Continual Learning for Personalized Co-Speech Gesture Generation

Chaitanya Ahuja[1], Pratik Joshi[1], Ryo Ishii[2] & Louis-Philippe Morency[1]
[1]Language Technologies Institute, CMU & [2]NTT Human Informatics Laboratories
ahujachaitanya@gmail.com, pratikmjoshi123@gmail.com, ryoct.ishii@ntt.com, morency@cs.cmu.edu

Abstract

Co-speech gestures are a key channel of human communication, making them important for personalized chat agents to generate. In the past, gesture generation models assumed that data for each speaker is available all at once, and in large amounts. However in practical scenarios, speaker data comes sequentially and in small amounts as the agent personalizes with more speakers, akin to a continual learning paradigm. While more recent works have shown progress in adapting to low-resource data, they catastrophically forget the gesture styles of initial speakers they were trained on. Also, prior generative continual learning works are not multimodal, making this space less studied. In this paper, we explore this new paradigm and propose C-DiffGAN: an approach that continually learns new speaker gesture styles with only a few minutes of per-speaker data, while retaining previously learnt styles. Inspired by prior continual learning works, C-DiffGAN encourages knowledge retention by 1) generating reminiscences of previous low-resource speaker data, then 2) crossmodally aligning to them to mitigate catastrophic forgetting. We quantitatively demonstrate improved performance and reduced forgetting over strong baselines through standard continual learning measures, reinforced by a qualitative user study that shows that our method produces more natural, style-preserving gestures. Code and videos can be found at https://chahuja.com/cdiffgan

1. Introduction

Human communication technologies for both verbal (e.g. spoken language) and nonverbal (e.g. co-speech gestures) have seen significant improvements which have made generative models for human communication more natural and semantically relevant [32]. Advancements in speech-based personal assistants such as Cortana, Alexa, Siri, and more recently in text based conversational agents such as chatGPT[1] [34], Meena [1], and Xiaoice [48] have paved the way for embodied personal assistants. As embodied agents have both verbal and nonverbal communicative channels, one technical challenge is to be able to generate personalized visual co-speech gestures (i.e. nonverbal) using spoken language (i.e. verbal) [46, 2]. Previous works in co-speech gesture generation learn unique speaker styles in a static paradigm, where training is done on a fixed dataset consisting of data for all speakers available all at once. However, in practical scenarios of embodied agents learning in the wild, the agent would receive small amounts of training data sequentially - also known as a continual learning paradigm. The main goal of the paper is to learn a unified co-speech gesture generation model with the ability to generate gestures in multiple different styles (see Figure 1). This goal is achieved in a challenging and practical continual learning setting where the model only has access to limited training data.

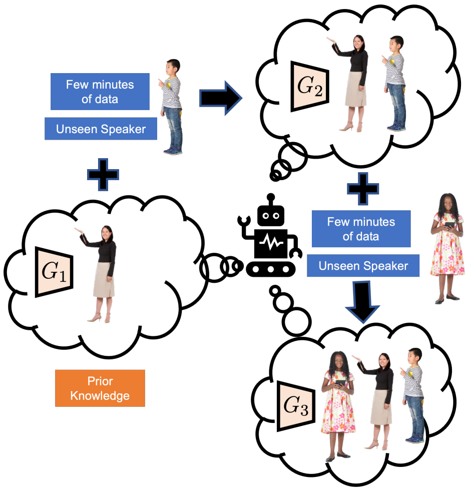


Figure 1: Overview of the continual learning paradigm of co-speech gesture personalization task. We start with a source model G_1 pre-trained on a source speaker. We personalize G_1 to target models G_2, G_3 and so on in a sequential manner using low-resource data (∼10 minutes) for each of the target speaker.

This problem setting brings a unique technical challenge, typically not studied in generative continual learning settings: *crossmodal catastrophic forgetting*. Due to the crossmodal

[1] https://openai.com/blog/chatgpt

nature of our task, crossmodal catastrophic forgetting refers to the forgetting of the crossmodal grounding relationships between input spoken language modalities and output gesture modality of speakers that the model interacted with earlier. For example, consider a virtual agent that has the knowledge of generating gestures for one speaker. As it starts to interact with the new speakers in the world, it experiences new crossmodal grounding relationships between gestures and spoken language. While the gestures are heavily dependent on the spoken language, they are also heavily influenced by the new speakers' idiosyncrasies. A practical challenge here is that these interactions are often short, hence creating a low-resource setting for this agent. Another challenge is that the agent sequentially receives new data as it interacts with multiple new speakers over time. The goal of this virtual agent is to learn to generate personalized gestures for many different speakers without forgetting the crossmodal grounding of the speakers that it interacted with earlier in its life. The agent should be able to achieve these goals with the practical constraints of low-resource data, limited storage space, and faster training.

In this paper, we propose an approach, named C-DiffGAN, that can efficiently personalize co-speech gesture generation models from a high-resource source speaker to multiple low-resource target speakers. To the best of our knowledge, this is the first approach that is able to learn a personalized model for multiple speakers with only 2-10 minutes each of speaker data (i.e. as opposed to 10 hours [11, 2, 20, 14]) in a continual learning setting. Our C-DiffGAN approach requires access to only 2-10 minutes of the input data (i.e. language and speech) for the prior speakers and 2-10 minutes of paired data (i.e. language, speech, and gestures) for the new speaker. For continually learning new speakers' behaviors while not forgetting the prior speakers', C-DiffGAN follows two steps: First, it directly identifies shifts in crossmodal grounding relationships along with the shifts in the output domain from the pretrained source model. Based on these identified distribution shifts, C-DiffGAN updates a few necessary parameters in a single layer of the source model, allowing efficient adaptation with low resources. Second, it utilizes the low-resource input data of prior speakers in tandem to prevent the model from drifting from the prior speakers' crossmodal grounding, hence preventing crossmodal catastrophic forgetting. This is done via a novel proposed objective term $\mathcal{L}_{ccf}$. Our experiments study the effectiveness of our C-DiffGAN approach on a diverse publicly available dataset and is substantiated through a myriad of quantitative and qualitative studies, which show that our proposed methodology significantly outperforms prior approaches for low resource continual learning of nonverbal grounding and personalization of gesture generation models.

2. Related Work

Co-speech gesture generation Gesture generation is the task of imbibing nonverbal communicative behaviors that humans use [12, 21] into virtual agents, making them more engaging and informative [18]. Co-speech gesture generation [33, 24] involves generating gestures accompanying speech utterances. Data-driven approaches [8, 19] here have shown to produce more diverse and natural gestures [33] than rule-based techniques [26, 37]. Prior work that posits that co-speech gestures are idiosyncratic [28, 45] has motivated gesture synthesis conditioned on speaker-specific style. Ginosar *et al*. [11] trains speaker-specific models that adversarially discriminate to produce style-consistent gestures. Ahuja *et al*. [4] generates speaker-specific gestures by learning style embeddings. Most of these works, however, require large amounts (5-10 hrs) of initial training data.

Gesture generation with low-resource speaker data It is practically infeasible to collect hours of multimodal speaker data, making the low-resource setting crucial to gesture generation. Prior works focus primarily on **low-resource adaptation**, where a previously trained model transfers to the style of low-resourced new speaker data. Methods include: 1) pre-training with high resource data, then a cross-modally grounded adaptation phase on low-resource data [3], 2) data augmentation [43], 3) zero-shot style adaptation via a learnt style encoder [7]. While effective in creating a speaker-specific model, **these models catastrophically forget** [27, 36, 9] **the source speakers' style**, and only generate gestures of the target style post adaptation. In contrast, we investigate the scenario of **low-resource continual learning**, where the aim is to incorporate the new speaker styles without forgetting previously learnt styles. Recent works in zero-shot style control [10] cannot incrementally incorporate new speaker information, making them strongly dependent on the diversity (number of speaker styles) and quality of source training data. In our paper, we learn new styles even when source training data is of a single speaker's style.

Continual Learning for Generative Settings Continual learning previously has been mainly applied for discriminative tasks [23, 25]. The aim is to learn classification tasks sequentially, without degrading earlier task performance [27, 36, 9]. Generative continual learning involves incremental class-conditional generation [35], where different classes of, say images (e.g. classes are digits 0-9 for MNIST [22], or locations for LSUN [47]) are sequentially trained on, and the model must finally be able to conditionally generate images of all classes. Prior works in generative continual learning involve using "replay", where a buffer of datapoints of previous classes are stored [6] or generated [39, 44] to augment the current training sequence. Other techniques use elastic

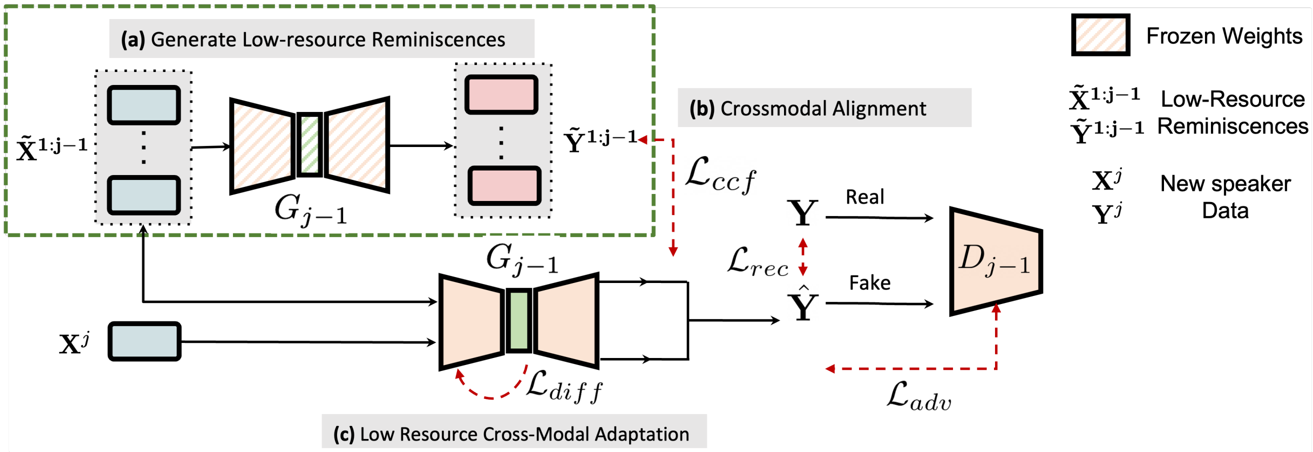

(a) Generate Low-resource Reminiscences
$\tilde{\mathbf{X}}^{1:j-1}$
G_{j-1}
$\tilde{\mathbf{Y}}^{1:j-1}$
(b) Crossmodal Alignment
$\mathcal{L}_{ccf}$
$\mathbf{Y}$
Real
$\mathcal{L}_{rec}$
$\hat{\mathbf{Y}}$
Fake
D_{j-1}
G_{j-1}
$\mathbf{X}^j$
$\mathcal{L}_{diff}$
$\mathcal{L}_{adv}$
(c) Low Resource Cross-Modal Adaptation
Frozen Weights
$\tilde{\mathbf{X}}^{1:j-1}$ $\tilde{\mathbf{Y}}^{1:j-1}$ Low-Resource Reminiscences
$\mathbf{X}^j$ $\mathbf{Y}^j$ New speaker Data

Figure 2: **C-DiffGAN: Overview of the key components**. (a) and (b) constitute the steps in mitigating crossmodal catastrophic forgetting in Sec. 4.1, and (c) refers to the low-resource adaptation in Sec. 4.2.

weight consolidation [17] on GANs to slow down gradient updates that cause forgetting [38]. However, these works are largely unimodal. To our knowledge, we are the first to tackle continual learning of gesture synthesis.

Prior works similary focus on pre-training a source model on large source data, then adapting it to a low-resource setting. [30, 41] introduce new parameters in the model, whereas [16, 42] fine-tunes the complete model on the target data or to specific layers or modules are applied [31, 29]. wang2020minegan, Li2020FewshotIG utilize importance sampling to transform the original latent space of the source to a space which is more relevant to the target. While this approach can be effective when the source distribution and the target distributions share support, it may not be well-generalizable when their supports are disjoint. To address this concern, ojha2021few introduces a contrastive learning approach to preserve the similarities and differences in the source, and then adapting to the target domain. These methods focus on adapting only the output domain of unimodal generative models (i.e. generate one modality with noise or a small set of discrete classes as the input). However, we believe that for crossmodal generative modeling tasks, we need to explicitly model complex relationships between the input modalities and the generated output modality, both of which have a spatial and/or temporal structure.

3. Problem Statement

We are given a set of training datasets for each speaker (or experiences) $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_M\}$. Here $\mathcal{S}_j = \{(j, \mathbf{X}_i^j, \mathbf{Y}_i^j)\}_{i=1}^{N^j}$ is paired data of input language and speech $\mathbf{X}_i^j$ and output sequence of body poses $\mathbf{Y}_i^j$ for speaker j. A goal here is to learn a sequence of gesture generation models, represented as Generator-Discriminator pairs [13], that are adapted through $(G_1, D_1) \rightarrow (G_2, D_2) \rightarrow \ldots \rightarrow (G_M, D_M)$ by sequentially training on speakers 1 through M in a continual learning setting. Here G_j is a model that is able to generate personalized gestures of speakers 1 through j that are driven by both language and speech. D_j is a model that can distinguish between real and fake sequence of gestures for the jth experience. In addition to sequential training, the number of training samples N^j is often small in practical scenarios which emulates a low-resource continual learning setting. In this setup, only the final version of the model G_M is deployed.

4. Continual-DiffGAN

We propose a new approach, Continual-DiffGAN (or C-DiffGAN), that learns a target model G_j for speaker (or experience) j while not forgetting the crossmodal grounding knowledge of speakers 1 through $j-1$ via the low-resource adaptation from the source model G_{j-1}. This approach is sequential and applies for speakers 1 through M. C-DiffGAN has three components: (1) *Generative Modeling* (Section 4.3) which learns to generate gestures that are driven by input spoken language along with personalizing to multiple speakers through a loss function $\mathcal{L}_{gen}$, (2) *Crossmodal Adaptation* (Section 4.2) which learns to adapt from a source model to a target model through a loss function $\mathcal{L}_{diffgan}$, and (3) *Crossmodal Catastrophic Forgetting* (Section 4.1) that prevents the new target model G_j from catastrophically forgetting the crossmodal grounding knowledge earlier speakers (i.e. 1 through $j-1$) through a loss function $\mathcal{L}_{ccf}$. Optimization of the combined loss function describes the complete model,

$$G_j^* = \mathbb{E}_{\mathcal{S}_j^{exp}} \underset{G_{j-1}}{\operatorname{argmin}} \max_{D_{j-1}} \mathcal{L}_{gen} + \mathcal{L}_{diffgan} + \mathcal{L}_{ccf} \qquad (1)$$

where $\mathcal{S}_j^{exp}$ is a low-resource training dataset through the jth experience. A diagram of the model is shown in Fig. 2.

4.1. Crossmodal Catastrophic Forgetting

A key technical challenge in continual learning paradigms is the phenomenon of catastrophic forgetting. Neural networks typically forget previously learnt domain knowledge when they are fine-tuned on new experiences [27, 36, 9]. This challenge becomes even more complex as our task is generative crossmodal, which means that both input and output modalities are part of a large and continuous representation space. To tackle this *crossmodal catastrophic forgetting* challenge, one possible approach is to fine-tune the source model with previous experiences' training data [35] along with the new experience's training data. But this approach is not scalable, as the memory and computational footprint of the continual learning models will linearly increase with the number of experiences. To mitigate the scalability challenge along with *crossmodal catastrophic forgetting*, we propose a two step approach: (1) Reminiscence and (2) Crossmodal Alignment.

Low-resource Reminiscences We leverage the source model G_{j-1} to create an extended dataset that contains both real training data of the j^{th} experience as well as memory reminiscences of the previous experiences. It is defined as $\mathcal{S}_j^{exp} = \mathcal{S}_j \cup \tilde{\mathcal{S}}_{1:j-1}$, where $\tilde{\mathcal{S}}_{1:j-1} = \bigcup_{e=1}^{j-1} \tilde{\mathcal{S}}_e$. The set $\tilde{\mathcal{S}}_e$ for a given experience e is constructed using input examples of all previous experiences and the generated gestures by source model G_{j-1} as $\bigcup_{i=1}^{N_e}\{(e, \mathbf{X}_i^e, G_{j-1}(\mathbf{X}_i^e, j))\}$. Due to the low-resource and crossmodal nature of this task, N_j is typically quite small ranging from 2-10 minutes of language and speech inputs. This makes the utilization of memory reminiscences especially challenging when compared to memory replays [44] where unlimited amount of replay data can be generated due to the unimodal nature of the tasks. With the constructed memory reminiscence, the target model G_j now has access to the previous experiences.

Crossmodal Alignment These previous experiences are essential to providing the the target model G_j with information on crossmodal relationships between spoken language and gestures for speakers 1 through $j-1$, hence preventing catastrophic forgetting. We propose an approach where the target model is encouraged to remember this knowledge explicitly through the loss function $\mathcal{L}_{ccf}$ defined as follows:

$$\mathcal{L}_{ccf} = \mathbb{E}_{(k,\tilde{\mathbf{X}},\tilde{\mathbf{Y}})\in\tilde{\mathcal{S}}_{1:j-1}} \|\tilde{\mathbf{Y}} - G_{j-1}(\tilde{\mathbf{X}}, k)\|_2 \tag{2}$$

This loss function is part of the overall loss function defined in Equation 1 which can be optimized by adapting the source model G_{j-1}. This constraint preserves the model's ability to generate gestures in the style of earlier speakers while leaving room to support the generation of gestures in the style of a new speaker. Furthermore, this loss also acts in a regularization capacity reducing the need of a larger dataset for learning to generate gestures of new speaker.

4.2. Low-resource Crossmodal Adaptation

A practical challenge in the continual learning paradigm is availability of training data for new experiences, often making it a low-resource paradigm. While the Crossmodal Alignment loss ($\mathcal{L}_{ccf}$) acts as a regularizer, learning a crossmodal generative model is still quite challenging in a low-resource setting. It is achievable if we learn a target model G_j for the new experience by drawing on from the knowledge of a source model G_{j-1} that is well trained on high-resource data. We adopt the two step approach for low-resource crossmodal adaptation from [3]. First, the model learns to identify the crossmodal grounding shifts through a loss function $\mathcal{L}_{diff}$ and low-resource target data. Second, the target model is encouraged to shift the output domain distribution to be closer to that of the targets' through the use of a loss function $\mathcal{L}_{shift}$. The combined loss function $\mathcal{L}_{diffgan} = \mathcal{L}_{diff} + \mathcal{L}_{shift}$ encourages low-resource crossmodal adaptation of our C-DiffGAN model.

This low-resource crossmodal adaptation in tandem with the mitigation of crossmodal catastrophic forgetting allows us to learn a well-trained target model G_j that is able to generate gestures of speakers from the first j experiences. Now, the role of G_j switches to a well-trained source model which can now be used to learn the new target model G_{j+1} with a new experience, continuing the training cycle.

4.3. Generative Modeling

The final challenge is to generate plausible gestures that correspond to the input spoken language for multiple speakers. As a first step we encourage the model to generate correct gestures via a reconstruction loss for every experience j,

$$\mathcal{L}_{rec} = \mathbb{E}_{(e,\mathbf{X},\mathbf{Y})\in\mathcal{S}_j} \|\mathbf{Y} - G_{j-1}(\mathbf{X}, e)\|_1 \tag{3}$$

To alleviate the challenge of overly smooth generation caused by L1 reconstruction in Equation 3, we use the generated pose sequence $\hat{\mathbf{Y}} = G_{j-1}(\mathbf{X}, j)$ as a signal for the adversarial discriminator D_{j-1}. The discriminator tries to classify the true pose $\mathbf{Y}$ from the generated pose $\hat{\mathbf{Y}}$, while the generator G_{j-1} is encouraged to fool the discriminator by generating realistic poses. The adversarial loss is written as [13],

$$\mathcal{L}_{adv} = \mathbb{E}_{(e,\mathbf{X},\mathbf{Y})\in\mathcal{S}_j} \log D_{j-1}(\mathbf{Y}) + \log\left(1 - D_{j-1}\left(G_{j-1}(\mathbf{X}, e)\right)\right) \tag{4}$$

In order to handle multiple modes, we explicitly learn multiple sub-generators as part of the main generator, following from Mix-StAGE [4]. We use two losses from this prior

Model	Accuracy		Forgetting	
	FID↓	PCK↑	FID↓	PCK↓
C-DiffGAN (Ours)	**56.2**	**0.35**	**13.2**	**0.01**
↳ **w/o** $\mathcal{L}_{ccf}$	242.7	0.25	258.5	0.12
↳ **w/o** $\mathcal{L}_{diffgan}$	70.8	0.34	14.0	**0.01**
MixSTAGe (Low Resource) [4]	323.7	0.27	-	-
MixStAGe [4]	22.0	0.40	-	-

Table 1: Ablations to our C-DiffGAN model, compared with joint training MixSTAGe baselines.

work, $\mathcal{L}_{mix}$ and $\mathcal{L}_{id}$, the former mitigating mode collapse and the latter to handle style disentanglement.

The combination of the loss functions in this section is defined as the generative modeling loss $\mathcal{L}_{gen} = \mathcal{L}_{rec} + \mathcal{L}_{adv} + \mathcal{L}_{mix} + \mathcal{L}_{id}$, which is trained together with the crossmodal adaptation and crossmodal catastrophic forgetting losses in Equation 1.

5. Experiments

Dataset: We use the PATS dataset [2, 4, 11] as the benchmark to measure performance. It consists of around 10 hours of aligned body pose, audio and transcripts for each of the 25 speakers. For all experiments we use a sequence of five randomly chosen speakers (`oliver`, `maher`, `chemistry`, `ytch_prof` and `lec_evol`) that have visually different gesture styles and diverse linguistic content for our experiments. Unless specified otherwise, we start with a MixStAGE model trained on a high-resource dataset of the speaker `oliver`. For speakers in the subsequent experiences, we limit the training data to 2 or 10 minutes[2] of to simulate a low-resource continual learning setting.

Baseline Models: To the best of our knowledge, this crossmodal continual low-resource generative modeling task has not been explored before, hence there are no baselines that are directly associated with this task. We use a family of strong baselines most relevant to the challenges posed by this task: **DiffGAN** [3] performs adaptation from a high-resource trained source model to a target model in a low-resource co-speech gesture generation setting. **MeRGAN-JTR** and **MeRGAN-RA** [44] performs continual learning (or CL) for unimodal generative modeling in a high-resource setting without the need to explicitly store training examples of the previous experiences. We modify it to work in our crossmodal task in a low-resource setting. **Buffer Replay** explicitly saves training examples from the previous experiences in a buffer, which becomes a part of the subsequent training cycles. This strong baseline requires extra storage memory and training time making it less scalable. **MixStAGe** [4] learns a common model for multiple speaker styles by jointly training (or JT) for multiple speakers in a high-resource setting. A relevant baseline arises from jointly training **MixStAGe** in a **low-resource** setting. Additionally, we ablate the different component losses of **C-DiffGAN**.

Quantitative Measures: In a continual learning setting we typically measure two performance criteria [44]. (1) Average Final Accuracy measures the average metrics of the final model over the examples of all the experiences and is defined as

$$\text{Accuracy}(R, M) = \frac{1}{M}\sum_{j=1}^{M} R_{M,j}, \qquad (5)$$

where $R_{M,j}$ is a metric R measured on a model at experience M on data from experience j, and M is the total number of experiences. It is a measure of the average performance of the final model G_M over all 1 through M experiences. A goal of this task is to be able to consistently generate relevant and diverse gestures corresponding to the input spoken language for all speakers the model was exposed to. (2) Average Forgetting measures the extent to which the final model has forgotten about the prior experiences and is defined as:

$$\text{Forgetting}(R, M) = \frac{\delta}{M-1}\sum_{j=1}^{M-1} \max_{j \le e \le M-1} R_{e,j} - R_{M,j}, \qquad (6)$$

where $\delta = +1$ if a higher value of metric R denotes better performance and $\delta = -1$ if a lower value of metric R denotes better performance. A model is said to perform better in a continual learning setting when the Average Forgetting is lower. A goal of this task is to be able to retain the knowledge of generating both relevant and diverse gestures corresponding to the input spoken language, especially for speakers that were seen earlier in the training cycle.

Metrics: We use two metrics that are useful to measure performance for co-speech gesture generation tasks: (a) **Probability of Correct Keypoints (PCK)** [5, 40] measures relevance and timing of gestures with respect to spoken language. Here the PCK values are averaged over $\alpha = 0.1, 0.2$ as suggested in [11]. (2) **Fréchet Inception Distance (FID)** is the distance between distributions of generated and ground truth poses [15, 2] which is used to measure the diversity in the generated gestures.

Qualitative Study: We conduct a human perceptual study on Amazon Mechanical Turk (AMT) that judges the model

[2] We achieve similar baseline results over both training data sizes, shown in the supplementary, and show mainly results for 10 minutes here.

Amount of Data (minutes)	Training	Buffer Memory	Models	Average Final Accuracy		Average Forgetting	
				FID↓	PCK↑	FID↓	PCK↓
10	CL	✗	DiffGAN [3]	613.6	0.16	674.5	0.18
	CL	✗	MeRGAN-JTR [44]	316.9	0.24	355.0	0.13
	CL	✗	MeRGAN-RA [44]	494.1	0.23	561.1	0.15
	CL	✗	**C-DiffGAN (Ours)**	**56.2**	**0.35**	**13.2**	**0.01**
	CL	✓	Buffer Replay	61.6	0.37	2.2	0.01
Full	JT	✓	MixStAGe [4]	22.0	0.40	-	-

Table 2: Comparison of our C-DiffGAN with prior work for low-resource continual learning (CL) and joint training (JT) for crossmodal generative modeling. We use the Average Final Accuracy and Average Forgetting as the continual learning metrics for FID and PCK. Buffer Memory indicates if the method requires additional storage memory.

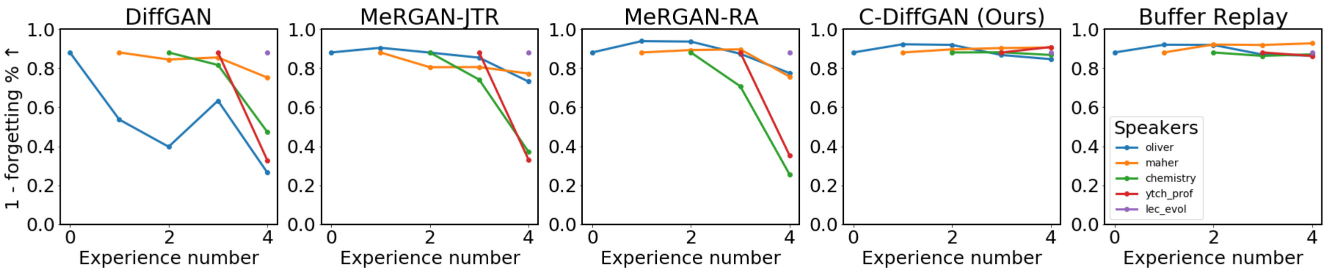


Figure 3: Comparing our C-DiffGAN with baselines on the measure of forgetting across number of experiences with 10 minutes of training data for each speaker. We plot (1-Forgetting)% for PCK for all speakers. Hence higher is better. The sudden dips of the measures for the older speakers indicate catastrophic forgetting and can be observed clearly in DiffGAN [2], MeRGAN-JTR [44] and MeRGAN-RA [44]. On the other hand, C-DiffGAN retains the performance over all the 5 experiences.

outputs on 5 criterion: Timing, Expressivity, Relevance, Naturalness, and Style with design principles adopted from [3]. More details of the study can be found in the supplementary section.

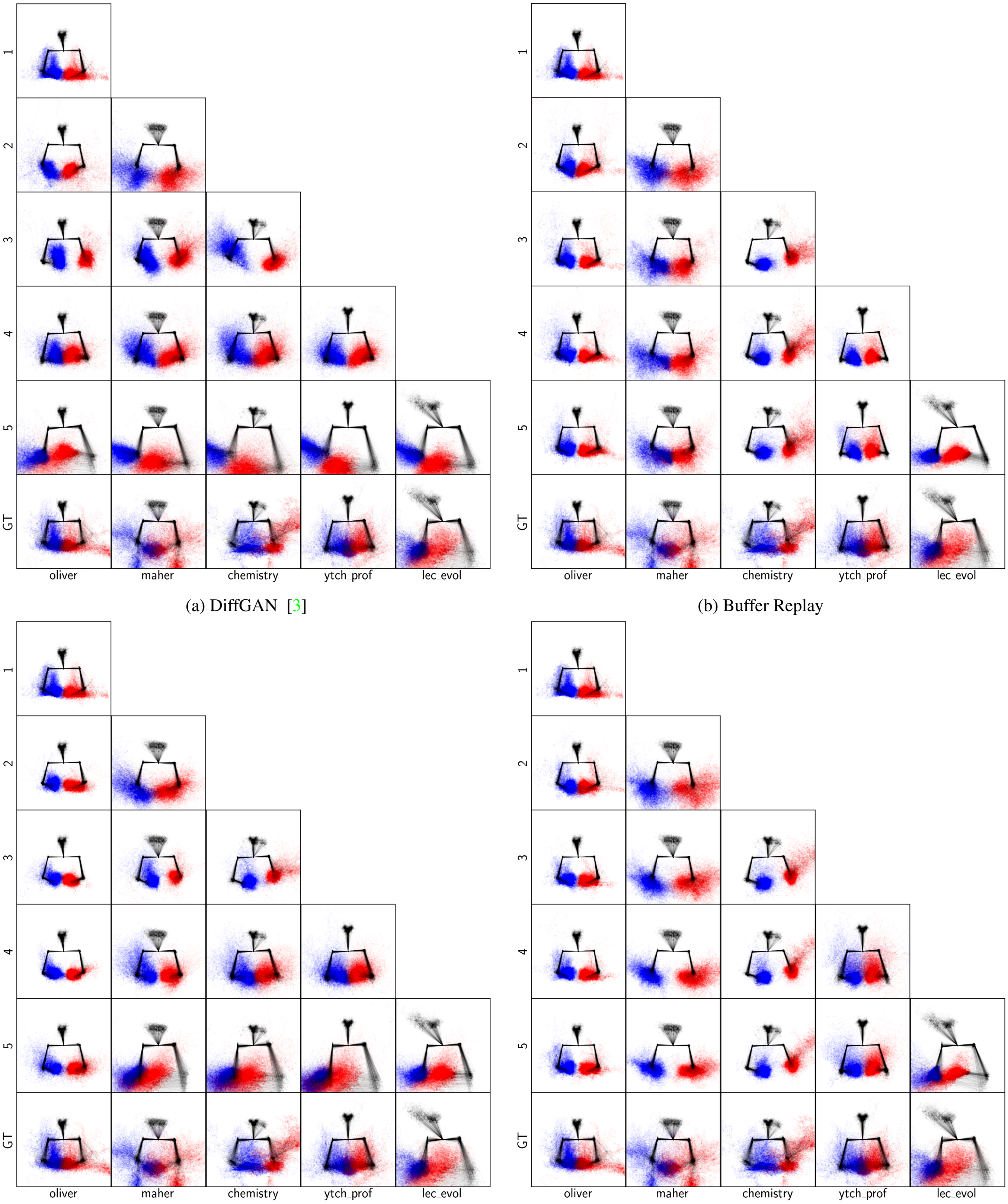
Implementation Details: For our pretrained source models, we use publicly available models by Ahuja *et al.* [2] for all experiments. We implement the continual learning baselines atop these source models. We trained all the baselines with the reported hyperparameters. All our models were trained for 4000 iterations with a batch size of 32. Either 2 minutes or 10 minutes of video recordings were used as the target data. To alleviate sample bias, each model was trained over three such randomly chosen target sets and quantitative metrics were averaged across these runs. We refer the readers to the supplementary materials for more implementation details.

6. Results and Discussion

In this section, we discuss both qualitative and quantitative experiments. To get a better idea of the generated videos, we refer the readers to the supplementary section.

Effective Crossmodal Adaptation With just 10 minutes of data per speaker, our C-DiffGAN model achieves significantly better PCK and FID Average Final Accuracy scores of **55.6** and **0.35** (Table 2), stark improvements over DiffGAN and MeRGAN baselines. We also yield a better FID score compared to Buffer Replay (61.6). This is indicative of the positive impact of crossmodal adaptation in learning new speaker personalizations in a continual learning setting. These results are consistent, irrespective of speaker sequence order (details in supplementary).

Reduced Catastrophic Forgetting Compared to the DiffGAN and MeRGAN baselines, C-DiffGAN reduces Average Forgetting by more than 15x for both PCK and FID; from around 400 and 0.15 to 12.7 and 0.01 (Table 2). This reinforces the benefit of low-resource reminiscences in retaining old speakers personalizations. In Figure 3, we observe that DiffGAN [3], MerGAN-JTR and MerGAN-RA [44] forget the crossmodal grounding relationships (i.e. PCK) by significant amounts over 1-2 new experiences. This is unlike our C-DiffGAN model which is completely able to retain the crossmodal grounding information in Figure



(a) DiffGAN [3]

(b) Buffer Replay

(c) MeRGAN-RA [44]

(d) C-DiffGAN (Ours)

Figure 4: Visual Histograms of generated gestures visually describe the distribution of hand gestures in space. Red and blue colors denote the left and right arms respectively. Each row represents the gesture distribution at the end of each continual learning experience. The final row denotes the true distribution of gestures for each speaker. The columns represent the speakers in the order they were exposed to the model in the continual learning paradigm. As we go from top to bottom in each column, we can see how the distribution of each speaker changes over the number of experiences. In Fig. (a) the baseline model DiffGAN [3] forgets the distribution of gestures for all the older speakers, while just retaining the information of the newest speaker. On the other hand, for our C-DiffGAN model in Fig. (d), we see that the model remembers the distribution of gestures for all the speakers through all the training experiences

Model	Timing	Expressivity	Relevance	Naturalness	Style
MixSTAGe (Low Resource) [4]	13.3 ± 3.4	14.4 ± 4.6	11.0 ± 3.1	12.3 ± 3.0	20.8 ± 4.8
DiffGAN [3]	15.6 ± 3.6	25.8 ± 3.0	14.8 ± 5.1	11.9 ± 2.9	30.6 ± 3.7
MeRGAN-JTR [44]	17.9 ± 3.8	**29.8 ± 4.8**	17.5 ± 3.1	14.0 ± 3.2	42.3 ± 3.3
MeRGAN-RA [44]	14.4 ± 4.6	24.8 ± 3.5	15.4 ± 3.9	10.2 ± 2.9	35.4 ± 3.3
C-DiffGAN (Ours)	**23.1 ± 4.0**	22.9 ± 3.2	**19.6 ± 3.2**	**18.5 ± 4.2**	**49.2 ± 3.2**
Buffer Replay	15.8 ± 3.7	19.6 ± 4.1	17.5 ± 3.4	14.0 ± 2.7	42.5 ± 4.7

Table 3: Human perceptual study comparing our model with prior work over five criterion measuring **Timing**, **Expressivity**, **Relevance**, **Naturalness** and **Style**. The preference scores along with 90% confidence intervals are reported for each model as compared to the ground truth gestures. Higher is better with 50% being the best possible score.

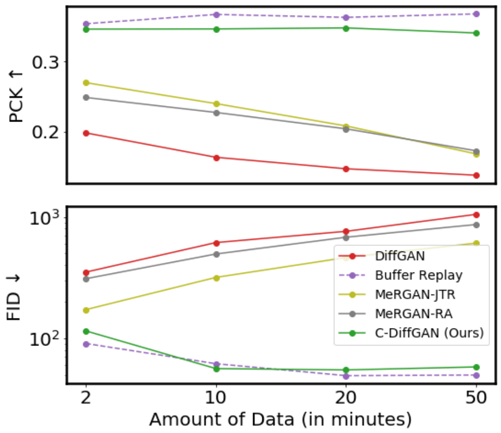


Figure 5: Trends of Average Final Accuracy (FID and PCK) vs Amount of Training Data for our C-DiffGAN when compared to other baselines. Lower is better for FID and higher is better for PCK.

3. The amount of knowledge lost could potentially be attributed to crossmodal grounding being a harder challenge and could benefit from more examples. This trend is qualitatively corroborated in the visual pose histograms of the last two rows of Figure 4d. Here, in row 5, DiffGAN [3] (Fig 4a) and MeRGAN-RA [44] (Fig 4c) forget the gesture styles of previous speakers. The red and blue hand gesture maps drift to the left which is characteristic of the newest speaker's (`lec_evol`) ground truth style, indicating that knowledge about previous experiences gets overwritten by new speaker data. C-DiffGAN (Fig 4d), however, is able to recreate the true distribution of the body poses for all 5 speakers through all the training experiences.

Style Preservation and Gesture Naturalness Table 3 shows results of the user study. Our C-DiffGAN model received the best scores for all factors except Expressivity compared to the other baseline models. For Style, C-DiffGAN scored **49.2**, at least 14 points more than DiffGAN and MeRGAN, and 7 points more than Buffer Replay. In terms of Naturalness, our model scored 18.5, at least 4 points more than the next best baseline. This reinforces how well our model preserves style across speakers, and generates the most natural gestures.

Impact of $\mathcal{L}_{ccf}$ on Accuracy and Forgetting Table 1 indicates that crossmodal catastrophic forgetting loss $\mathcal{L}_{ccf}$ has a significant effect on this continual learning setting, without which we observe a severe degradation in FID and PCK scores. FID Accuracy degrades from 56.2 to 242.7 (lower the better), a 4x degradation. Notably, forgetting becomes much worse, with FID Forgetting worsening from 13.2 to 258.5 (lower the better), and PCK Forgetting worsening tenfold from 0.01 to 0.12 (lower the better). This effect is also comparable to the overfitting of a jointly trained MixSTAGe in a low-resource data which has similar performance to a model without $\mathcal{L}_{ccf}$'s regularization effect.

Impact of $\mathcal{L}_{diffgan}$ on Accuracy and Forgetting In table 1, we observe that learning without $\mathcal{L}_{diffgan}$ degrades both crossmodal grounding (PCK) and output domain (FID), particularly worsening FID Accuracy from 56.2 to 70.8 (lower the better). This indicates that its impact is complimentary to the challenge of Crossmodal Catastrophic forgetting. Notably, degradation of Forgetting isn't as severe as for $\mathcal{L}_{ccf}$, emphasizing the respective roles that these losses have.

Impact of number of training examples on model gesture style and grounding As the amount of data increases, we observe in Figure 5 that our C-DiffGAN is able to model the output domain (FID) comparably much better before plateauing. The crossmodal grounding (PCK) remains fairly stable even with an increase in training data. In contrast, we observe an opposite effect for DiffGAN [3], MeRGAN-JTR and MeRGAN-RA [44] where the modeling ability of output domain (FID) and crossmodal grounding (PCK) consistently gets worse. This is indicative of these baselines easily learning to personalize to new speakers and just as easily forgetting old speakers. Adding more examples to the training experiences only speeds up this process.

Reminiscence vs Buffer Replay vs Joint Training trade-off *Buffer Replay* is not scalable in long-term continual learning settings. Instead, we use *Reminiscences* to reconstruct data for older experiences, bypassing the need for extra storage and compute. While this is successful in preventing catastrophic forgetting, explicitly storing examples in the buffer also can boost performance (Table 2). The extreme case of storing examples of all the speakers in the buffer (i.e. Joint Training for MixStAGe [4] in Table 2) can further boost the performance. The trade-off here is the need for extra storage, computational resources and training time for better performance. We advise readers to be aware of the trade-off and choose the methodology that better fits their scenario.

7. Conclusions

We studied the paradigm of low-resource gesture generation in a continual learning setting. We proposed C-DiffGAN, that efficiently leverages the continually arriving data to personalize the grounding and gesture style of the model to that of the new speakers. By generating low-resource reminiscences and a crossmodal catastrophic forgetting loss, the model is able to retain the grounding and style knowledge of the older speakers, the effectiveness of which we substantiate on a large scale publicly available dataset through quantitative and qualitative studies.

Acknowledgements: This material is based upon work partially supported by the National Science Foundation (Awards #1750439 #1722822), National Institutes of Health, NTT Japan, and, the InMind project. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of National Science Foundation or National Institutes of Health, and no official endorsement should be inferred.

References

[1] Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. Towards a human-like open-domain chatbot, 2020. 1

[2] Chaitanya Ahuja, Dong Won Lee, Ryo Ishii, and Louis-Philippe Morency. No gestures left behind: Learning relationships between spoken language and freeform gestures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1884–1895, 2020. 1, 2, 5, 6

[3] Chaitanya Ahuja, Dong Won Lee, and Louis-Philippe Morency. Low-resource adaptation for personalized co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 4, 5, 6, 7, 8

[4] Chaitanya Ahuja, Dong Won Lee, Yukiko I Nakano, and Louis-Philippe Morency. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. *Proceedings of the European Conference on Computer Vision*, 2020. 2, 4, 5, 6, 8, 9

[5] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 5

[6] Francisco Manuel Castro, Manuel J. Marín-Jiménez, Nicolás Guil Mata, Cordelia Schmid, and Alahari Karteek. End-to-end incremental learning. *ArXiv*, abs/1807.09536, 2018. 2

[7] Mireille Fares, Michele Grimaldi, Catherine Pelachaud, and Nicolas Obin. Zero-shot style transfer for gesture animation driven by text and speech using adversarial disentanglement of multimodal style encoding, 2022. 2

[8] Ylva Ferstl, Michael Neff, and Rachel McDonnell. Multi-objective adversarial gesture generation. In *Motion, Interaction and Games*, page 3. ACM, 2019. 2

[9] Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999. 2, 4

[10] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F Troje, and Marc-André Carbonneau. Zeroeggs: Zero-shot example-based gesture generation from speech. *arXiv preprint arXiv:2209.07556*, 2022. 2

[11] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2019. 2, 5

[12] Susan Goldin-Meadow. *Hearing Gesture: How Our Hands Help Us Think*. Harvard University Press, 2003. 2

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3, 4

[14] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 2

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017. 5

[16] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*, 2020. 3

[17] James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan,

John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114:3521 – 3526, 2016. 3

[18] Stefan Kopp, Paul A. Tepper, Kimberley Ferriman, Kristina Striegnitz, and Justine Cassell. Trading spaces: How humans and humanoids use speech and gesture to give directions. In T. Nishida, editor, *Conversational Informatics*, chapter 8, pages 133–160. John Wiley, 2007. 2

[19] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. Analyzing input and output representations for speech-driven gesture generation. *arXiv preprint arXiv:1903.03369*, 2019. 2

[20] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexanderson, Iolanda Leite, and Hedvig Kjellström. Gesticulator: A framework for semantically-aware speech-driven gesture generation. *arXiv preprint arXiv:2001.09326*, 2020. 2

[21] Jessica L. Lakin, Valerie E. Jefferis, Clara Michelle Cheng, and Tanya L. Chartrand. The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of Nonverbal Behavior*, 27:145–162, 2003. 2

[22] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86:2278–2324, 1998. 2

[23] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:2935–2947, 2016. 2

[24] Yu Liu, Gelareh Mohammadi, Yang Song, and Wafa Johal. Speech-based gesture generation for robots and embodied agents: A scoping review. In *Proceedings of the 9th International Conference on Human-Agent Interaction*, HAI '21, page 31–38, New York, NY, USA, 2021. Association for Computing Machinery. 2

[25] David Lopez-Paz and Marc' Aurelio Ranzato. Gradient episodic memory for continual learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2

[26] Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. Virtual character performance from speech. In *Symposium on Computer Animation*, pages 25–35, 2013. 2

[27] Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press, 1989. 2, 4

[28] David McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992. 2

[29] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze discriminator: A simple baseline for fine-tuning gans. *ArXiv*, abs/2002.10964, 2020. 3

[30] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 3

[31] Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2750–2758, 2019. 3

[32] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. A comprehensive review of data-driven co-speech gesture generation. *arXiv preprint arXiv:2301.05339*, 2023. 1

[33] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. A comprehensive review of data-driven co-speech gesture generation, 2023. 2

[34] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 1

[35] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, G. Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5533–5542, 2016. 2, 4

[36] ANTHONY ROBINS. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146, 1995. 2, 4

[37] Maha Salem, Stefan Kopp, Ipke Wachsmuth, and Frank Joublin. *Towards Meaningful Robot Gesture*, volume 6, pages 173–182. 11 2009. 2

[38] Ari Seff, Alex Beatson, Daniel Suo, and Han Liu. Continual learning in generative adversarial nets. *ArXiv*, abs/1705.08395, 2017. 3

[39] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *NIPS*, 2017. 2

[40] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1145–1153, 2017. 5

[41] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9332–9341, 2020. 3

[42] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 218–234, 2018. 3

[43] Jonathan Windle, Sarah Taylor, David Greenwood, and Iain Matthews. Pose augmentation: Mirror the right way. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, IVA '22, New York, NY, USA, 2022. Association for Computing Machinery. 2

[44] Chenshen Wu, Luis Herranz, Xialei Liu, Joost van de Weijer, Bogdan Raducanu, et al. Memory replay gans: Learning to generate new categories without forgetting. *Advances in Neural Information Processing Systems*, 31, 2018. 2, 4, 5, 6, 7, 8

[45] Jiang Xu, Patrick J Gannon, Karen Emmorey, Jason F Smith, and Allen R Braun. Symbolic gestures and spoken language are processed by a common neural system. *Proceedings of the National Academy of Sciences*, 106(49):20664–20669, 2009. 2

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

[47] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *ArXiv*, abs/1506.03365, 2015. 2

[48] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The design and implementation of XiaoIce, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93, 2020. 1