Low-Resource Adaptation for Personalized Co-Speech Gesture Generation

Chaitanya Ahuja[1], Dong Won Lee[2] & Louis-Philippe Morency[1]
[1]Language Technologies Institute, CMU & [2]Machine Learning Department, CMU
{cahuja, dongwonl}@andrew.cmu.edu, morency@cs.cmu.edu

Abstract

Personalizing an avatar for co-speech gesture generation from spoken language requires learning the idiosyncrasies of a person's gesture style from a small amount of data. Previous methods in gesture generation require large amounts of data for each speaker, which is often infeasible. We propose an approach, named DiffGAN, that efficiently personalizes co-speech gesture generation models of a high-resource source speaker to target speaker with just 2 minutes of target training data. A unique characteristic of DiffGAN is its ability to account for the crossmodal grounding shift, while also addressing the distribution shift in the output domain. We substantiate the effectiveness of our approach a large scale publicly available dataset through quantitative, qualitative and user studies, which show that our proposed methodology significantly outperforms prior approaches for low-resource adaptation of gesture generation. Code and videos can be found at https://chahuja.com/diffgan.

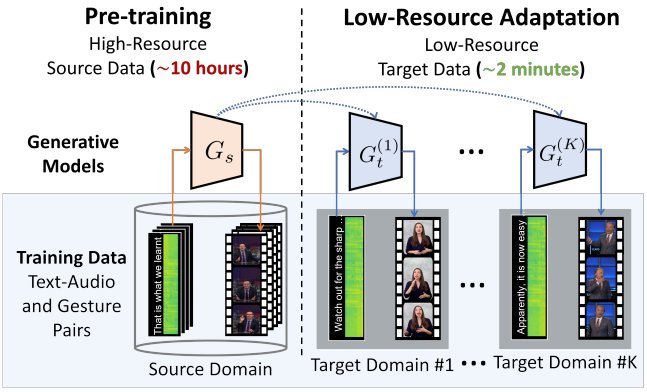


Figure 1. Overview of the co-speech gesture personalization task. On the left is a generative source model G_s pre-trained on a source speaker. We adapt G_s to multiple target models G_t using low-resource data for each of the target speaker.

1. Introduction

Technologies to assist human communication, both verbal (e.g. spoken language) and nonverbal (e.g. co-speech gestures), have gained more traction in the past decade. One promising direction is virtual reality [10, 17, 18, 29] which aims at creating a more realistic online communication platform for embodied virtual agents [6, 28] and remote avatars [38, 39]. These advancements could be seen as a normal progression to speech-based technologies such as intelligent personal assistants (e.g. Alexa, Siri, Cortana). These agents, in the future, could also communicate more naturally with a nonverbal embodiment that complements the verbal communication [33]. To enable this vision of immersive verbal and nonverbal communication through an avatar, one technical challenge is generating visual gestures based on input speech and language [2, 12, 15, 21]. An even more challenging task is the generation of personalized visual gestures, which reflects the idiosyncratic behaviours of a specific person [33]. The main goal of our paper is to create a personalized gesture generation model (e.g. as part of a personalized avatar) with limited data from a new speaker. In technical terms, this problem requires an adaptation of crossmodal generative models in a low-resource setting as illustrated in Figure 1. Leveraging an existing source model, pretrained on a large dataset of one speaker (i.e. source domain), our goal is to personalize to a new speaker (i.e. target domain) with only 2 minutes of the target data.

The problem setting brings a unique challenge, typically not studied in typical domain adaptation settings: *crossmodal grounding shift*. Due to the crossmodal nature of our task, crossmodal grounding shift refers to the distributional shift of the relationships between the input spoken language modalities and output gesture modality. For example, consider a speaker, *Aarti*, who waves their right hand while greeting a friend. These gestures are conditionally dependent on what the speaker says. We define such relationships between the gestures and spoken language as crossmodal grounding. While these relationships are conditioned on spoken language, they are also heavily influenced by the speaker's idiosyncrasies. Now, consider a new speaker, *Bob*, who chooses from a set of two gestures while greeting: a

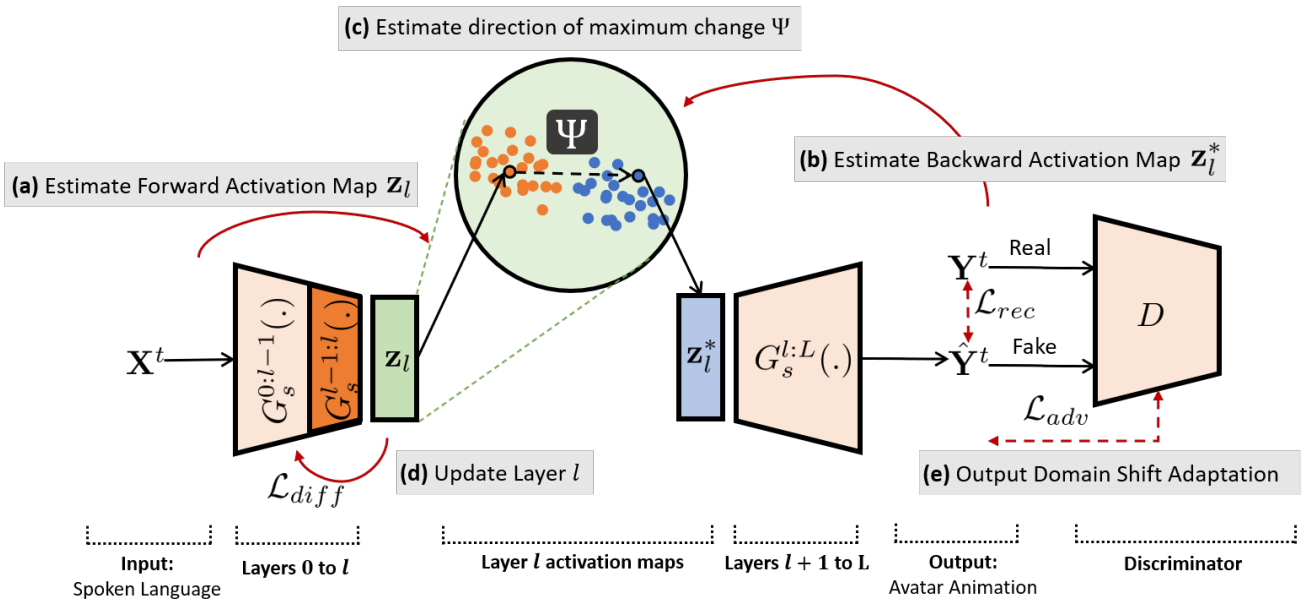


Figure 2. Overview of the key components of our proposed model DiffGAN. (a)-(d): Low-resource adaptation of the crossmodal grounding relationships from source to target domain. (e): Modeling output domain shift from source to target domain

left handed wave or waving both their hands vigorously. As is the case for this speaker, typically the conditional gesture spaces have a larger support (i.e. different kinds of gestures) for the same language context. Such differences between conditional gestures of source and target speakers are very common, especially because the conditional variable is language which is a very large space. These common, yet complex differences represent crossmodal grounding shift.

In this paper, we propose an approach, named DiffGAN, that can efficiently personalize co-speech gesture generation models from a high-resource source speaker to a low-resource target speaker. To the best of our knowledge, this is the first approach that is able to learn a personalized model with only 2 minutes of speaker data (i.e. as opposed to 10 hours [2, 12, 15, 21]). Our DiffGAN approach does not require access to source training data. Instead, DiffGAN directly identifies shifts in crossmodal grounding relationships along with the shifts in the output domain from the pre-trained source model. Based on these identified distribution shifts, DiffGAN updates a few necessary parameters in a single layer of the source model, allowing efficient adaptation with low resources. Our experiments study the effectiveness of our DiffGAN approach on a diverse publicly available dataset. As part of our evaluation methodology, we report that DiffGAN produces a consistent improvement of around 10% preference scores of human judgments over strong baselines among other quantitative improvements. Furthermore, DiffGAN extrapolates to gestures in the target distribution without ever having seen them in the source distribution.

2. Related Work

Language and speech for gesture generation A rule-based approach was proposed in an earlier study by Cassell et al. [7], where the behavior expression animation toolkit (BEAT) was developed to schedule behaviors, such as hand gestures, head nods and gaze. This approach was extended to utilize linguistic information from input text for decision making [22, 25, 26, 31, 49].

Rule based approaches were replaced by deep conditional neural fields [8, 9] and Hidden Markov Models for prosody-driven head motion generation [42] and body motion generation [23, 24]. These use a dictionary of predefined animations, limiting the diversity of generated gestures. Soon, neural network based models were introduced, using unimodal inputs, specifically speech, to generate a sequence of gestures [14], head motions [41] and body motions [1, 3, 11, 12, 43]. On the other hand, Yoon et al. [50] uses only a text input for gesture generation. More recently, multimodal models relying both speech and language were developed. Kucherenko et al. [21] uses early fusion to combine the two representations, Ahuja et al. [2] utilizes a cross-modal attention mechanism to account for correlations between speech and language. These approaches typically require many hours of multimodal data to train a single speaker-specific model. We propose an approach that can adapt a single-speaker generative model to a new speaker domain after being exposed to only a few minutes of data.

Figure 3. Qualitative comparison of our DiffGAN with prior work over shape of generated gestures. With **maher** as the source domain, target model outputs over the target domain are superimposed over ground truth video frames for easy comparison.

Low-resource adaptation of generative models Prior works similary focus on pre-training a source model on large source data, then adapting it to a low-resource setting. Nichol et al. [35], Wang et al. [47] introduce new parameters in the model, whereas Karras et al. [20], Wang et al. [46] fine-tunes the complete model on the target data or to specific layers or modules are applied [34, 36]. Li et al. [27], Wang et al. [47] 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, Ojha et al. [37] 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 pretrained gesture generation model of a source speaker as a generator-discriminator pair (G_s, D_s) [13] that is trained on a large gesture source dataset $\mathcal{D}_s = \{\mathbf{X}_i^s, \mathbf{Y}_i^s\}_{i=1}^N$. It generates gestures as a sequence of body poses $\mathbf{Y}_i^s$ that is driven by both language and speech as the input modalities $\mathbf{X}_i^s$. A goal is to adapt parameters of the pretrained generator G_s to a target model G_t by using a much smaller target dataset of the target speaker $\mathcal{D}_t = \{\mathbf{X}_i^t, \mathbf{Y}_i^t\}_{i=1}^P$ where $P << N$.

4. Method

We propose a new approach, DiffGAN, that learns a target model G_t by adapting a pre-trained source model G_s in a low-resource setting. This approach is a two-step process illustrated in Figure 2. First, in section 4.1, the model learns to identify the crossmodal grounding shifts through a novel loss function $\mathcal{L}_{diff}$ and low-resource target data. Second, in section 4.2, we discuss use of a loss function $\mathcal{L}_{shift}$, which encourages the target model to shift the output domain distribution to be closer to that of the target's. Optimization of the combined loss function describes the complete model,

$$G_t^* = \mathbb{E}_{\mathbf{X},\mathbf{Y}\in\mathcal{D}_t} \underset{G_s,\theta_{l-1:l}}{\operatorname{argmin}} \max_{D_s} \mathcal{L}_{diff}(\theta_{l-1:l}) + \mathcal{L}_{shift}(G_s, D_s) \tag{1}$$

where $\theta_{l-1:l}$ are parameters of a layer l in G_s (discussed in Section 4.1).

4.1. Crossmodal Grounding Shift

The crossmodal grounding relationships between input and outputs spaces of source data are already encoded in the source model, G_s. Instead of modifying all of these relationships in the source model, we first discover the shift in grounding from the source to the the target dataset. This is followed by the adaptation of the model parameters which account for only the shifted relationships in the target model. This approach has a two key advantages. First, adapting to only the differences suggested by the discrepancies between target and source data, allows the model to retain the previously learnt essential grounding relationships intact. Second, in a low-resource setting, updating only a few layers instead of the complete model is sufficient [5, 34]. Doing so allows the target model to learn new grounding relationships while preventing overfitting.

Notations: Source G_s and target G_t models both have a total of L layers. Function $G_s^{l:m}(.)$ represents layers l through m in G_s. For example, $G_s^{l:m}(.)$ takes activation maps of layer l (or $\mathbf{z}_l$) in G_s as the input and returns activation maps of layer m (or $\mathbf{z}_m$) in G_s as the output. Parameters of layers l through m can be explicitly specified in function $G_s^{l:m}(.;\theta_{l:m})$, but may be skipped for brevity.

Discovering shift in crossmodal grounding relationships: Crossmodal grounding represents relationships between the input and output modalities. For the source data. these relationships are already encoded in the latent spaces [3, 19, 51] of the source model, G_s. In the adapted target model G_t, this

Amount of data (minutes)	Pre-trained G_s	Models	Gesture Quality: Naturalness	Gesture Quality: Expressivity	Crossmodal Grounding: Timing	Crossmodal Grounding: Relevance	Output Domain: Style
2	✗	**AISLe [2]**	7.3 ± 2.9	15.3 ± 7.6	9.2 ± 2.4	8.3 ± 4.1	16.3 ± 5.3
	✓	**TGAN [46]**	9.4 ± 3.8	12.7 ± 4.6	12.1 ± 2.3	10.8 ± 4.0	13.7 ± 2.9
	✓	**MineGAN [47]**	13.0 ± 2.9	16.6 ± 4.8	16.0 ± 2.3	14.1 ± 4.2	29.6 ± 7.3
	✓	**ConsistentGAN [37]**	9.0 ± 1.9	17.7 ± 3.1	10.9 ± 1.9	9.3 ± 1.6	17.6 ± 6.3
2	✓	**DiffGAN (Ours)**	**21.9 ± 2.5**	**27.6 ± 6.5**	**26.2 ± 2.1**	**23.9 ± 4.8**	**46.3 ± 9.2**
	✓	**DiffGAN w/o $\mathcal{L}_{diff}$**	**19.8 ± 1.3**	**24.4 ± 5.0**	**22.1 ± 3.3**	21.0 ± 3.0	**47.8 ± 4.8**
	✓	**DiffGAN w/o $\mathcal{L}_{shift}$**	12.3 ± 4.1	20.0 ± 5.9	15.8 ± 3.9	13.6 ± 3.8	26.5 ± 6.4

Table 1. Human perceptual study comparing our model with prior work and strong baselines over five criteria measuring **quality**, **crossmodal grounding** and **output domain shift** of generated gestures. We report the preference scores of a model as compared to the ground truth gestures. Confidence intervals reported as standard deviation across experiments on all source-target pairs. Higher is better with 50 % being the best possible score. Scores in green are the best and orange are the second best but lie in the confidence interval of the best.

latent space will shift creating new grounding relationships. More concretely, this latent space represents the activation maps at layer l for source and target models, or z_l and z_l^* respectively. To estimate the direction along which the grounding relationships have shifted, we compute the element-wise difference $\Psi = |z_l - z_l^*|$ between the activation maps at layer l. We can now update the parameters of layer l in the direction Ψ to produce the required grounding shift.

Computing direction of grounding shift Ψ: To compute $\Psi = |z_l - z_l^*|$, we need both z_l and z_l^*. As z_l is an activation map of the source model with a target sample $\mathbf{X}^t$ as input, we can compute it as $z_l = G_s^{0:l}(\mathbf{X}^t)$ as shown in Figure 2a. Estimating z_l^* is tricky as the target model is not available yet. As we are only updating the parameters of layer l in this step, the parameters of $G_s^{l:L}$ do not change. As a result, we can use values of the target output modality $\mathbf{Y}^t$ to optimize $\text{argmin}_{\mathbf{z}} \|G_s^{l:L}(\mathbf{z}) - \mathbf{Y}^t\|_2$ as shown in Figure 2b. This minimization objective serves as an accurate estimate of z_l^*, and consequently an accurate estimate of the direction of grounding shift Ψ. To prevent overfitting due to limited amount of data, we concentrate the gradient update to the directions (i.e. channel dimensions) with top-k grounding shifts represented as Ψ_k. The criteria for choosing l and k is discussed in Section 6.

Updating Crossmodal Grounding in layer l: To encourage generation of the shifted latent space z_l^* for target domain inputs $\mathbf{X}_t$, we update the weights of only layer l (or $\theta_{l-1:l}$) through an L2 loss. Furthermore, as Ψ_k is the measure of grounding shift for each parameter, we use it as a weighting function to guide parameter updates of layer l,

$$\mathcal{L}_{diff} = \|\Psi_k \odot \mathbf{z}_l^* - \Psi_k \odot G_s^{l-1:l}\left(G_s^{0:l-1}(\mathbf{X}^t); \theta_{l-1:l}\right)\|_2, \tag{2}$$

where $\odot$ is element-wise product. As the training progresses, both Ψ_k and z_l^* are re-estimated based on the updated parameters of the source model. Hence, as the latent space of the adapted source model shifts closer to that of the target domain, Ψ_k will re-adjust until convergence, arriving at the target model. Please note that while this approach is described for a single layer l, it can easily be adjusted to update any sequence of layers l through m without loss of generality.

4.2. Output Domain Shift

The second step is to shift the output domain of the source model G_s toward that of the target gesture distribution. We follow the fine-tuning approach suggested in [37, 46] of optimizing for the adversarial loss function L_{adv},

$$\mathcal{L}_{adv} = \mathbb{E}_{\mathbf{X}^t, \mathbf{Y}^t \in \mathcal{D}_t} \log D_s\left(\mathbf{Y}^t\right) + \log\left(1 - D_s\left(G_s\left(\mathbf{X}^t\right)\right)\right), \tag{3}$$

where the discriminator $D_s(.)$ measures domain correctness of the output modality. This adversarial loss encourages the model to generate gesture sequences whose structure represents the target distribution. We would also like to encourage generation of output sequences that temporally match the target ground truth sequences $\mathbf{Y}^t \in \mathcal{D}_t$ for which we use a reconstruction loss [2, 3, 12],

$$\mathcal{L}_{rec} = \mathbb{E}_{\mathbf{X}^t, \mathbf{Y}^t \in \mathcal{D}_t} \|\mathbf{Y}^t - \hat{\mathbf{Y}}^t\|_1, \tag{4}$$

where $\hat{\mathbf{Y}}^t = G_t(\mathbf{X}^t)$. The combination of the adversarial and reconstruction loss, $\mathcal{L}_{adv} + \mathcal{L}_{rec}$, is defined as $\mathcal{L}_{shift}$ and encourages the output domain to shift toward the target distribution (see Figure 2e).

5. Experiments

Dataset: We use the PATS dataset [2, 3, 12] 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. We choose five speakers (`oliver`, `maher`, `chemistry`, `ytch_prof` and `lec_evol`) with visually different gesture styles and diverse linguistic content for our experiments, in which `source` $\rightarrow$ `target` denotes the source and target domain. For speakers in the target domain, we simulate a low-resource setting by randomly sampling 2 or 10 minutes of data for all experiments.

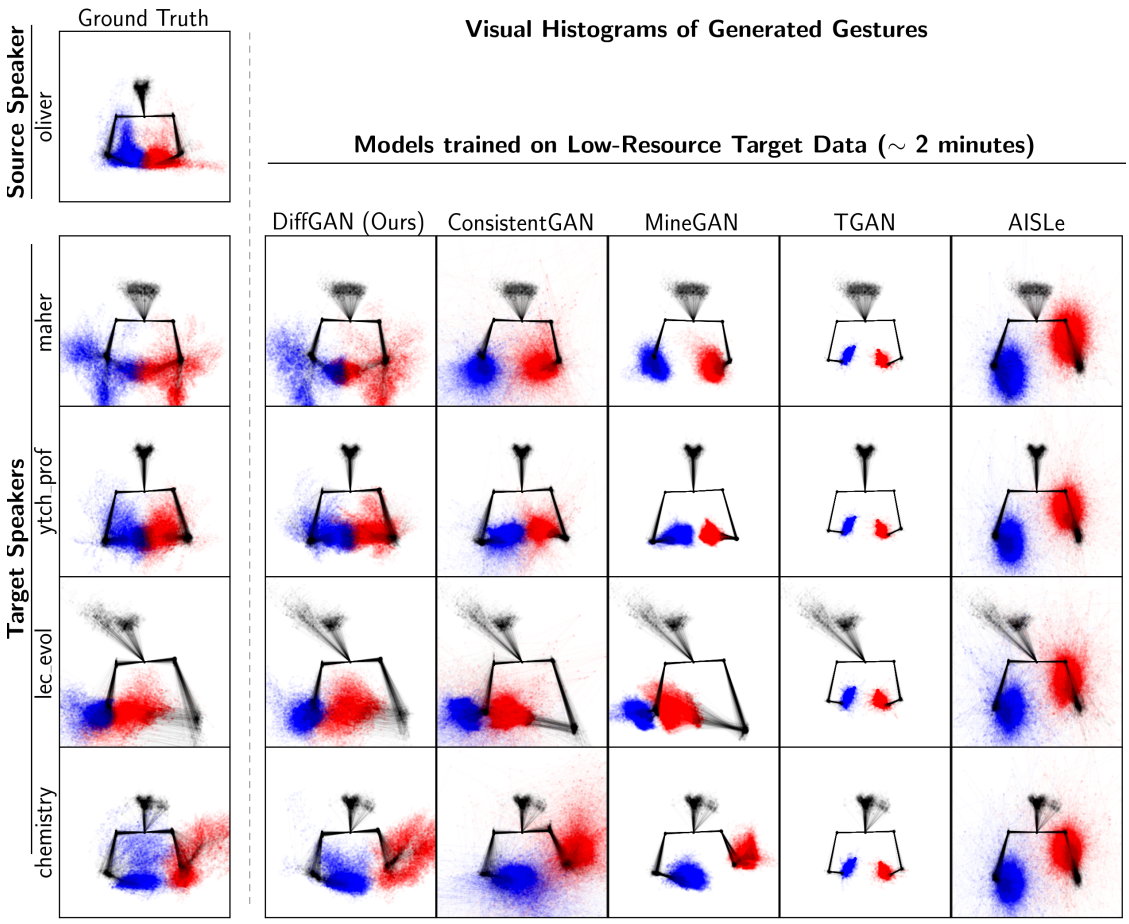


Figure 4. Visual Histograms of generated gestures visually describe the distribution of hand gestures in space. Red and blue colors denote the left and right arms respectively. First row is the source speaker, below which we have all the target speakers. Each column denotes a model which adapts output distribution of the source domain to the target domain. Qualitatively, DiffGAN is successful in modeling the distribution of the source speaker with just 2 minutes of data.

Baseline Models: We compare our proposed model with a family of baselines that adapts the same source model to a target domain in a low-resource setting. (a) **TGAN** [46] fine-tunes all layers of source model with the low-resource target data, (b) **MineGAN** [47] projects the source latent space of the noise input onto a latent space representative of the target data, (c) **ConsistentGAN** [37] uses a cross-domain consistency loss which regularizes the tuning process and a patch discriminator which encourages different levels of realism over different image patches, and (d) **AISLe** [2] learns the target model without a pretrained source model. We also run ablation studies on two versions of our model (e) **DiffGAN w/o** $\mathcal{L}_{\mathbf{shift}}$ and (f) **DiffGAN w/o** $\mathcal{L}_{\mathbf{diff}}$.

Human Perceptual Study: We conduct a human perceptual study on Amazon Mechanical Turk (AMT) to measure human preference towards generated animations. Given a pair of videos, one of which is from the ground truth and the other is generated by a model, the annotators have to choose either one of the videos based on the five criterion: gesture quality (**naturalness** and **expressivity**), crossmodal grounding (**timing** and **relevance**) and output domain shift (**style**) of generated gestures. The correctness of output domain shift is measured by the reflection of the true gesture style of a target speaker in the gestures generated by a target model [3]. We report the average preference % of human annotators as a score for comparison. We refer the readers to the appendix for more detailed definitions and setup.

Quantitative Metrics: (a) To measure relevance and timing of gestures with respect to spoken language we use two metrics, **Probability of Correct Keypoints (PCK)** [4, 44] where values are averaged over $\alpha = 0.1, 0.2$ as suggested in [12] and **L1 distance** between generated and ground truth gestures. To measure the distribution of the output domain we use **Fréchet Inception Distance (FID)** which is the dis-

Amount of data (minutes)	Pre-trained G_s	Models	source ↓ target	FID ↓ maher ↓ oliver	FID ↓ maher ↓ chemistry	FID ↓ oliver ↓ maher	FID ↓ oliver ↓ chemistry	PCK ↑ maher ↓ oliver	PCK ↑ maher ↓ chemistry	PCK ↑ oliver ↓ maher	PCK ↑ oliver ↓ chemistry
2	✗	**AISLe [2]**		49.2 ± 0.8	84.5 ± 3.1	83.7 ± 2.6	84.5 ± 3.1	0.18 ± 0.01	0.2 ± 0.0	0.18 ± 0.0	0.2 ± 0.0
	✓	**TGAN [46]**		57.5 ± 2.4	184.3 ± 5.4	339.1 ± 1.2	323.5 ± 1.9	0.31 ± 0.01	0.23 ± 0.0	0.2 ± 0.0	0.25 ± 0.0
	✓	**MineGAN [47]**		42.5 ± 2.5	157.5 ± 10.6	290.3 ± 7.2	302.5 ± 6.8	0.38 ± 0.03	0.26 ± 0.02	0.21 ± 0.01	**0.31 ± 0.01**
	✓	**ConsistentGAN [37]**		61.0 ± 3.2	194.2 ± 15.6	320.1 ± 11.5	325.6 ± 44.3	0.39 ± 0.01	0.27 ± 0.01	0.21 ± 0.01	0.25 ± 0.01
	✓	**DiffGAN (Ours)**		**25.0 ± 3.7**	**42.8 ± 5.1**	**47.9 ± 25.5**	**48.2 ± 15.9**	**0.45 ± 0.02**	**0.31 ± 0.01**	**0.26 ± 0.01**	**0.29 ± 0.01**
10	✗	**AISLe [2]**		47.1 ± 0.2	85.0 ± 1.9	80.3 ± 0.8	85.0 ± 1.9	0.18 ± 0.0	0.21 ± 0.0	0.19 ± 0.0	0.21 ± 0.0
	✓	**TGAN [46]**		62.9 ± 1.8	191.7 ± 1.2	341.5 ± 0.4	326.8 ± 1.6	0.3 ± 0.01	0.22 ± 0.01	0.2 ± 0.0	0.24 ± 0.0
	✓	**MineGAN [47]**		41.4 ± 3.1	145.0 ± 14.1	293.5 ± 12.7	318.2 ± 5.4	0.4 ± 0.01	0.24 ± 0.03	0.22 ± 0.02	**0.31 ± 0.01**
	✓	**ConsistentGAN [37]**		63.1 ± 1.9	188.2 ± 17.0	322.2 ± 2.4	327.0 ± 18.7	0.41 ± 0.03	**0.28 ± 0.04**	0.22 ± 0.01	0.27 ± 0.01
	✓	**DiffGAN (Ours)**		**15.0 ± 5.2**	**31.7 ± 3.8**	**30.3 ± 6.9**	**24.3 ± 4.2**	**0.46 ± 0.01**	**0.32 ± 0.02**	**0.26 ± 0.01**	**0.3 ± 0.01**
Full	✗	**AISLe [2]**		16.1	8.7	10.2	8.7	0.49	0.39	0.27	0.39

Table 2. Comparison of our DiffGAN with prior work for low-resource crossmodal generative modeling from source to target speakers to evaluate output domain shift (i.e. FID) and crossmodal grounding (i.e. PCK)

tance between distributions of generated and ground truth poses [16]

Qualitative Visualization: To judge the quality of the generated spatio-temporal outputs, we would encourage the readers to see the supplementary video. Other than that, we qualitatively visualize three key properties of gestures [32, 40] (1) distribution, (2) velocities and (3) shapes of gestures

Implementation Details: For our pretrained source models, we use publicly available models by Ahuja et al. [2] for all experiments. 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. 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 the qualitative, quantitative and user study results from our experiments.

Comparison with prior work When evaluating generative models human judgements are often seen as a de facto evaluation [2, 48]. The results of out human perceptual study are summarized in in Table 1. We see a 10% larger preference, if not more, for our model DiffGAN as compared to the baseline models across all five criteria.

Crossmodal grounding shift is evaluated by measuring the relevance, timing and correctness of the generated gesture in context of spoken language. In the human perceptual study in Table 1, higher preference scores for DiffGAN over *relevance* and *timing* criteria are indicative of the positive impact of modeling the grounding shift explicitly which is not the case for the other baselines. Qualitatively, in Figure 3, we observe gesture shapes that are closer to the ground truth

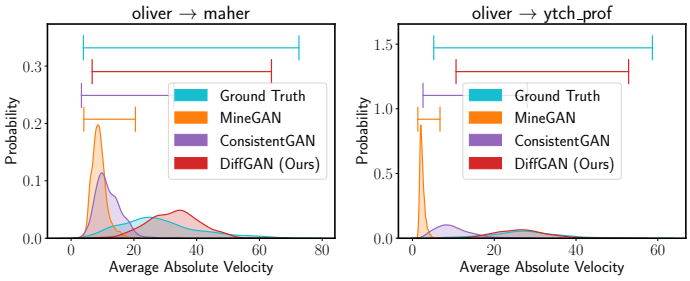


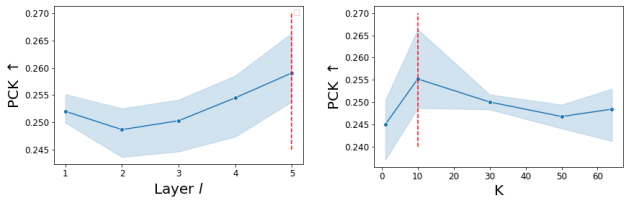
Figure 5. Distribution of the generated gestures with average absolute velocity as the statistic for source to target domain adaptation. The support (or coverage) of the distribution is denoted with the colour coded lines at the top of each plot. Larger overlap of a model's distribution with the ground truth distribution is desirable.

for DiffGAN than the other baselines, further indicating that the generated gesture shapes are more relevant to the input modality for DiffGAN. This is further corroborated by significantly higher PCK values for DiffGAN when compared with TGAN [46], MineGAN [47] and, ConsistentGAN [37] in Table 2.

Output domain shift (i.e. gesture style) of the generated gestures was especially convincing, as DiffGAN was preferred by human annotators (in Table 1) over the ground truth motion 46% of the time. A similar trend is also seen qualitatively in Figure 4 where we are comparing the pose histograms for both ground truth of the target speaker and the generated animations. DiffGAN is able to adapt the source model such that it generates a distribution of gestures similar to the target domain. Even though source `oliver` typically has gestures close to his body with hands moving up and down, DiffGAN is able to extrapolate to a target model for `maher` with his prototypical side to side arm movements. For TGAN [46] and MineGAN [47], the hand positions are concentrated in only one region indicating reduced diversity

Amount of data (minutes)	Pre-trained G_s	Models	source ↓ target	FID ↓ oliver ↓ maher	FID ↓ oliver ↓ chemistry	FID ↓ maher ↓ oliver	FID ↓ maher ↓ chemistry	L1 ↓ oliver ↓ maher	L1 ↓ oliver ↓ chemistry	L1 ↓ maher ↓ oliver	L1 ↓ maher ↓ chemistry
10	✓	**DiffGAN (Ours)**		**30.3 ± 6.9**	**24.3 ± 4.2**	**15.0 ± 5.2**	**31.7 ± 3.8**	1.48 ± 0.01	1.36 ± 0.03	**0.53 ± 0.02**	0.88 ± 0.03
	✓	**DiffGAN w/o $\mathcal{L}_{diff}$**		**25.9 ± 4.3**	**27.5 ± 6.7**	**19.0 ± 7.7**	**29.0 ± 1.6**	1.57 ± 0.03	1.43 ± 0.01	0.61 ± 0.02	0.96 ± 0.02
	✓	**DiffGAN w/o $\mathcal{L}_{shift}$**		119.1 ± 7.0	131.0 ± 5.1	22.4 ± 2.1	54.1 ± 1.9	**1.33 ± 0.01**	**1.26 ± 0.01**	**0.53 ± 0.01**	**0.84 ± 0.0**
Full	✗	**AISLe [2]**		10.2	8.7	16.1	8.7	0.91	0.73	0.71	0.73

Table 3. **Evaluating impact of loss functions**: DiffGAN and its ablations are trained on 10 minuites of low-resource data. The metrics measure the impact of $\mathcal{L}_{shift}$ and $\mathcal{L}_{diff}$ on both output domain shift (i.e. FID) and crossmodal grounding (i.e. L1).



(a) Choice of Layer l (b) Choice of k in Grounding Shift

Figure 6. (a) Impact of choice of Layer l on crossmodal grounding shift (i.e. PCK ↑). Layer numbers are in increasing order from input to output. Layers corresponding to indices on the X-axis are defined in supplementary. (b) Impact of choice of k in grounding shift direction Ψ_k on crossmodal grounding shift (i.e. PCK ↑).

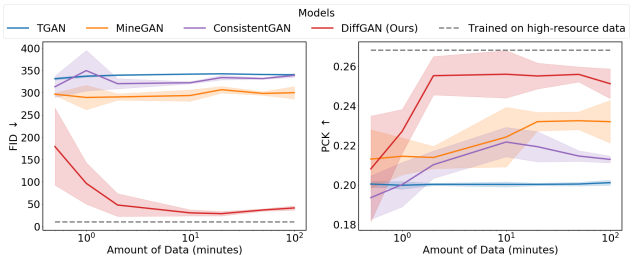


Figure 7. Impact of amount of data on crossmodal grounding shift (i.e. PCK ↑) and output domain shift (i.e. FID ↓) in crossmodal generative models. Note that X-Axis is logarithmic and error bars are standard deviation over three randomly sampled training sets.

in the generated gestures and therefore a mode collapse. In column three of Figure 4, ConsistentGAN [37] is able to learn the correct rest pose for the target speakers, potentially due to its approach of encouraging distance consistency in the output domain. But the distribution of output gestures does not correctly match the distribution of true distribution of target speakers illustrated in column one of Figure 4. These trends are further corroborated by significantly better values of FID for DiffGAN when compared with TGAN [46], MineGAN [47] and ConsistentGAN [37] in Table 2.

The distribution of gesture velocities is another statistic with which we can examine the correctness of output domain shift [40]. In Figure 5, we find that our model DiffGAN (┣┫) is closely able to generate a velocity distribution that is similar to the true distribution of the target domain. However, distribution modes of MineGAN [47] and, ConsistentGAN [37] are close to zero indicating that the generated gestures are have very little or no movements.

Impact of $\mathcal{L}_{diff}$ on crossmodal grounding shift In Table 1, we observe that the removal of $\mathcal{L}_{diff}$ from DiffGAN reduces human preference for the gestures generated by the model with respect to *timing* and *relevance* criteria. We observe a similar trend in the accuracy metric L1, which significantly worsens in Table 3. This supports our hypothesis that optimizing $\mathcal{L}_{diff}$ can improve the discovery of new grounding relationships in a low-resource setting.

Impact of $\mathcal{L}_{shift}$ on output domain shift On the other hand, removal of $\mathcal{L}_{shift}$ reduces correctness of the style of generated gestures (i.e. output domain shift) as indicated by human preference in Table 1. This is most likely due to the adversarial component of $\mathcal{L}_{shift}$ which adapts the output domain with the help of gesture sequences from the target data. In parallel, FID values in Table 3 undergo the same effect indicating that the target model does not generate a large variety of gestures without $\mathcal{L}_{shift}$, even though the gestures are well-grounded in the input. We note here that human annotators' judgements can be strongly influenced by the naturalness of generated gestures [48]. This is a likely reason for decrease in preference of crossmodal grounding metrics for DiffGAN w/o $\mathcal{L}_{shift}$, however it is still preferred more often than other baseline models in Table 1.

Impact of number of training examples on model gesture style and grounding The amount of data has a large part to play in generative modeling and adaptation [45, 52]. We vary the amount of training data from 30 seconds to 100 minutes in Figure 7. We observe for our model DiffGAN, the output domain shift and crossmodal grounding adapts faster than all the baselines. The variance of these metrics decreases with increasing amount of data indicates a more stable training. Our choice of low-resource datasets is completely random.

Which weights should be updated? For a successful low-resource adaptation, a challenge is to select the best weights to update. Our DiffGAN approach requires a choice of trainable layer l and a choice of k number of channels that get updated in each iteration. Through hyperparameter tuning, we observe that layers closer to the output are typically able to model a better crossmodal grounding shift as seen in Figure 6a. The ideal choice of k is trickier. We want to update enough parameters to model the grounding shifts, but not so many that the model would overfit on the target data. For our choice of pretrained source models, we conduct an experiment with varying number of ks in Figure 6b. At $k = 0$ we see a drop in performance, likely due to the inactivity of $\mathcal{L}_{diff}$. At $k = 64$, we find the performance saturates indicating an overfitted model. We find a balance somewhere in the middle at $k = 10$.

Visualizing crossmodal grounding shift For the same input, we probe the output spaces of the source and target model in Figure 8. We find that distribution gestures corresponding to the input can be sparse (i.e. visually different gestures) or dense (i.e. visually similar gestures). In other words, a verbal concept such as a greeting has a single way of gesturing when the conditional output distribution is dense, but has multiple possible gestures if the conditional output distribution is sparse. As the output distribution is conditioned not only on the input but also on the speaker, we can visually observe crossmodal grounding shift in form of expansion or contraction as we traverse from the source to target output space.

Limitations and future work: While our method generates compelling results, it is not without limitations. Our choice of source models were trained on a single source domain (i.e. speaker), which may can sometimes have a smaller overlap with the target domain. This poses a trade-off between the complexity of the source model and the amount of target data. Another challenge with our approach, is that the choice of layer(s) l and k grounding shift channels can potentially change depending on the choice of the pretrained model architecture. Hence, these hyperparameters may need some tuning.

7. Conclusions

In this paper, we studied low-resource adaptation of crossmodal generative models for gesture generation. We introduced a new generative model DiffGAN, that can efficiently address the shift in crossmodal grounding and the output distribution from the source to target speaker with only a few minutes of data. We benchmarked the effectiveness of our approach on a publicly available dataset through quantitative, qualitative and human studies. To our knowledge, this is the first approach that is able to learn a personalized gesture generation model with only 2 minutes of speaker data.

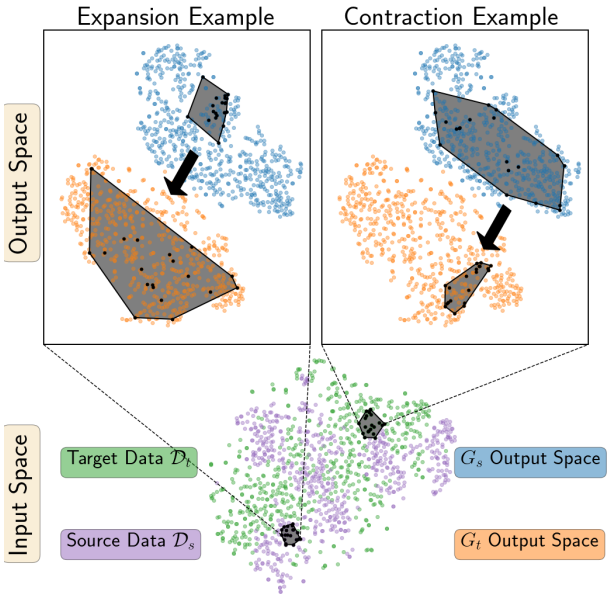


Figure 8. At the bottom, we display a t-SNE [30] plot of the input space for both the target and source data. We choose a region which contains both source and target input samples. At the top, we display the t-SNE plots corresponding to where these samples map to in the output space (indicated in black). On top left, similar inputs produces a compact source output space (i.e. visually similar gestures) but a sparse target output space (i.e. visually different gestures). On top right, the opposite effect is observed. These contractions and expansions of the output space conditioned on the input space represent crossmodal grounding shift.

Broader Impact: Our work enables for low-resource generation of personalized avatar animation, which typically requires many hours of training. It allows for the generation of gestures for new speakers with low resources. We developed this technology to improve the naturalness of an AI agent, which would improve human-to-AI communication as nonverbal behavior plays a key role in communication. These gestures in the form of skeletal keypoints alone could not be used to fully impersonate others, but it could be maliciously used to enhance the naturalness of deepfakes when combined with other generative modeling technology. Realistic deepfakes can further enable misinformation spread, abuse and stolen identities. As a potential measure to deter such behaviour, we release our code under an ethical license which prevent the usage of the code by any party that support or contribute to hate speech or false impersonation (Do No Harm, Nonviolent Public or Hippocratic License)

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] Chaitanya Ahuja, Shugao Ma, Louis-Philippe Morency, and Yaser Sheikh. To react or not to react: End-to-end visual pose forecasting for personalized avatar during dyadic conversations. In *2019 International Conference on Multimodal Interaction*, pages 74–84. ACM, 2019. 2

[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, 4, 5, 6, 7

[3] 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, 3, 4, 5

[4] 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

[5] David Bau, Steven Liu, Tongzhou Wang, Jun-Yan Zhu, and Antonio Torralba. Rewriting a deep generative model. In *European Conference on Computer Vision*, pages 351–369. Springer, 2020. 3

[6] Justin Cassell. More than just another pretty face: Embodied conversational agents. *Communications of the ACM*, 43(4), 2001. 1

[7] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. BEAT: the Behavior Expression Animation Toolkit. In *the 28th annual conference on Computer graphics and interactive techniques (SIGGRAPH '01)*, pages 477–486, 2001. doi: https://doi.org/10.1145/383259.383315. 2

[8] Chung-Cheng Chiu and Stacy Marsella. Gesture generation with low-dimensional embeddings. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 781–788, 2014. 2

[9] Chung Cheng Chiu, Louis Philippe Morency, and Stacy Marsella. Predicting co-verbal gestures: A deep and temporal modeling approach. In *Proceedings of the 15th international conference on Intelligent virtual agents (IVA2015)*, volume 9238, pages 152–166, 2015. ISBN 9783319219950. 2

[10] Pietro Cipresso, Irene Alice Chicchi Giglioli, Mariano Alcañiz Raya, and Giuseppe Riva. The past, present, and future of virtual and augmented reality research: a network and cluster analysis of the literature. *Frontiers in psychology*, 9:2086, 2018. 1

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

[12] 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. 1, 2, 4, 5

[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

[14] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. Evaluation of Speech-to-Gesture Generation Using Bi-Directional LSTM Network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents (IVA18)*, pages 79–86, 2018. 2

[15] 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. 1, 2

[16] 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. 6

[17] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics (ToG)*, 36(6):1–14, 2017. 1

[18] Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. Dynamic 3d avatar creation from hand-held video input. *ACM Transactions on Graphics (ToG)*, 34 (4):1–14, 2015. 1

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 3

[20] 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

[21] 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. 1, 2

[22] Jina Lee and Stacy Marsella. Nonverbal behavior generator for embodied conversational agents. In *Proceedings of the 6th international conference on Intelligent virtual agents (IVA2006)*, pages 243–255, 2006. 2

[23] Sergey Levine, Christian Theobalt, and Vladlen Koltun. Real-time prosody-driven synthesis of body language. *ACM Trans. Graph.*, 28(5):172:1–172:10, December 2009. ISSN 0730-0301. 2

[24] Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. Gesture controllers. *ACM Trans. Graph.*, 29(4):124:1–124:11, July 2010. ISSN 0730-0301. 2

[25] Margot Lhommet and Stacy Marsella. From embodied metaphors to metaphoric gestures. *CogSci*, pages 788–793, 2016. 2

[26] Margot Lhommet, Yuyu Xu, and Stacy Marsella. Cerebella: Automatic Generation of Nonverbal Behavior for Virtual Humans. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 4303–4304, 2015. 2

[27] Y. Li, Richard Zhang, Jingwan Lu, and E. Shechtman. Few-shot image generation with elastic weight consolidation. *ArXiv*, abs/2012.02780, 2020. 3

[28] Tze Wei Liew and Su-Mae Tan. Exploring the effects of specialist versus generalist embodied virtual agents in a multi-product category online store. *Telematics and Informatics*, 35(1):122–135, 2018. 1

[29] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics (TOG)*, 37(4): 68, 2018. 1

[30] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 8

[31] 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. ISBN 9781450321327. doi: 10.1145/2485895.2485900. 2

[32] David McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago Press, 1992. 6

[33] David McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992. 1

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

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

[36] 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

[37] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. *arXiv preprint arXiv:2104.06820*, 2021. 3, 4, 5, 6, 7

[38] Minna Pakanen, Paula Alavesa, Niels van Berkel, Timo Koskela, and Timo Ojala. "nice to see you virtually": Thoughtful design and evaluation of virtual avatar of the other user in ar and vr based telexistence systems. *Entertainment Computing*, 40:100457, 2022. 1

[39] Ye Pan and Anthony Steed. The impact of self-avatars on trust and collaboration in shared virtual environments. *PloS one*, 12(12):e0189078, 2017. 1

[40] Catherine Pelachaud. Studies on gesture expressivity for a virtual agent. *Speech Communication*, 51(7):630–639, 2009. 6, 7

[41] Najmeh Sadoughi and Carlos Busso. Novel realizations of speech-driven head movements with generative adversarial networks. pages 6169–6173, 04 2018. doi: 10.1109/ICASSP.2018.8461967. 2

[42] Mehmet E. Sargin, Yucel Yemez, Engin Erzin, and Ahmet M. Tekalp. Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30:1330–1345, 2008. doi: 10.1109/TPAMI.2007.70797. 2

[43] Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher. Audio to body dynamics. *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer*

Society Conference on Computer Vision and Pattern Recognition, 06 2018. 2

[44] 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

[45] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 7

[46] 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, 4, 5, 6, 7

[47] 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, 4, 5, 6, 7

[48] Pieter Wolfert, Nicole Robinson, and Tony Belpaeme. A review of evaluation practices of gesture generation in embodied conversational agents. *arXiv preprint arXiv:2101.03769*, 2021. 6, 7

[49] Yuyu Xu, Catherine Pelachaud, and Stacy Marsella. Compound gesture generation: A model based on ideational units. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8637 LNAI:477–491, 2014. ISSN 16113349. doi: 10.1007/978-3-319-09767-1_58. 2

[50] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4303–4309. IEEE, 2019. 2

[51] 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. 3

[52] Xiangxin Zhu, Carl Vondrick, Charless C Fowlkes, and Deva Ramanan. Do we need more training data? *International Journal of Computer Vision*, 119(1):76–92, 2016. 7