a length of 20s. Each video contains four baselines and our method. After watching each video, users need to choose which result is more physic-correctness, diverse, aligned with the speech, and the overall preference. 42 subjects (15 females and 27 males, excluding authors) participated in the study. As shown in Figure 6, our method outperforms baseline methods in diversity with a large margin (36.4% vs. 24.8%). For physic-correctness and alignment, we also perform better than the highest baseline (25.4% vs. 23.1% and 26.7% vs. 21.9%). The results of overall preference (33.6% vs. 23.6%) demonstrate improving the quality of low frequent motion will result in higher human preference.

5.4 Ablation Study

For our final method, we gradually control one of the variables to perform ablation experiments. Our baseline model is CNN + LSTM + MLP structure. The performance of baseline is shown in Table 3.

5.4.1 Limitation of Resampling. We first prove that simple resampling has limitations based on class-level resampling with L2 regularization. The result in Table 3 shows that lower resampling can increase the diversity of motion content but reduce the speech-motion synchrony. Resampling causes the generated motions to deviate from the original data distribution, leading to a higher FGD. Besides, due to the destruction of the balance of rhythm, the model

Figure 7: Subjective Comparisons. Each sub-figure summarizes the 40s generation motion from Trinity dataset sequence 30, our method generates more diverse gestures comparing with previous methods. From *Left* to *Right*: results from Speech2Gesture [28], StyleGestures [4], SpeechDrivenTemplate [59], Audio2Gestures [47], Ours and Ground Truth. (video results are in Supp.)

tends to fit certain rhythm classes, leading to a decrease in synchrony. This result proves that resampling motion contents cannot improve the overall performance of the model. Other baseline models use resampling will get the same conclusion.

5.4.2 Effectiveness of Disentanglement. Our framework disentangles the motion latent features into implicit content and rhythm features. Table 3 shows that by disentangling motions, the model can ensure diversity while increasing synchrony. After disentangling motion, we can use different sampling strategies in different branches to integrate the advantage of resampling. Results show the combination of resampling and disentanglement achieves higher diversity while decreasing the synchrony slightly as the current model has not considered the inherent mapping between rhythm and content explicitly.

5.4.3 Effectiveness of Diversity Networks. We present a framework that learns the mapping between motion rhythm and content features considering the content diversity. We design DIN where DN and IN networks are for improving the content diversity and modelling the inherent mapping, respectively. As shown in Table 3, we evaluate the effectiveness of the DN by averaging the score of generated motion from the uniform and reverse candidates. The results demonstrate that by setting up multiple sub-networks in the content branch, our model further enhances the ability to fit the unbalanced content distribution.

5.4.4 Effeciveness of The Inclusion Network. Inclusion network models the inherent mapping between rhythm and content features. Besides, it increases the multimodality and controllability of the model. We study the multimodality and controllability of generated motion from the same speech audio, and our experiments are based on the first 60s of tenth sequence on the Trinity [24] dataset. As shown in Figure 8, we test three content-candidate selection strategies: all from uniform, original IN's outputs and all from reverse sampling. Experimental results show that the score of diversity and alignment depends on the selected candidates. Thus, IN plays a role to balance diversity and alignment. Furthermore, the results demonstrate our method allows the generated motion content candidates to be manually selected by users, *i.e.*, by replacing the outputs of IN with a predefined value. The evaluation of co-speech gestures is very subjective, and users also have different requirements for the frequency of different motions according to different scenes. This proves to be a certain advantage for our model's controllability.

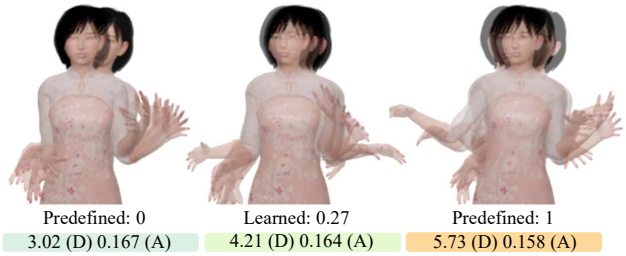


Figure 8: Visualization of controllability. 0, 1 represents selecting all candidates from the uniform and reverse sampling network, respectively, which ignores IN's outputs. In contrast, 0.27 is the probability of selecting reverse sampling candidates calculated by averaging the outputs of IN. *Bottom*: Diversity score for 60 seconds.

5.5 Limitation

One limitation of DisCo is that the speech-motion synchrony does not explicitly include the semantic relevance between speech and motion. People can understand the semantic relevance of motions and speech since they have prior knowledge of the meaning of the text corresponding to the speech. Therefore, it is not easy to model a network that can extract the inherent semantic relevance through this data-driven model using only speech as input. Instead, the inherent semantic relevance should be included in the framework by transferring prior knowledge to specific speech and text, such as word embedding.

6 CONCLUSION

We present DisCo and Diversity-and-Inclusion Network to improve the diversity of the generated motions by distribution balanced training, which demonstrates that balancing the data distribution can significantly improve the model's performance. In the future, other data balancing/data augmentation methods, *e.g.*, MixUp [70] based methods, could be also explored in this field.

7 ACKNOWLEDGMENTS

This work was conducted during Haiyang Liu and Zihao Zhu's internship at Digital Human Lab, Tokyo Research Center, Huawei Technologies.

REFERENCES

[1] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. 2020. Unpaired motion style transfer from video to animation. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 64–1.

[2] Kfir Aberman, Rundi Wu, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. 2019. Learning character-agnostic motion for motion retargeting in 2D. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–14.

[3] Chaitanya Ahuja, Dong Won Lee, Yukiko I Nakano, and Louis-Philippe Morency. 2020. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *European Conference on Computer Vision*. Springer, 248–265.

[4] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 487–496.

[5] Andreas Aristidou, Daniel Cohen-Or, Jessica K Hodgins, Yiorgos Chrysanthou, and Ariel Shamir. 2018. Deep motifs and motion signatures. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–13.

[6] Andreas Aristidou, Daniel Cohen-Or, Jessica K Hodgins, and Ariel Shamir. 2018. Self-similarity analysis for motion capture cleaning. In *Computer graphics forum*, Vol. 37. Wiley Online Library, 297–309.

[7] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271* (2018).

[8] Kirsten Bergmann, Volkan Aksu, and Stefan Kopp. 2011. The relation of speech and gestures: Temporal synchrony follows semantic synchrony. In *Proceedings of the 2nd Workshop on Gesture and Speech in Interaction (GeSpIn 2011)*.

[9] Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, and Dinesh Manocha. 2021. Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2027–2036.

[10] Elif Bozkurt, Shahriar Asta, Serkan Özkul, Yücel Yemez, and Engin Erzin. 2013. Multimodal analysis of speech prosody and upper body gestures using hidden semi-markov models. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 3652–3656.

[11] Elif Bozkurt, Yücel Yemez, and Engin Erzin. 2016. Multimodal analysis of speech and arm motion for prosody-driven synthesis of beat gestures. *Speech Communication* 85 (2016), 29–42.

[12] Elif Bozkurt, Yücel Yemez, and Engin Erzin. 2020. Affective synthesis and animation of arm gestures from speech prosody. *Speech Communication* 119 (2020), 1–11.

[13] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks* 106 (2018), 249–259.

[14] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence* 43, 1 (2019), 172–186.

[15] Justine Cassell. 2000. Embodied conversational interface agents. *Commun. ACM* 43, 4 (2000), 70–78.

[16] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*. 413–420.

[17] Agisilaos Chartsias, Giorgos Papanastasiou, Chengjia Wang, Scott Semple, David E Newby, Rohan Dharmakumar, and Sotirios A Tsaftaris. 2020. Disentangle, align and fuse for multimodal and semi-supervised image segmentation. *IEEE transactions on medical imaging* 40, 3 (2020), 781–792.

[18] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[19] Chung-Cheng Chiu and Stacy Marsella. 2011. How to train your avatar: A data driven approach to gesture generation. In *International Workshop on Intelligent Virtual Agents*. Springer, 127–140.

[20] Chung-Cheng Chiu and Stacy Marsella. 2014. Gesture generation with low-dimensional embeddings. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. 781–788.

[21] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9268–9277.

[22] Emily L Denton et al. 2017. Unsupervised learning of disentangled representations from video. *Advances in neural information processing systems* 30 (2017).

[23] Chris Drummond, Robert C Holte, et al. 2003. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, Vol. 11. Citeseer, 1–8.

[24] Ylva Ferstl and Rachel McDonnell. 2018. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 93–98.

[25] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2020. Adversarial gesture generation with realistic gesture phasing. *Computers & Graphics* 89 (2020), 117–130.

[26] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2021. ExpressGesture: Expressive gesture generation from speech through database matching. *Computer Animation and Virtual Worlds* (2021), e2016.

[27] Yuan Gao, Xingyuan Bu, Yang Hu, Hui Shen, Ti Bai, Xubin Li, and Shilei Wen. 2018. Solution for large-scale hierarchical object detection datasets with incomplete annotation and data imbalance. *arXiv preprint arXiv:1810.06208* (2018).

[28] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3497–3506.

[29] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. 2021. Learning Speech-driven 3D Conversational Gestures from Video. *arXiv preprint arXiv:2102.06837* (2021).

[30] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*. Springer, 878–887.

[31] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. 2018. Evaluation of speech-to-gesture generation using bi-directional LSTM network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 79–86.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[33] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. 2020. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–14.

[34] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[35] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. 2016. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5375–5384.

[36] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. 2019. Decoupling Representation and Classifier for Long-Tailed Recognition. In *International Conference on Learning Representations*.

[37] Kenan Kasarcı, Elif Bozkurt, Yücel Yemez, and Engin Erzin. 2016. Real-time speech driven gesture animation. In *2016 24th Signal Processing and Communication Application Conference (SIU)*. IEEE, 1917–1920.

[38] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. 2017. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems* 29, 8 (2017), 3573–3587.

[39] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[40] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Sixteenth annual conference of the international speech communication association*.

[41] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012).

[42] Taras Kucherenko, Dai Hasegawa, Naoshi Kaneko, Gustav Eje Henter, and Hedvig Kjellström. 2021. Moving fast and slow: Analysis of representations and post-processing in speech-driven automatic gesture generation. *International Journal of Human–Computer Interaction* (2021), 1–17.

[43] 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.

[44] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. 2020. Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision* 128, 10 (2020), 2402–2417.

[45] Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. 2010. Gesture controllers. In *ACM SIGGRAPH 2010 papers*. 1–11.

[46] Sergey Levine, Christian Theobalt, and Vladlen Koltun. 2009. Real-time prosody-driven synthesis of body language. In *ACM SIGGRAPH Asia 2009 papers*. 1–10.

[47] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. 2021. Audio2Gestures: Generating Diverse Gestures from Speech Audio with Conditional Variational Autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11293–11302.

[48] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13401–13412.

[49] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. 2019. Selective kernel networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 510–519.

[50] Jongin Lim, Hyung Jin Chang, and Jin Young Choi. 2019. PMnet: Learning of Disentangled Pose and Movement for Unsupervised Motion Retargeting.. In *BMVC*. 136.

[51] Haiyang Liu and Jihan Zhang. 2021. Improving Ultrasound Tongue Image Reconstruction from Lip Images Using Self-supervised Learning and Attention Mechanism. *arXiv preprint arXiv:2106.11769* (2021).

[52] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022. BEAT: A Large-Scale Semantic and Emotional Multi-Modal Dataset for Conversational Gestures Synthesis. *arXiv preprint arXiv:2203.05297* (2022).

[53] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. 2019. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2537–2546.

[54] JinHong Lu, TianHang Liu, ShuZhuang Xu, and Hiroshi Shimodaira. 2021. Double-DCCCAE: Estimation of Body Gestures From Speech Waveform. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 900–904.

[55] Soroosh Mariooryad and Carlos Busso. 2012. Generating human-like behaviors using joint, speech-driven models for conversational agents. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 8 (2012), 2329–2340.

[56] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, Vol. 8. Citeseer, 18–25.

[57] S Ozkul, Elif Bozkurt, Shahriar Asta, Yücel Yemez, and Engin Erzin. 2012. Multimodal analysis of upper-body gestures, facial expressions and speech. In *Proceedings of the 4th International Workshop on Corpora for Research on Emotion Sentiment and Social Signals*.

[58] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7753–7762.

[59] Shenhan Qian, Zhi Tu, Yihao Zhi, Wen Liu, and Shenghua Gao. 2021. Speech Drives Templates: Co-Speech Gesture Synthesis with Learned Templates. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11077–11086.

[60] Gregory Rogez and Cordelia Schmid. 2016. MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild. In *Advances in Neural Information Processing Systems (NIPS)*. Curran Associates, 3108–3116.

[61] Li Shen, Zhouchen Lin, and Qingming Huang. 2016. Relay backpropagation for effective learning of deep convolutional neural networks. In *European conference on computer vision*. Springer, 467–482.

[62] Alon Shoshan, Nadav Bhonker, Igor Kviatkovsky, and Gerard Medioni. 2021. Gan-control: Explicitly controllable gans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14083–14093.

[63] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[64] Kenta Takeuchi, Souichirou Kubota, Keisuke Suzuki, Dai Hasegawa, and Hiroshi Sakuta. 2017. Creating a gesture-speech dataset for speech-based automatic gesture generation. In *International Conference on Human-Computer Interaction*. Springer, 198–202.

[65] Luan Tran, Xi Yin, and Xiaoming Liu. 2017. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1415–1424.

[66] Bowen Wu, Carlos Ishi, Hiroshi Ishiguro, et al. 2021. Probabilistic human-like gesture synthesis from speech using GRU-based WGAN. In *GENEA: Generation and Evaluation of Non-verbal Behaviour for Embodied Agents Workshop 2021*.

[67] Bowen Wu, Chaoran Liu, Carlos Toshinori Ishi, and Hiroshi Ishiguro. 2021. Modeling the conditional distribution of co-speech upper body gesture jointly using conditional-GAN and unrolled-GAN. *Electronics* 10, 3 (2021), 228.

[68] Fanglu Xie, Go Irie, and Tatsushi Matsubayashi. 2021. Disentangling Subject-Dependent/-Independent Representations for 2D Motion Retargeting. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4200–4204.

[69] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–16.

[70] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.

[71] Bo Zheng, Yibiao Zhao, Joey C Yu, Katsushi Ikeuchi, and Song-Chun Zhu. 2013. Beyond point clouds: Scene understanding by reasoning geometry and physics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3127–3134.

[72] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. 2020. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9719–9728.