Adversarial Training for Multi-Channel Sign Language Production

Ben Saunders
b.saunders@surrey.ac.uk
Necati Cihan Camgöz
n.camgoz@surrey.ac.uk
Richard Bowden
r.bowden@surrey.ac.uk

Centre for Vision, Speech
and Signal Processing
University of Surrey
Guildford, UK

Abstract

Sign Languages are rich multi-channel languages, requiring articulation of both manual (hands) and non-manual (face and body) features in a precise, intricate manner. Sign Language Production (SLP), the automatic translation from spoken to sign languages, must embody this full sign morphology to be truly understandable by the Deaf community. Previous work has mainly focused on manual feature production, with an under-articulated output caused by regression to the mean.

In this paper, we propose an *Adversarial Multi-Channel* approach to SLP. We frame sign production as a minimax game between a transformer-based *Generator* and a conditional *Discriminator*. Our adversarial discriminator evaluates the realism of sign production conditioned on the source text, pushing the generator towards a realistic and articulate output. Additionally, we fully encapsulate sign articulators with the inclusion of non-manual features, producing facial features and mouthing patterns.

We evaluate on the challenging RWTH-PHOENIX-Weather-2014**T** (PHOENIX14**T**) dataset, and report state-of-the art SLP back-translation performance for manual production. We set new benchmarks for the production of multi-channel sign to underpin future research into realistic SLP.

1 Introduction

Sign languages, the principal communication of the Deaf community, are rich multi-channel languages. Communication is expressed through manual articulations of hand shape and motion, in combination with diverse non-manual features including mouth gestures, facial expressions and body pose [48]. The combination of manual and non-manual features is subtle and complicated, requiring a detailed articulation to fully express the desired meaning. Sign Language Production (SLP), the translation from spoken language input to sign language output, is therefore required to encompass the full sign morphology in order to generate an accurate and understandable production.

Although sign languages are inherently multi-channel languages, deep learning based SLP approaches have, to date, focused solely on the manual features of sign [42, 46, 62], producing only the hand and body articulators. Ignoring non-manual features discards the

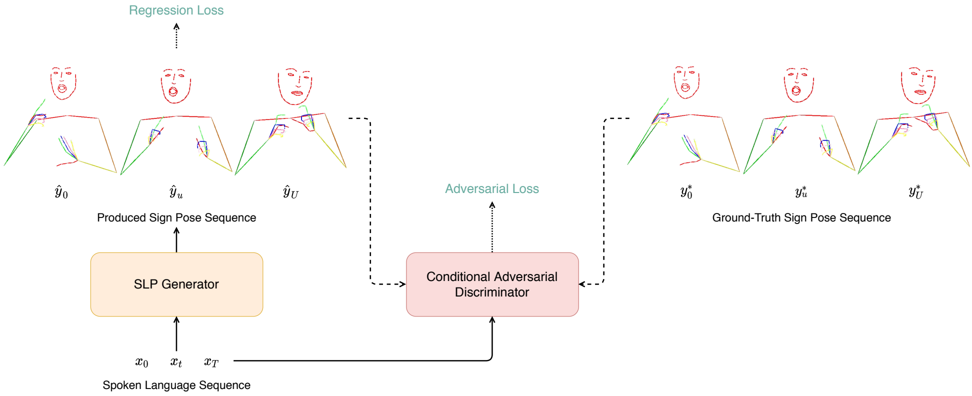

Figure 1: Adversarial Multi-Channel SLP overview, with a Conditional Adversarial Discriminator measuring the realism of Sign Pose Sequences produced by an SLP Generator.

contextual and grammatical information that is required to fully understand the meaning of the produced sign [51]. Mouthing, in particular, is vital to the comprehension of most sign languages, differentiating signs that may otherwise be homophones. Previous SLP models have also been trained using a regression loss [42, 62], which results in an under-articulated production due to the problem of regression to the mean. Specifically, an average sign pose is generated, with non-expressive hand shape and body motion.

In this paper, we propose adversarial training for multi-channel SLP, implementing a discriminator model conditioned on the source spoken language sentence, and expanding production to non-manual features. We frame SLP as a minimax game between a progressive transformer **Generator** that produces a sequence of sign poses from input text, and a conditional **Discriminator** that evaluates and promotes the realism of sign production. Building on the increase in discriminative production, we expand SLP to include **Non-Manual Features**, producing the head motion and mouthing patterns alongside the hands and body for a more expressive output. An overview of our approach is shown in Figure 1.

We evaluate on the RWTH-PHOENIX-Weather-2014**T** (PHOENIX14**T**) dataset using a back translation evaluation, achieving state-of-the-art results for the production of manual features and setting new benchmarks for non-manual and multi-channel production. We provide qualitative examples, demonstrating the impact of adversarial training in increasing the articulation of sign production. The contributions of this paper can be summarised as:

- The first application of conditional adversarial training to SLP, to produce expressive and articulate sign pose sequences
- The first SLP model to fully encapsulate sign articulators through the production of non-manual features
- State-of-the-art SLP results on the PHOENIX14**T** dataset, with baselines for multi-channel sign production

The rest of this paper is organised as follows: We outline the previous work in SLP and adversarial training in Section 2, and the background on machine translation and transformer models in Section 3. We present our Adversarial Multi-Channel approach for SLP in Section 4, with quantitative and qualitative evaluation provided in Section 5. Finally, we conclude the paper in Section 6 by discussing our findings and future work.

2 Related Work

Sign Language Recognition & Translation The goal of vision-based sign language research is to develop systems capable of recognition, translation and production of sign languages [4]. Although studied for the last three decades [43, 49], previous work has mainly focused on Sign Language Recognition (SLR) [6, 26, 27]. These early works relied on manual features to understand sign, but as further linguistic aspects of sign were understood [38, 55], focus shifted to more than just the hands. Subsequent tackling of the modalities of face [24, 53], head pose [34] and mouthings [1, 25] have aided recognition performance.

Recently, Camgoz *et al.* introduced the first end-to-end Sign Language Translation (SLT) approach [7], learning a translation from sign videos to spoken language rather than just recognising the sequence of signs. SLT is more demanding than SLR due to sign language possessing different linguistic rules and grammatical syntax from spoken language [44]. Neural Machine Translation (NMT) networks are predominantly used in SLT [7, 23, 60], translating directly to spoken language or via gloss[1] intermediary. Transformer based models are the state-of-the-art in SLT, jointly learning the recognition and translation tasks [8, 9].

Sign Language Production Previous work into SLP has focused on avatar-based [3, 21, 32, 65] or Statistical Machine Translation (SMT) [19, 28] methods, requiring expensive motion capture or post-processing, with output limited to pre-recorded phrases. Non-manual features have been included in avatar production, such as mouthings [56] and head positions [13], but are often viewed as "stiff and emotionless" with an "absence of mouth patterns" [22].

More recently, there have been approaches to automatic SLP via deep learning [58, 62]. However, these works focus on the production of isolated signs of a set length and order without realistic transitions, resulting in robotic and non-realistic animations that are poorly received by the Deaf [4]. Stoll *et al.* [45, 46] use Generative Adversarial Networks (GANs) to generate a sign language video of a human signer, as opposed to a skeleton pose. Even though the output video is visually pleasing, the approach still relies on the concatenation of isolated signs, which disregards the grammatical syntax of sign.

The closest work to this paper is that of Saunders *et al.* [42], who use a progressive transformer architecture to produce continuous 3D sign pose sequences, utilising a counter decoding to predict sequence length and drive generation. However, the use of regression-based training, even with multiple data augmentation techniques, suffers from the known problem of regression to the mean, resulting in an under-expressed sign production.

All previous deep learning based SLP works produce only manual features, ignoring the important non-manuals. The expansion to non-manual features is challenging due to the requirement of temporal coherence with manual features and the intricacies of facial movement. We expand production to non-manual features via the use of adversarial training to synchronise manual features and produce natural, expressive sign.

Adversarial Training Since being introduced by Goodfellow *et al.* [16], GANs have been used extensively to generate images of increasing realism, pairing a generator and discriminator model in an adversarial training setup. GANs have produced impressive results when applied to image generation [17, 39, 64] and, more recently, video generation tasks [50, 54]. Conditional GANs [36] extend GANs to a dependent setting, enabling generation conditioned on specific external data inputs.

[1] Glosses are a written representation of sign language, and defined as minimal lexical items.

There has been recent progress in using GANs for natural language tasks [31, 61, 63]. Specific to NMT, Wu *et al*. designed Adversarial-NMT [57], complimenting the original NMT model with a Convolutional Neural Network (CNN) based adversary, and Yang *et al*. [59] proposed a GAN setup with translation conditioned on the input sequence.

Specific to human pose generation, adversarial discriminators have been used for the production of realistic pose videos [5, 11, 40]. Ginosar *et al*. show that the task of generating skeleton motion suffers from regression to the mean, and adding an adversarial discriminator can improve the realism of gesture production [14]. Lee *et al*. utilise a conditioned discriminator to produce smooth and diverse human dancing motion from music [30].

3 Background

In this section, we provide a brief background on NMT sequence-to-sequence models, focusing on the recent transformer networks and their application to SLP. The goal of machine translation is to learn the conditional probability $P(Y|X)$ of generating a target sequence $Y = (y_1,...,y_U)$ of U tokens, given a source sequence $X = (x_1,...,x_T)$ with T tokens.

Recurrent Neural Networks (RNNs) were first introduced for sequence-to-sequence tasks, mapping between sequences of different lengths using an iterative hidden state computation [18]. The encoder-decoder architecture was later developed, encoding the source sentence into a "context" vector used to decode the target sequence [12, 47]. However, this context introduced an information bottleneck and long term dependency issues. Attention mechanisms overcame this by expanding the context to a soft-search over the entire source sequence, conditioning each target prediction with a learnt weighting of the encoded tokens [2, 33].

Building on attention mechanisms, Vaswani *et al*. introduced the transformer network, a feed-forward model that replaces recurrent modules with self-attention and positional encoding [52]. Within each encoder and decoder stack, Multi-Headed Attention (MHA) layers perform multiple projections of self-attention, learning complementary representations of each sequence. The decoder utilises a further MHA sub-layer to combine these representations, learning the mapping between source and target sequences in an auto-regressive manner.

Progressive Transformer Model Sign languages are inherently continuous, encompassing fluid motions of hand shape, body pose and facial expressions. As SLP represents sign with continuous joint positions [45, 62], classic symbolic NMT architectures, such as transformers, cannot be applied directly without modification. To tackle this, Saunders *et al*. proposed a progressive transformer architecture [42], an alternative formulation of transformer decoding for continuous sequences. The model employs a counter decoding mechanism that drives generation and enables a prediction of the sequence end, alleviating the need for the classic end of sequence token found in symbolic NMT. Multiple MHA sub-layers are applied over both the source, $x_{1:T}$, and target, $y_{1:U}$, sequences separately, with a final MHA layer used to learn the translation mapping between them. This can be formalised as:

$$[\hat{y}_{u+1}, \hat{c}_{u+1}] = \text{ProgressiveTransformer}(y_u \mid y_{1:u-1}, x_{1:T}) \tag{1}$$

where $\hat{y}_{u+1}$ and $\hat{c}_{u+1}$ are the produced joint positions and counter value respectively, given the source sentence, $x_{1:T}$, and previously predicted target poses, $y_{1:u-1}$. The model can be trained end-to-end using a regression loss of Mean Squared Error (MSE) between the ground truth, y_i^*, and produced, $\hat{y}_i$, sign pose sequences:

$$\mathcal{L}_{Reg} = \frac{1}{U}\sum_{i=1}^{U}(y_i^* - \hat{y}_i)^2 \tag{2}$$

In this paper, we build upon the progressive transformer architecture, employing a conditional adversarial discriminator that supplements the regression loss with an adversarial loss. This mitigates the effect of regression to the mean and prediction drift found in the original architecture. To further improve sign comprehension, we also include production of the non-manual sign features of facial expressions and mouthings.

4 Adversarial Training for Multi-Channel SLP

In this section, we introduce our **Adversarial Training** scheme for **Multi-Channel SLP**, learning to distinguish between real and fake sign pose sequences to ensure the production of realistic and expressive multi-modal sign language. Our objective is to learn a conditional probability $P(Y|X)$ of generating a target sign pose sequence $Y = (y_1,...,y_U)$ of U time steps, given a source spoken language sentence $X = (x_1,...,x_T)$ with T words.

Realistic sign consists of subtle and precise movements of both manuals and non-manuals. However, SLP models often suffer from regression to the mean resulting in under-articulated output, producing average hand shapes due to the high variability of joint positions. To address the under-articulation of sign production, we propose an adversarial training mechanism for SLP. We utilise the previously described progressive transformer architecture (Section 3) as a **Generator**, G, to produce sign pose sequences from input text. To ensure realistic and expressive sign production, we introduce a conditional adversarial **Discriminator**, D, which learns to differentiate real and generated sign pose conditioned on the input spoken language. These models are co-trained in an adversarial manner, with mutually improved performance. The adversarial training scheme for SLP can thus be formalised as a minimax game, with G aiming to minimise the following equation, whilst D maximises it:

$$\min_G \max_D \mathcal{L}_{GAN}(G,D) = \mathbb{E}[\log D(Y^* \mid X)] + \mathbb{E}[\log(1 - D(G(X) \mid X))] \tag{3}$$

where $Y^* = y^*_{1:U}$ is the ground truth sign pose sequence, $G(X)$ equates to the produced sign pose sequence, $\hat{Y} = \hat{y}_{1:U}$, and X is the source spoken language.

In addition to the adversarial training, we incorporate **Non-Manual Feature** production to create a more realistic signer output. Non-manual features are essential in the understanding of sign language, providing grammatical syntax, context and emphasis [38]. In this paper, we model the facial landmarks of the signer, expanding sign pose sequences, Y, to include head nods, mouthings and eyebrow motion. The facial landmarks of a signer can be represented as coordinates, similar to the manuals, enabling a direct regression.

4.1 Generator

Our **Generator**, G, learns to produce sign pose sequences given a source spoken language sequence, integrating the progressive transformer into a GAN framework. Contrary to the standard GAN implementation, we require sequence generation to be conditioned on a specific source input. Therefore, we remove the traditional noise input [16], and generate a sign pose sequence conditioned on the source sequence, taking inspiration from conditional GANs [36].

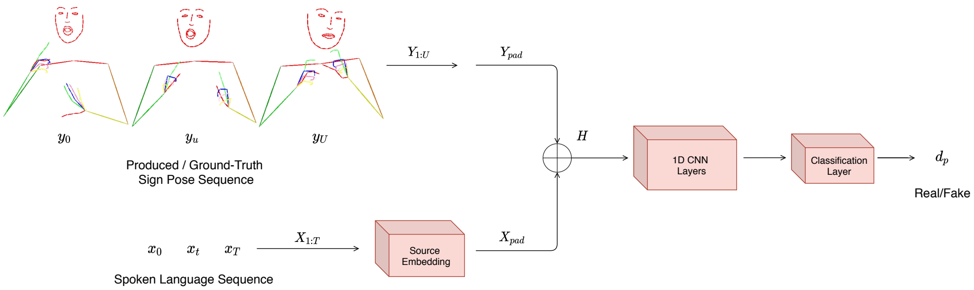

Figure 2: Architecture details of our Conditional Adversarial Discriminator. Sign pose, $Y_{1:U}$, is concatenated with source spoken language, $X_{1:T}$, and projected to a single scalar, d_p.

We propose training G using a combination of loss functions, namely regression loss, $\mathcal{L}_{Reg}$ (Equation 2), and adversarial loss, $\mathcal{L}^{G}_{GAN}$ (Equation 3), with the total loss function as:

$$\mathcal{L}^{G} = \lambda_{Reg}\mathcal{L}_{Reg}(G) + \lambda_{GAN}\mathcal{L}^{G}_{GAN}(G,D) \tag{4}$$

where $\mathcal{L}^{G}_{GAN}$ is the latter component of Equation 3 and λ_{Reg}, λ_{GAN} determines the importance of each loss function during training. The regression loss provides specific details about how to produce the given input, whereas the adversarial loss ensures a realistic signer motion. These losses work in tandem to create both an accurate and expressive sign production.

4.2 Discriminator

We present a conditional adversarial **Discriminator**, D, used to differentiate generated sign sequences, $\hat{Y}$, and ground-truth sign sequences, Y^{*}, conditioned on the source spoken language sequence, X. The aim of D is to measure the realism of sign production, prompting G towards an expressive and articulate output. In parallel, conditioning on the source sequence allows D to concurrently measure the translation accuracy of source-target sequence pair, (X,Y). Figure 2 shows an overview of the discriminator architecture.

For each pair of source-target sequences, (X,Y), of either generated or real sign pose, the aim of the discriminator is to produce a single scalar, $d_p \in (0,1)$, representing the probability that the sign pose sequence originates from the data, Y^{*}:

$$d_p = P(Y = Y^{*} \mid X,Y) \in (0,1) \tag{5}$$

Due to the variable frame lengths of the sign sequences, we apply padding to transform them to a fixed length, U_{max}, the maximum frame length of target sequences found in the data:

$$Y_{pad} = [Y_{1:U}, \varnothing_{U:U_{max}}] \tag{6}$$

where Y_{pad} is the sign pose sequence padded with zero vectors, $\varnothing$, enabling convolutions upon the now fixed size tensor. In order to condition the discriminator on the source spoken language, we first embed the source tokens via a linear embedding layer. Again dealing with variable sequence lengths, these embeddings are also padded to a fixed length T_{max}, the maximum source sequence length:

$$X_{pad} = [W^{X} \cdot X_{1:T} + b^{X}, \varnothing_{T:T_{max}}] \tag{7}$$

where W^X and b^X are the weight and bias of the source embedding respectively and $\varnothing$ is zero padding. As shown in the centre of Figure 2, the source representation is then concatenated with the padded sign pose sequence, to create the conditioned features, H:

$$H = [Y_{pad}, X_{pad}] \tag{8}$$

To determine the realism of the sign pose sequence, the discriminator extracts meaningful representations through multiple 1D CNN layers. These convolutional filters are passed over the sign pose at the sequence level, analysing the local context to determine the temporal continuity of the signing motion. This is more effective than a frame level discriminator at determining realism, as a mean hand shape is a valid pose for a single frame, but not consistently over a large temporal window. Leaky ReLU activation [35] is applied after each layer, promoting healthy gradients during training. A final feed-forward linear layer and sigmoid activation projects the combined features down to the single scalar, d_p, representing the probability that the sign pose sequence is real.

We train the discriminator by maximising the likelihood of producing $d_p = 1$ for real sign sequences and $d_p = 0$ for generated sequences. This objective can be formalised as maximising Equation 3, resulting in the loss function $\mathcal{L}^D = \mathcal{L}^D_{GAN}(G,D)$.

5 Experiments

In this section, we report quantitative and qualitative experimental results. Dataset and evaluation details are provided, with an evaluation of our adversarial SLP model to follow.

5.1 Implementation Details

Dataset: We evaluate our approach on the publicly available PHOENIX14**T** dataset introduced by Camgoz et al. [7]. The corpus provides 8257 German sentences and sign gloss translations alongside parallel sign pose videos of a combined 835,356 frames. We train our adversarial model to generate sign pose sequences of skeleton joint positions. Manual features of each video are extracted in 2D using OpenPose [10], and lifted to 3D using the skeletal model estimation improvements presented in [62]. For non-manual features, we represent facial landmarks as 2D coordinates, again extracted using OpenPose [10]. The face coordinates are scaled to a consistent size and then centered around the nose joint. Each frame is then represented by the normalised joints of the signer, as x, y and z coordinates.

Implementation setup: We setup our adversarial training with a progressive transformer generator built with 2 layers, 4 heads and a 512 embedding size. Our discriminator consists of 3 1D convolution layers, each with a feature size of 64 and a filter size of 10. We jointly train G and D by providing batches of source spoken language and target sign pose sequences, updating the model weights simultaneously with their respective loss functions $\mathcal{L}^G$ and $\mathcal{L}^D$. Experimentally, we find the best generator loss weights to be $\lambda_{Reg} = 100$ and $\lambda_{GAN} = 0.001$.

During testing, we drop D and use the trained G to produce sign pose sequences given an input text. All parts of our network are trained with Xavier initialisation [15] and Adam optimization [20], with a learning rate of 10^{-3}. Our code is based on Kreutzer et al.'s NMT toolkit, JoeyNMT [29], and implemented using PyTorch [37].

	DEV SET					TEST SET				
Configuration:	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE
Regression [42]	11.93	15.08	20.50	32.40	34.01	10.43	13.51	19.19	31.80	32.02
Adversarial (Ours)	12.63	15.83	21.37	32.94	35.11	11.63	14.78	20.49	32.70	33.47
Conditional Adv. (Ours)	**12.74**	**15.97**	**21.68**	**33.95**	**35.83**	**11.70**	**14.95**	**20.86**	**33.51**	**33.64**

Table 1: Adversarial Training results on the Gloss to Pose (G2P) task

	DEV SET					TEST SET				
Configuration:	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE
Regression [42]	11.82	14.80	19.97	31.41	33.18	10.51	13.54	**19.04**	**31.36**	32.46
Conditional Adv. (Ours)	**12.65**	**15.61**	**20.58**	**31.84**	**33.68**	**10.81**	**13.72**	18.99	30.93	**32.74**

Table 2: Adversarial Training results on the Text to Pose (T2P) task

Evaluation: We use the back translation evaluation metric for SLP introduced by Saunders *et al.* [42], employing a pre-trained SLT model [9] to translate the produced sign pose sequences back to spoken language. This is likened to the use of inception score for generative models [41], using a pre-trained classifier. BLEU and ROUGE scores are computed against the original input, with BLEU n-grams from 1 to 4 provided for completeness. The SLP evaluation protocols on the PHOENIX14**T** dataset, set by [42], are as follows: **Gloss to Pose (G2P)** is the production of sign pose from gloss intermediary, evaluating the sign production capabilities; **Text to Pose (T2P)** is the production of sign pose directly from spoken language, requiring both a translation to sign representation and a subsequent production of sign pose.

5.2 Adversarial Training

We start with evaluation of our proposed adversarial training regime, initially producing only manual features to isolate the effect of the adversarial loss. We first conduct experiments on the **Gloss2Pose (G2P)** task, evaluating the production capabilities of our network. As shown in Table 1, our adversarial training regime improves performance over Saunders *et al.*, a model trained solely with a regression loss [42]. This shows that the inclusion of a discriminator model increases the comprehension of sign production. We believe this is due to the discriminator pushing the generator towards a more expressive and articulate production, in order to deceive the adversary. This, in turn, increases the sign content contained in the generated sequence, leading to a more understandable output.

We next experiment with conditioning the discriminator on the source input, to provide discrimination upon both the raw translation and the realism of sign production. As shown, the additional conditioning on the source input improves performance even further. We believe this is due to the generator now requiring a more accurate translation to fool the discriminator, improving the mapping between source input and sign pose.

Our next experiment evaluates the performance of our adversarial training approach for the **Text2Pose (T2P)** task. Table 2 demonstrates that our adversarial model again achieves state-of-the-art results, further showcasing the effect of adversarial training. As the discriminator is conditioned upon the source text, the generator is prompted to accomplish both the accurate translation and realistic production tasks simultaneously.

5.3 Multi-Channel Sign Production

Our final experiment evaluates the production of non-manual features, either independently (Non-M), or in combination with manual features (M + Non-M). We first produce sign using a sole regression loss and subsequently add the proposed adversarial loss, with G2P

	DEV SET					TEST SET				
Configuration:	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE
Regression (Non-M)	7.19	9.13	12.93	23.31	25.01	6.51	8.50	12.44	23.85	24.38
Adversarial (Non-M)	7.39	9.38	13.35	24.38	25.65	7.12	9.10	13.02	24.40	25.16
Regression (M + Non-M)	12.12	15.38	20.97	32.67	35.21	11.54	14.53	20.05	31.63	**34.22**
Adversarial (M + Non-M)	**13.16**	**16.52**	**22.42**	**34.09**	**36.75**	**12.16**	**15.31**	**20.95**	**32.41**	34.19

Table 3: Non-Manual production results on the G2P task (Non-M: Non-Manual, M: Manual)

results shown in Table 3. The sole production of non-manual features contains less signing information than manuals, shown by the relatively low BLEU-4 score of 7.39. This is because facial features complement the manual communication of the hands, providing contextual syntax to emphasise meaning as opposed to independently delivering content.

However, the combination of manual and non-manual feature production significantly increases performance to the highest BLEU-4 score of 13.16. Even the regression model improves performance compared to the manual production alone, highlighting the isolated effect. We believe the multi-channel sign production allows the communication of complementary information, with non-manuals providing further context to increase comprehension. This results in an articulate sign production, moving the field of SLP closer towards a more understandable output. The addition of adversarial training further improves the performance of both non-manual and manual feature production, indicating the ability of our approach to capture the full content of the sign and its morphology.

5.4 Qualitative Experiments

Figure 3 shows example frames of multi-channel sign pose sequences produced by our proposed adversarial training approach, compared against Saunders *et al.* [42]. The examples show an increase in articulation and realism, with a highlight on the importance of non-manual production. Specific to non-manual features, we find a close correspondence to the ground truth video alongside accurate mouthings and head movements.

Figure 4 shows the isolated effect of adversarial training compared to a pure regression approach. Viewed alongside ground truth frames, the produced sign pose demonstrates accurate manual and non-manual production. We find that the addition of adversarial training produces sequences of increased articulation, with a smoother production. Hand shapes can

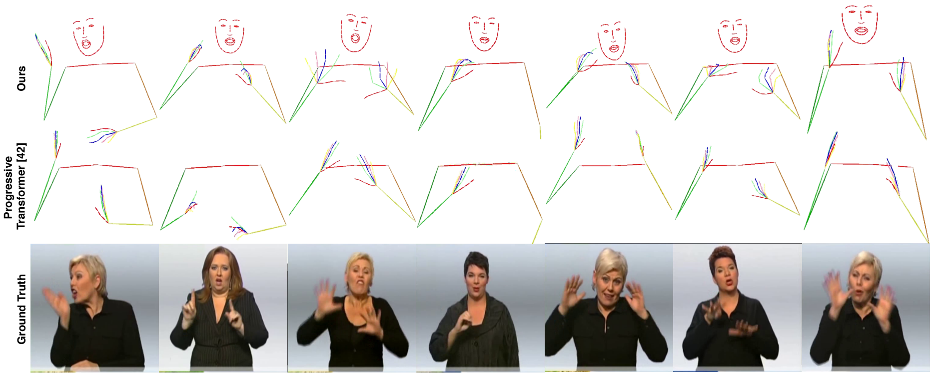

Figure 3: Produced sign pose examples from our proposed model (top) compared to that of [42] (middle), alongside the ground truth frame (bottom)

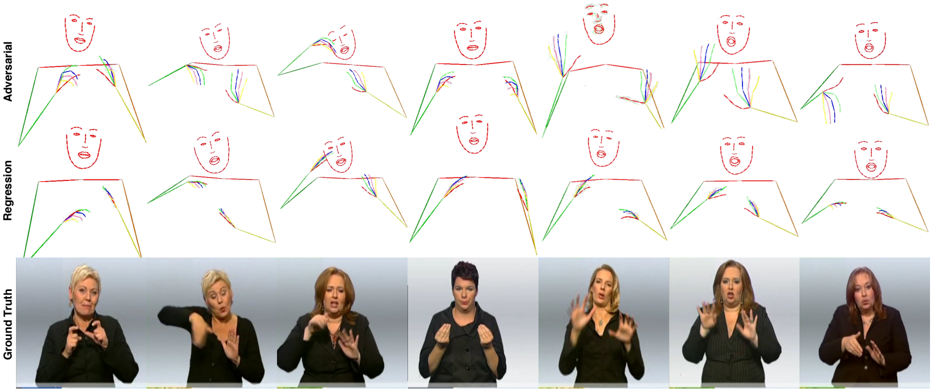

Figure 4: Produced sign pose examples from our proposed adversarial model (top) and a regression model comparison (middle), alongside the ground truth frame (bottom)

be seen to be more expressive and meaningful, an important result for sign comprehension and understandable SLP. Further examples are available in the supplementary materials.

6 Conclusion

Sign languages are visual multi-channel languages and the principal form of communication of the Deaf. Sign Language Production (SLP) requires the production of the full sign morphology in an articulate manner in order to be understood by the Deaf community. Previous deep learning based SLP work has generated only manual features, in an under-expressed production due to the problem of regression to the mean.

In this paper, we proposed an adversarial multi-channel approach for SLP. Framing SLP as a minimax game, we presented a conditional adversarial discriminator that measures the realism of generated sign sequences and pushes the generator towards an articulate production. We also introduced non-manual feature production to fully encapsulate the sign language articulators. We evaluated on the PHOENIX14**T** dataset, showcasing the effectiveness of our adversarial approach by reporting state-of-the-art results for manual production and setting baselines for non-manuals.

As future work, we would like to further increase the realism of sign production by generating photo-realistic human signers, using GAN image-to-image translation models [11, 17, 64] to expand from the skeleton representation. Furthermore, user studies in collaboration with the Deaf are required to evaluate the reception of the produced sign pose sequences.

7 Acknowledgements

This work received funding from the SNSF Sinergia project 'SMILE' (CRSII2 160811), the European Union's Horizon2020 research and innovation programme under grant agreement no. 762021 'Content4All' and the EPSRC project 'ExTOL' (EP/R03298X/1). This work reflects only the authors view and the Commission is not responsible for any use that may be made of the information it contains. We would also like to thank NVIDIA Corporation for their GPU grant.

References

[1] Epameinondas Antonakos, Vassilis Pitsikalis, Isidoros Rodomagoulakis, and Petros Maragos. Unsupervised Classification of Extreme Facial Events using Active Appearance Models Tracking for Sign Language Videos. In *19th IEEE International Conference on Image Processing (ICIP)*, 2012.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[3] J Andrew Bangham, SJ Cox, Ralph Elliott, JRW Glauert, Ian Marshall, Sanja Rankov, and Mark Wells. Virtual Signing: Capture, Animation, Storage and Transmission – an Overview of the ViSiCAST Project. In *Speech and Language Processing for Disabled and Elderly People*, 2000.

[4] Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, and et al. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, 2019.

[5] Haoye Cai, Chunyan Bai, Yu-Wing Tai, and Chi-Keung Tang. Deep Video Generation, Prediction and Completion of Human Action Sequences. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[6] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. SubUNets: End-to-end Hand Shape and Continuous Sign Language Recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[7] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural Sign Language Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[8] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Multichannel Transformers for Multi-articulatory Sign Language Translation. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, 2020.

[9] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[10] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody Dance Now. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, 2019.

[12] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[13] Stephen Cox, Michael Lincoln, Judy Tryggvason, Melanie Nakisa, Mark Wells, Marcus Tutt, and Sanja Abbott. TESSA, a System to Aid Communication with Deaf People. In *Proceedings of the ACM International Conference on Assistive Technologies*, 2002.

[14] 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 (CVPR)*, 2019.

[15] Xavier Glorot and Yoshua Bengio. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2014.

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] Nal Kalchbrenner and Phil Blunsom. Recurrent Continuous Translation Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

[19] Dilek Kayahan and Tunga Güngör. A Hybrid Translation System from Turkish Spoken Language to Turkish Sign Language. In *IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA)*, 2019.

[20] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[21] Michael Kipp, Alexis Heloir, and Quan Nguyen. Sign Language Avatars: Animation and Comprehensibility. In *International Workshop on Intelligent Virtual Agents (IVA)*, 2011.

[22] Michael Kipp, Quan Nguyen, Alexis Heloir, and Silke Matthes. Assessing the Deaf User Perspective on Sign Language Avatars. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*, 2011.

[23] Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. Neural Sign Language Translation based on Human Keypoint Estimation. *Applied Sciences*, 2019.

[24] Oscar Koller, Jens Forster, and Hermann Ney. Continuous Sign Language Recognition: Towards Large Vocabulary Statistical Recognition Systems Handling Multiple Signers. *Computer Vision and Image Understanding (CVIU)*, 2015.

[25] Oscar Koller, Hermann Ney, and Richard Bowden. Deep Learning of Mouth Shapes for Sign Language. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2015.

[26] Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.

[27] Oscar Koller, Necati Cihan Camgoz, Richard Bowden, and Hermann Ney. Weakly Supervised Learning with Multi-Stream CNN-LSTM-HMMs to Discover Sequential Parallelism in Sign Language Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

[28] Dimitris Kouremenos, Klimis S Ntalianis, Giorgos Siolas, and Andreas Stafylopatis. Statistical Machine Translation for Greek to Greek Sign Language Using Parallel Corpora Produced via Rule-Based Machine Translation. In *IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 2018.

[29] Julia Kreutzer, Joost Bastings, and Stefan Riezler. Joey NMT: A Minimalist NMT Toolkit for Novices. *To Appear in EMNLP-IJCNLP 2019: System Demonstrations*, 2019.

[30] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to Music. In *Advances in Neural Information Processing Systems (NIPS)*, 2019.

[31] Kevin Lin, Dianqi Li, Xiaodong He, Zhengyou Zhang, and Ming-Ting Sun. Adversarial Ranking for Language Generation. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[32] Pengfei Lu and Matt Huenerfauth. Data-Driven Synthesis of Spatially Inflected Verbs for American Sign Language Animation. *ACM Transactions on Accessible Computing (TACCESS)*, 2011.

[33] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015.

[34] Marcos Luzardo, Matti Karppa, Jorma Laaksonen, and Tommi Jantunen. Head Pose Estimation for Sign Language Video. In *Scandinavian Conference on Image Analysis*, 2013.

[35] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2013.

[36] Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Nets. *arXiv preprint arXiv:1411.1784*, 2014.

[37] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic Differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017.

[38] Roland Pfau, Josep Quer, et al. *Nonmanuals: Their Grammatical and Prosodic Roles*. 2010.

[39] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv preprint arXiv:1511.06434*, 2015.

[40] Xuanchi Ren, Haoran Li, Zijian Huang, and Qifeng Chen. Music-oriented Dance Video Synthesis with Pose Perceptual Loss. *arXiv preprint arXiv:1912.06606*, 2019.

[41] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[42] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Progressive Transformers for End-to-End Sign Language Production. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[43] Thad Starner and Alex Pentland. Real-time American Sign Language Recognition from Video using Hidden Markov Models. *Motion-Based Recognition*, 1997.

[44] William C Stokoe. Sign Language Structure. *Annual Review of Anthropology*, 1980.

[45] Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. Sign Language Production using Neural Machine Translation and Generative Adversarial Networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[46] Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. Text2Sign: Towards Sign Language Production using Neural Machine Translation and Generative Adversarial Networks. *International Journal of Computer Vision (IJCV)*, 2020.

[47] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2014.

[48] Rachel Sutton-Spence and Bencie Woll. *The Linguistics of British Sign Language: An Introduction*. Cambridge University Press, 1999.

[49] Shinichi Tamura and Shingo Kawasaki. Recognition of Sign Language Motion Images. *Pattern Recognition*, 1988.

[50] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing Motion and Content for Video Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[51] Clayton Valli and Ceil Lucas. *Linguistics of American Sign Language: an Introduction*. Gallaudet University Press, 2000.

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[53] Christian Vogler and Siome Goldenstein. Facial Movement Analysis in ASL. *Universal Access in the Information Society*, 2008.

[54] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating Videos with Scene Dynamics. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[55] Ronnie B Wilbur. Phonological and Prosodic Layering of Nonmanuals in American Sign Language. *The Signs of Language Revisited: An Anthology to Honor Ursula Bellugi and Edward Klima*, 2000.

[56] Rosalee Wolfe, Thomas Hanke, Gabriele Langer, Elena Jahn, Satu Worseck, Julian Bleicken, John C McDonald, and Sarah Johnson. Exploring Localization for Mouthings in Sign Language Avatars. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2018.

[57] Lijun Wu, Yingce Xia, Li Zhao, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. Adversarial Neural Machine Translation. In *Proceedings of The Asian Conference on Machine Learning (ACML)*, 2017.

[58] Qinkun Xiao, Minying Qin, and Yuting Yin. Skeleton-based Chinese Sign Language Recognition and Generation for Bidirectional Communication between Deaf and Hearing People. In *Neural Networks*, 2020.

[59] Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. Improving Neural Machine Translation with Conditional Sequence Generative Adversarial Nets. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (ACL)*, 2017.

[60] Kayo Yin and Jesse Read. Attention is All You Sign: Sign Language Translation with Transformers. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshop on Sign Language Recognition, Translation and Production (SLRTP)*, 2020.

[61] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[62] Jan Zelinka and Jakub Kanis. Neural Sign Language Synthesis: Words Are Our Glosses. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[63] Yizhe Zhang, Zhe Gan, and Lawrence Carin. Generating Text via Adversarial Training. In *Neural Information Processing Systems (NIPS) workshop on Adversarial Training*, 2016.

[64] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[65] Inge Zwitserlood, Margriet Verlinden, Johan Ros, and Sanny Van Der Schoot. Synthetic Signing for the Deaf: Esign. In *Proceedings of the Conference and Workshop on Assistive Technologies for Vision and Hearing Impairment (CVHI)*, 2004.