Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology (SLTAT 7), pages 95–102
Language Resources and Evaluation Conference (LREC 2022), Marseille, 20-25 June 2022

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.

Keywords: Sign Language Production (SLP), Graph Neural Network (GNN), Computational Sign Language

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 (Sutton-Spence and Woll, 1999). 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 (Pfau et al., 2010). 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 (Saunders et al., 2021a; Stoll et al., 2018; Zelinka and Kanis, 2020).

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 (Stokoe, 1980), with the continuous nature and inherent spatial structure of sign requiring sign-dependent architectures. Saunders *et al* (Saunders et al., 2020c) 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. GNNs are neural models used to capture graphical relationships, and predominantly operate on a high-level graphical structure (Bruna et al., 2014), 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 (Yun et al., 2019; Dwivedi and Bresson, 2020; Veličković et al., 2017), 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.

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

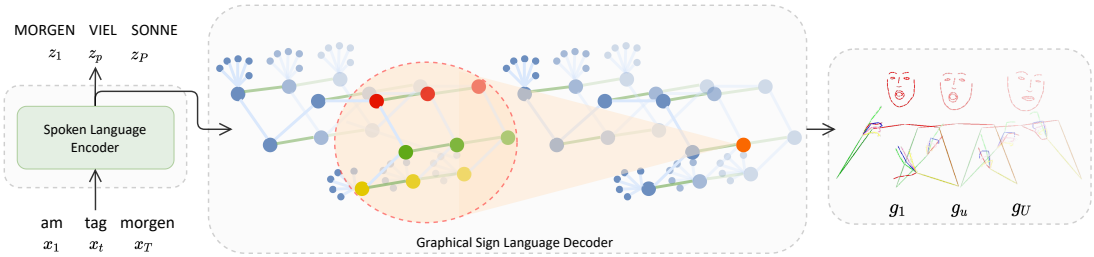

Figure 1: An overview of our proposed SLP network, showing an initial translation from a spoken language sentence using a text encoder, with gloss supervision. A subsequent skeletal graphical structure is formed, with multiple proposed Skeletal Graph Self-Attention layers applied to embed a skeleton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

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 (Wilson and Anspach, 1993). Early work focused on isolated Sign Language Recognition (SLR) (Grobel and Assan, 1997), with a subsequent move to continuous SLR (Camgoz et al., 2017). The task of Sign Language Translation (SLT) was introduced by Camgoz *et al* (Camgoz et al., 2018) and has since become a prominent research area (Yin, 2020; Camgoz et al., 2020a). Sign Language Production (SLP), the automatic translation from spoken language sentences to sign language sequences, was initially tackled using avatar-based technologies (Elliott et al., 2008). The rule-based Statistical Machine Translation (SMT) achieved partial success (Kouremenos et al., 2018), albeit with costly, labour-intensive pre-processing.

Recently, there have been many deep learning approaches to SLP proposed (Zelinka and Kanis, 2020; Stoll et al., 2018; Saunders et al., 2020b), with Saunders *et al* achieving state-of-the-art results with gloss supervision (Saunders et al., 2021b). These works predominantly represent sign languages as sequences of skeletal frames, with each frame encoded as a vector of joint coordinates (Saunders et al., 2021a) that disregards any spatio-temporal structure available within a skeletal representation. In addition, these models apply standard spoken language architectures (Vaswani et al., 2017), 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) (Bruna et al., 2014) apply neural layers on these graphical structures to learn representations (Zhou et al., 2020), classify nodes (Yan et al., 2018; Yao et al., 2019) or generate new data (Li et al., 2018). A skeleton pose representation can be structured as a graph, with joints as $\mathcal{J}$ and natural limb connections as $\mathcal{E}$ (Straka et al., 2011; Shi et al., 2019). GNNs have been proposed for operating on such dynamic skeletal graphs, in the context of action recognition (Yan et al., 2018; Shi et al., 2019) and human pose estimation (Straka et al., 2011). Attention networks can be formalised as a fully connected GNN, where the adjacency between each word, $\mathcal{E}$, is a weighting learnt using self-attention. Expanding this, Graph Attention Networks (GATs) (Veličković et al., 2017) define explicit weighted adjacency between nodes, achieving state-of-the-art results across multiple domains (Kosaraju et al., 2019). Recently, there have been multiple graphical transformer architectures proposed (Yun et al., 2019; Dwivedi and Bresson, 2020), which have been extended to the spatio-temporal domain for applications such as multiple object tracking (Chu et al., 2021) and pedestrian tracking (Yu et al., 2020).

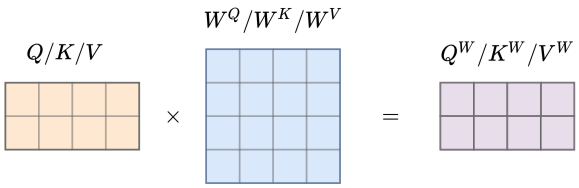

Figure 2: Weighted calculation of Queries, Q, Keys, K and Values, V, for global self-attention.

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 (de Amorim et al., 2019; Flasiński and Myśliński, 2010). 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 (Bahdanau et al., 2015), particularly with the introduction of transformers (Vaswani et al., 2017). Although proposed with global context (Bahdanau et al., 2015), more recent works have selectively restricted attention to a local context (Yang et al., 2018) or the top-k tokens(Zhao et al., 2019), 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 (Bahdanau et al., 2015). Transformers (Vaswani et al., 2017) 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 (Vaswani et al., 2017), 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:

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

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 (Saunders et al., 2021a; Stoll et al., 2018; Zelinka and Kanis, 2020) 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 (Saunders et al., 2020c). However, this approach removes all spatial information contained within the skeletal data, 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 (Bollobás, 2013), $\mathcal{G}$ 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).

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

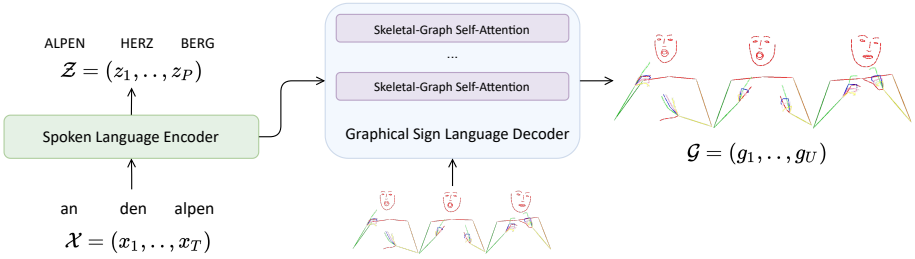

Figure 3: Overview of the proposed model architecture, detailing the Spoken Language Encoder (Sec. 4.1) and the Graphical Sign Language Decoder (Sec. 4.2). We propose novel Skeletal Graph Self-Attention layers to operate on the sign language skeletal graphs, $\mathcal{G}$.

As outlined in Sec. 3, classical self-attention operates with global context over all 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 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 (Vaswani et al., 2017) 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 (Saunders et al., 2021b), 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:

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

where $\text{Con}(i,j) = True$ 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 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:

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

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

Self-loops and Normalisation To account for information loops back to the same joint (Bollobás, 2013), 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 (Bollobás, 2013), 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 is formulated as:

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

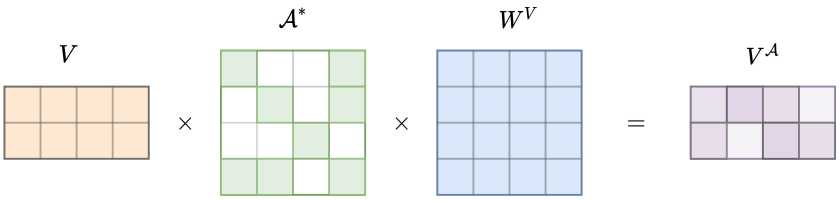

Figure 4: Skeletal Graph Self-Attention: Weighted calculation of Values, V, masked with a spatio-temporal adjacency matrix $\mathcal{A}^*$ to embed a skeleton inductive bias.

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

Skeletal Graph Self-Attention We apply $\mathcal{A}^*$ as a sparsely weighted mask over the weighted value calculation, $V^W = V \cdot W^V$, (Eq. 1), ensuring that values used in the weighted context for each node are 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 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 and subsequent layer norm 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 (Saunders et al., 2021b) and decoder skeleton pose losses.

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. (Camgoz et al., 2018), 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 (Zhang et al., 2016; Efthimiou and Fotinea, 2007), or contain larger domains of discourse that currently prove difficult for the SLP field (Camgoz et al., 2021). 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 (Cao et al., 2017), with the manuals lifted to 3D using the skeletal model estimation model proposed in (Zelinka and Kanis, 2020). 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 (Saunders et al., 2020c)). Adjacency information, $\mathcal{A}$, is defined as the natural limb connections of 3D body, hand and face joints, as in (Zelinka and Kanis, 2020), 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*

Skeletal Graph Layers, $\mathcal{L}$:	DEV SET					TEST SET				
	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE	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.

al (Saunders et al., 2020c), we apply Gaussian noise augmentation with a noise rate of 5. We train all parts of our network with Xavier initialisation, Adam optimization with default parameters and a learning rate of 10^{-3}. Our code is based on Kreutzer et al.'s NMT toolkit, JoeyNMT, and implemented using PyTorch.

Evaluation We use the back translation metric (Saunders et al., 2020c) for evaluation, which employs a pre-trained SLT model (Camgoz et al., 2020b) 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 (Saunders et al., 2020c). 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 experiments, we examine the impact of the temporal adjacency distance, $\mathcal{N}$, (Sec. 4.2). We set $\mathcal{N}$ by analysing 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 relying on this being learnt.

Table 2 shows results of our temporal adjacency evaluation, ranging from an infinite adjacency (no constraint) to $\mathcal{N} \in [1, 5]$. 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 (Saunders et al., 2021b), 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 (Saunders et al., 2020c), which applied the classical transformer architecture to sign language production. 2) Adversarial training (Saunders et al., 2020a), which utilised an adversarial discriminator to prompt more expressive productions, 3) Mixture Density Networks (MDNs) (Saunders et al., 2021a), which modelled the variation found in sign language using multiple distributions to parameterise the entire prediction subspace, and 4) Mixture of Motion Primitives

Temporal Adjacency, $\mathcal{N}$:	DEV SET					TEST SET				
	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE	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

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	11.82	14.80	19.97	31.41	33.18	10.51	13.54	19.04	31.36	32.46
Adversarial Training	12.65	15.61	20.58	31.84	33.68	10.81	13.72	18.99	30.93	32.74
Mixture Density Networks	11.54	14.48	19.63	30.94	33.40	11.68	14.55	19.70	31.56	33.19
Mixture of Motion Primitives	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.

(MOMP) (Saunders et al., 2021b), 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.

8. Bibliographical References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Bollobás, B. (2013). *Modern graph theory*. Springer Science & Business Media.

Bruna, J., Zaremba, W., Szlam, A., and LeCun, Y. (2014). Spectral Networks and Locally Connected Networks on Graphs. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Camgoz, N. C., Hadfield, S., Koller, O., and Bowden, R. (2017). SubUNets: End-to-end Hand Shape and Continuous Sign Language Recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., and Bowden, R. (2018). Neural Sign Language Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Camgoz, N. C., Koller, O., Hadfield, S., and Bowden, R. (2020a). Multi-channel Transformers for Multi-articulatory Sign Language Translation. In *Assistive Computer Vision and Robotics Workshop (ACVR)*.

Camgoz, N. C., Koller, O., Hadfield, S., and Bowden, R. (2020b). 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)*.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chu, P., Wang, J., You, Q., Ling, H., and Liu, Z. (2021). TransMOT: Spatial-Temporal Graph Transformer for Multiple Object Tracking. *arXiv preprint arXiv:2104.00194*.

de Amorim, C. C., Macêdo, D., and Zanchettin, C. (2019). Spatial-Temporal Graph Convolutional Networks for Sign Language Recognition. In *International Conference on Artificial Neural Networks*.

Dwivedi, V. P. and Bresson, X. (2020). A Generalization of Transformer Networks to Graphs. *arXiv preprint arXiv:2012.09699*.

Elliott, R., Glauert, J. R., Kennaway, J., Marshall, I., and Safar, E. (2008). Linguistic Modelling and Language-Processing Technologies for Avatar-based Sign Language Presentation. *Universal Access in the Information Society*.

Flasiński, M. and Myśliński, S. (2010). On The Use of Graph Parsing for Recognition of Isolated Hand Postures of Polish Sign Language. *Pattern Recognition*.

Grobel, K. and Assan, M. (1997). Isolated Sign Language Recognition using Hidden Markov Models. In *IEEE International Conference on Systems, Man, and Cybernetics*.

Kosaraju, V., Sadeghian, A., Martín-Martín, R., Reid, I.,

Rezatofighi, S. H., and Savarese, S. (2019). Social-BiGAT: Multimodal Trajectory Forecasting using Bicycle-GAN and Graph Attention Networks. In *Advances in Neural Information Processing Systems (NIPS)*.

Kouremenos, D., Ntalianis, K. S., Siolas, G., and Stafylopatis, A. (2018). 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)*.

Li, Y., Vinyals, O., Dyer, C., Pascanu, R., and Battaglia, P. (2018). Learning Deep Generative Models of Graphs. *arXiv preprint arXiv:1803.03324*.

Pfau, R., Quer, J., et al. (2010). *Nonmanuals: Their Grammatical and Prosodic Roles*. Cambridge University Press.

Saunders, B., Camgoz, N. C., and Bowden, R. (2020a). Adversarial Training for Multi-Channel Sign Language Production. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Saunders, B., Camgoz, N. C., and Bowden, R. (2020b). Everybody Sign Now: Translating Spoken Language to Photo Realistic Sign Language Video. *arXiv preprint arXiv:2011.09846*.

Saunders, B., Camgoz, N. C., and Bowden, R. (2020c). Progressive Transformers for End-to-End Sign Language Production. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Saunders, B., Camgoz, N. C., and Bowden, R. (2021a). Continuous 3D Multi-Channel Sign Language Production via Progressive Transformers and Mixture Density Networks. In *International Journal of Computer Vision (IJCV)*.

Saunders, B., Camgoz, N. C., and Bowden, R. (2021b). Mixed SIGNals: Sign Language Production via a Mixture of Motion Primitives. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Shi, L., Zhang, Y., Cheng, J., and Lu, H. (2019). Skeleton-based Action Recognition with Directed Graph Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Stokoe, W. C. (1980). Sign Language Structure. *Annual Review of Anthropology*.

Stoll, S., Camgoz, N. C., Hadfield, S., and Bowden, R. (2018). Sign Language Production using Neural Machine Translation and Generative Adversarial Networks. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Straka, M., Hauswiesner, S., Rüther, M., and Bischof, H. (2011). Skeletal Graph Based Human Pose Estimation in Real-Time. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Sutton-Spence, R. and Woll, B. (1999). *The Linguistics of British Sign Language: An Introduction*. Cambridge University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention Is All You Need. In *Advances in Neural Information Processing Systems (NIPS)*.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph Attention Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Wilson, B. J. and Anspach, G. (1993). Neural Networks for Sign Language Translation. In *Applications of Artificial Neural Networks IV*. International Society for Optics and Photonics.

Yan, S., Xiong, Y., and Lin, D. (2018). Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yang, B., Tu, Z., Wong, D. F., Meng, F., Chao, L. S., and Zhang, T. (2018). Modeling Localness for Self-Attention Networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yao, L., Mao, C., and Luo, Y. (2019). Graph Convolutional Networks for Text Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yin, K. (2020). Sign Language Translation with Transformers. *arXiv preprint arXiv:2004.00588*.

Yu, C., Ma, X., Ren, J., Zhao, H., and Yi, S. (2020). Spatio-Temporal Graph Transformer Networks for Pedestrian Trajectory Prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Yun, S., Jeong, M., Kim, R., Kang, J., and Kim, H. J. (2019). Graph Transformer Networks. In *Advances in Neural Information Processing Systems (NIPS)*.

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

Zhao, G., Lin, J., Zhang, Z., Ren, X., Su, Q., and Sun, X. (2019). Explicit Sparse Transformer: Concentrated Attention Through Explicit Selection. *arXiv preprint arXiv:1912.11637*.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. (2020). Graph Neural Networks: A Review of Methods and Applications.

9. Language Resource References

Camgoz, Necati Cihan and Saunders, Ben and Rochette, Guillaume and Giovanelli, Marco and Inches, Giacomo and Nachtrab-Ribback, Robin and Bowden, Richard. (2021). *Content4All Open Research Sign Language Translation Datasets*.

Efthimiou, Eleni and Fotinea, Stavroula-Evita. (2007). *GSLC: Creation and Annotation of a Greek Sign Language Corpus for HCI*.

Zhang, Jihai and Zhou, Wengang and Xie, Chao and Pu, Junfu and Li, Houqiang. (2016). *Chinese Sign Language Recognition with Adaptive HMM*.