Sign Language Production using Neural Machine Translation and Generative Adversarial Networks

Stephanie Stoll
s.m.stoll@surrey.ac.uk
Necati Cihan Camgoz
n.camgoz@surrey.ac.uk
Simon Hadfield
s.hadfield@surrey.ac.uk
Richard Bowden
r.bowden@surrey.ac.uk

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

Abstract

We present a novel approach to automatic Sign Language Production using state-of-the-art Neural Machine Translation (NMT) and Image Generation techniques. Our system is capable of producing sign videos from spoken language sentences. Contrary to current approaches that are dependent on heavily annotated data, our approach requires minimal gloss and skeletal level annotations for training. We achieve this by breaking down the task into dedicated sub-processes. We first translate spoken language sentences into sign gloss sequences using an encoder-decoder network. We then find a data driven mapping between glosses and skeletal sequences. We use the resulting pose information to condition a generative model that produces sign language video sequences. We evaluate our approach on the recently released PHOENIX14**T** Sign Language Translation dataset. We set a baseline for text-to-gloss translation, reporting a BLEU-4 score of 16.34/15.26 on dev/test sets. We further demonstrate the video generation capabilities of our approach by sharing qualitative results of generated sign sequences given their skeletal correspondence.

1 Introduction

Sign Languages are the dominant form of communication used by the deaf and hard of hearing. Like spoken languages they have their own grammatical rules and linguistic structures. To facilitate easy and clear communication between the hearing and the Deaf, it is vital to build robust systems that can translate spoken languages into sign languages and vice versa. This two way process can be facilitated using sign language recognition and production. Research into recognition focusses on mapping sign to spoken language typically providing a text transcription of the sequence of signs, such as [8], and [24]. This is due to the misconception that deaf people are comfortable with reading spoken language and therefore do not require translation into sign language. However, there is no guarantee that someone who's first language is, for example, British Sign Language, is familiar with written English, as the two are completely separate languages. Furthermore, generating sign language from spoken language is a complicated task that cannot be accomplished with a simple one-to-one mapping. Unlike spoken languages, sign languages employ multiple asynchronous channels to convey information. These channels include both the manual (i.e. upper body motion, hand shape and trajectory) and non-manual (i.e. facial expressions, mouthings, body posture) features.

The problem of sign language production is generally tackled using animated avatars, such as [6], [10], and [19]. When driven using motion capture data, avatars can produce life-like signing, however this approach is limited to pre-recorded phrases, and the production of motion capture data is costly. Another method relies on translating the spoken language into sign glosses[1], and connecting each entity to a parametric representation, such as the hand shape and motion needed to animate the avatar. However, there are several problems with this method. Translating a spoken sentence into sign glosses is a non-trivial task, as the ordering and number of glosses does not match the words of the spoken language sentence. Additionally, by treating sign language as a concatenation of isolated glosses, any context and meaning conveyed by non-manual features is lost. This results in at best crude, and at worst incorrect translations.

To advance the field of sign language production, we propose a new method, harnessing recent developments in both NMT, and neural network based image/video generation. The proposed method is capable of generating a Sign Language video, given a written or spoken language sentence. An encoder-decoder network translates written sentences into gloss sequences. This is then followed by mapping glosses to skeletal sequences. These sequences are then used to condition a generative model to produce videos containing sign translations of the input sentences (see Figure 1).

The contributions of this paper can be summarised as 1) an NMT-based continuous-text-to-gloss network with subsequent conversion into pose sequences, 2) a generative network conditioned on pose and appearance without the need for a refinement network and 3) to the best of our knowledge, the first end-to-end spoken language-to-sign-language video translation system without the need for costly motion capture or an avatar.

The rest of this paper is organised as follows: Section 2 presents recent developments in Conditional Image Generation and Neural Machine Translation. In Section 3 we introduce our spoken-language-to-sign-language translation system. In section 4 we evaluate our system both quantitatively and qualitatively, before concluding in Section 5.

2 Related Work

Sign language production inherently requires visual content generation and can be treated as a translation problem. Therefore, in this section we review the recent developments in the conditional image generation and NMT fields.

Conditional Image Generation: With the advancements in deep learning, the field of image and video generation has seen various approaches utilising neural-network based architectures. Chen and Koltun [4] used Convolutional Neural Network (CNN) based cascaded refinement networks to produce photographic images given semantic label maps. Similarly, van den Oord et. al. [28] developed PixelCNN, which produces images conditioned on a vector, that can be image tags or feature embeddings provided by another network. Gregor

[1] Glosses are lexical entities that represent individual signs.

et. al. [12] and Van den Oord et. al. [27] also explored the use of Recurrent Neural Networks (RNNs) for image generation and completion. All these approaches rely on either rich semantic and spatial information as input, such as semantic label maps, or they suffer from being blurry and spatially incoherent.

Since the advent of Generative Adversarial Networks (GANs) [11], they have been used extensively for the task of image generation. Soon after their emergence, Mirza and Osindero [20] developed a conditional GAN model, by feeding the conditional information to both the Generator and Discriminator. Radford et. al. [22] proposed Deep Convolutional GAN (DCGAN) which combines the general architecture of a conditional GAN with a set of architectural constraints, such as replacing deterministic spatial pooling with strided convolutions. These changes made the system more stable to train and well-suited for the task of generating realistic and spatially coherent images. Many conditional image generation models have been built by extending the DCGAN model. Notably Reed et. al. [23] have built a system to generate images of birds that are conditioned on positional information and text description, using text embedding and binary pose heat maps. Isola et. al. [13, 30] applied conditional adversarial nets to the related field of image-to-image translation.

An alternative to GAN-based image generation models is provided by Variational Auto-Encoders (VAEs) [14]. Similar to classical auto-encoders, VAEs consist of two networks, an encoder and a decoder. However, VAEs constrain the encoding network to follow a unit gaussian distribution. Yan et. al. developed a conditional VAE [29], that is capable of generating spatially coherent, but blurry images, a tendency of most VAE-based approaches.

Recent work has looked at combining GANs and VAEs to create robust and versatile image generation models. Makhzani et. al. introduced Adversarial Auto-encoders and applied them to problems in supervised, semi-supervised and unsupervised learning [18]. Larsen et. al. have combined VAEs and GANs that can encode, generate and compare samples in an unsupervised fashion [15]. Perarnau et. al. developed Invertible Conditional GANs that use an encoder to learn a latent representation of an input image and a feature vector to change the attributes of human faces [21].

Most relevant to our work, VAE/GAN hybrid models have been used for human pose conditioned image generation. Ma et. al. use a two-stage process to synthesise images of people in arbitrary poses [17]. They achieve this by decoding an input image of a person and combining it with a pose heat map, before using a second network to refine the image. Siarohin et. al. also use two networks to generate human pose guided images. The first is to learn the the appearance of a person, the second learns the affine transformations of body parts using skip connections [25].

Neural Machine Translation NMT utilises RNN based sequence-to-sequence (seq2seq) architectures which learn a statistical model to translate between different languages. Seq2seq [26] [8] has seen success in translating between spoken languages. It consists of two RNNs, an encoder and a decoder, that learn to translate a source sequence to a target sequence. More recently, Camgoz et. al. modified the standard seq2seq framework to translate sign language videos to text [2], which is the inverse of our problem.

Using NMT methods to translate text to poses is a relatively unexplored and open problem. Ahn et. al. use an RNN-based encoder-decoder model to produce upper body pose sequences of human actions from text and map them onto a Baxter robot [1]. However, their results are purely qualitative and rely on human interpretation. In this paper, we simplify the problem by translating text to gloss first and then defining a mapping between glosses and skeletal pose sequences.

3 Text to Sign Language Translation

Our text-to-sign-language translation system consists of three stages: A text-to-gloss NMT network, a learned lookup table to generate motion from gloss sequences, and a pose-conditioned sign generation network consisting of a VAE/GAN hybrid (see Figure 1). We will now discuss each part of our system in detail.

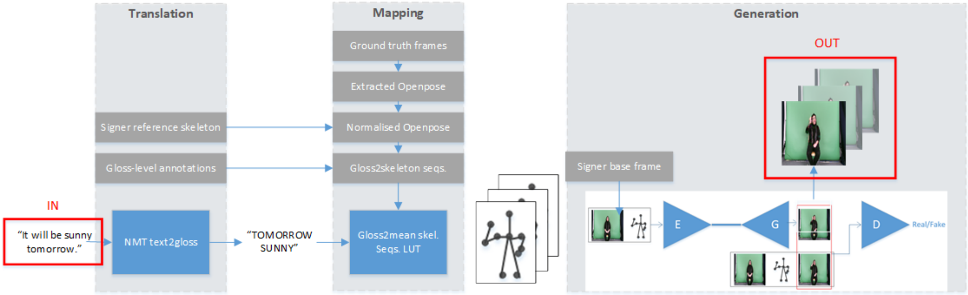


Figure 1: Full System Overview. A spoken language sentence is translated into a representative skeletal pose sequence. This sequence is fed into the generative network frame by frame, in order to generate the input sentence's sign language translation.

3.1 Text to Gloss Translation

We employ state-of-the-art RNN based machine translation methods, namely attention based NMT approaches, to realize spoken language sentence to sign language gloss sequence translation. We use an encoder-decoder architecture [26] with Luong attention [16].

Given a spoken language sentence, $S^N = \{w_1, w_2, ..., w_N\}$, with N number of words, our encoder maps the sequence into a latent representation as in:

$$o_{1:N}, h_S = \text{Encoder}(S^N) \tag{1}$$

where $o_{1:N}$ is the output of the encoder for each word w, and h_S is the hidden representation of the provided sentence. This hidden representation and the encoder outputs are then passed to our decoder, which utilizes an attention mechanism to generate sign gloss sequences, one gloss at a time, as in:

$$gloss_m = \underset{gloss}{\text{argmax}}\, \text{Decoder}(h_S, \alpha(o_{1:N}), gloss_{m-1}) \tag{2}$$

where $\alpha(.)$ is the attention function and $gloss_m$ is the gloss produced at the time step m. The reason we utilize an attention based approach instead of a vanilla sequence-to-sequence based architecture is to tackle the long term dependency issues by providing additional information to the decoder. During training we use cross entropy loss over generated glosses for each time step while for testing we use a beam search decoder.

3.2 Gloss to Skeletal Pose Mapping

To convert each gloss to a sequence of upper-body motions, we build a lookup-table that provides a mapping between sign glosses and 2D skeletal pose sequences. OpenPose [3] was used to extract skeletal joint coordinates from sign videos. We ignore the lower-body joints, as they play no role in sign language and use 10 joints describing the upper body (head, neck, shoulders, elbows, wrists, and hips). Each joint is defined as a pixel in the image with coordinates x and y. To utilise multiple datasets from different domains we perform the following normalisation:

All skeletons are aligned at the neck joint to a chosen reference skeleton:

$$Skel_T = Skel(x,y) + (N_{ref}(x,y) - N_{in}(x,y)). \tag{3}$$

where N_{ref} and N_{in} are the neck joints of the reference and the input skeletons. $Skel_T$ is the translated input skeleton after alignment. We then calculate a scaling factor f using the reference and input skeleton's shoulder-to-shoulder distances:

$$f = \frac{abs(Sl_{ref}(x,y) - Sr_{ref}(x,y))}{abs(Sl_{in}(x,y) - Sr_{in}(x,y))}, \tag{4}$$

where Sl_{ref} and Sr_{ref} are the reference skeleton's and Sl_{in} and Sr_{in} are the input skeleton's left and right shoulder joints, respectively. We finally calculate the normalised skeleton, $Skel_{norm}$, using the previously obtained translated skeleton, $Skel_T$, and the scaling factor, f:

$$Skel_{norm} = N_{ref}(x,y) + (Skel_T(x,y) - N_{ref}(x,y)) * f. \tag{5}$$

To build the gloss to pose sequence lookup-table we group all skeletal sequences by glosses using their gloss annotations. We then align all skeletal sequences for each gloss using dynamic time warping, before combining them into one representative mean skeleton sequence:

$$Skel_{gloss} = \frac{1}{i} \sum_{j=1}^{i} DTW(Skel_{norm_j}, Skel_{norm_0}), \tag{6}$$

where i is the number of example sequences for a gloss, and $Skel_{gloss}$ is the representative mean sequence for a gloss.

3.3 Pose-Conditioned Sign Generation Network

The pose-conditioned sign generation network is a convolutional image encoder followed by a Generative Adversarial Network (GAN). A GAN consists of two models that are trained in conjunction: A generator G that creates new data instances, and a discriminator D that evaluates whether these belong to the same data distribution as the training data. In training, G's aim is to maximise the likelihood of D falsely predicting a sample generated by G to be part of the training data, while D tries to correctly identify samples to be either fake or real. Using this minmax game setup, the generator learns to produce more and more realistic samples, ideally to the point where D cannot separate them from the ground truth.

For our system we adapted the design rules proposed by [22] to build a conditional DCGAN and convolutional image encoder E. E takes an image of a signer in a base pose (not signing) and translates it into its latent representation. It is then decoded by the generator G taking into account the skeletal pose information provided to the network. The discriminator D assesses the output of G using the skeletal information and base pose. See Figure 1 for an overview. We will now provide more detail on all parts of this image generation system.

3.3.1 Image Encoder and Generator

In the encoder, the input image goes through five convolution stages before using two fully connected layers to get a vector representation of the image. The skeletal information is given into the network as a 128x128x10 binary heat map. It is resized and then concatenated to the input image before each convolution step and the first fully connected layer. The original-size heat map then goes through a fully connected layer, with the resulting vector representation being concatenated to the vector representation of the input image. This provides the input for the generator which can be seen as a decoder. It uses up-convolution and resize-convolution to decode the latent vector back into an image using the embedded skeletal information. Additionally, skip connections between the encoder and generator encourage it to produce an output that is close to the input but with the desired spatial differences. The generator is described in more detail in Figure 2.

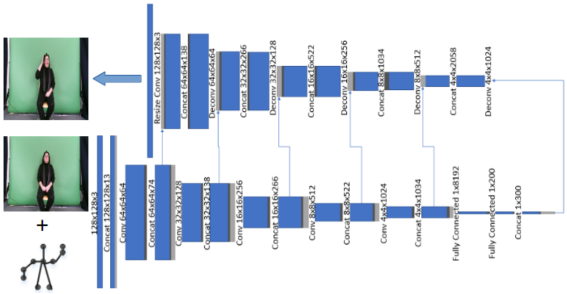


Figure 2: Detailed Generator Description.

3.3.2 Discriminator

The discriminator receives either a tuple of the generated synthetic image or ground truth, the skeletal pose heat map, and base pose input image as input. It decides on image's authenticity. Given that the system is trained on multiple signers, the base pose image is needed to establish whether the generated image resembles the desired signer. The skeletal information is used to assess if the generated image has the desired joint configuration.

3.3.3 Loss

We use adversarial loss, as well as a pixel loss between generated and ground truth images to train our network. The encoder E is treated as part of the generator G, and trained conjointly with G. The discriminator loss is purely adversarial and is defined as:

$$L_{dis} = L_{ce}(D(I_{bp}, I_{tp}, H_{sk}, 1)) + L_{ce}(D(I_{bp}, \hat{I}_{tp}, H_{sk}), 0), \tag{7}$$

where I_{bp} is the base-pose image, I_{tp} is the target-pose ground truth image, $\hat{I}_{tp}$ is the generated target-pose image, and H_{sk} is the skeleton key point heat map. L_{ce} denotes cross-entropy loss. The generator's adversarial loss is defined as:

$$L_{ga} = L_{ce}(D(I_{bp}, \hat{I}_{tp}, H_{sk}), 1). \tag{8}$$

Its pixel loss is the absolute pixel difference between ground truth target-pose and generated target-pose image:

$$L_{pix} = abs(I_{tp} - \hat{I}_{tp}). \tag{9}$$

The generator's total loss is a combination of adversarial and pixel loss:

$$L_{gen} = L_{ga} + L_{pix}. \tag{10}$$

4 Experiments

To evaluate the performance of the approach, we conduct experiments focusing on each component and the system as a whole. We first describe the datasets we have used for training our networks. We then share both quantitative and qualitative results on the text-to-gloss translation network and compare it against the gloss-to-text network performance of [2], as there is no baseline result for spoken language to sign gloss translation. Finally, we conclude this section with a qualitative ablation study evaluating each components' effect on sign production performance.

4.1 Datasets

In order to realise spoken language to sign video generation, we require a large scale dataset, which provides sign language sequences and their spoken language translations.

Although there is vast quantities of broadcast and linguistic data, they lack spoken language sentence to sign sequence (i.e. topic-comment) alignment. However, recently Camgoz et al. released RWTH-PHOENIX-Weather 2014**T** (PHOENIX14**T**) [2], which is the extended version of the continuous sign language recognition benchmark dataset PHOENIX-2014 [9]. PHOENIX14**T** consists of German Sign Language (DGS) interpretations of weather broadcasts. It contains 8257 sequences being performed by 9 signers. It has a sign gloss and spoken language vocabulary of 1066 and 2887, respectively. Each sequence is annotated with both the sign glosses and spoken language translations.

We trained our spoken language to sign gloss network using PHOENIX14**T**. However, due to the limited number of signers in the dataset, we utilised another large scale dataset to train the sign generation network, namely the SMILE Sign Language Assessment Dataset [7]. The SMILE dataset contains 42 signers performing 100 isolated signs for three repetitions in Swiss German Sign Language (DSGS). Although the SMILE dataset is multi-view, we only used the Kinect colour stream.

Using two datasets is motivated by the fact that there is no single dataset that provides both text-to-sign translations and a broad range of signers of different appearance. Using two datasets from different subject domains demonstrates the robustness and flexibility of our method, as it allows us to transfer knowledge between specialised datasets. This makes the approach suitable for translating between different spoken and signed languages, as well as other problems, such as text-conditioned image and video generation.

4.2 German to Gloss Translation

As described in Section 3.1, we utilised a state-of-the-art NMT architecture for spoken language to sign gloss translation. Both our encoder and decoder networks have 4 layers with 1000 Gated Recurrent Units (GRUs) each. As an attention mechanism we use Luong et al.'s approach as it utilises both encoder and decoder outputs during context vector calculation. We trained our network using Adam optimisation with a learning rate of 10^{-5} for 30 epochs. We also employed dropout with 0.2 probability on GRUs to regularise training.

To measure the translation performance of our approach, we used BLEU and ROGUE (ROGUE-L F1) score, which are the most popular metrics in the machine translation domain. We measure the BLEU scores on different n-gram granularities, namely BLEU 1, 2, 3 and 4, to give readers a better perspective of the translation performance.

	DEV SET					TEST SET				
Approach:	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE
Gloss2Text [2]	44.64	31.71	**24.31**	**19.68**	44.91	44.47	31.00	**23.37**	**18.75**	43.88
Text2Gloss (Ours)	**50.15**	**32.47**	22.30	16.34	**48.42**	**50.67**	**32.25**	21.54	15.26	**48.10**

Table 1: BLEU and ROUGE scores for PHOENIX-2014T dev and test data. Our network performs Text2Gloss translation. Gloss2Text scores are provided as a reference.

Our results, as seen in Table 1, show that Text2Gloss performs comparably with the Gloss2Text network presented in [2], which is the opposite task of translating sign glosses to spoken language sequences. While Gloss2Text achieves a higher BLEU-4 score, our Text2Gloss surpasses its performance on BLEU scores with smaller n-gram and ROUGE scores. We believe this is due to shorter length of sign gloss sequences and their smaller

GT Text:	am samstag ist es wieder unbestaendig . (on saturday it is changing again .)
GT Gloss:	SAMSTAG WECHSELHAFT (SATURDAY CHANGING)
Text2Gloss:	SAMSTAG WECHSELHAFT (SATURDAY CHANGING)
GT Text:	am freundlichsten ist es noch im nordosten sowie in teilen bayerns .(It is friendliest still in the north-east as well as parts of Bavaria .)
GT Gloss:	BESONDERS FREUNDLICH NORDOST BISSCHEN BEREICH (ESPECIALLY FRIENDLY NORTH-EAST LITTLE-BIT AREA)
Text2Gloss:	BESONDERS FREUNDLICH NORDOST (ESPECIALLY FRIENDLY NORTH-EAST)
GT Text:	am sonntag ab und an regenschauer teilweise auch gewitter . (on sunday rain on and off and partly thunderstorms .)
GT Gloss:	SONNTAG REGEN TEIL GEWITTER (SUNDAY RAIN PART THUNDER-STORM)
Text2Gloss:	SONNTAG WECHSELHAFT REGEN GEWITTER (SUNDAY CHANGING RAIN THUNDER-STORM)
GT Text:	im suedosten regnet es teilweise laenger . (In the south-east it partially rains longer .)
GT Gloss:	SUEDOST DURCH REGEN (SOUTH-EAST THROUGH RAIN)
Text2Gloss:	SUED LANG REGEN (SOUTH LONG RAIN)

Table 2: Translations from our NMT network. (GT: Ground Truth)

vocabulary. The challange is further exacerbated by the fact that sign languages employ a spatio-temporal grammar which is challenging to represent in text.

We also provide qualitative results by examining sample text-to-gloss translations (see Table 2). Our experiments indicate that, the network is able to produce gloss sequences from text that are close to the gloss ground truth. Even when the predicted gloss sequence does not exactly match the ground truth, the network chooses glosses that are close in meaning.

4.3 Sign Video Generation

The Pose-Conditioned Sign Generation Network was trained on 40 different signers over 90,000 iterations. Out of these signers, one signer was chosen, and the network fine-tuned for another 10,000 iterations on the appearance of this signer. We started training with a learning rate of $2*10^{-4}$ using Adam optimisation, which was lowered to $2*10^{-5}$ after the first 20,000 iterations. The system generates sign video from spoken language with a speed of approx. 4 sec/frame of video.

We evaluate our sign production system at three different stages, comparing it to the PHOENIX14**T** ground truth frames. First we take the extracted and normalised PHOENIX-14**T** skeletal data and pass it to the sign generation network, to produce sign video. This is labelled as "GAN" in Table 3. We next use the PHOENIX14**T** ground truth gloss sequences and pass them through the gloss-to-pose lookup table and the resulting skeleton sequence into the sign generation network. In Table 3 this is called "LUT + GAN". Finally, we test the whole system by passing a German sentence to our text-to-gloss network, its output to the lookup table and then to the sign generation network. This is called "FULL" in Table 3.

We analyse four representative sequences. We display every tenth frame of each video. The text input, ground truth gloss, as well as the predicted gloss is provided for each.

Sequences 1 and 2 represent generated sign videos from a long input. For both cases the NMT manages to predict a sequence of glosses that is close in meaning to the ground truth, albeit with small differences. This is mirrored in the resulting video sequences. Sequence 1 stays close to the ground truth frames for all three test cases, except for number 50 and 70, where both LUT + GAN and FULL diverge slightly from GAN. This is likely due to differences in timing between the gloss generated sequences and the one generated from raw skeletal input. However the fact that LUT + GAN and FULL stay close to each other, demonstrate a successful video generation from the gloss sequences. The gloss translation for Sequence 2 is less close to the ground truth gloss than for Sequence 1. However the produced NMT Gloss sequence manages to stay close to the ground truth for most of the video. At frame number 50, GAN fails to predict the subject's left hand correctly, which is probably due to OpenPose failing to locate it. Both LUT + GAN and FULL succeed in predicting both hands, as they use mean sequences for each gloss, making them more robust to errors in the skeletal annotation.

Sequence 3 and 4 represent video generations from short inputs. Sequence 3's gloss prediction matches the ground truth gloss exactly. This is mirrored in the resulting video sequences for LUT + GAN and FULL, which stay consistent with one another throughout the sequence, and close to the ground truth frames. Again, GAN suffers from some undetected hand key points, which is not carried over into the sequences generated from gloss. Similar results can be reported for Sequence 4. The NMT predicts an extra gloss, but all three sequences stay close to the ground truth frames.

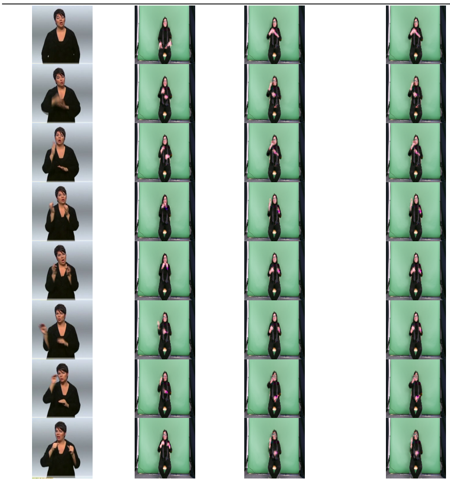
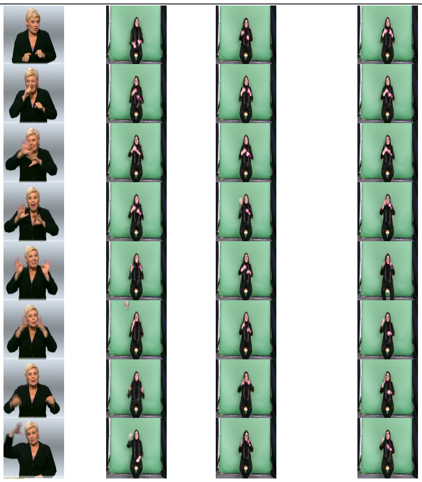
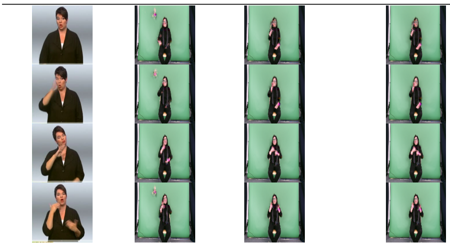
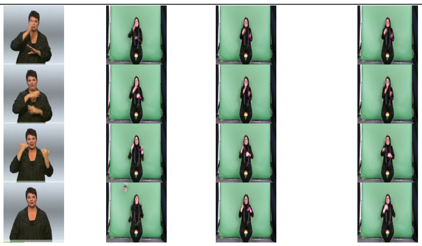
Sequence 1	**Sequence 2**
Text in: im norden maessiger wind an den kuesten weht er teilweise frisch. (in the north moderate winds at the coast it blows partly fresh.) **Gloss GT in:** NORD WIND MAESSIG KUESTE KOENNEN FRISCH WIND (NORTH WIND MODERATE COAST CAN FRESH WIND) **Gloss NMT in:** NORD WIND MAESSIG KUESTE TEIL WIND (NORTH WIND MODERATE COAST PART WIND)	**Text in:** sobald sich der nebel am tag lichtet scheint die sonne. (as soon as the fog lifts during the day the sun shines.) **Gloss GT in:** WENN NEBEL TAG AUFLOESEN SONNE (WHEN FOG DAY DESOLVE SUN) **Gloss NMT in:** NEBEL VERSCHWINDEN IX SONNE (FOG DISSAPPEAR IS SUN)
GT Frames GAN LUT + GAN **FULL**	GT Frames GAN LUT + GAN **FULL**
	
Sequence 3	**Sequence 4**
Text in: richtung norden und westen ist es recht freundlich . (In the north and west direction it is fairly friendly.) **Gloss GT in:** NORDWEST FREUNDLICH (NORTH-WEST FRIENDLY) **Gloss NMT in:** NORDWEST FREUNDLICH (NORTH-WEST FRIENDLY)	**Text in:** im westen ist es freundlich . (It is friendly in the west.) **Gloss GT in:** WEST FREUNDLICH (WEST FRIENDLY) **Gloss NMT in:** WEST MEHR FREUNDLICH (WEST MORE FRIENDLY)
GT Frames GAN LUT + GAN **FULL**	GT Frames GAN LUT + GAN **FULL**
	

Table 3: Four example sequences.

5 Conclusions

In this paper, we presented the first end-to-end spoken language-to-sign language video translation system. While other approaches rely on motion capture data and/or the complex animation of avatars, our deep learning approach combines an NMT-based text-to-gloss model with a sign generation network capable of producing video frames directly which are conditioned on pose and appearance.

For text-to-gloss translation we achieve faithful, and consistent results, employing state-of-the-art NMT methods. Our generative network is capable of producing video sequences faithful to its skeletal pose and appearance conditioning. We are aware that much more work needs to be done to compete with existing avatar approaches. However, we believe that our approach has the potential to provide continuous, end-to-end realistic sign language synthesis, using minimal annotation. This means it will be possible to train our system on various domains and languages, something current approaches lack.

For future work we will focus on increasing the resolution of our video output, and increasing sharpness of the face and hands.

Acknowledgements

This work was funded by the SNSF Sinergia project "Scalable Multimodal Sign Language Technology for Sign Language Learning and Assessment" (SMILE) grant agreement number CRSII2 160811 and the European Union's Horizon2020 research and innovation programme under grant agreement no. 762021 (Content4All). We would also like to thank NVIDIA Corporation for their GPU grant.

References

[1] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. *CoRR*, abs/1710.05298, 2017. URL http://arxiv.org/abs/1710.05298.

[2] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *CoRR*, abs/1611.08050, 2016. URL http://arxiv.org/abs/1611.08050.

[4] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. *CoRR*, abs/1707.09405, 2017. URL http://arxiv.org/abs/1707.09405.

[5] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078, 2014. URL http://arxiv.org/abs/1406.1078.

[6] 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 fifth international ACM conference on Assistive technologies*, pages 205–212. ACM, 2002.

[7] Sarah Ebling, Necati Cihan Camgoz, Penny Braem, Katja Tissi, Sandra Sidler-Miserez, Stephanie Stoll, Simon Hadfield, Tobias Haug, Richard Bowden, Sandrine Tornay, Marzieh Razavi, and Mathew Magimai-Doss. Smile swiss german sign language dataset, 02 2018.

[8] Mohamed Elwazer. Kintrans. 2018. URL http://www.kintrans.com/.

[9] Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In *Language Resources and Evaluation*, pages 1911–1916, Reykjavik, Island, May 2014.

[10] JRW Glauert, R Elliott, SJ Cox, J Tryggvason, and M Sheard. Vanessa–a system for communication between deaf and hearing people. *Technology and Disability*, 18(4): 207–216, 2006.

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.

[12] Karol Gregor, Ivo Danihelka, Alex Graves, and Daan Wierstra. DRAW: A recurrent neural network for image generation. *CoRR*, abs/1502.04623, 2015. URL http://arxiv.org/abs/1502.04623.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016. URL http://arxiv.org/abs/1611.07004.

[14] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. URL http://arxiv.org/abs/1312.6114.

[15] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *CoRR*, abs/1512.09300, 2015. URL http://arxiv.org/abs/1512.09300.

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

[17] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *CoRR*, abs/1705.09368, 2017. URL http://arxiv.org/abs/1705.09368.

[18] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian J. Goodfellow. Adversarial autoencoders. *CoRR*, abs/1511.05644, 2015. URL http://arxiv.org/abs/1511.05644.

[19] John McDonald, Rosalee Wolfe, Jerry Schnepp, Julie Hochgesang, Diana Gorman Jamrozik, Marie Stumbo, Larwan Berke, Melissa Bialek, and Farah Thomas. An automated technique for real-time production of lifelike animations of american sign language. *Universal Access in the Information Society*, 15(4):551–566, 2016.

[20] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. URL http://arxiv.org/abs/1411.1784.

[21] Guim Perarnau, Joost van de Weijer, Bogdan Raducanu, and Jose M. Álvarez. Invertible conditional gans for image editing. *CoRR*, abs/1611.06355, 2016. URL http://arxiv.org/abs/1611.06355.

[22] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015. URL http://arxiv.org/abs/1511.06434.

[23] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 217–225. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6111-learning-what-and-where-to-draw.pdf.

[24] Zsolt Robotka. Signall. 2018. URL http://www.signall.us/.

[25] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe. Deformable GANs for Pose-based Human Image Generation. *ArXiv e-prints*, December 2018.

[26] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to Sequence Learning with Neural Networks. In *Advances in neural information processing systems*, 2014.

[27] Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *CoRR*, abs/1601.06759, 2016. URL http://arxiv.org/abs/1601.06759.

[28] Aäron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixelcnn decoders. *CoRR*, abs/1606.05328, 2016. URL http://arxiv.org/abs/1606.05328.

[29] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. *CoRR*, abs/1512.00570, 2015. URL http://arxiv.org/abs/1512.00570.

[30] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017. URL http://arxiv.org/abs/1703.10593.