Skeletal Graph Self-Attention: Embedding a Skeleton Inductive Bias into Sign Language Production

Ben Saunders, Necati Cihan Camgoz, Richard Bowden
University of Surrey
{b.saunders, n.camgoz, r.bowden}@surrey.ac.uk

Abstract

Recent approaches to Sign Language Production (SLP) have adopted spoken language Neural Machine Translation (NMT) architectures, applied without sign-specific modifications. In addition, these works represent sign language as a sequence of skeleton pose vectors, projected to an abstract representation with no inherent skeletal structure.

In this paper, we represent sign language sequences as a skeletal graph structure, with joints as nodes and both spatial and temporal connections as edges. To operate on this graphical structure, we propose Skeletal Graph Self-Attention (SGSA), a novel graphical attention layer that embeds a skeleton inductive bias into the SLP model. Retaining the skeletal feature representation throughout, we directly apply a spatio-temporal adjacency matrix into the self-attention formulation. This provides structure and context to each skeletal joint that is not possible when using a non-graphical abstract representation, enabling fluid and expressive sign language production.

*We evaluate our Skeletal Graph Self-Attention architecture on the challenging RWTH-PHOENIX-Weather-2014**T** (PHOENIX14**T**) dataset, achieving state-of-the-art back translation performance with an 8% and 7% improvement over competing methods for the dev and test sets.*

1. Introduction

Sign languages are rich visual languages, the native languages of the Deaf communities. Comprised of both manual (hands) and non-manual (face and body) features, sign languages can be visualised as spatio-temporal motion of the hands and body [58]. When signing, the local context of motions is particularly important, such as the connections between fingers in a sign, or the lip patterns when mouthing [44]. Although commonly represented via a graphical avatar, more recent deep learning approaches to Sign Language Production (SLP) have represented sign as a continuous sequence of skeleton poses [51, 56, 71].

Due to the recent success of Neural Machine Translation (NMT), computational sign language research often naively applies spoken language architectures without sign-specific modifications. However, the domains of sign and spoken language are drastically different [55], with the continuous nature and inherent spatial structure of sign requiring sign-dependent architectures. Saunders *et al*. [49] introduced *Progressive Transformers*, an SLP architecture specific to a continuous skeletal representation. However, this still projects the skeletal input to an abstract feature representation, losing the skeletal inductive bias inherent to the body, where each joint upholds its own spatial representation. Even if spatio-temporal skeletal relationships can be maintained in an latent representation, a trained model may not correctly learn this complex structure.

Graphical structures can be used to represent pairwise relationships between objects in an ordered space. Graph Neural Networks (GNNs) are neural models used to capture graphical relationships, and predominantly operate on a high-level graphical structure [4], with each node containing an abstract feature representation and relationships occurring at the meta level. Conversely, skeleton pose sequences can be defined as spatio-temporal graphical representations, with both intra-frame spatial adjacency between limbs and inter-frame temporal adjacency between frames. In this work, we employ attention mechanisms as global graphical structures, with each node attending to all others. Even though there have been attempts to combine graphical representations and attention [15, 61, 70], there has been no work on graphical self-attention specific to a spatio-temporal skeletal structure.

In this paper, we represent sign language sequences as spatio-temporal skeletal graphs, the first SLP model to operate with a graphical structure. As seen in the centre of Figure 1, we encode skeletal joints as nodes, $\mathcal{J}$ (blue dots), and natural limb connections as edges, $\mathcal{E}$, with both spatial (blue lines) and temporal (green lines) relationships. Operating on a graphical structure explicitly upholds the skeletal representation throughout, learning deeper and more informative features than using an abstract representation.

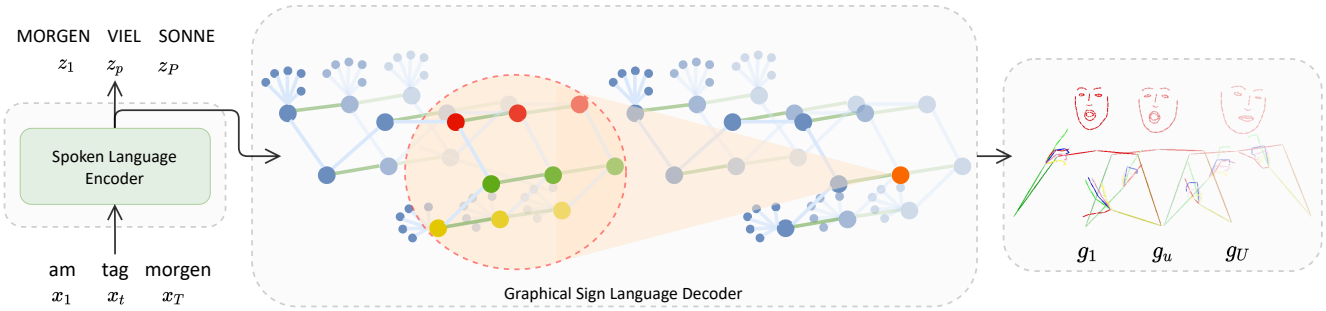

Figure 1. An overview of our proposed SLP network, showing an initial translation from a spoken language sentence using a text encoder, with gloss supervision. A subsequent skeletal graphical structure is formed, with multiple proposed Skeletal Graph Self-Attention layers applied to embed a skeleton inductive bias and produce expressive sign language sequences.

Additionally, we propose Skeletal Graph Self-Attention ($SGSA$), a novel spatio-temporal graphical attention layer that embeds a hierarchical body inductive bias into the self-attention mechanism. We directly mask the self-attention by applying a sparse adjacency matrix to the weights of the value computation, ensuring a spatial information propagation. To the best of our knowledge, ours is the first work to embed a graphical structure directly into the self-attention mechanism. In addition, we expand our model to the spatio-temporal domain by modelling the temporal adjacency only on $\mathcal{N}$ neighbouring frames.

Our full SLP model can be seen in Figure 1, initially translating from spoken language using a spoken language encoder with gloss supervision. The intermediary graphical structure is then processed by a graphical sign language decoder containing our proposed $SGSA$ layers, with a final output of sign language sequences. We evaluate on the challenging RWTH-PHOENIX-Weather-2014**T** (PHOENIX14**T**) dataset, performing spatial and temporal ablation studies of the proposed $SGSA$ architecture. Furthermore, we achieve state-of-the-art back translation results for the text to pose task, with an 8% and 7% performance increase over competing methods for the development and test sets respectively.

The contributions of this paper can be summarised as:

- The first SLP system to model sign language as a spatio-temporal graphical structure, applying both spatial and temporal adjacency.
- A novel Skeletal Graph Self-Attention ($SGSA$) layer, that embeds a skeleton inductive bias into the model.
- State-of-the-art Text-to-Pose SLP results on the PHOENIX14**T** dataset.

2. Related Work

Sign Language Production The past 30 years has seen extensive research into computational sign language [35, 62]. Early work focused on isolated Sign Language Recognition (SLR) [20,36], with a subsequent move to continuous SLR [6, 28]. The task of Sign Language Translation (SLT) was introduced by Camgoz *et al*. [7] and has since become a prominent research area [8, 46, 68]. Sign Language Production (SLP), the automatic translation from spoken language sentences to sign language sequences, was initially tackled using avatar-based technologies [17, 38]. The rule-based Statistical Machine Translation (SMT) achieved partial success [26,31], albeit with costly, labour-intensive pre-processing.

Recently, there have been many deep learning approaches to SLP proposed [23, 42, 48, 50, 52, 56, 63, 71], with Saunders *et al*. achieving state-of-the-art results with gloss supervision [52]. These works predominantly represent sign languages as sequences of skeletal frames, with each frame encoded as a vector of joint coordinates [51] that disregards any spatio-temporal structure available within a skeletal representation. In addition, these models apply standard spoken language architectures [60], disregarding the structural format of the skeletal data. Conversely, in this work we propose a novel spatio-temporal graphical attention layer that injects an inductive skeletal bias into SLP.

Graph Neural Networks A graph is a data structure consisting of nodes, $\mathcal{J}$, and edges, $\mathcal{E}$, where $\mathcal{E}$ defines the relationships between $\mathcal{J}$. Graph Neural Networks (GNNs) [4] apply neural layers on these graphical structures to learn representations [45, 74], classify nodes [64, 67] or generate new data [34, 66]. A skeleton pose representation can be structured as a graph, with joints as $\mathcal{J}$ and natural limb connections as $\mathcal{E}$ [53, 57]. GNNs have been proposed for operating on such dynamic skeletal graphs, in the context of action recognition [25, 40, 53, 64] and human pose estimation [57].

Attention networks can be formalised as a fully connected GNN, where the adjacency between each word,

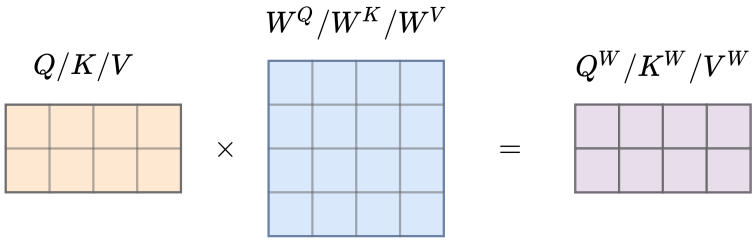

Figure 2. Weighted calculation of Queries, Q, Keys, K and Values, V, for global self-attention.

$\mathcal{E}$, is a weighting learnt using self-attention. Expanding this, Graph Attention Networks (GATs) [61] define explicit weighted adjacency between nodes, achieving state-of-the-art results across multiple domains [5, 30, 54]. Recently, there have been multiple graphical transformer architectures proposed [15, 29, 33, 70], which have been extended to the spatio-temporal domain for applications such as multiple object tracking [13] and pedestrian tracking [69].

However, there has been no work on graphical attention mechanisms where the features of each time step holds a relevant graphical structure. We build a spatio-temporal graphical architecture that operates on a skeletal representation per frame, explicitly injecting a skeletal inductive bias into the model. There have been some applications of GNNs in computational sign language in the context of SLR [14, 18, 24, 41, 59]. We extend these works to the SLP domain with our proposed Skeletal Graph Self-Attention architecture.

Local Attention Attention mechanisms have demonstrated strong Natural Language Processing (NLP) performance [2], particularly with the introduction of transformers [60]. Although proposed with global context [2], more recent works have selectively restricted attention to only a local context [12, 37, 65] or the top-k tokens [73], often due to computational issues or to enable long-range dependencies. In this paper, we propose using local attention to represent temporal adjacency within our graphical skeletal structure.

3. Background

In this section, we provide a brief background on self-attention. Attention mechanisms were initially proposed to overcome the information bottleneck found in encoder-decoder architectures [2, 39]. Transformers [60] apply multiple scaled self-attention layers in both encoder and decoder modules, where the input is a set of queries, $Q \in \mathbb{R}^{d_k}$, and keys, $K \in \mathbb{R}^{d_k}$, and values, $V \in \mathbb{R}^{d_v}$. Self-attention aims to learn a context value for each time-step as a weighted sum of all values, where the weight is determined by the relationship of the query with each corresponding key. An associated weight vector, $W^{Q/K/V}$, is first applied to each input, as shown in Figure 2, as:

$$Q^W = Q \cdot W^Q, \quad K^W = K \cdot W^K, \quad V^W = V \cdot W^V \tag{1}$$

where $W^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W^K \in \mathbb{R}^{d_{model} \times d_k}$ and $W^V \in \mathbb{R}^{d_{model} \times d_v}$ are weights related to each input variable and d_{model} is the dimensionality of the self-attention layer. Formally, scaled self-attention (SA) outputs a weighted vector combination of values, V^W, by the relevant queries, Q^W, keys, K^W, and dimensionality, d_k, as:

$$\text{SA}(Q, K, V) = \text{softmax}(\frac{Q^W (K^W)^T}{\sqrt{d_k}}) V^W \tag{2}$$

Multi-Headed Attention (MHA) applies h parallel attention mechanisms to the same input queries, keys and values, each with different learnt parameters. In the initial architecture [60], the dimensionality of each head is proportionally smaller than the full model, $d_h = d_{model}/h$. The output of each head is then concatenated and projected forward, as:

$$\begin{aligned} \text{MHA}(Q, K, V) &= [head_1, ..., head_h] \cdot W^O, \\ &\text{where } head_i = \text{SA}(Q^W, K^W, V^W) \end{aligned} \tag{3}$$

where $W^O \in \mathbb{R}^{d_{model} \times d_{model}}$. In this paper, we introduce Skeletal Graph Self-Attention layers that inject a skeletal inductive bias into the self-attention mechanism.

4. Methodology

The ultimate goal of SLP is to automatically translate from a source spoken language sentence, $\mathcal{X} = (x_1, ..., x_T)$ with $\mathcal{T}$ words, to a target sign language sequence, $\mathcal{G} = (g_1, ..., g_U)$ of $\mathcal{U}$ time steps. Additionally, an intermediary gloss[1] sequence representation can be used, $\mathcal{Z} = (z_1, ..., z_P)$ with P glosses. Current approaches [51, 56, 71] predominantly represent sign language as a sequence of skeletal frames, with each frame containing a vector of body joint coordinates. In addition, they project this skeletal structure to an abstract representation before being processed by the model [49]. However, this approach removes all spatial information contained within the skeletal data,

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

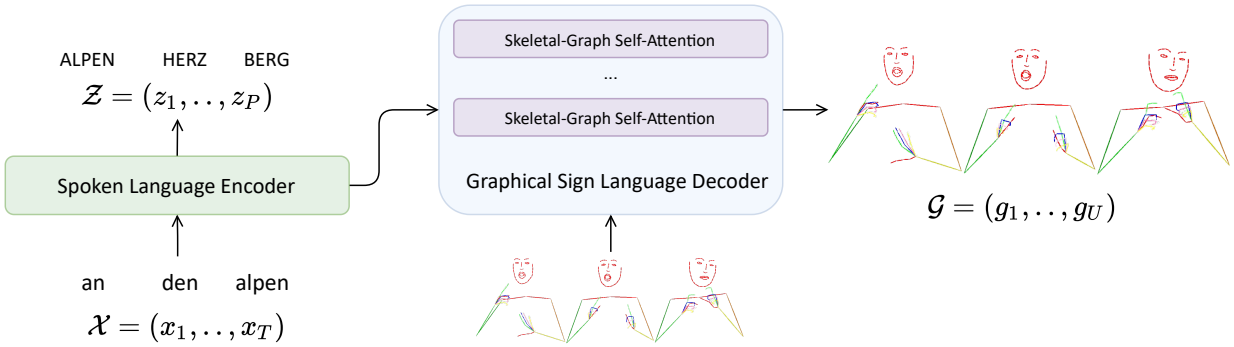

Figure 3. Overview of the proposed model architecture, detailing the Spoken Language Encoder (Sec. 4.1) and the Graphical Sign Language Decoder (Sec. 4.2). We propose novel Skeletal Graph Self-Attention layers to operate on the sign language skeletal graphs, $\mathcal{G}$.

restricting the model to only learning the internal relationships within a latent representation.

Contrary to previous work, in this paper we represent sign language sequences as spatio-temporal skeletal graphs, $\mathcal{G}$, as in the centre of Figure 1. As per graph theory [3], $\mathcal{G}$ can be formulated as a function of nodes, $\mathcal{J}$ and edges, $\mathcal{E}$. We define $\mathcal{J}$ as the skeleton pose sequence of temporal length $\mathcal{U}$ and spatial width $\mathcal{S}$, with each node representing a single skeletal joint coordinate from a single frame (blue dots in Fig. 1). $\mathcal{S}$ is therefore the dimensionality of the skeleton representation of each frame. $\mathcal{E}$ can be represented as a spatial adjacency matrix, $\mathcal{A}$, defined as the natural limb connections between skeleton joints both of its own frame (blue lines) and of neighbouring frames (green lines).

As outlined in Sec. 3, classical self-attention operates with global context over all sequence time-steps. However, a skeletal inductive bias can be embedded into a model by restricting attention to only the natural limb connections within the skeleton. To embed a skeleton inductive bias into self-attention, we propose a novel Skeletal Graph Self-Attention ($SGSA$) layer that operates with sparse attention. Modeled within a transformer decoder, $SGSA$ retains the original skeletal structure throughout multiple deep layers, ensuring the processing of spatio-temporal information contained in skeletal pose sequences. In-built adjacency matrices of both intra- and inter-frame relationships provide structure and context directly to each skeletal joint that is not possible when using a non-graphical abstract representation.

In the rest of this section, we outline the full SLP model, containing a spoken language encoder and a graphical sign language decoder, with an overview shown in Figure 3.

4.1. Spoken Language Encoder

As shown on the left of Figure 3, we first translate from a spoken language sentence, $\mathcal{X}$, of dimension $\mathcal{E} \times \mathcal{T}$, where $\mathcal{E}$ is the encoder embedding size, to a sign language representation, $\mathcal{R} = (r_1, ..., r_U)$ (Fig. 1 Left). We build a classical transformer encoder [60] that applies self-attention using the global context of a spoken language sequence. $\mathcal{R}$ is represented with a spatio-temporal structure, containing identical temporal length, $\mathcal{U}$, and spatial shape, $\mathcal{S}$, as the final skeletal graph, $\mathcal{G}$. This structure enables a graphical processing by the proposed sign language decoder. Additionally, as proposed in [52], we employ a gloss supervision to the intermediate sign language representation. This prompts the model to learn a meaningful latent sign representation for the ultimate goal of sign language production.

4.2. Graphical Sign Language Decoder

Given the intermediary sign language representation, $\mathcal{R} \in$, we build an auto-regressive transformer decoder containing our novel Skeletal Graph Self-Attention (SGSA) layers (Figure 3 middle). This produces a graphical sign language sequence, $\hat{\mathcal{G}}$, of spatial shape, $\mathcal{S}$, and temporal length, $\mathcal{U}$.

Spatial Adjacency We define a spatial adjacency matrix, $\mathcal{A} \in \mathbb{R}^{\mathcal{S} \times \mathcal{S}}$, expressed as a sparse attention map, as seen in Figure 4. $\mathcal{A}$ contains a spatial skeleton adjacency structure, modelled as the natural skeletal limb connections within a frame (blue lines in Fig. 1). $\mathcal{A}$ can be formalised as:

$$\mathcal{A}_{i,j} = \begin{cases} 1, & \textit{if } \mathrm{Con}(i,j) \\ 0, & \textit{otherwise} \end{cases} \tag{4}$$

where $\mathrm{Con}(i,j) = True$ if joints i and j are connected. For example, the skeletal elbow joint is connected to the skeletal wrist joint. We use an undirected graph representation, defining $\mathcal{E}$ as bidirectional edges.

Temporal Adjacency We expand the spatial adjacency matrix to the spatio-temporal domain by modelling the inter-frame edges of the skeletal graph structure (green lines

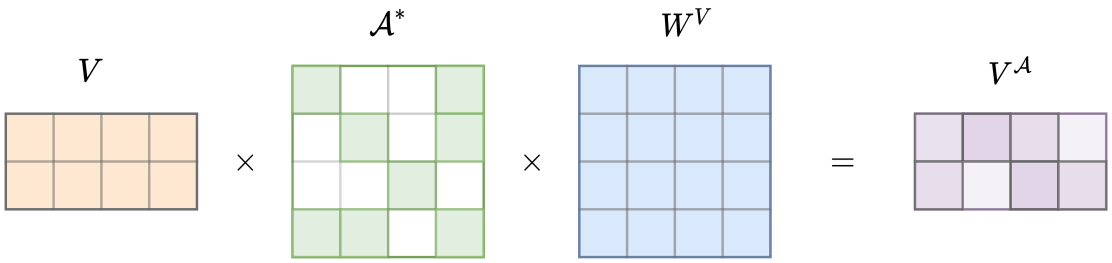

Figure 4. Skeletal Graph Self-Attention: Weighted calculation of Values, V, masked with a spatio-temporal adjacency matrix $\mathcal{A}^*$ to embed a skeleton inductive bias.

in Fig. 1). The updated spatial-temporal adjacency matrix can be formalised as $\mathcal{A} \in \mathbb{R}^{\mathcal{S}\times\mathcal{S}\times\mathcal{U}}$. We set $\mathcal{N}$ as the temporal distance that defines 'adjacent', where edges are established as both same joint connections and natural limb connections between the $\mathcal{N}$ adjacent frames. In the standard attention shown in Sec. 3, each time-step can globally attend to all others, which can be modelled as $\mathcal{N} = \infty$. We formalise our spatio-temporal adjacency matrix, as:

$$\mathcal{A}_{i,j,t} = \begin{cases} 1, & \textit{if } \mathrm{Con}(i,j) \textit{ and } t \leq \mathcal{N} \\ 0, & \textit{otherwise} \end{cases} \tag{5}$$

where t is the temporal distance from the reference frame, $t = u - u_{\text{ref}}$.

Self-loops and Normalisation To account for information propagation loops back to the same joint [3], we add self-loops to $\mathcal{A}$ using the identity matrix, $\mathcal{I} \in \mathbb{R}^{\mathcal{S}\times\mathcal{S}}$. In practice, due to our multi-dimensional skeletal representation, we add self-loops from each coordinate of the joint both to itself and all other coordinates of the same joint, which we define as $\mathcal{I}^* \in \mathbb{R}^{\mathcal{S}\times\mathcal{S}}$. Furthermore, to prevent numerical instabilities and exploding gradients [3], we normalise the adjacency matrix by inversely applying the degree matrix, $\mathcal{D} \in \mathbb{R}^{\mathcal{S}}$. $\mathcal{D}$ is defined as the numbers of edges a node is connected to. Normalisation can be formulated as:

$$\mathcal{A}^* = \mathcal{D}^{-1}(\mathcal{A} + \mathcal{I}^*) \tag{6}$$

where $\mathcal{A}^*$ is the normalised adjacency matrix.

Skeletal Graph Self-Attention We apply $\mathcal{A}^*$ as a sparsely weighted mask onto the weighted value calculation, $V^W = V \cdot W^V$, of Eq. 1, ensuring that the values used in the weighted context for each node is only impacted by the adjacent nodes of the previous layer:

$$V^{\mathcal{A}} = V \cdot \mathcal{A}^* \cdot W^V \tag{7}$$

where Figure 4 shows a visual representation of the sparse adjacent matrix $\mathcal{A}^*$ containing spatio-temporal connections, applied as a mask to the weighted calculation. With a value matrix containing a skeletal structure, $V \in \mathbb{R}^{\mathcal{S}}$, $\mathcal{A}^*$ restricts the information propagation of self-attention layers only through the spatial and temporal skeletal edges, $\mathcal{E}$, and thus embeds a skeleton inductive bias into the attention mechanism.

We formally define a Skeletal Graph Self-Attention ($SGSA$) layer by plugging both the weighted variable computation of Eq. 1 and the adjacent weighted computation of Eq. 7 into the self-attention Eq. 2, as:

$$SGSA(Q,K,V,A) = \\ \text{softmax}(\frac{Q \cdot W^Q (K \cdot W^K)^T}{\sqrt{d_k}}) V \cdot \mathcal{A}^* \cdot W^V \tag{8}$$

where $d_{model} = \mathcal{S}$. This explicitly retains the spatial skeletal shape, $\mathcal{S}$, throughout the sign language decoder, enabling a spatial structure to be extracted.

To extend this to a multi-headed transformer decoder, we replace self-attention in Eq. 3 with our proposed $SGSA$ layers. To retain the spatial skeletal representation within each head, the dimensionality of each head is kept as the full model dimension, $d_h = d_{model} = \mathcal{S}$, with the final projection layer enlarged to $h \times \mathcal{S}$.

We build our auto-regressive sign language decoder with $\mathcal{L}$ multi-headed $SGSA$ sub-layers, interleaved with fully-connected layers and a final feed-forward layer, each with a consistent spatial dimension of $\mathcal{S}$. A residual connection [22] and subsequent layer norm [1] is employed around each of the sub-layers, to aid training. As shown on the right of Figure 3, the final output of our sign language decoder module is a graphical skeletal sequence, $\hat{\mathcal{G}}$, that contains $\mathcal{U}$ frames of skeleton pose, each with a spatial shape of $\mathcal{S}$.

We train our sign language decoder using the Mean Squared Error (MSE) loss between the predicted sequence, $\hat{\mathcal{G}}$, and the ground truth sequence, $\mathcal{G}^*$. This is formalised as $\mathcal{L}_{\mathcal{MSE}} = \frac{1}{U}\sum_{i=1}^{u} \hat{g}_{1:U} - {g^*}_{1:U})^2$, where $\hat{g}$ and g^* represent the frames of the produced and ground truth sign language sequences, respectively. We train our full SLP model end-to-end with a weighted combination of the encoder gloss supervision [52] and decoder skeleton pose losses.

Skeletal Graph Layers, $\mathcal{L}$:	DEV SET BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE	TEST SET BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE
0 (4 SA)	14.25	17.73	23.47	34.79	37.65	13.64	17.03	23.09	35.03	36.59
1	14.37	17.67	23.13	33.95	36.98	13.63	17.08	23.17	35.39	37.05
2	14.50	18.14	24.10	35.96	38.09	13.85	17.23	23.14	34.93	37.33
3	14.53	18.02	24.00	35.71	37.62	13.72	17.23	23.10	34.45	36.99
4	14.68	18.30	**24.31**	**36.16**	38.51	14.05	17.59	23.73	35.63	37.47
5	**14.72**	**18.39**	24.29	35.79	**38.72**	**14.27**	**17.79**	**23.79**	**35.72**	**37.79**

Table 1. Impact of Skeletal Graph Self-Attention layers, $\mathcal{L}$, on model performance.

4.3. Sign Language Output

Generating a sign language video from the produced graphical skeletal sequence, $\hat{\mathcal{G}}$, is then a trivial task, animating each frame in temporal order. Frame animation is done by connecting the nodes, $\mathcal{J}$, using the natural limb connections defined by $\mathcal{E}$, as seen in Fig. 1.

5. Experiments

Dataset We evaluate our approach on the PHOENIX14**T** dataset introduced by Camgoz et al. [7], containing parallel sequences of 8257 German sentences, sign gloss translations and sign language videos. Other available sign datasets are either simple sentence repetition tasks of non-natural signing not appropriate for translation [16, 72], or contain larger domains of discourse that currently prove difficult for the SLP field [10, 21]. We extract 3D skeletal joint positions from the sign language videos to represent our spatio-temporal graphical skeletal structure. Manual and non-manual features of each video are first extracted in 2D using OpenPose [11], with the manuals lifted to 3D using the skeletal model estimation model proposed in [71]. We normalise the skeleton pose and set the spatial skeleton shape, $\mathcal{S}$, as 291, with 290 joint coordinates and 1 counter decoding value (as in [49]). Adjacency information, $\mathcal{A}$, is defined as the natural limb connections of 3D body, hand and face joints, as in [71], where each coordinate of a joint is adjacent to both the coordinates of its own joint and all connected joints. We define the counter value as global adjacency, with connections to all joints.

Implementation Details We setup our SLP model with a spoken language encoder of 2 layers, 4 heads and an embedding size, $\mathcal{E}$, of 256, and a graphical sign language decoder of 5 layers, 4 heads and an embedding size of $\mathcal{S}$. Our best performing model contains 9M trainable parameters. As proposed by Saunders *et al.* [49], we apply Gaussian noise augmentation with a noise rate of 5. We train all parts of our network with Xavier initialisation [19], Adam optimization [27] with default parameters and a learning rate of 10^{-3}. Our code is based on Kreutzer et al.'s NMT toolkit, JoeyNMT [32], and implemented using PyTorch [43].

Evaluation We use the back translation metric [49] for evaluation, which employs a pre-trained SLT model [9] to translate the produced sign pose sequences back to spoken language. We compute BLEU and ROUGE scores against the original input, with BLEU n-grams from 1 to 4 provided. The SLP evaluation protocols on the PHOENIX14**T** dataset have been set by [49]. We share results on the *Text to Pose (T2P)* task which constitutes the production of sign language sequences directly from spoken language sentences, the ultimate goal of an SLP system. We omit Gloss to Pose evaluation to focus on the more important spoken language translation task.

Skeletal Graph Self-Attention Layers We start our experiments on the proposed Skeletal Graph Self-Attention layers, evaluating the effect of stacking multiple $SGSA$ layers, $\mathcal{L}$, each with a multi-head size, h, of 4. We first ablate the effect of using no $SGSA$ layers, and replacing them with 4 standard self-attention layers, as described in Section 3. We then build our graphical sign language decoder with 1 to 5 $SGSA$ layers, with each model retaining a constant spoken language encoder size and a global temporal adjacency.

Table 1 shows that using standard self-attention layers achieves the worst performance of 14.25 BLEU-4, showing the benefit of our proposed $SGSA$ layers. Increasing the number of $SGSA$ layers, as expected, increases model performance to a peak of 14.72 BLEU-4. A larger number of layers enables a deeper representation of the skeletal graph and thus provides a stronger skeleton inductive bias to the model. In lieu of this, for the rest of our experiments we build our sign language decoder with five $SGSA$ layers.

Temporal Adjacency In our next set of experiments, we examine the impact of the temporal adjacency distance, $\mathcal{N}$, defined in Sec. 4.2. In order to logically set $\mathcal{N}$, we analyse the trained temporal attention matrix of the best performing decoder evaluated above. We notice that the attention predominantly falls on the last 3 frames, as the model learns to attend to the local temporal context of skeletal motion. Manually restricting the temporal attention provides this information as an inductive bias into the model, rather than

Temporal Adjacency, $\mathcal{N}$:	DEV SET BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE	TEST SET BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE
∞	14.72	18.39	24.29	35.79	38.72	14.27	17.79	23.79	35.72	37.79
1	**15.15**	18.67	24.47	35.88	38.44	**14.33**	17.77	23.72	35.26	37.96
2	15.09	18.51	24.43	36.17	38.04	14.07	17.62	23.91	**36.28**	37.82
3	15.08	**18.84**	24.89	36.66	38.95	14.32	**17.95**	**24.04**	36.10	**38.38**
5	14.90	18.81	**25.30**	**37.31**	**39.55**	14.21	17.79	23.98	35.88	38.44

Table 2. Impact of Temporal Adjacency, $\mathcal{N}$, on $SGSA$ model performance

relying on this being learnt.

Table 2 shows results of our temporal adjacency evaluation, ranging from an infinite adjacency (no constraint) to $\mathcal{N} \in [1, 5]$. It can be seen that a temporal adjacency distance of one achieves the best BLEU-4 performance. Note: Although we report BLEU of n-grams 1-4 for completeness, we use BLEU-4 as our final evaluation metric to enable a clear result. Although counter-intuitive to the global self-attention utilised by a transformer decoder, we believe this is modelling the Markov property, where future frames only depend on the current state. Due to the intermediary gloss supervision [52], the defined sign language representation, $\mathcal{R}$, should contain all frame-level information relevant to a sign language translation. The sign language decoder then has the sole task of accurately animating each skeletal frame. Therefore, a single temporal adjacency in the graphical decoder makes sense, as no new information is required to be learnt from temporally distant frames.

Baseline Comparisons We compare the performance of the proposed Skeletal Graph Self-Attention architecture against 4 baseline SLP models: 1) Progressive transformers [49], which applied the classical transformer architecture to sign language production. 2) Adversarial training [47], which utilised an adversarial discriminator to prompt more expressive productions, 3) Mixture Density Networks (MDNs) [51], which modelled the variation found in sign language using multiple distributions to parameterise the entire prediction subspace, and 4) Mixture of Motion Primitives (MOMP) [52], which split the SLP task into two distinct jointly-trained sub-tasks and learnt a set of motion primitives for animation.

Table 3 presents *Text to Pose* results, showing that $SGSA$ achieves 15.15/14.33 BLEU-4 for the development and test sets respectively, an 8/7% improvement over the state-of-the-art. These results highlight the significant success of our proposed $SGSA$ layers. We have shown that representing sign pose skeletons in a graphical skeletal structure and embedding a skeletal inductive bias into the self-attention mechanism enables a fluid and expressive sign language production.

6. Conclusion

In this paper, we proposed a skeletal graph structure for SLP, with joints as nodes and both spatial and temporal connections as edges. We proposed a novel graphical attention layer, Skeletal Graph Self-Attention, to operate on the graphical skeletal structure. Retaining the skeletal feature representation throughout, we directly applied a spatio-temporal adjacency matrix into the self-attention formulation, embedding a skeleton inductive bias for expressive sign language production. We evaluated $SGSA$ on the challenging PHOENIX14**T** dataset, achieving state-of-the-art back translation performance with an 8% and 7% improvement over competing methods for the dev and test set. For future work, we aim to apply $SGSA$ layers to the wider computational sign language tasks of SLR and SLT.

7. Acknowledgements

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

Approach:	DEV SET BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE	TEST SET BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE
Progressive Transformers [49]	11.82	14.80	19.97	31.41	33.18	10.51	13.54	19.04	31.36	32.46
Adversarial Training [47]	12.65	15.61	20.58	31.84	33.68	10.81	13.72	18.99	30.93	32.74
Mixture Density Networks [51]	11.54	14.48	19.63	30.94	33.40	11.68	14.55	19.70	31.56	33.19
Mixture of Motion Primitives [52]	14.03	17.50	23.49	35.23	37.76	13.30	16.86	23.27	**35.89**	36.77
Skeletal Graph Self-Attention	**15.15**	**18.67**	**24.47**	**35.88**	**38.44**	**14.33**	**17.77**	**23.72**	35.26	**37.96**

Table 3. Baseline comparisons on the PHOENIX14**T** dataset for the *Text to Pose* task.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer Normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 3

[3] Béla Bollobás. *Modern graph theory*. Springer Science & Business Media, 2013. 4, 5

[4] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral Networks and Locally Connected Networks on Graphs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 1, 2

[5] Dan Busbridge, Dane Sherburn, Pietro Cavallo, and Nils Y Hammerla. Relational Graph Attention Networks. *arXiv preprint arXiv:1904.05811*, 2019. 3

[6] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. SubUNets: End-to-end Hand Shape and Continuous Sign Language Recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[7] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural Sign Language Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 6

[8] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Multi-channel Transformers for Multi-articulatory Sign Language Translation. In *Assistive Computer Vision and Robotics Workshop (ACVR)*, 2020. 2

[9] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[10] Necati Cihan Camgoz, Ben Saunders, Guillaume Rochette, Marco Giovanelli, Giacomo Inches, Robin Nachtrab-Ribback, and Richard Bowden. Content4All Open Research Sign Language Translation Datasets. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2021. 6

[11] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6

[12] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating Long Sequences with Sparse Transformers. *arXiv preprint arXiv:1904.10509*, 2019. 3

[13] Peng Chu, Jiang Wang, Quanzeng You, Haibin Ling, and Zicheng Liu. TransMOT: Spatial-Temporal Graph Transformer for Multiple Object Tracking. *arXiv preprint arXiv:2104.00194*, 2021. 3

[14] Cleison Correia de Amorim, David Macêdo, and Cleber Zanchettin. Spatial-Temporal Graph Convolutional Networks for Sign Language Recognition. In *International Conference on Artificial Neural Networks*, 2019. 3

[15] Vijay Prakash Dwivedi and Xavier Bresson. A Generalization of Transformer Networks to Graphs. *arXiv preprint arXiv:2012.09699*, 2020. 1, 3

[16] Eleni Efthimiou and Stavroula-Evita Fotinea. GSLC: Creation and Annotation of a Greek Sign Language Corpus for HCI. In *International Conference on Universal Access in Human-Computer Interaction*, 2007. 6

[17] Ralph Elliott, John RW Glauert, JR Kennaway, Ian Marshall, and Eva Safar. Linguistic Modelling and Language-Processing Technologies for Avatar-based Sign Language Presentation. *Universal Access in the Information Society*, 2008. 2

[18] Mariusz Flasiński and Szymon Myśliński. On The Use of Graph Parsing for Recognition of Isolated Hand Postures of Polish Sign Language. *Pattern Recognition*, 2010. 3

[19] Xavier Glorot and Yoshua Bengio. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010. 6

[20] Kirsti Grobel and Marcell Assan. Isolated Sign Language Recognition using Hidden Markov Models. In *IEEE International Conference on Systems, Man, and Cybernetics*, 1997. 2

[21] Thomas Hanke, Lutz König, Sven Wagner, and Silke Matthes. DGS Corpus & Dicta-Sign: The Hamburg Studio Setup. In *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010), Valletta, Malta*, 2010. 6

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[23] Wencan Huang, Wenwen Pan, Zhou Zhao, and Qi Tian. Towards Fast and High-Quality Sign Language Production. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 2

[24] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. Skeleton Aware Multi-Modal Sign Language Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021. 3

[25] Jiun-Yu Kao, Antonio Ortega, Dong Tian, Hassan Mansour, and Anthony Vetro. Graph Based Skeleton Modeling for Human Activity Analysis. In *2019 IEEE International Conference on Image Processing (ICIP)*, 2019. 2

[26] Dilek Kayahan and Tunga Güngör. A Hybrid Translation System from Turkish Spoken Language to Turkish Sign Language. In *IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA)*, 2019. 2

[27] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 6

[28] Oscar Koller, Jens Forster, and Hermann Ney. Continuous Sign Language Recognition: Towards Large Vocabulary Statistical Recognition Systems Handling Multiple Signers. *Computer Vision and Image Understanding (CVIU)*, 2015. 2

[29] Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. Text Generation from

Knowledge Graphs with Graph Transformers. *arXiv preprint arXiv:1904.02342*, 2019. 3

[30] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, S Hamid Rezatofighi, and Silvio Savarese. Social-BiGAT: Multimodal Trajectory Forecasting using Bicycle-GAN and Graph Attention Networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2019. 3

[31] Dimitris Kouremenos, Klimis S Ntalianis, Giorgos Siolas, and Andreas Stafylopatis. Statistical Machine Translation for Greek to Greek Sign Language Using Parallel Corpora Produced via Rule-Based Machine Translation. In *IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 2018. 2

[32] Julia Kreutzer, Joost Bastings, and Stefan Riezler. Joey NMT: A Minimalist NMT Toolkit for Novices. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. 6

[33] Devin Kreuzer, Dominique Beaini, William L Hamilton, Vincent Létourneau, and Prudencio Tossou. Rethinking Graph Transformers with Spectral Attention. *arXiv preprint arXiv:2106.03893*, 2021. 3

[34] Yujia Li, Oriol Vinyals, Chris Dyer, Razvan Pascanu, and Peter Battaglia. Learning Deep Generative Models of Graphs. *arXiv preprint arXiv:1803.03324*, 2018. 2

[35] Rung-Huei Liang and Ming Ouhyoung. A Sign Language Recognition System using Hidden Markov Model and Context Sensitive Search. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, 1996. 2

[36] Kian Ming Lim, Alan Wee Chiat Tan, Chin Poo Lee, and Shing Chiang Tan. Isolated Sign Language Recognition using Convolutional Neural Network Hand Modelling and Hand Energy Image. *Multimedia Tools and Applications*, 2019. 2

[37] Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating Wikipedia by Summarizing Long Sequences. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[38] Pengfei Lu and Matt Huenerfauth. Collecting a Motion-Capture Corpus of American Sign Language for Data-Driven Generation research. In *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies*, 2010. 2

[39] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015. 3

[40] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning Trajectory Dependencies for Human Motion Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[41] Lu Meng and Ronghui Li. An Attention-Enhanced Multi-Scale and Dual Sign Language Recognition Network Based on a Graph Convolution Network. *Sensors*, 2021. 3

[42] Taro Miyazaki, Yusuke Morita, and Masanori Sano. Machine Translation from Spoken Language to Sign Language using Pre-trained Language Model as Encoder. In *Proceedings of the LREC2020 Workshop on the Representation and Processing of Sign Languages*, 2020. 2

[43] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic Differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017. 6

[44] Roland Pfau, Josep Quer, et al. *Nonmanuals: Their Grammatical and Prosodic Roles*. Cambridge University Press, 2010. 1

[45] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning Human-Object Interactions by Graph Parsing Neural Networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[46] Jefferson Rodriguez and Fabio Martínez. How Important is Motion in Sign Language Translation? *IET Computer Vision*, 2021. 2

[47] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Adversarial Training for Multi-Channel Sign Language Production. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2020. 7

[48] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Everybody Sign Now: Translating Spoken Language to Photo Realistic Sign Language Video. *arXiv preprint arXiv:2011.09846*, 2020. 2

[49] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Progressive Transformers for End-to-End Sign Language Production. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 3, 6, 7

[50] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. AnonySign: Novel Human Appearance Synthesis for Sign Language Video Anonymisation. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG) (To Appear)*, 2021. 2

[51] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Continuous 3D Multi-Channel Sign Language Production via Progressive Transformers and Mixture Density Networks. In *International Journal of Computer Vision (IJCV)*, 2021. 1, 2, 3, 7

[52] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Mixed SIGNals: Sign Language Production via a Mixture of Motion Primitives. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2, 4, 5, 7

[53] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based Action Recognition with Directed Graph Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[54] Weiping Song, Zhiping Xiao, Yifan Wang, Laurent Charlin, Ming Zhang, and Jian Tang. Session-Based Social Recommendation via Dynamic Graph Attention Networks. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, 2019. 3

[55] William C Stokoe. Sign Language Structure. *Annual Review of Anthropology*, 1980. 1

[56] Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. Sign Language Production using Neural

Machine Translation and Generative Adversarial Networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 1, 2, 3

[57] Matthias Straka, Stefan Hauswiesner, Matthias Rüther, and Horst Bischof. Skeletal Graph Based Human Pose Estimation in Real-Time. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2011. 2

[58] Rachel Sutton-Spence and Bencie Woll. *The Linguistics of British Sign Language: An Introduction*. Cambridge University Press, 1999. 1

[59] MF Tolba, Ahmed Samir, and Magdy Abul-Ela. A Proposed Graph Matching Technique for Arabic Sign Language Continuous Sentences Recognition. In *2012 8th International Conference on Informatics and Systems (INFOS)*, 2012. 3

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 2, 3, 4

[61] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph Attention Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. 1, 3

[62] Beth J Wilson and Gretel Anspach. Neural Networks for Sign Language Translation. In *Applications of Artificial Neural Networks IV*. International Society for Optics and Photonics, 1993. 2

[63] Qinkun Xiao, Minying Qin, and Yuting Yin. Skeleton-based Chinese Sign Language Recognition and Generation for Bidirectional Communication between Deaf and Hearing People. In *Neural Networks*, 2020. 2

[64] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2

[65] Baosong Yang, Zhaopeng Tu, Derek F Wong, Fandong Meng, Lidia S Chao, and Tong Zhang. Modeling Localness for Self-Attention Networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. 3

[66] Carl Yang, Peiye Zhuang, Wenhan Shi, Alan Luu, and Pan Li. Conditional Structure Generation through Graph Variational Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2019. 2

[67] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph Convolutional Networks for Text Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 2

[68] Kayo Yin. Sign Language Translation with Transformers. *arXiv preprint arXiv:2004.00588*, 2020. 2

[69] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. Spatio-Temporal Graph Transformer Networks for Pedestrian Trajectory Prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3

[70] Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. Graph Transformer Networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2019. 1, 3

[71] Jan Zelinka and Jakub Kanis. Neural Sign Language Synthesis: Words Are Our Glosses. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020. 1, 2, 3, 6

[72] Jihai Zhang, Wengang Zhou, Chao Xie, Junfu Pu, and Houqiang Li. Chinese Sign Language Recognition with Adaptive HMM. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, 2016. 6

[73] Guangxiang Zhao, Junyang Lin, Zhiyuan Zhang, Xuancheng Ren, Qi Su, and Xu Sun. Explicit Sparse Transformer: Concentrated Attention Through Explicit Selection. *arXiv preprint arXiv:1912.11637*, 2019. 3

[74] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph Neural Networks: A Review of Methods and Applications. *AI Open*, 2020. 2