Noname manuscript No.
(will be inserted by the editor)

Text2Sign: Towards Sign Language Production using Neural Machine Translation and Generative Adversarial Networks

Stephanie Stoll · Necati Cihan Camgoz · Simon Hadfield · Richard Bowden

Received: date / Accepted: date

Abstract We present a novel approach to automatic Sign Language Production (SLP) using recent developments in Neural Machine Translation (NMT), Generative Adversarial Networks (GANs), and motion generation. 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 pose sequences by combining an NMT network with a Motion Graph (MG). The resulting pose information is then used to condition a generative model that produces photo realistic sign language video sequences. This is the first approach to continuous sign video generation that does not use a classical graphical avatar. We evaluate the translation abilities of our approach on the 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 for both multi-signer and high-definition settings qualitatively and quantitatively using broadcast quality assessment metrics.

Keywords Generative Adversarial Networks · Neural Machine Translation · Sign Language Production

1 Introduction

According to the World Health Organization there are around 466 million people in the world that are deaf or suffer from disabling hearing loss. This equates to 5% of the world population relying on sign languages as their primary form of communication.

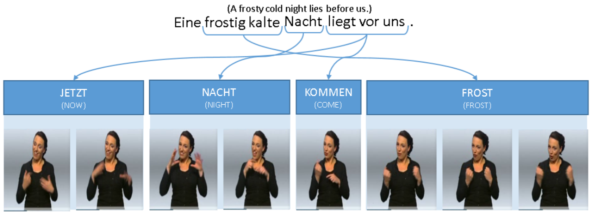

(A frosty cold night lies before us.)
Eine frostig kalte Nacht liegt vor uns .
JETZT
(NOW)
NACHT
(NIGHT)
KOMMEN
(COME)
FROST
(FROST)

Fig. 1 Translating from spoken language text into sign language video. Glosses are used as an intermediate representation. There is often no direct mapping between spoken language and sign language sentences.

Like spoken languages, sign languages have their own grammatical rules and linguistic structures. This makes the task of translating between spoken and signed languages a complex problem, as it is not simply an exercise of mapping text to gestures word-by-word (see Figure 1 which demonstrates that both the tokenization of the languages and their ordering is different). It requires machine translation methods to find a mapping between a spoken and signed language, that takes into account both their language models.

S. Stoll
Centre for Vision, Speech, and Signal Processing
E-mail: s.m.stoll@surrey.ac.uk

N. C. Camgoz
Centre for Vision, Speech, and Signal Processing
E-mail: n.camgoz@surrey.ac.uk

S. Hadfield
Centre for Vision, Speech, and Signal Processing
E-mail: s.hadfield@surrey.ac.uk

R. Bowden
Centre for Vision, Speech, and Signal Processing
E-mail: r.bowden@surrey.ac.uk

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 (SLR) and Sign Language Production (SLP), (see Figure 2).

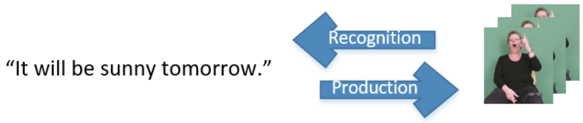


Fig. 2 Sign Language Recognition vs. Production.

Commercial applications for sign language primarily focus on SLR, by mapping sign to spoken language, typically providing a text transcription of the sequence of signs, such as [17], and [48]. 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 (referred to as articulators in linguistics) 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.

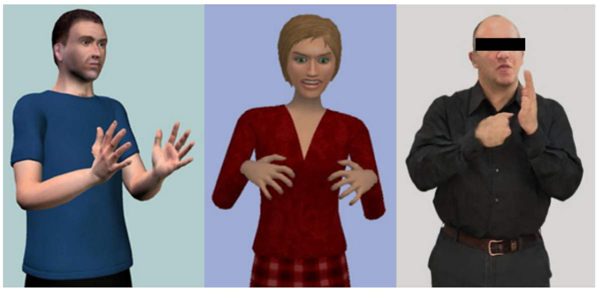


Fig. 3 Traditional avatar-based approaches to SLP compared to our deep generative approach, left: eSign (2005) [57], middle: DictaSign (2012) [16], and right: our photo-realistic approach.

The problem of SLP is generally tackled using animated avatars, such as [12], [20], and [38]. When driven using motion capture data, avatars can produce lifelike 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 (see Figure 1). 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, and results in the indicative 'robotic' motion seen in many avatar based approaches.

To advance the field of SLP, we propose a new approach, harnessing methods from NMT, computer graphics, 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 provides a sequence of gloss probabilities from spoken language text input, that is used to condition a Motion Graph (MG) to find a pose sequence representing the input. Finally, this sequence is used to condition a GAN to produce a video containing sign translations of the input sentence (see Figure 4). The contributions of this paper can be summarised as:

- an NMT-based network combined with a motion graph that allows for continuous-text-to-pose translation.
- a generative network conditioned on pose and appearance.
- To our knowledge the first spoken language to sign language video translation system without the need for costly motion capture or an avatar.

A preliminary version of this work was presented in [51]. This extended manuscript contains an improved pipeline and additional formulation. We introduce an MG into the process, that combined with the NMT network is capable of text-to-pose (text2pose) translations. Furthermore, we demonstrate the generation of multiple signers of varying appearance. We also investigate high-definition (HD) sign generation. Extensive new quantitative as well as qualitative evaluation is provided, exploring the capabilities of our approach. Figure 3 gives a comparison of the output of our approach (right) to other avatar based approaches (left and middle).

The rest of this paper is organised as follows: Section 2 gives an overview of recent developments in NMT as well as traditional SLP using avatars. We explain the

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

concept of motion graphs, before describing recent advancements in generative image models. Section 3 introduces all parts of our approach. In Section 4 we evaluate our system both quantitatively and qualitatively, before concluding in Section 5.

2 Related Work

We treat Sign Language Production (SLP), as a translation problem from spoken into signed language. We therefore first review recent developments in the field of Neural Machine Translation (NMT). However, SLP is different from traditional translation tasks, in that it inherently requires visual content generation. Normally this is performed by animating a 3D avatar. We will therefore give an overview of past and current sign avatar technology. Finally, we cover the concept of Motion Graphs (MGs), a technique used in computer graphics to dynamically animate characters, and the field of conditional image generation.

2.1 Neural Machine Translation

NMT utilises Recurrent Neural Network (RNN) based sequence-to-sequence (seq2seq) architectures which learn a statistical model to translate between different languages. Seq2seq [52, 10] 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. To tackle longer sequences Long Short-Term Memory (LSTM) [23] or Gated Recurrent Units (GRU) [11] are used as RNN cells. Both architectures have mechanisms that allow each cell to pass only the relevant information to the next time step, hence improving translation performance over longer-term dependencies.

To further improve the translation of long sequences Bahdanau et al. [3] introduced the attention mechanism. It provides additional information to the decoder by allowing it to peek at the encoder's hidden states. This mechanism was later improved by Luong et al. [35].

Camgoz et. al. combine a standard seq2seq framework with a Convolutional Neural Network (CNN) to translate sign language videos to spoken language sentences [6], by first extracting features from video using the CNN before translating to text. This can be seen as a kind of inverse to our problem, of translating text to pose.

More recently non-RNN based NMT methods have been explored. ByteNet [27] performs translation using dilated convolutions, and Vaswani et al. [53] introduced the transformer, which is a purely attention-based translation method.

Using NMT methods to translate text to pose 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. For our work we first translate from text to gloss using a seq2seq architecture with Luong attention [35] and GRUs [11], similar to [6]. However, as we are translating text to pose we do not use a CNN as an initial step. In contrast, we use the probabilities produced by the decoder at each time step to solve a Motion Graph (MG) of sign language pose data, to obtain the text to pose translation.

2.2 Avatar Approaches for Sign Language Production

Sign avatars can either be driven directly from motion capture data, or rely on a sequence of parametrised glosses. Since the early 2000s there have been several research projects exploring avatars animated from parametrised glosses, e.g. VisiCast [4], eSign [57], Tessa [12], dicta-sign [16], and JASigning [25]. All of these approaches rely on sign video data to be annotated using a transcription language, such as HamNoSys [45] or SigML [28]. Whilst these avatars are capable of producing sign sequences, they are not popular with the Deaf community. This is due to under-articulated and unnatural movements, but mostly due to missing non-manuals, such as eye gaze and facial expressions (see Figure 3). Important meaning and context is lost this way, making the avatars difficult to understand. Furthermore, the robotic motion of the aforementioned avatars can make viewers uncomfortable, due to the uncanny valley[2][41]. Recent work has begun to integrate non-manuals into the annotation and animation process [14], [15]. However, the correct alignment and articulation of these features poses an unsolved problem, that limit recent avatars such as [38] and [30].

To make avatars both easier to understand, and increase viewer acceptance, recent sign avatars rely on data collected from motion capture. One example of a motion capture driven avatar is the Sign3D project by MocapLab [19]. Given the richness of motion capture data, this approach provides highly realistic results, but is limited to a very small set of phrases, as collecting

[2] The uncanny valley is a concept aimed at explaining the sense of unease people often experience when confronted with simulations that closely resemble humans, but are not quite convincing enough.

and annotating data is expensive, time consuming and requires expert knowledge. So, although these avatars are better received by the Deaf community, they do not provide a scalable solution. The uncanny valley also still remains a large hurdle. To make synthetic signings more realistic, scalable and avoid the aforementioned problems of 3D avatars, we propose to directly generate sign video from weakly annotated data using the latest developments in machine translation, generative image models and Motion Graphs (MGs).

2.3 Motion Graphs

Motion Graphs (MGs) are used in computer graphics to dynamically animate characters, and can be formulated as a directed graph that is constructed from motion capture data. It allows new *lifelike* sequences to be generated that satisfy specific goals at runtime. MGs were independently introduced by Kovar et al. [31], Arikan et al. [2], and Lee et al. [33]. Kovar et al. [31] define the distance between two frames by calculating the distance between two point clouds. For creating the transitions themselves, the motions are aligned and positions are linearly interpolated between joint rotations. As a search strategy, branch and bound is used. Arikan et al. [2] use the difference between joint positions and velocities and the difference between torso velocities and accelerations, to define how close or distant two frames are. A smoothing function is applied to the discontinuity between two clips. The graph is searched by first summarizing it and then performing a random search over the summaries. Lee et al. [33] chose a two layer approach to represent motion data. In the lower layer all data is modelled as a first-order Markov process, where the Markov process is represented by a matrix holding the transition probabilities between frames. The probabilities are derived from measuring the distances of weighted joint angles and velocities. Transitions of low probability are pruned. For blending transitions the hierarchical motion fitting algorithm is used [34]. The higher layer generalises the motion preserved in the lower layer by performing cluster analysis, to make it easier to search. Each cluster represents similar motion, but to capture connections between frames a cluster tree is formed at each motion frame. The whole higher layer is called a cluster forest.

We build an MG for sign language pose data, by splitting continuous sign sequences into individual glosses, and grouping all motion sequences by gloss. These motion sequences populate the nodes of our MG. We then use the probabilities provided by our NMT decoder at each time step to transition between nodes.

2.4 Conditional Image Generation

With the advancements in deep learning, the field of image generation has seen various approaches utilising neural-network based architectures. Chen and Koltun [9] used CNN based cascaded refinement networks to produce photographic images given semantic label maps. Similarly, van den Oord et. al. [42] developed PixelCNN, which produces images conditioned on a vector, that can be image tags or feature embeddings provided by another network. Gregor et. al. [22] and Van den Oord et. al. [43] also explored the use of 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 GANs [21], they have been used extensively for the task of image generation. Soon after their emergence, Mirza and Osindero [40] developed a conditional GAN model, by feeding the conditional information to both the Generator and Discriminator. Radford et. al. [46] 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. [47] 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.

An alternative to GAN-based image generation models is provided by Variational Auto-Encoders (VAEs) [29]. 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 [56], 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 [37]. Larsen et. al. have combined VAEs and GANs that can encode, generate and compare samples in an unsupervised fashion [32]. 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 [44].

VAE/GAN hybrid models have proven particularly popular for generating images conditioned on human pose, as done by Ma et al. [36] and Siarohin et al. [50]. Ma et al. synthesize images of people in arbitrary poses in a two-stage process by fusing an input image of a person for appearance and a heat map providing pose information into a new image in one network, before refining it in a second network. Siarohin et al. use a similar method, but additionally use affine transformations to help change the position of body parts.

In the sub-field of image-to-image translation, Isola et al. introduced pix2pix [24] a conditional GAN, which given its information-rich input and avoidance of fully connected layers was also among the first contenders for generating high definition image content. Building on the success of pix2pix and architecture proposed by Johnson et al. [26], Wang et al. recently presented pix2pixHD [54], a network capable of producing 2048x1024 images from semantic label maps, using a generator and multi-scale discriminator architecture: A global generator consisting of a convolutional encoder, a set of residual blocks and a convolutional decoder, and a local enhancer network of similar architecture, in conjunction provide high resolution images from semantic label maps. Three discriminators are used at different scales to differentiate real from generated images.

For our work, we follow two strands of conditional image generation techniques: We build a multi-person sign generation network conditioned on human appearance and pose, similar to the works of Ma et al. [36] and Siarohin et al. [50]. In addition we also investigate single-signer HD sign generation by building on the work of Wang et al. [54].

3 Text to Sign Language Translation

Our text-to-sign-language (text2sign) translation system consists of two stages: We train an NMT network to obtain a sequence of gloss probabilities that is used to solve a Motion Graph (MG) which generates human pose sequences (text2pose in Figure 4).Then a pose-conditioned sign generation network with an encoder-decoder-discriminator architecture produces the output sign video (see pose2video in Figure 4). We will now discuss each part of our system in detail.

3.1 Text to Pose Translation

We employ recent 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 [52] with Luong attention [35] (see Figure 5).

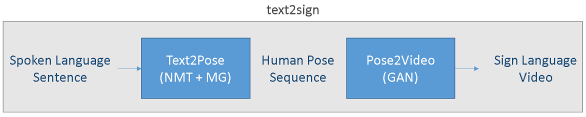


Fig. 4 Full System Overview. A spoken language sentence is translated into a representative skeletal pose sequence. This sequence is fed into the generative network frame by frame, in order to generate the input sentence's sign language translation.

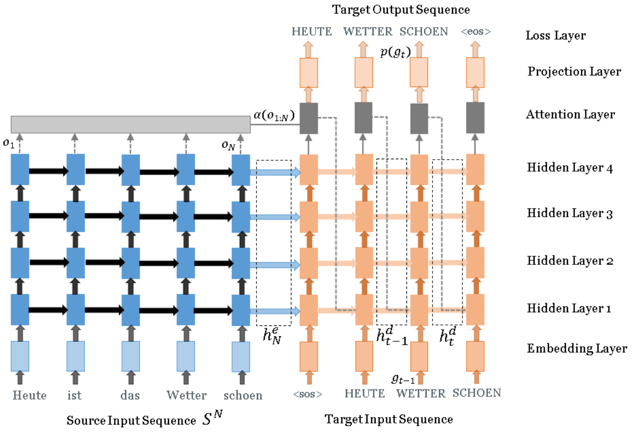


Fig. 5 Our NMT-based encoder-decoder architecture [52] with Luong attention [35].

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:

$$o_{1:N}, h_N^e = \text{Encoder}(S^N) \quad (1)$$

where $o_{1:N}$ is the output of the encoder for each word w, and h_N^e is the hidden representation of the encoded sentence. In Figure 5 the encoder is depicted in blue. This hidden representation and the encoder outputs are then passed to our decoder, which utilizes an attention mechanism and generates a probability distribution over glosses:

$$p(g_t) = \text{Decoder}(g_{t-1}, h_{t-1}^d, \alpha(o_{1:N})) \quad (2)$$

where $\alpha(.)$ is the attention function, g_t is the gloss produced at the time step t and h_{t-1}^d is the hidden state of the decoder passed from the previous time step. At the beginning of the decoding, i.e. $t = 1$, h_{t-1}^d is set as the encoded representation of the input sentence, i.e. $h_0^d = h_N^e$. See Figure 5 for a visualisation.

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. To train our NMT network, we use cross entropy loss over the gloss probabilities at each time step.

We build a Motion Graph (MG) that allows a sequence of 2D skeletal poses to be generated for a given gloss sequence. An MG is a Markov process that can be used to generate new motion sequences that are representative of real motion but fulfil the objectives of the animator e.g. getting from A to B using a specific style of motion. A standard formalisation of an MG is as a finite directed graph of motion primitives [39]: $MG = (V, E)$, where node $v_i \in V$ in the graph corresponds to one or more sequences of motion (motion primitives) and a prior distribution function $p(x_i)$ over those motion primitives (x_i). Each motion primitive for a node is an example of the style of motion the node represents. It is therefore possible to have a variable number of motion primitives in a node, the minimum being one. An edge $e_{i,j} \in E$, which represents an allowable transition from node v_i to v_j, stores a morphable function $\mathbf{Y_{i,j}} = M(x_i, x_j)$ that enables blending between motions, and a probability distribution $p(x_j|x_i)$ over the motion primitives x_j at node v_j, given a chosen motion primitive x_i at node v_i. See Figure 6 for a visualisation.

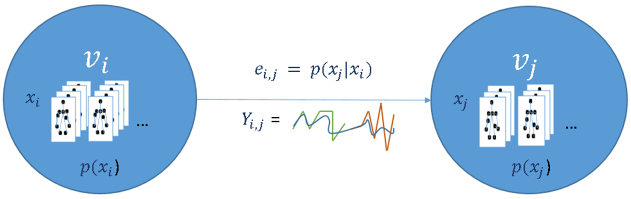


Fig. 6 The graph nodes v_i and v_j contain one or more motion primitives (depicted by skeletal pose maps) x_i and x_j, and a prior distribution $p(x_i)$ and $p(x_j)$. We define the transition probability between nodes v_i and v_j as the probability of motion primitive x_j given x_i. $\mathbf{Y_{i,j}}$ smooths between motion primitives.

The motion primitives need to be extracted from a larger set of motion capture data. This can be done by identifying key frames in the motion data that are at the transition points between motions e.g. the left foot impacting the floor for walking sequences. These key-frames are then used to *cut* the data up into a larger set of motion primitives x_i, where each motion primitive is a continuous motion between two key-frames. For more complex datasets of motion, a typical approach is to define a distance metric between skeletal poses which can be used to identify possible transition points as those that fall below a given threshold. The threshold being set to be small enough such that interpolation between two poses will not cause visual disturbance in the fluidity of motion. For our application, we use the gloss boundaries to automatically *cut* the pose sequences into individual signs so $|V|$ is equal to the gloss vocabulary size and x_g contains all examples of sign gloss g.

In a graphics context E is learned directly from the data by looking at the transitions between nodes in the graph present in the original data. However, in our case, E is generated at each time step by the decoder network, given the previously generated glosses and encoded sequence, as in:

$$\begin{aligned} e_{t-1,t} &= p(x_t|x_{t-1}) \\ &= p(g_t|g_{1:t-1}, S^N) \\ &= \text{Decoder}(g_{t-1}, h^d_{t-1}, \alpha(o_{1:N})). \end{aligned} \tag{3}$$

The purpose of $\mathbf{Y_{i,j}}$ is to allow smooth transition between different motion primitives. In our case it is constant for all nodes in the graph. We use a Savitzky-Golay filter [49] to create smooth transitions. This is done dynamically as the graph is searched. The Savitzky-Golay filter smooths between motion primitives by fitting a low-order polynomial to adjacent data points. We use a window size of five and a polynomial order of two to smooth between the last five frames of the current motion primitive and the first five frames of the upcoming primitive. This allows us to preserve the articulation of each motion primitive, but avoid discontinuities and artefacts at transition points.

To find the most probable motion sequence given a spoken language sentence, we employ beam search over our motion graph. We start generating our sequence from the special x_0 = <bos> (beginning of sequence) node. At each motion step, we consider a list of hypotheses, $\mathcal{H}^B = \{H_1, ..., H_b, ..., H_B\}$ where B denotes our beam width.

At each step we expand our hypotheses with a new motion as in:

$$H^t_b = \{H^{t-1}_b, x^*_t\}, \tag{4}$$

where H^t_b denotes the set of motions in H_b at step t. We choose x^*_t by:

$$x^*_t = \underset{x}{\operatorname{argmax}}\, p(x|x_{t-1}), \tag{5}$$

where $x_{t-1} \in H^{t-1}_b$. We continue expanding our hypotheses until all of them reach to the special $x_{\_}$ = <eos> (end of sequence) node. We then choose the most probable motion sequence $\mathcal{H}^*$ by:

$$\mathcal{H}^* = \underset{H_b}{\operatorname{argmax}} \prod_{i=1}^{|H_b|} p(x_i|x_{i-1}). \tag{6}$$

3.2 Pose to Video Translation

The pose-to-video (pose2video) network combines a convolutional image encoder and a Generative Adversarial Network (GAN), see Figure 7 for an overview. 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. During training, G aims 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.

G is an encoder-decoder, conditioned on human pose and appearance, with a latent space. This latent space can either be a fixed-size one-dimensional vector, or a variable-size residual block. A fixed size 1D vector latent space using a fully connected layer allows generation of images with both large appearance and spatial change and is employed for multi-signer (MS) output. However, the ability to generate spatial change, and the requirement for fully connected layers increasing memory consumption limits the output size of the generated images. In contrast, A fully convolutional latent space, such as a number of residual layers, allows for changes in appearance, like changing from a pose label map to an image of a human being in that pose, but does not allow for large spatial changes. This enables the network capable of style transfer similar to pix2pixHD by Wang et al. [54] or Chan et al. [8]. However, due to the avoidance of fully connected layers and with the use of an additional enhancing network, it is capable of producing sharp high definition outputs. We investigate this second formulation for generating high-definition (HD) sign video.

3.2.1 Image Generator

As input to the generator we concatenate P_t and I_a in the depth dimension, where P_t is a human pose label map. For MS generation I_a is an image of an arbitrary human subject in a resting position (base pose). The HD sign generator cannot be conditioned on a base pose, as it does not allow for large spatial changes. Instead it is conditioned on the generated image from the previous time step. On top of helping with appearance this enforces temporal consistency.

The input to the generator is pushed through the convolutional encoder part of the generator and encoded into the latent space. The decoder part of the generator uses up-convolution and resize-convolution to decode from the latent space back into an image using the embedded skeletal information provided by the label map P_t. This produces an image $G(P_t, I_a)$ of the signer in the pose indicated by P_t (see Figure 7).

In the HD sign variant, an enhancer network En is used to upscale and refine the output images produced by the generator G. Its architecture is very similar to G, consisting of a convolutional encoder, a residual block and an up-convolutional decoder. G is first trained individually, followed by En, before training both networks in conjunction.

3.2.2 Discriminator

The discriminator D receives either a tuple of the generated synthetic image $G(P_t, I_a)$ or ground truth I_t, and the pose label map P_t as input. In the MS case, D is also provided with I_a (see Figure 7). D decides on image's authenticity. In the MS case, given that the system is trained on multiple signers, I_a is used to establish whether the generated image resembles the desired signer. The skeletal information provided by P_t is used to assess if the generated image has the desired joint configuration. For the HD sign case, like Wang et al. [54] we use a multi-scale discriminator with three scales (in our case 1080x720, 540x360, and 270x180).

3.2.3 Loss

We use the GAN's adversarial loss, as well as an L1 loss between generated and ground truth images to train our networks. See Figure 7 for a visualisation. The overall loss is therefore defined as:

$$\mathcal{L} = L_{GAN} + \delta L_1, \tag{7}$$

where δ weighs the influence of L_1.

For MS generation we give I_a to the generator and the discriminator to distinguish between signers. The adversarial loss is thus defined as:

$$\begin{aligned} L_{GAN_{ms}}(G, D) = & \underset{(P_t, I_t, I_a)}{\mathbb{E}} [logD(P_t, I_t, I_a)] + \\ & \underset{(P_t, I_a)}{\mathbb{E}} [log(1 - D(P_t, G(P_t, I_a), I_a))]. \end{aligned} \tag{8}$$

The MS L1 loss is defined as the sum of absolute pixel difference between ground truth and generated

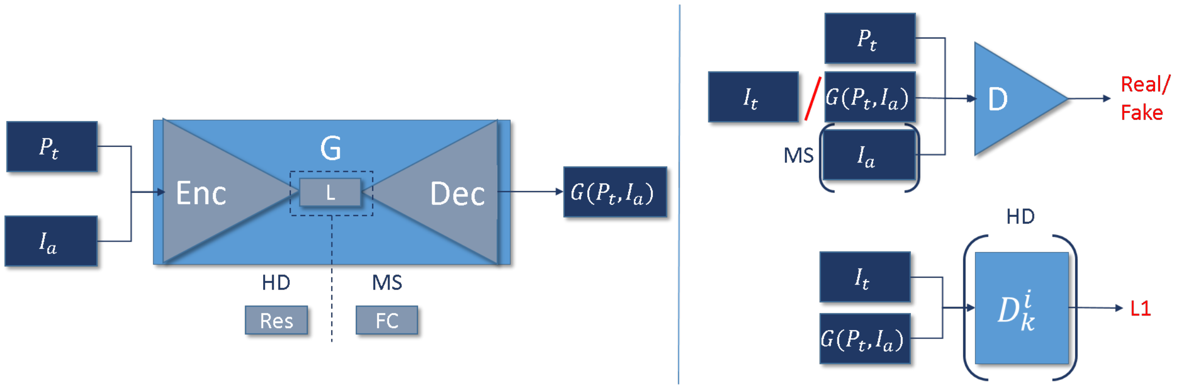


Fig. 7 Our sign generator G has an encoder-decoder structure. It can be conditioned on human pose and appearance. For this we use a human pose label map P_t and a frame of the signer I_a. The latent space Res for the HD generator is a block of residual layers, whereas the the latent space FC for the MS generator is encoded in a 1-dimensional vector using a fully connected layer. We employ two losses: an adversarial loss using the discriminator D, and an L1 loss. For the MS case we take a pixel-based L1 loss, whereas for the HD case we match extracted features from multiple layers of D and calculate the L1 distance between features.

image:

$$L_{1_{ms}}(G) = \sum |(I_t - G(P_t, I_a))|. \tag{9}$$

For HD generation the adversarial loss is defined as:

$$\begin{aligned} L_{GAN_{hd}}(G, D_k) = & \mathop{\mathbb{E}}_{(P_t,I_t)}[log D_k(P_t, I_t)] + \\ & \mathop{\mathbb{E}}_{(P_t)}[log(1 - D_k(P_t, G(P_t, I_a)))], \end{aligned} \tag{10}$$

where k is the number of discriminator scales. To combine the adversarial losses of all D_k, we sum:

$$L_{GAN_{hd}}(G, D) = \sum_{k=1,2,3} L_{GAN_{hd}}(G, D_k). \tag{11}$$

For HD generation the L1 loss is based on the feature matching loss presented in [54]. Features extracted from multiple stages of the discriminator are matched, rather than pixels:

$$\begin{aligned} L_{1_{hd}}(G, D_k) = \mathop{\mathbb{E}}_{(P_t,I_t)} \sum_{i=1}^{T} \frac{1}{N_i} \Big[\sum & |D_k^{(i)}(P_t, I_t) \\ & - D_k^{(i)}(P_t, G(P_t, I_a))| \Big], \end{aligned} \tag{12}$$

where T is the total number of layers in D_k, i is the current layer of D_k, N_i is the total number of elements per layer, and $D_k^{(i)}$ is the ith layer feature extractor of D_k. Again we sum the $L1$ losses of all D_k to obtain the overall $L1$ loss:

$$L_{1_{hd}}(G, D) = \sum_{k=1,2,3} L_{1_{hd}}(G, D_k). \tag{13}$$

4 Experiments

We first introduce the datasets used and any necessary pre-processing steps, before evaluating all sub-parts of our system both quantitatively and qualitatively. We show results for translating spoken language text to gloss sequences and pose sequences, and for generating multi-signer (MS) and high-definition (HD) sign video, using broadcast quality assessment metrics. A set of qualitative examples showcases the current state of the full preliminary translation pipeline.

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 data available and many linguistically annotated datasets, they lack spoken language sentence to sign sequence (i.e. topic-comment) alignment. However, recently Camgoz et al. released RWTH-PHOENIX-Weather2014**T** (PHOENIX14**T**) [6], which is the extended version of the continuous sign language recognition benchmark dataset PHOENIX-2014 [18]. 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 pose network using PHOENIX14**T**. However, due to the limited number of signers in the dataset, we utilised another large scale dataset to train the multi-signer (MS) generation network, namely the SMILE Sign Language Assessment Dataset [13]. 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, without any depth information or the Kinect's built-in pose estimations.

We trained the HD sign generation network on 1280x720 HD dissemination material acquired by the Learning to Recognise Dynamic Visual Content from Broadcast Footage (Dynavis) project [5]. It consists of multiple videos featuring the same subject performing continuous British Sign Language (BSL) sequences. There is no alignment between spoken language sentences to sign sequences.

Using multiple datasets is motivated by the fact that there is no single dataset that provides text-to-sign translations, a broad range of signers of different appearance, and high definition signing content. Using datasets from different subject domains and languages 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.

4.1.1 Data Pre-Processing

In order to perform translation from spoken language to sign pose, we need to find pose sequences that represent the appropriate glosses. We split the continuous samples of the PHOENIX14**T** dataset by gloss using a forced alignment approach. Then, for each gloss we perform a normalisation over all example sequences for each gloss. First, we have to relate the different body shapes and sizes of all signers to that of a selected target subject. Additionally we have to time-align all example sequences, before we can find an average representation for each frame of the sequence. To align different signers' skeletons to that of a target subject, we use OpenPose [7] to extract upper body key points for each frame in the sequence and for a reference frame of the target subject. We align the skeletons at the neck joint and scale by the shoulder width. We use dynamic time warping to time align sequences, before taking the mean of each joint per frame over all example sequences to generate a representative mean sequence. These mean sequences form the nodes of our MG. We decided to use mean sequences rather than raw example sequences, as they provide a more stable performance. We found that corruptions in the gloss boundary information obtained by forced alignment produced an immense variability in quality and correctness for the samples per node in the graph. Whilst we agree that taking the mean sequence instead is not the ideal solution to this problem, it allows us to estimate the overall effectiveness of our approach.

4.2 German to Pose Translation

We provide results for translating German sentences into their intermediate gloss representation, and show how this combined with a MG can be used to generate human pose sequences from spoken language sentences.

As described in Section 3.1, we utilised an encoder-decoder 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. During inference the width B of the beam decoder is set to three, meaning the three top hypotheses are kept per time step. We found this number to be a good trade-off between translation quality and computational complexity. For text2pose generation we report an average time of 0.79 seconds per translated gloss using a Intel Core i7-6700 CPU (3.40 GHz, 8MB cache), where the majority of time is taken up by generating the pose maps (0.77 seconds/gloss).

4.2.1 Translating German to Gloss

To measure the translation performance of our approach we used BLEU and ROUGE (ROUGE-L F1) score as well as Word Error Rate (WER), which are amongst 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.

We compare our Text2Gloss performance against the Gloss2Text network of Camgoz et al. [6], which is the opposite task of translating sign glosses to spoken language sequences. We do this as to our knowledge there is no other text-to-gloss translation approach for a direct comparison. We aim to give the reader context, rather than claiming to supersede the Gloss2Text approach [6]. Our results, as seen in Table 1, show that Text2Gloss performs comparably with the Gloss2Text

	DEV SET						TEST SET					
Approach:	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	WER	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	WER
Gloss2Text [6]	44.64	31.71	**24.31**	**19.68**	44.91	9.90	44.47	31.00	**23.37**	**18.75**	43.88	9.31
Text2Gloss (Ours)	**50.15**	**32.47**	22.30	16.34	**48.42**	**4.83**	**50.67**	**32.25**	21.54	15.26	**48.10**	**4.53**

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

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)
GT Text:	der tag beginnt ganz im osten noch freundlich spter zeigt sich dann auch im nordwesten hufiger die sonne sonst berwiegen die wolken . (The day begins friendly right in the west, later the sun shows itself more often in the north-west, othwerwise clouds prevail .)
GT Gloss:	OST REGION ANFANG FREUNDLICH SPAETER NORDWEST AUCH SONNE SONST REGION WOLKE (EAST REGION BEGINNING FRIENDLY LATER NORTHWEST ALSO SUN OTHERWISE RAIN CLOUD)
Text2Gloss:	MORGEN OST REGION FREUNDLICH WEST REGION SONNE HABEN REGION UEBERWIEGEND WOLKE (TOMORROW EAST REGION FRIENDLY WEST REGION SUN HAVE REGION MOSTLY CLOUD)
GT Text:	besonders im osten deutschlands kann es ein wenig regnen oder schneien . (Especially in the east of Germany it can rain a bit or now .)
GT Gloss:	BESONDERS OST DEUTSCH LAND MEHR REGEN ODER SCHNEE (ESPECIALLY EAST GERMAN LAND MORE RAIN OR SNOW)
Text2Gloss:	ABER OST SUEDOST DOCH ANFANG REGEN SCHNEE (BUT EAST SOUTHEAST HOWEVER BEGINNING RAIN SNOW)
GT Text:	schwacher bis miger wind aus nord bis west . (Weak up to moderate wind from the north to the west .)
GT Gloss:	WIND SCHWACH MAESSIG WEHEN (WIND WEAK MODERATE BLOW)
Text2Gloss:	SCHWACH MAESSIG WEHEN (WEAK MODERATE BLOW)

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

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

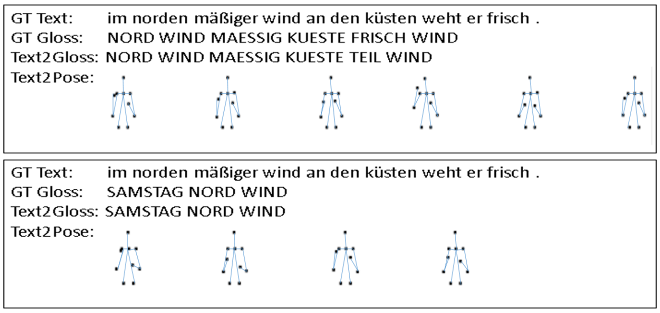
We also provide qualitative results by examining sample Text2Gloss translations (see Table 2). Our experiments indicate that the network is able to produce gloss sequences from text that are close to the gloss ground truth. Even when the predicted gloss sequence does not exactly match the ground truth, the network chooses glosses that are close in meaning.

After reporting these promising intermediate results, we will now show how this approach can be extended to generate human pose maps that encode the motion of signs.

4.2.2 Translating German to Pose

We solve a MG using the NMT's beam search probabilities to generate human pose maps from German sentences. We give a qualitative evaluation of translating German sentences into human pose sequences by solving a MG using the NMTs beam search. Figure 8 shows two examples. In both cases we show key frames that are indicative of the translated glosses. It is interesting to note that both sequences contain the gloss WIND, twice in the top sequence and once in the bottom sequence. The relevant key frames for each occurrence (key frame 2 and 6 for the top sequence, key frame 4 for the bottom sequence) are very similar, showing the conditioning of poses on a specific gloss.

The poses are encoded as 128x128x10 binary label maps, where each joint inhabits one of the 10 depth channels. This type of map is used to generate sign language video in Section 4.3 and 4.4.



GT Text: im norden mäßiger wind an den küsten weht er frisch .
GT Gloss: NORD WIND MAESSIG KUESTE FRISCH WIND
Text2Gloss: NORD WIND MAESSIG KUESTE TEIL WIND
Text2Pose:

GT Text: im norden mäßiger wind an den küsten weht er frisch .
GT Gloss: SAMSTAG NORD WIND
Text2Gloss: SAMSTAG NORD WIND
Text2Pose:

Fig. 8 We show results for translating text to human pose sequences by solving an MG using the NMT's beam search. We show pose label maps for the MS generator. For visualisation we have condensed the 10 depth channel pose maps into 1 depth channel binary images, and inverted colour channels, where each joint is now represented by a black dot. For better interpretability we added bones connecting the joints in blue.

4.3 Multi-Signer Generation of Isolated Signs

This section presents results using the generated label maps to condition a GAN that generates sign video for multi-signer (MS) video generation. We test using isolated signs from the SMILE dataset. When testing on a GeForce GTX TITAN X we report an average time of 1.71 seconds per generated image.

We generate synthetic sign video from previously unseen label data. To evaluate the quality of the generated output, we use the Structural Similarity Index Measurement (SSIM) [55], Peak Signal-to-Noise Ratio (PSNR), and Mean Squared Error (MSE), three well-known metrics for assessing image quality.

SSIM is a metric used to assess the perceptual degradation of images and video in broadcast, by comparing a corrupted image to its original. We adapt this approach to compare the generated synthetic image $G(P_t, I_a)$ to its ground truth image I_t.

For ease of notation we define:

$$\hat{I}_t = G(P_t, I_a). \tag{14}$$

$$SSIM(\hat{I}_t, I_t) = [l(\hat{I}_t, I_t)]^{\alpha} \cdot [c(\hat{I}_t, I_t)]^{\beta} \cdot [s(\hat{I}_t, I_t)]^{\gamma}, \tag{15}$$

where $l(\hat{I}_t, I_t)$ is a luminance term:

$$l(\hat{I}_t, I_t) = \frac{2\mu_{\hat{I}_t}\mu_{I_t} + C_1}{\mu^2_{\hat{I}_t} + \mu^2_{I_t} + C_1}, \tag{16}$$

$c(\hat{I}_t, I_t)$ is a contrast term:

$$c(\hat{I}_t, I_t) = \frac{2\sigma_{\hat{I}_t}\sigma_{I_t} + C_2}{\sigma^2_{\hat{I}_t} + \sigma^2_{I_t} + C_2}, \tag{17}$$

and $s(\hat{I}_t, I_t)$ is a structural term:

$$s(\hat{I}_t, I_t) = \frac{\sigma_{\hat{I}_t I_t} + C_3}{\sigma_{\hat{I}_t}\sigma_{I_t} + C_3}, \tag{18}$$

with $\mu_{\hat{I}_t}$, and μ_{I_t} being the means, $\sigma_{\hat{I}_t}$, and σ_{I_t} the standard deviations and $\sigma_{\hat{I}_t I_t}$ the cross-covariance for images $\hat{I}_t$ and I_t. $C_1 = (k_1 L)^2$ and $C_2 = (k_2 L)^2$, where L is the dynamic range of pixel values, and $k_1 = 0.01$ and $k_2 = 0.03$. C_3 is set to equal $C_2/2$.

With default values of $\alpha, \beta, \gamma, = 1$ the expression for SSIM simplifies to:

$$SSIM(\hat{I}_t, I_t) = \frac{(2\mu_{\hat{I}_t}\mu_{I_t} + C_1)(2\sigma_{\hat{I}_t I_t} + C_2)}{(\mu^2_{\hat{I}_t} + \mu^2_{I_t} + C_1)(\sigma^2_{\hat{I}_t} + \sigma^2_{I_t} + C_2)}. \tag{19}$$

The calculated SSIM ranges from −1 to 1, with 1 indicating the images are identical.

PSNR and MSE are metrics used to assess the quality of compressed images compared to their original. We use MSE to calculate the average squared error between a synthetic image $\hat{I}_t$ and its ground truth image I_t, by:

$$MSE = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}[I_t(m,n) - \hat{I}_t(m,n)]^2 \tag{20}$$

where N and M are the number of columns and rows respectively.

In contrast PSNR measures the peak error in dB, using the MSE:

$$PSNR = 10log_{10}\left(\frac{R^2}{MSE}\right), \tag{21}$$

where R is the maximum possible value of the input data, in this case 255 for 8-bit unsigned integers.

The MS generation network was trained on 40 different signers from the SMILE dataset over 90,000 iterations. Out of these signers, several signers were chosen, and the network fine-tuned for another 10,000 iterations on the appearance of those signers. The pose label maps were generated from running OpenPose on the full-size SMILE ground truth footage of 1920x1080 pixels, and then downsampled to 128x128 pixels. The original SMILE footage was then also downsampled to 128x128 pixels to function as input to our network.

We test for three different signers, over a 1000 frames each. We report the mean SSIM, PSNR, and MSE (see Table 3). The results indicate that the images produced of all three signers are very close to their ground truth, with SSIM values close to 1. Signer 1 has slightly worse scores than signer 2, and 3, which is due to a corrupted sequence in the gathered data.

	SSIM	PSNR	MSE
Signer 1	0.9378	23.697	191.4903
Signer 2	0.9449	26.4280	154.7963
Signer 3	0.9444	26.6884	153.5546

Table 3 Mean SSIM, PSNR, and MSE values over the test set, comparing synthetic images to their ground truth. For SSIM the range is -1 to +1, with +1 indicating identical images. The lower the MSE between two images, the more alike they are, whereas we want to maximise the PSNR between two images.

Qualitative results in Figure 9 and 10 show that the synthetic sequences generated by our network stay close to their ground truth in terms of both motion and appearance. Details for hands and faces are largely preserved, however the network can struggle to form both arms and hands fully, especially when held in front of the chest and face. This is likely due to the similarity in colour, which also could have led to errors in the key point extraction process.

The results also highlight the power of our data-driven approach to capture natural variations in sign. Signer 2 is left-handed, whereas signer 1 and 3 are right-handed. There are also noticeable discrepancies in speed and size of motion amongst the signers. Linguistically, these are very important factors that can have a significant impact on the meaning of a sign. They convey additional information such as emotion and intent, for example haste, anger, or uncertainty.

Overall our experiments show that our MS generation network is capable of synthesizing sign language videos that are highly realistic and variable in terms of motion and appearance for multiple signers. The limiting factor to this approach is the small aspect ratio of 128x128 pixels. We therefore investigate a different variant of our network to produce HD sign videos in Section 4.5.

4.4 Spoken Language to Sign Language Translation

In this section we test the full translation pipeline: going from spoken language sentences to sign language video translations. We translate from German to German Sign Language (DGS). Our test data is taken from the PHOENIX14**T** test dataset. Our Motion Graph (MG) is built from extracting OpenPose skeletal information from the PHOENIX14**T** training set. The obtained OpenPose extraction was prone to errors due to the small resolution of the PHOENIX14**T** data (260x210 pixels). It did not scale to the 1080x720 resolution required for conditioning the HD generator. We therefore only test our full pipeline using the MS generator, as it is better aligned in scale with the PHOENIX14**T** data.

We depict results for four translations. For all cases the input for our translation is a German sentence. The resulting gloss and sign video translations are given in Figure 11, 12, 13 and 14. The beam search over the MG provides the motion sequences that incorporate the translation from spoken language text. These pose sequences then condition the sign generation network. Transitions between sequences are added dynamically. We give representative frames for the generated sequences, indicating which glosses they belong to.

For sequence 1 in Figure 11 the NMT network correctly translates to a German gloss sequence which corresponds to the ground truth. The overall motion of the arms and hands is consistent with the video ground truth. The signers' appearances are clearly distinguishable from one another. Signer 1 stays closest to the ground truth, having the most developed arms and hands. Signer 2 struggles to fully form the right arm at times, this might be due to the fact that this signer was a left-handed signer in the original dataset and therefore less right handed motion was observed during training. Signer 3 has under-developed hands, something that is consistent across frames and sequences, indicating a failure in conditioning.

Sequence 2 (see Figure 12) also correctly translates the original input sentence. On top of the observations made for sequence 1, we notice a failure case for the gloss WECHSELHAFT (CHANGING). The sign for this gloss in DGS is a repeated left to right motion of both arms in front of the body (see the last three frames of the ground truth in Figure 12). In the generated sequence, the arms are in front of the body, but remain in the centre. We assume that this is due to a failure in the time alignment leading to key points of motion to the left and right resulting in hands positioned in the centre.

Results for sequence 3 in Figure 13 are in accordance with the first sequence. The sequence of glosses predicted matches the ground truth. Again signer 1 stays closest to the original motion sequence, even encapsulating the subtle difference in right hand position (not hand shape) between glosses GUT (GOOD) and LIEB (DEAR).

Sequence 4 is longer than previous examples, and contains one translation error (see Figure 14). The positions of arms and hands are consistent with the ground truth for the first four glosses, before encountering the error in gloss prediction. The motion for the last gloss WIND (WIND) is slightly under-articulated in contrast to the ground truth.

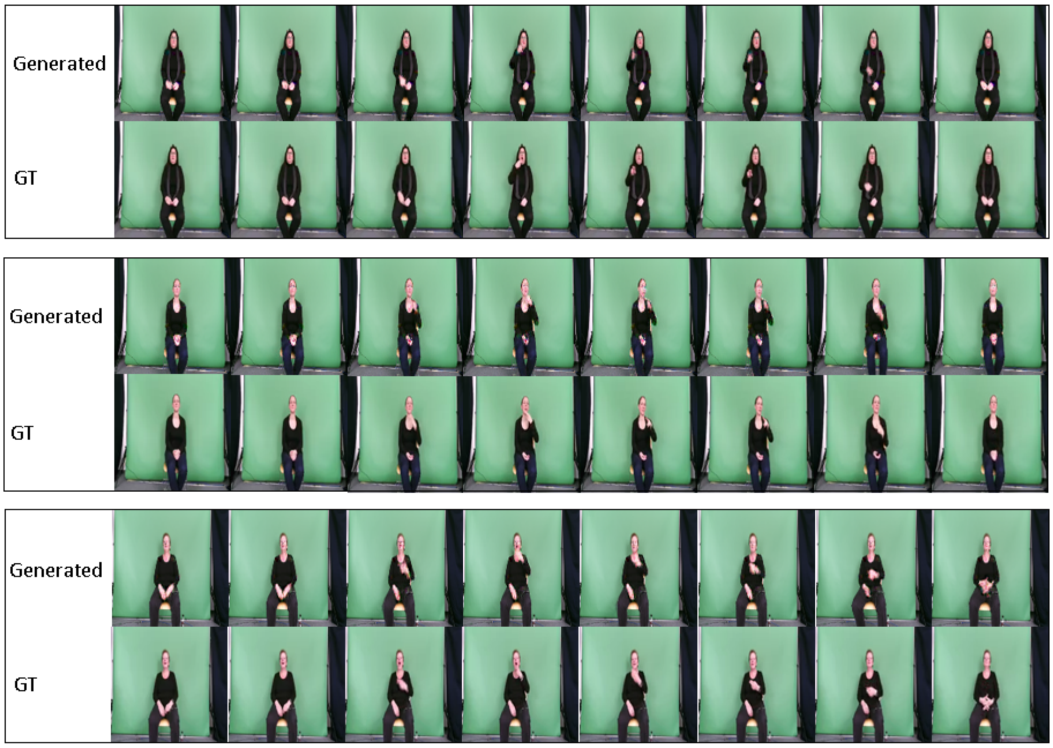


Fig. 9 Synthetic productions of the gloss ANTWORT (ANSWER), for signer 1 (top), signer 2 (middle), and signer 3 (bottom). Every 5th frame is shown. We can see that all three generated sequences are very close to their ground truth. It is also interesting to note that with our data-driven approach it is easy to account for natural variations, such as left-handed vs. right-handed signing.

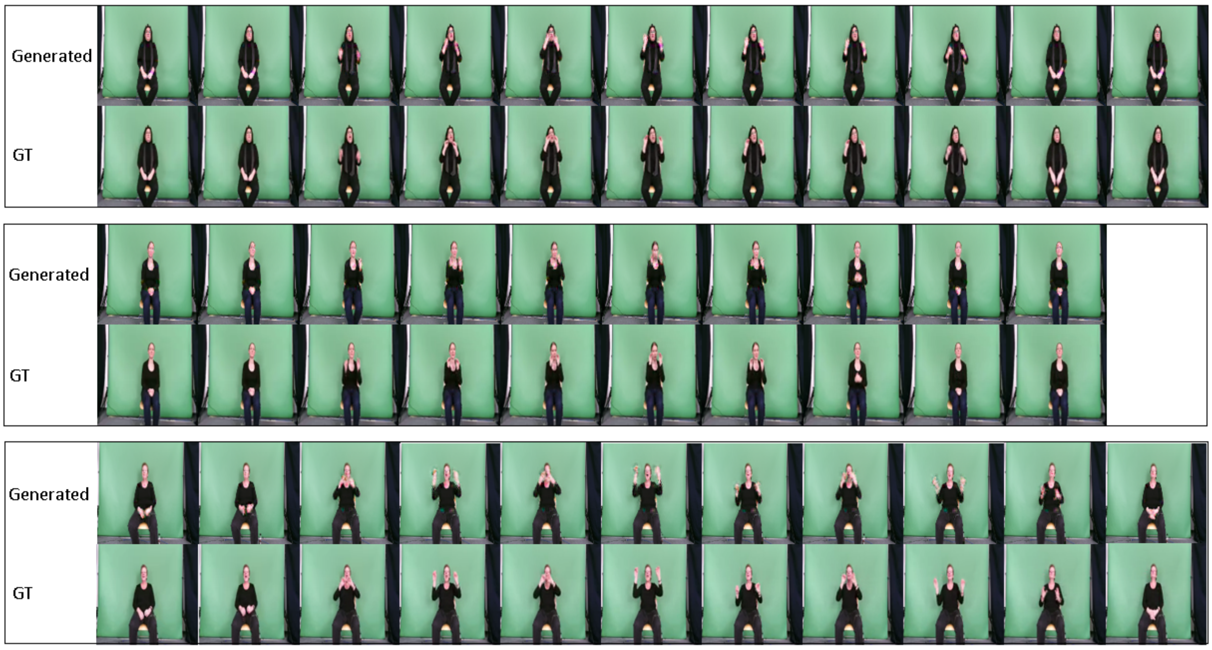


Fig. 10 Synthetic productions of the gloss ERKLAEREN (EXPLAIN), for signer 1 (top), signer 2 (middle), and signer 3 (bottom). Every 10th frame is shown. Generated sequences stay close to their ground truth throughout, however for signer 2 the network fails to form the right arm in frame 20. Additionally, this example shows two more forms of natural variation in sign language: speed and size of movement. These factors can have a significant impact, as they convey additional information such as emotion or intent.

Overall, the movement of signers is smooth and consistent with the glosses they represent, but not as expressive as the ground truth. We suspect that the limited motion stems from the averaging of all example sequences for a gloss to generate one mean sequence. To our knowledge the timing information for all glosses was automatically extracted from the PHOENIX14**T** data by the creators of the dataset using a Forced Alignment approach. It is therefore reasonable to assume that the provided timings contain errors, which negatively affect the generated mean sequence. Additionally, for most signs more than one variation exists, but this is not annotated in the dataset, neither is the use of left or right as the dominant hand. This further diminishes the motion of the mean sequences.

For future experiments an averaging and data cleaning process needs to be developed that pays consideration to variability in speed, expression, and left vs. right-handed signing. To improve the quality of extracted pose information, and add additional conditioning for hands and faces we need datasets of high image resolution. For translation we require sign language datasets that have topic-comment alignment. If both is combined, it would be possible to avoid the heavy cost of manually annotating details in sign motion such as facial expression and still get rich, natural translations.

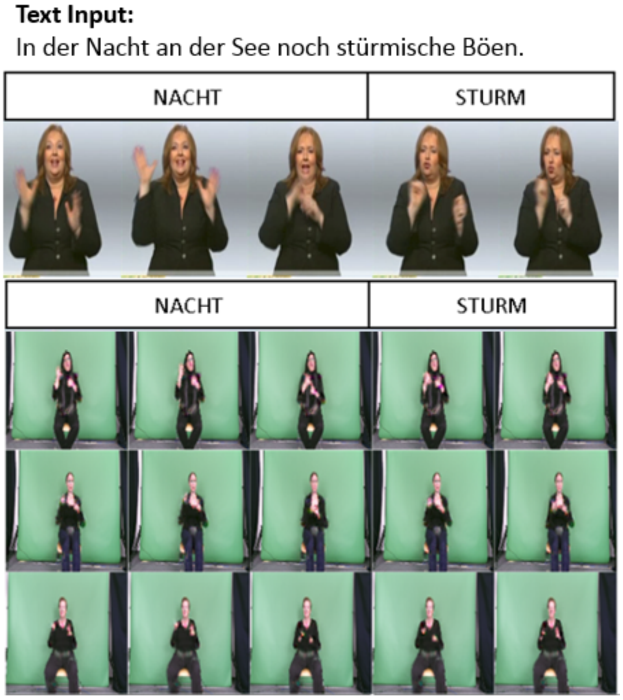


Fig. 11 Translation results for "In der Nacht an der See noch stuermische Boeen ." (In the night still storms near the sea.). The ground truth gloss and video is given in the top row. Below we see the gloss translation and synthetic video generated.

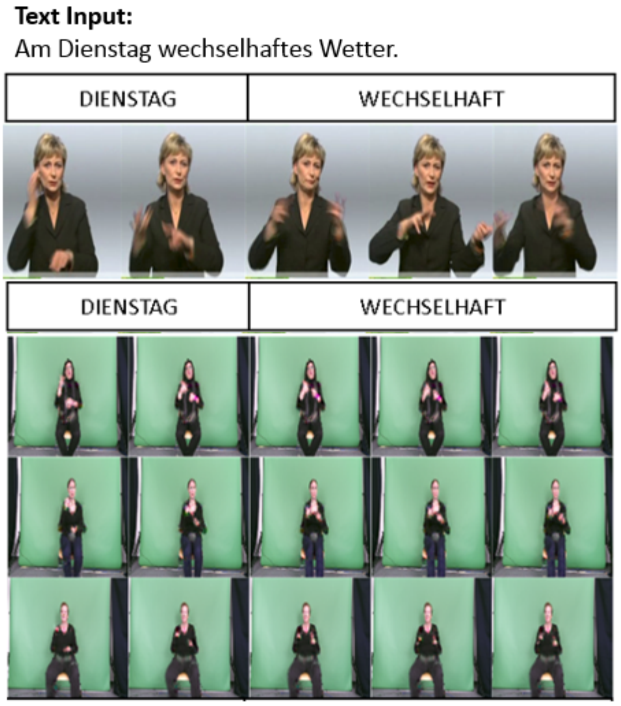


Fig. 12 Translation results for "Am Dienstag wechselhaftes Wetter ." (On Tuesday changing weather.). The ground truth gloss and video is given in the top row. Below we see the gloss translation and synthetic video generated.

4.5 High Definition Continuous Sign Generation

To improve the resolution and sharpness of our sign generation we generate HD continuous sign language video using the HD signing network. The network is conditioned with semantic label maps encoding human poses. We evaluate two configurations: A network conditioned only on 15 upper body joints (as was used in the MS network), and a network conditioned on the same 15 joints, plus 21 key points for each hand, and 68 key points for the face. For details see Figure 15. We trained for 16 epochs over 19,850 frames and corresponding label maps. For both models we report an average time of 0.42 seconds per image generated during inference using a GeForce GTX TITAN X.

Quantitative as well as qualitative results are provided. As with MS generation, we report the mean SSIM, PSNR, and MSE, this time over a test set of 500 frames, for just the pose input (HDSp) and pose, hands, and face input (HDSphf) in Table 4. The results indicate that more detailed conditioning with pose, hands, and face key points produces synthetic images that are closer to the ground truth. However, the difference in scores is not as significant as might be expected.

Looking at example frames we can see that both HDSp and HDSpfh create synthetic images that closely

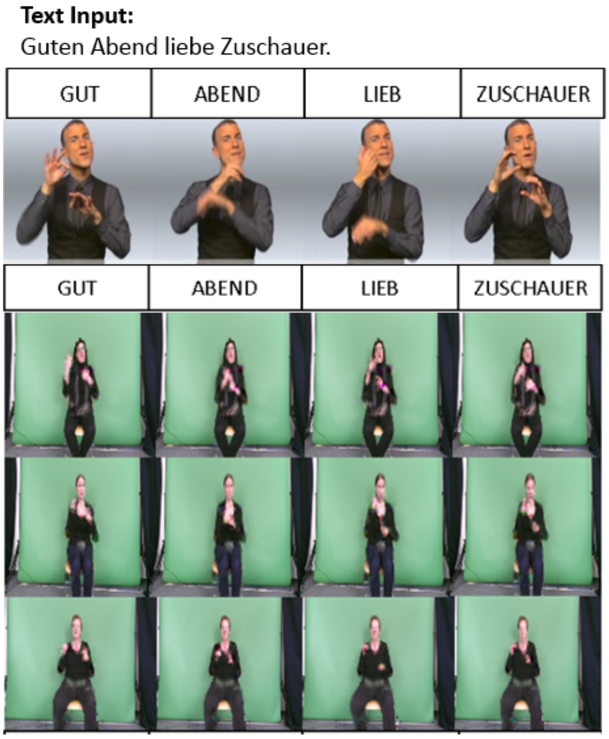


Fig. 13 Translation results for "Guten Abend liebe Zuschauer." (Good evening dear viewers). The ground truth gloss and video is given in the top row. Below we see the gloss translation and synthetic video generated.

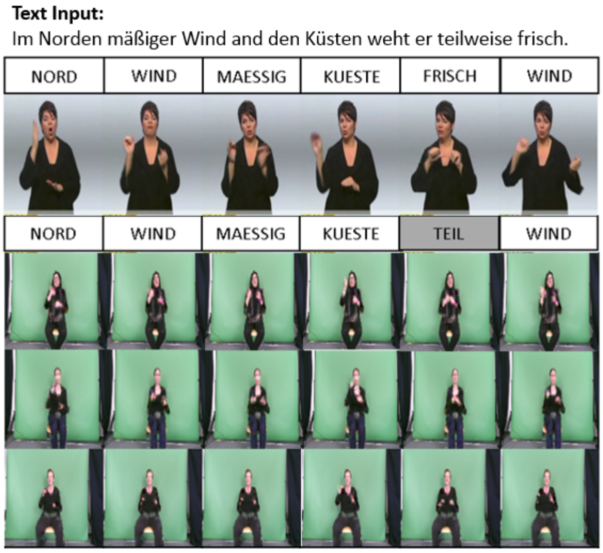


Fig. 14 Translation results for "Im Norden maeiger Wind an den Kuesten weht er teilweise frisch." (Mild winds in the north, at the coast it blows fresh in parts). The ground truth gloss and video is given in the top row. Below we see the gloss translation and synthetic video generated.

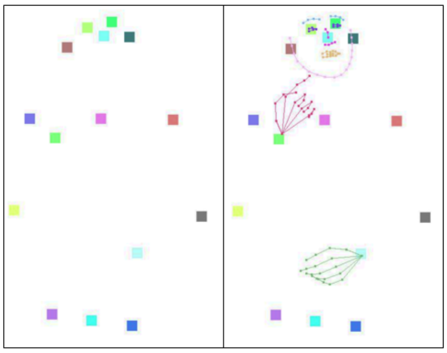


Fig. 15 Example semantic label maps for conditioning on pose only, containing 15 upper body pose key points (left), and conditioning on pose, hands and face, containing 15 upper body pose key points, plus 21 key points for each hand, and 68 facial key points (right). Each key point is assigned to a separate class using a different pixel value. The hand key points are grouped into two classes representing left and right hand, the facial key points are grouped as contour, left eye, right eye, nose, and mouth. Colour channels are inverted for visualisation purposes.

	SSIM	PSNR	MSE
HDSp	0.9332	23.1649	331.0621
HDSphf	**0.9338**	**23.5652**	**303.0417**

Table 4 Mean SSIM, PSNR, and MSE values over the test set, comparing synthetic images to their ground truth. For SSIM the range is -1 to +1, with +1 indicating identical images. The lower the MSE between two images, the more alike they are, whereas we want to maximise the PSNR between two images.

resemble their ground truth in both overall structure and detail such as clothing, overall facial expression and hand shape (see Figure 16). However, HDSpfh surpasses HDSp clearly for details of the generated hands and facial features. Whereas both networks learn to generate realistic hands and faces, HDSp can generate the wrong hand shape (see middle column in Figure 16 and 17), as it does not receive the positional information for all the finger joints, but merely an overall position of the hand.

Overall our results indicate that it is possible to generate highly realistic and detailed synthetic sign language videos, given sufficient positional information. A compromise can be found that keeps the annotation effort minimal (like using an automatic pose detector), whilst maintaining realism and expressiveness in the synthetic sign video.

5 Conclusions

In this paper, we presented the first 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 network with a Motion Graph (MG) to produce human pose sequences. This conditions a sign generation network capable of producing sign video frames.

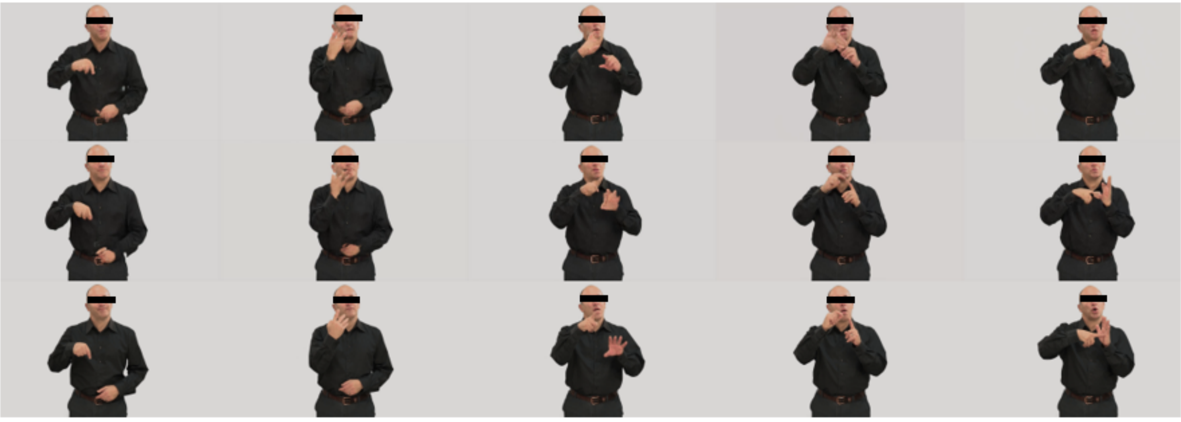


Fig. 16 Synthetic example frames for HDSp (top), and HDSpfh (middle), compared to ground truth frames (bottom).

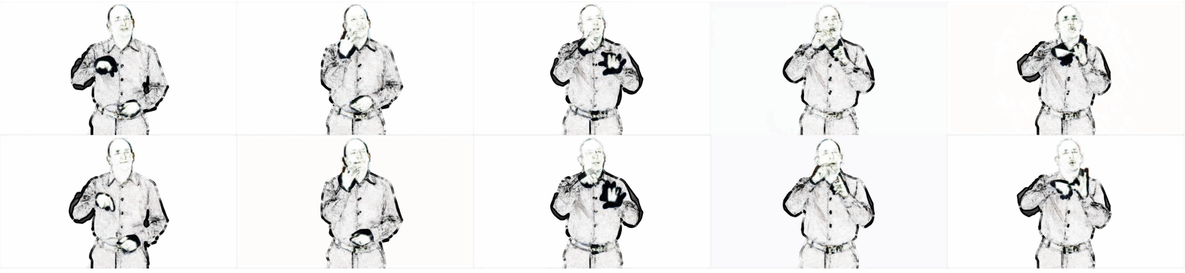


Fig. 17 Local SSIM values comparing HDSp to the ground truth (top) and HDSpfh to the ground truth (bottom). Both seem capable of generating faces close to the ground truth, but HDSpfh seems to outperform HDSp for generating correct hand shapes.

The NMT network's predictions can successfully be used to solve the MG, resulting in consistent text2pose translations. We show this by analysing example text2pose sequences, and by providing qualitative and quantitative results for an intermediate text2gloss representation. With our multi-signer (MS) generator we are able to produce multiple signers of different appearance. We show this for isolated signs, and as part of our text2sign translation approach.

Additionally we investigated the generation of HD continuous sign language video. Our results indicate that it is possible to produce photo-realistic video representations of sign language, by conditioning on key points extracted from training data. The accuracy and fidelity of key points seems to play a vital role, reinforcing the need for datasets of sufficient resolution.

Currently our text2sign translation system cannot compete with existing avatar approaches. Due to the low resolution of our translation training data, our results do not have the output resolution and expressiveness obtained by motion capture and avatar-based approaches. However, we have outlined that continuous, realistic sign language synthesis is possible, using minimal annotation. For training we only require text and gloss-level annotations, as skeletal pose information can be extracted from video automatically using an off-the-shelf solution such as OpenPose [7]. In contrast, avatar-based approaches require detailed annotations using task-specific transcription languages, which can only be carried out by expert linguists. Animating the avatar itself often involves a considerable amount of hand-engineering, and the results thus far remain robotic and under-articulated. Motion capture-based approaches require high-fidelity data, which needs to be captured, cleaned, and stored at considerable cost, limiting the amount of data available, hence making this approach unscalable. We believe that in time our approach will enable highly-realistic, and cost-effective translation of spoken languages to sign languages, improving equal access for the Deaf and Hard of Hearing.

For future work, our goal is to combine the MS and HD sign generation capabilities to synthesize highly detailed sign video, with signers of arbitrary appearance. The MS's ability to account for spatial and appearance changes, in combination with the high resolution of the HD generator would enable us to synthesize highly realistic and expressive sign language video. Additionally, we plan to improve our current MG by developing a data-processing strategy, that pays attention to the intricate features of sign language data, such as size of motion, and speed. This means replacing the current use of mean sequences with a more thoughtful approach that takes into account the likelihood of an example sequence being correct, the skeletal composition of different signers, and their dominant hand. We further plan to train our text2sign system end-to-end, and develop a performance metric to further quantitatively analyse the performance of our SLP system. Thus far we are only able to quantitatively analyse the translation capability and image quality separately. We hope to utilize and further developments in Sign Language Recognition (SLR) and synthetic image quality assessment in the future. As we want to be able to produce synthetic sign language with its natural fidelity, we continue collecting sign language training data of high resolution, and work closely with sign language experts both Hearing and Deaf in order to best serve the needs of the Deaf community.

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, the European Union's Horizon2020 research and innovation programme under grant agreement no. 762021 (Content4All) and the EPSRC project ExTOL (EP/R03298X/1). We would also like to thank Oscar Koller at Microsoft, and NVIDIA Corporation for their GPU grant.

References

1. Ahn, H., Ha, T., Choi, Y., Yoo, H., Oh, S.: Text2action: Generative adversarial synthesis from language to action
2. Arikan, O., Forsyth, D.A.: Interactive motion generation from examples. In: Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '02, pp. 483–490 (2002)
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
4. Bangham, J.A., Cox, S.J., Elliott, R., Glauert, J.R.W., Marshall, I., Rankov, S., Wells, M.: Virtual signing: capture, animation, storage and transmission-an overview of the visicast project. In: IEE Seminar on Speech and Language Processing for Disabled and Elderly People (Ref. No. 2000/025), pp. 6/1–6/7 (2000)
5. Bowden, R., Zisserman, A., Hogg, D., Magee, D.: Learning to recognise dynamic visual content from broadcast footage. URL https://cvssp.org/projects/dynavis/index.html
6. Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
7. Cao, Z., Simon, T., Wei, S., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 00, pp. 1302–1310 (2017)
8. Chan, C., Ginosar, S., Zhou, T., Efros, A.A.: Everybody dance now. CoRR **abs/1808.07371** (2018)
9. Chen, Q., Koltun, V.: Photographic image synthesis with cascaded refinement networks. In: ICCV, pp. 1520–1529. IEEE Computer Society (2017)
10. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724–1734. Association for Computational Linguistics (2014)
11. Chung, J., Gülçehre, Ç., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR **abs/1412.3555** (2014). URL http://arxiv.org/abs/1412.3555
12. Cox, S., Lincoln, M., Tryggvason, J., Nakisa, M., Wells, M., Tutt, M., Abbott, S.: Tessa, a system to aid communication with deaf people. In: Proceedings of the fifth international ACM conference on Assistive technologies, pp. 205–212. ACM (2002)
13. Ebling, S., Camgoz, N.C., Braem, P., Tissi, K., Sidler-Miserez, S., Stoll, S., Hadfield, S., Haug, T., Bowden, R., Tornay, S., Razavi, M., Magimai-Doss, M.: Smile swiss german sign language dataset (2018)
14. Ebling, S., Glauert, J.: Exploiting the full potential of jasigning to build an avatar signing train announcements (2013)
15. Ebling, S., Huenerfauth, M.: Bridging the gap between sign language machine translation and sign language animation using sequence classification. In: SLPAT@Interspeech (2015)
16. Efthimiou, E.: The Dicta-Sign Wiki: Enabling Web Communication for the Deaf (2012)
17. Elwazer, M.: Kintrans (2018). URL http://www.kintrans.com/
18. Forster, J., Schmidt, C., Koller, O., Bellgardt, M., Ney, H.: Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In: Language Resources and Evaluation, pp. 1911–1916. Reykjavik, Island (2014)
19. Gibet, S., Lefebvre-Albaret, F., Hamon, L., Brun, R., Turki, A.: Interactive editing in french sign language dedicated to virtual signers: Requirements and challenges. Univers. Access Inf. Soc. **15**(4), 525–539 (2016)
20. Glauert, J., Elliott, R., Cox, S., Tryggvason, J., Sheard, M.: Vanessa–a system for communication between deaf and hearing people. Technology and Disability **18**(4), 207–216 (2006)
21. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: 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, pp. 2672–2680 (2014)
22. Gregor, K., Danihelka, I., Graves, A., Rezende, D., Wierstra, D.: Draw: A recurrent neural network for image generation. In: F. Bach, D. Blei (eds.) Proceedings of the 32nd International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 37, pp. 1462–1471. PMLR, Lille, France (2015)
23. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**, 1735–80 (1997). DOI 10.1162/neco.1997.9.8.1735
24. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5967–5976 (2017)
25. JASigning: Virtual humans research for sign language animation (2017). URL http://vh.cmp.uea.ac.uk/index.php/Main$_{Page}$

26. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision (2016)
27. Kalchbrenner, N., Espeholt, L., Simonyan, K., van den Oord, A., Graves, A., Kavukcuoglu, K.: Neural machine translation in linear time. CoRR **abs/1610.10099** (2016). URL http://arxiv.org/abs/1610.10099
28. Kennaway, R.: Avatar-independent scripting for real-time gesture animation. CoRR **abs/1502.02961** (2013)
29. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. CoRR **abs/1312.6114** (2013). URL http://arxiv.org/abs/1312.6114
30. Kipp, M., Héloir, A., Nguyen, Q.: Sign language avatars: Animation and comprehensibility. In: IVA (2011)
31. Kovar, L., Gleicher, M., Pighin, F.: Motion graphs. In: Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '02, pp. 473–482 (2002)
32. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16, pp. 1558–1566 (2016)
33. Lee, J., Chai, J., Reitsma, P.S.A., Hodgins, J.K., Pollard, N.S.: Interactive control of avatars animated with human motion data. In: Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '02, pp. 491–500 (2002)
34. Lee, J., Shin, S.Y.: A hierarchical approach to interactive motion editing for human-like figures. In: SIGGRAPH (1999)
35. Luong, T., Pham, H., Manning, C.D.: Effective Approaches to Attention-based Neural Machine Translation. In: Conference on Empirical Methods in Natural Language Processing (EMNNLP) (2015)
36. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Van Gool, L.: Pose guided person image generation. In: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (eds.) Advances in Neural Information Processing Systems 30, pp. 406–416. Curran Associates, Inc. (2017)
37. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I.: Adversarial autoencoders. In: International Conference on Learning Representations
38. McDonald, J., Wolfe, R., Schnepp, J., Hochgesang, J., Jamrozik, D.G., Stumbo, M., Berke, L., Bialek, M., Thomas, F.: 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)
39. Min, J., Chai, J.: Motion graphs++: A compact generative model for semantic motion analysis and synthesis. ACM Trans. Graph. **31**(6), 153:1–153:12 (2012)
40. Mirza, M., Osindero, S.: Conditional generative adversarial nets. CoRR **abs/1411.1784** (2014). URL http://arxiv.org/abs/1411.1784
41. Mori, M., MacDorman, K., Kageki, N.: The uncanny valley [from the field] **19**, 98–100 (2012)
42. van den Oord, A., Kalchbrenner, N., Espeholt, L., kavukcuoglu, k., Vinyals, O., Graves, A.: Conditional image generation with pixelcnn decoders. In: D.D. Lee, M. Sugiyama, U.V. Luxburg, I. Guyon, R. Garnett (eds.) Advances in Neural Information Processing Systems 29, pp. 4790–4798. Curran Associates, Inc. (2016)
43. Oord, A.V., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. In: M.F. Balcan, K.Q. Weinberger (eds.) Proceedings of The 33rd International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 48, pp. 1747–1756. PMLR, New York, New York, USA (2016)
44. Perarnau, G., van de Weijer, J., Raducanu, B., Álvarez, J.M.: Invertible conditional gans for image editing. CoRR **abs/1611.06355** (2016). URL http://arxiv.org/abs/1611.06355
45. Prillwitz, S.: HamNoSys Version 2.0. Hamburg Notation System for Sign Languages: An Introductory Guide. Intern. Arb. z. Gebärdensprache u. Kommunik
46. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. CoRR **abs/1511.06434** (2015). URL http://arxiv.org/abs/1511.06434
47. Reed, S.E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H.: Learning what and where to draw. In: D.D. Lee, M. Sugiyama, U.V. Luxburg, I. Guyon, R. Garnett (eds.) Advances in Neural Information Processing Systems 29, pp. 217–225. Curran Associates, Inc. (2016)
48. Robotka, Z.: Signall (2018). URL http://www.signall.us/
49. Savitzky, A., Golay, M.J.E.: Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry **36**(8), 1627–1639 (1964)
50. Siarohin, A., Sangineto, E., Lathuilière, S., Sebe, N.: Deformable gans for pose-based human image generation. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
51. Stoll, S., Camgoz, N.C., Hadfield, S., Bowden, R.: Sign language production using neural machine translation and generative adversarial networks. In: British Machine Vision Conference (BMVC) (2018)
52. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to Sequence Learning with Neural Networks. In: Advances in neural information processing systems (2014)
53. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. CoRR **abs/1706.03762** (2017). URL http://arxiv.org/abs/1706.03762
54. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
55. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
56. Yan, X., Yang, J., Sohn, K., Lee, H.: Attribute2image: Conditional image generation from visual attributes. In: ECCV (4), *Lecture Notes in Computer Science*, vol. 9908, pp. 776–791. Springer (2016)
57. Zwitserlood, I., Verlinden, M., Ros, J., Schoot, S.V.D.: Synthetic signing for the deaf: esign (2004)