Gloss Semantic-Enhanced Network with Online Back-Translation for Sign Language Production

Shengeng Tang
tsg1995@mail.hfut.edu.cn
Hefei University of Technology
Key Laboratory of Knowledge Engineering with Big Data (HFUT), Ministry of Education
Hefei, China

Richang Hong*
Dan Guo*
hongrc.hfut@gmail.com
guodan@hfut.edu.cn
Hefei University of Technology
Key Laboratory of Knowledge Engineering with Big Data (HFUT), Ministry of Education
Hefei, China

Meng Wang
eric.mengwang@gmail.com
Hefei University of Technology
Key Laboratory of Knowledge Engineering with Big Data (HFUT), Ministry of Education
Hefei, China

ABSTRACT

Sign Language Production (SLP) aims to generate the visual appearance of sign language according to the spoken language, in which a key procedure is to translate sign Gloss to Pose (G2P). Existing G2P methods mainly focus on regression prediction of posture coordinates, namely closely fitting the ground truth. In this paper, we provide a new viewpoint: a Gloss semantic-Enhanced Network is proposed with Online Back-Translation (GEN-OBT) for G2P in the SLP task. Specifically, GEN-OBT consists of a gloss encoder, a pose decoder, and an online reverse gloss decoder. In the gloss encoder based on the transformer, we design a learnable gloss token without any prior knowledge of gloss, to explore the global contextual dependency of the entire gloss sequence. During sign pose generation, the gloss token is aggregated onto the existing generated poses as gloss guidance. Then, the aggregated features are interacted with the entire gloss embedding vectors to generate the next pose. Furthermore, we design a CTC-based reverse decoder to convert the generated poses backward into glosses, which guarantees the semantic consistency during the processes of gloss-to-pose and pose-to-gloss. Extensive experiments on the challenging PHOENIX14T benchmark demonstrate that the proposed GEN-OBT outperforms the state-of-the-art models. Visualization results further validate the interpretability of our method.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Natural language processing**; *Machine learning*.

KEYWORDS

Sign Language Production, Gloss Semantic Enhancement, Online Back-Translation, Deep Learning

ACM Reference Format:
Shengeng Tang, Richang Hong, Dan Guo, and Meng Wang. 2022. Gloss Semantic-Enhanced Network with Online Back-Translation for Sign Language Production. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3547830

1 INTRODUCTION

Sign Language Production (SLP) is an emerging and challenging task in the vision-language field. Specifically, SLP is the inverse process of Sign Language Recognition (SLR) and Sign Language Translation (SLT), which converts a text sentence into the visual representation of sign language. This task requires the model to understand textual semantics and generate a corresponding sign pose or appearance sequence. It refers to many prevalent techniques, such as natural language processing [47, 49], human pose estimation [1, 17], video generation [41, 43], *etc.*

Sign gloss is defined as a minimal lexical item in sign linguistics; it plays a crucial textual representational unit in the process of SLP. As shown in Figure 1, the SLP system usually translates the text language into a gloss sequence (T2G) [14, 16], and then converts the gloss sequence into a series of sign poses (G2P) [25, 26, 30]. Since the T2G can be well-solved by Neural Machine Translation (NMT, language-to-language) based method [22] and rule-based method [20], G2P (a cross-modal task in essence) becomes an urgent need to be addressed. In this work, we focus on the G2P task of SLP, the key procedure of SLP.

Early works utilize **avatar-based** method [13] and **Statistical Machine Translation (SMT)** method [14], which require an expensive cost of pose pre-acquisition and struggle to cope with a large number of unseen phrases. Recent efforts toward SLP attempt to use deep neural networks for text-to-vision mapping [18, 42, 46]. Inspired by the merit of Generative Adversarial Networks (GANs) on generative tasks, some **conditional GANs**-based SLP methods have emerged [25, 28]. These methods discriminate real poses from the fakes (*i.e.*, original or generated) to ensure the realistic production of poses. Meanwhile, some **Non-AutoRegressive models** are explored to address the error propagation problem of pose generation in SLP [11, 12]. Nowadays, a new common practice is to use the **Transformer** framework [37] to decode a pose sequence [26, 45].

The above-mentioned work makes effort to build accurate pose coordinates as well as the ground truth. In other words, they pay

[*] Corresponding authors.

MM '22, October 10–14, 2022, Lisboa, Portugal
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547830

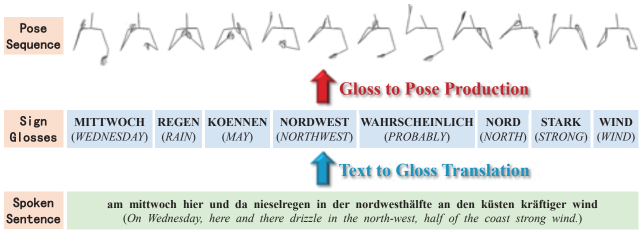

Pose Sequence
Gloss to Pose Production
Sign Glosses
MITTWOCH (WEDNESDAY)
REGEN (RAIN)
KOENNEN (MAY)
NORDWEST (NORTHWEST)
WAHRSCHEINLICH (PROBABLY)
NORD (NORTH)
STARK (STRONG)
WIND (WIND)
Text to Gloss Translation
Spoken Sentence
am mittwoch hier und da nieselregen in der nordwesthälfte an den küsten kräftiger wind
(On Wednesday, here and there drizzle in the north-west, half of the coast strong wind.)

Figure 1: The pipeline of SLP: text to gloss transformation (T2G) and gloss to pose production (G2P). The T2G can be well-performed by NMT-based approaches [14, 16], whereas the G2P is challenging due to poor quality of vision production. In this work, we focus on the G2P procedure of the SLP.

attention to the accuracy of 3D coordinates of pose consistently. In this work, besides the pose constraint, we investigate semantic preservation in the process of gloss-to-pose. We explore a guidance effect of gloss during pose generation and constraint the semantic consistency of original glosses (gloss-to-pose) and online back-translated glosses (pose-to-gloss).

To this end, as shown in Figure 2, we propose a novel Gloss semantic-Enhanced Network with Online Back-Translation (GEN-OBT), which enhances the original linguistics in terms of gloss modeling, translation of gloss-to-pose, and back-translation of pose-to-gloss. The GEN-OBT is a transformer-based model, which consists of a gloss encoder, a pose decoder, and an online reverse gloss decoder. First, the GEN-OBT introduces a learnable token (named gloss token [*GLO*]) into the gloss encoder. [*GLO*] explicitly captures the global semantics of the original gloss sequence. Next, [*GLO*] is used as a guidance term to decode the pose sequence. The pose decoder is a recurrent transformer, which performs interactive attention between the entire gloss sequence and existing decoded poses to predict the next pose. Finally, a reverse decoder is designed to translate the above generated poses backward into a reproduced gloss sequence. We apply the theory of Connectionist Temporal Classification (CTC) [8] into the gloss decoder; we parse the input pose sequence in the gloss vocabulary and find an available pose-to-gloss alignment path with max probability. In our work, the CTC-based decoder is used to restrict the linguistic consistency during the processes of gloss-to-pose and pose-to-gloss back.

Compared with previous methods, our proposed GEN-OBT has two distinctive characteristics: (1) GEN-OBT introduces a learnable token term (*i.e.*, gloss token) to represent the global semantics of glosses. The token is used to guide the pose generation in a recurrent transformer architecture; (2) an online gloss decoder is designed to back-translate the generated poses into glosses. We evaluate the semantic re-productivity of glosses. We expect our work will inspire the related work in the fields of cross-modal machine translation and production. Our main contributions are summarized as follows:

- We propose a novel Gloss semantic-Enhanced Network with Online Back-Translation (GEN-OBT) for SLP, which explores the gloss context in terms of gloss modeling, translation of gloss-to-pose, and back-translation of pose-to-gloss.
- We design a learnable gloss token without any prior knowledge of gloss and embed it in a gloss transformer (encoder) to explore the global contextual dependency of the entire gloss sequence.
- We perform a recurrent pose transformer. For each time stamp for pose generation, the gloss token is aggregated onto existing the generated poses (deemed as a gloss guidance). Then, the aggregated features is interacted with the entire gloss embedding vectors to generate next pose.
- We further design a CTC-based reverse decoder to convert the generated poses backward into glosses. Through path merging and item de-redundancy in CTC, we consider all the possible decoding paths of original gloss sequence and the output path with maximum probability, which guarantees the semantic consistency during the processes of gloss-to-pose and pose-to-gloss.
- Extensive experiments on the challenging PHOENIX14T [3] dataset demonstrate the superiority of the proposed method. Ablation studies and qualitative visualizations verify the contribution of each component.

2 RELATED WORK

2.1 Sign Language Production (SLP)

Over the past decades, sign language research has developed from isolated Sign Language Recognition (SLR) [6, 10, 40] and continuous Sign Language Translation (SLT) [3, 9, 33] to Sign Language Production (SLP) [5, 11, 12, 26, 29, 30]. The tasks of SLR&SLT and SLP provide mutually inverse sign language solutions, in which SLP is more challenging due to poor quality of vision production.

Previous SLP works focus on avatar-based method [7, 13] and Statistical Machine Translation (SMT) method [14, 16]. These methods aim to generate realistic sign gestures. However, they heavily depended on the rule-based lookup of phrases in the pre-captured action database, requiring an expensive cost of action collection and being limited to the predefined phrases. Recently, deep learning-based methods have emerged for SLP, such as the RNN-based model [5, 45, 46], Generative Adversarial Network (GAN) [18, 30, 31, 36], Variational Auto Encoder (VAE) [12, 42], and transformer [11, 19, 25, 26, 28, 29].

Early deep learning models aim to translate textual descriptions into photo-realistic sign video (TG2V) as well as classic methods. They struggle to handle both details of gesture and finger [5, 48] but have unsatisfactory performance. A classic solution [30, 31] is proposed to divide the challenging SLP task into three sub-tasks, namely Text-to-Gloss translation (T2G), Gloss-to-Pose generation (G2P), and Pose-to-Video synthesis (P2V). T2G refers to the scope of natural language understanding, which can be well solved by the NMT-based approach [22] or the rule-based approach [20]. P2V is a back-end pure computer vision (CV) task, involving video synthesis techniques [21, 27, 38]. Among these sub-tasks, the G2P task of SLP is particularly crucial in modeling the skeletal gesture and pose, which remains the core semantics of sign linguistics. In this work, we focus on the G2P task.

2.2 Gloss-to-Pose generation (G2P)

For G2P, on the one hand, Saunders *et al.* propose a motion primitives network, which produces an infinite number of sign poses based on a Mixture-of-Expert (MoE) architecture [29]. To avoid the

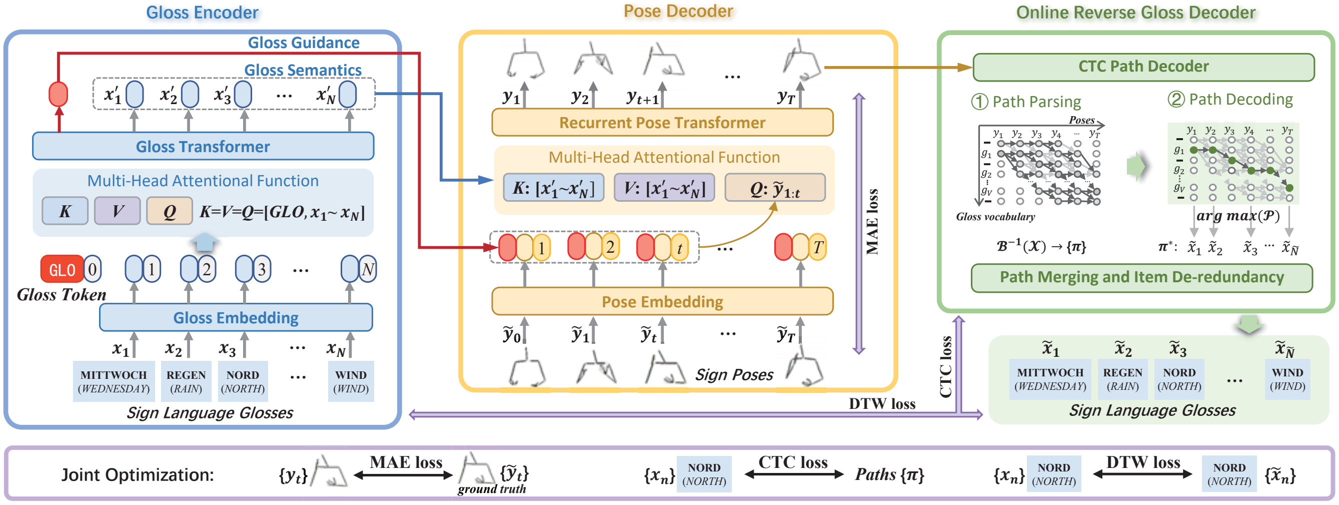


Figure 2: Overview of the proposed framework - GEN-OBT. It consists of a gloss encoder, a pose decoder, and an online reverse gloss decoder. Thereinto, in the gloss encoder, a token $[GLO]$ is used to learn the global semantics of a gloss sequence $\mathcal{X} = \{x_{1:N}\}$. Then, $[GLO]$ is added up to existing generated poses $\{y_{1:t}\}$ and then interacted with the gloss sequence to predict next pose y_{t+1}. After T time stamps, based on the pose sequence $\mathcal{Y} = \{y_{1:T}\}$, a connectionist temporal classification (CTC) optimization is applied to back-translate a new gloss sequence $\tilde{\mathcal{X}} = \{\tilde{x}_{1:\tilde{N}}\}$. For optimization, the MAE loss $\mathcal{L}_{MAE}$ calculates the accuracy of pose coordinates, and the CTC loss $\mathcal{L}_{CTC}$ and DTW loss $\mathcal{L}_{DTW}$ constraint the semantic consistency of reproduced glosses to original glosses ($\tilde{\mathcal{X}} = \{\tilde{x}_{1:\tilde{N}}\}$: pose-to-gloss, the path alignment of poss and gloss vocabulary).

error accumulation of regression prediction of pose coordinates, Hwang *et al.* build a Gaussian space to learn the spatial distribution of each pose and adopt a non-AutoRegressive model to map the language sentence into the target pose distribution [11]. Besides, Huang *et al.* propose a spatial-temporal GNN generator to smooth the variation of generated poses in sequence [11]. On the other hand, inspired by the great success of the transformer [23, 34, 44], researchers have generalized it into the field of vision-language learning [24, 32]. There are some new transformer-based models for G2P [11, 25, 26, 28, 29, 39, 45]. Saunders *et al.* design a progressive transformer to generate sign poses in an end-to-end manner [26]. Going a step further, they introduce a adversarial training scheme into the transformer, which learns to distinguish real from fake pose sequences to ensure the realistic production of poses [25]. Zelinka *et al.* devise a feed-forward transformer and employ a soft non-monotonic attention mechanism to convert the Czech language into skeletal sequences [45]. Viegas *et al.* propose a dual encoder transformer to generate both facial expression and sign pose from both the text word and gloss annotations [39]. Different from the above work focusing on the accuracy of pose generation, we explore semantic preservation by the manner of translation and back-translation between gloss and pose for G2P.

3 OUR METHOD

Given a sign gloss sentence with N glosses $\mathcal{X} = \{x_1, x_2, \cdots, x_N\}$, the SLP system is required to generate a pose sequence $\mathcal{Y} = \{y_1, y_2, \cdots, y_T\}$, where T is the number of generated poses. The task is flexible in which T is usually unequal to N.

3.1 Overall Pipeline

As illustrated in Figure 2, in this paper, we propose a Gloss semantic-Enhanced Network with Online Back-Translation (GEN-OBT), including three modules: a Gloss Encoder (see Section 3.2), a Pose Decoder (see Section 3.3), and an online reverse Gloss Decoder (see Section 3.4). We apply the transformer architecture [37] as the network backbone. In the *Gloss Encoder*, after gloss embedding, we use a learnable token named 'gloss token' to capture the global semantics of the gloss sequence. Then, in the *Pose Decoder*, we add up gloss token to the pose embedding vectors and take it as $Query$, and further take the gloss embedding sequence as both Key and $Value$ in a recurrent transformer. In other words, we leverage previous poses and the entire gloss sequence to predict the next pose. In the reverse *Gloss Decoder* (existing in the training stage), we calculate the probability of each generated pose over the gloss vocabulary and decode the alignment paths of pose-to-gloss by Connectionist Temporal Classification (CTC) optimization [8]. As shown in Figure 2, loss $\mathcal{L}_{MAE}$ is proposed to constraint the coordinate consistency of generated poses and the ground-truth; $\mathcal{L}_{CTC}$ optimizes all the alignments of pose-to-gloss during back-translation; while $\mathcal{L}_{DTW}$ measures the matching score of the reproduced glosses with the original gloss sequences. These losses guarantee the semantic preservation of both pose and gloss.

3.2 Gloss Encoder

To explore the semantics of gloss, we propose a gloss transformer as the encoder. At first, we map all the glosses into a high-dimensional feature space using a linear embedding layer. Then, we introduce a gloss token $[GLO]$ concatenated with the gloss sequence as follows:

$$\begin{aligned} x_n^e &= W^e \cdot x_n + b^e; \\ x_{0:N}^e &= \{[GLO], x_1^e, x_2^e, \cdots, x_N^e\} \in \mathbb{R}^{(N+1)\times d_x}, \end{aligned} \tag{1}$$

where x_n is a one-hot vector of the n-th gloss over the gloss vocabulary $\mathcal{V}$, $[GLO]$ is randomly initialized, and W^e and b^e represent the weight and bias respectively. Here is $x_0^e = [GLO]$.

Similarly to the sequential learning of natural language, in order to encode the temporal information, we apply a positional encoding layer to replenish the order of gloss:

$$x_n'^e = x_n^e + PE(n), \tag{2}$$

where PE is conducted by the sine and cosine functions on the temporal gloss order as in [37].

Until now, we have obtained the positional gloss representation $\{x_{0:N}'^e\}$ and will feed it into a gloss transformer to capture the global semantics of the glosses. The gloss transformer consists of L transformer blocks, where each block includes a Multi-Head Attention layer (MHA), a Normalization Layer (NL), and a Feed-forward Layer (FL). The encoding process can be expressed as:

$$\begin{aligned} & x_{0:N}' = GlossFormer(x_{0:N}'^e) \\ \Leftrightarrow & \begin{cases} z_0 = x_{0:N}'^e; \\ z_l = FL(MHA(NL(z_l)) + z_{l-1}), l \in [1, L]; \\ x_{0:N}' = NL(z_L). \end{cases} \end{aligned} \tag{3}$$

To be specific, MHA plays a key role in tackling contextual dependencies in sequence. We learn the token [GLO] in the MHA mechanism (namely implementing MHA on $z_0 = x_{0:N}'^e$, where $x_0'^e$ refers to [GLO]). As well as known, MHA performs scaled dot-product attention, which learns the relationship in the gloss sequence by using a series of variables - $Query\ Q$, $Key\ K$ and $Value\ V$.

$$Attention(Q, K, V) = softmax(\frac{QK^\top}{\sqrt{d}})V \tag{4}$$

where d is a scaling factor and Q, K and V refer to z_l consistently.

In this work, we realize the MHA with M heads as follows:

$$\begin{cases} MHA(Q, K, V)|_{Q=K=V} = [h_1, \cdots, h_M] \cdot W_O; \\ h_m = Attention(z_l W_Q^m, z_l W_K^m, z_l W_V^m), \end{cases} \tag{5}$$

where W_Q^m,W_K^m,W_V^m and W_O are learnable parameters.

At last, we disassemble the final output $\{x_n'\}_{n=0}^N$ into two original parts: token $[\widehat{GLO}]$=x_0' and the new gloss representation $\{x_n'\}_{n=1}^N$.

3.3 Pose Decoder

In this work, we aim to generate fine-grained 3D poses. At each time stamp, the pose data contains 50 joint points, *i.e.*, 8 points of body skeleton and 42 points of both left and right hands covering the finger skeleton. Thus, the dimension of 3D coordinates of each pose is d_{pos}=50 × 3=150. Similar to the gloss encoding, we encode the coordinates of each pose into a high-dimensional feature representation. We use a linear layer and another positional encoding layer PE. The calculation is formulated as follows:

$$\begin{aligned} y_t^p &= W^p \cdot y_t + b^p; \\ y_t'^p &= y_t^p + PE(t), \end{aligned} \tag{6}$$

where $y_t \in \mathbb{R}^{d_{pos}}$ denotes the pose's coordinates at t-th time stamp; W^p and b^p denote the learnable weight and bias respectively.

The proposed pose decoder is built based on a recurrent transformer, which aggregates all the gloss vectors and previously generated poses $\{y_{1:t}\}$, to predict the next pose y_{t+1}. It is worth noting that we take the gloss token $[GLO]$ as a global gloss semantic vector to guide the pose generation.

Step 1: We aggregate the gloss token onto each pose feature as below:

$$y_t^{\text{glo}} = [\widehat{GLO}] + y_t'^e \tag{7}$$

where $[\widehat{GLO}] \in \mathbb{R}^{d_x}$, $y_t'^e \in \mathbb{R}^{d_y}$ and we set $d_x = d_y$.

Step 2: We implement a recurrent pose transformer, which differs from the gloss transformer in two ways: (1) we build a progressive transformer for pose generation; (2) we realize the MHA with an interactive attention mechanism in the transformer. The recurrent transformer can be expressed as:

$$\begin{aligned} & y_{t+1}'^{\text{glo}} = PoseFormer(y_{1:t}^{\text{glo}}, x_{1:N}') \\ \Leftrightarrow\ & y_{t+1}'^{\text{glo}} = FL(MHA(y_{1:t}^{\text{glo}}, x_{1:N}') + y_t'^{\text{glo}}), t \in [1, T]. \end{aligned} \tag{8}$$

The interactive MHA is performed in the way that we take the encoded glosses as both Key and $Value$ (*i.e.*, K=V=$x_{1:N}'$) and all the previous poses as $Query$ (*i.e.*, Q=$y_{1:t}'^{\text{glo}}$). We predict the next pose $y_{t+1}'^{\text{glo}}$ as below:

$$\begin{cases} MHA(Q, K, V)|_{Q,\ K=V} = [h_1, \cdots, h_M] \cdot W_O'; \\ h_m = Attention(y_{1:t}^{\text{glo}} W_Q'^m, x_{1:N}' W_K'^m, x_{1:N}' W_V'^m), \end{cases} \tag{9}$$

where $W_Q'^m$,$W_K'^m$,$W_V'^m$ and W_O' are learnable parameters.

Step 3: After T time stamps, we have obtain the pose representation $y_{1:T}'^{\text{glo}}$. A linear layer is used to map each $y_t'^{\text{glo}}$ into the 3D coordinates as follows:

$$y_t = W'^p \cdot y_t'^{\text{glo}} + b'^p \in \mathbb{R}^{d_{pos}}, \tag{10}$$

where W'^p and b'^p denote the weight and bias respectively.

Pose Optimization: In the training stage, the Mean Absolute Error (MAE) loss is used to constraint the consistency of the produced poses $\mathcal{Y} = \{y_t\}_{t=1}^T$ and the ground truth $\widetilde{\mathcal{Y}} = \{\tilde{y}_t\}_{t=1}^T$.

$$\mathcal{L}_{MAE} = \frac{1}{T} \sum_{t=1}^{T} |y_t - \tilde{y}_t|. \tag{11}$$

3.4 Online Reverse Gloss Decoder

In this part, in order to ensure the semantic preservation in the process of gloss-to-pose, we back-translate the produced pose sequence $\mathcal{Y} = \{y_t\}_{t=1}^T$ to a new reproduced gloss sequence $\widetilde{\mathcal{X}}$. In practice, we adopt a combination of Multi-Layer Perception (MLP) and CTC optimizer [8] as an online back-translator. MLP maps each pose y_t over the gloss vocabulary $\mathcal{V}$ to obtain all the probability scores.

$$\mathcal{P} = \{p_t\}_{t=1}^T = softmax[MLP(\{y_t\}_{t=1}^T)], \tag{12}$$

where $\mathcal{P} \in \mathbb{R}^{T\times|\mathcal{V}|}$ and $|\mathcal{V}|$ is the vocabulary size. $\mathcal{V}$ is set as the collection of all the glosses in the training set and a blank gloss '-'.

Based on the probability matrix $\mathcal{P}$={$p_{1:T}$}, we calculate all the possible alignment paths between pose sequence $\mathcal{Y}$ and gloss vocabulary $\mathcal{V}$. For either path π, there is a many-to-one mapping operation $\mathcal{B}$, which merges the repetition and deletes the blank gloss in path π. If $\mathcal{B}(\pi) = \mathcal{X}$, we take π as an accessible alignment. In the CTC, the probability sum of all the possible paths $\{\pi\}$ is formulated as follows:

$$\mathrm{Pr}^{\pi} = \sum_{\pi \in \mathcal{B}^{-1}(\mathcal{X})} \mathcal{P}(\pi | p_t), \tag{13}$$

where $\mathcal{B}^{-1}(\mathcal{X}) = \{\pi | \mathcal{B}(\pi) = \mathcal{X}\}$ involves all the possible paths $\{\pi\}$. The probability of π is defined as follows:

$$\mathcal{P}(\pi | p_t) = \prod\nolimits_{t=1}^{T} p_{\pi_t}, \tag{14}$$

where p_{π_t} denotes the probability of t-th gloss in path π.

Gloss Optimization I: We implement the CTC [8] to parse all the possible alignment paths. The objective loss of CTC is to maximize the probability sum of all the possible alignment paths, which is formulated as follows:

$$\mathcal{L}_{CTC} = \sum_{\pi \in \mathcal{B}^{-1}(\mathcal{X})} - \log \mathrm{Pr}^{\pi} = - \sum_{\pi \in \mathcal{B}^{-1}(\mathcal{X})} \sum\nolimits_{t=1}^{T} p_{\pi_t}. \tag{15}$$

Gloss Optimization II: Here, we output an accessible path for the final reproduced gloss $\mathcal{X}$. We select the one with the maximum probability as follows:

$$\begin{aligned} \pi^* &= \arg\max_{\pi}(\mathcal{P}); \\ \widetilde{\mathcal{X}} &= \mathcal{B}(\pi^*), \end{aligned} \tag{16}$$

where $\widetilde{\mathcal{X}} = \{\tilde{x}_{1:\tilde{N}}\}$, please note that the gloss number $\tilde{N}$ may be unequal to N by using the many-to-one mapping operation $\mathcal{B}$. We measure the distance between the two sequences $\mathcal{X}$ and $\widetilde{\mathcal{X}}$ by using Dynamic Time Warping (DTW) [2]. This distance objective is defined as $\mathcal{L}_{DTW}$=DTW($\mathcal{X}, \widetilde{\mathcal{X}}$).

In the end, to train the model in an end-to-end manner, the full objective function in this work is given as follows:

$$\mathcal{L} = \alpha \mathcal{L}_{MAE} + \beta \mathcal{L}_{CTC} + \gamma \mathcal{L}_{DTW}, \tag{17}$$

where α, β, and γ are the hyper-parameters to balance the loss terms.

3.5 Discussion

Why introduce a gloss token into the pose representation in the pose decoder? Compared with the glosses already existing in the input sequence, the gloss token is randomly initialized and has no prior knowledge; it can 'fairly' model the interaction with each gloss in the sequence. Specifically, the self-attention in the gloss transformer calculates the relation between each two-elements in the sequence. The token [GLO] obverses the entire gloss sequence and absorbs the importance of each gloss by using the weighted average of multi-layer attention. Thus, we deem that the update of [GLO] is on behalf of the gloss sentence. The aggregation of the token [GLO] onto each pose representation is used to enhance the semantic guidance of the gloss context during pose generation.

Why design a CTC-based reverse gloss decoder? Connectionist Temporal Classification (CTC) [8] is a sequential learning model, which excels at solving the element-wise alignment between two unequal-length sequences. In this work, the task aims to translate a gloss sequence with length N into a pose sequence with length T, where N is not equal to T. The CTC theory is eminently suitable for the task. We employ it with two purposes: (1) we explore all the possible alignments of pose-to-gloss to recover the original gloss $\mathcal{X}$, and maximize the probability sum of all these alignments; (2) we output only one alignment $\widetilde{\mathcal{X}}$ with the maximum probability and restrict it close to the original gloss $\mathcal{X}$. Based on the CTC, we use the two-stream restrictions to ensure the semantic consistency of glosses. Different from previous work that focuses on pose accuracy, we consider the semantic preservation in the way of gloss-to-pose translation and the backward pose-to-gloss translation.

4 EXPERIMENTS

4.1 Experimental setup

Dataset. We experiment on the RWTH-PHOENIX-Weather2014T (PHOENIX14T) dataset [3], a publicly available German sign language corpus. The PHOENIX14T is a challenging dataset, which provides 8257 samples containing spoken sentences and corresponding gloss sequences and sign videos. Specifically, the corpus covers 2887 German words and the gloss vocabulary contains 1066 glosses.

Evaluation metrics. We evaluate the method with common metrics *BLEU*, *ROUGE* and *Word Error Rate* (*WER*); for *BLEU*, we provide n-grams from 1 to 4 for evaluating phase completeness. Additionally, we report the DTW distance [2] between the predicted poses and the ground truth of each sample, denoted as *DTW-P*.

Implementation details. For data prepossessing, following the convention in SLP [11, 29], we use OpenPose [4] to extract 2D coordinates of joint points of each signer from the original video, and apply a skeletal correction model [45] to convert 2D into 3D coordinates. In this task, the obtained 3D coordinates are regarded as the ground-truth pose label. In addition, in previous work [11, 12, 26, 29, 39], a SLT model named NSLT [3] is offline used as a translation evaluation tool, which translates the generated poses into a gloss sequence and a spoken language sentence. For a fair comparison, we follow the previous work and have retrained NSLT [11, 12] on PHOENIX14T. The NSLT is merely used for evaluation in the following experiments.

For model parameters, all the transformer modules in our GEN-OBT method are built with 2 layers and 4 heads in the embedding size of 512 consistently (*i.e.*, $L = 2$, $M = 4$, and $d_x = d_y = 512$). We apply a Gaussian noise onto pose coordinates in the pose embedding phase and the noise rate is set to 5. During training, we set $\alpha = \beta = \gamma = 1.0$ with Adam optimizer [15] and the learning rate of 1×10^{-3}. Experiments are performed with PyTorch on NVIDIA GeForce GTX 1080 Ti GPU.

4.2 Ablation Study

Empirical parameters in transformer. The transformer in the gloss encoder and the pose decoder contains two main hyper-parameters: **layer number** L and **head number** M. As depicted in Figure 3, the GEN-OBT achieves the best performances at $L = 2$ and $M = 4$. In GEN-OBT, stacking L=2 layers achieves the best performance, whereas $L > 2$ accords with the performance drop. The

Table 1: Ablation studies of token [GLO] and OBT (gloss decoder) in the GEN-OBT on PHOENIX14T dataset.

Methods	DEV						TEST					
	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	ROUGE↑	DTW-P↓	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	ROUGE↑	DTW-P↓
GEN-OBT w/o Gloss-Token	18.42	10.39	7.14	5.41	18.04	10.87	17.66	10.63	7.48	5.77	18.63	10.74
GEN w/o OBT	18.86	11.10	7.68	5.77	19.43	10.85	18.71	11.53	8.09	6.20	19.79	10.71
GEN-OBT (Ours)	**24.92**	**15.72**	**11.20**	**8.68**	**25.21**	**10.37**	**23.08**	**14.91**	**10.48**	**8.01**	**23.49**	**10.07**

Table 2: Ablation studies of token aggregation in the pose transformer on PHOENIX14T dataset.

Positions	DEV						TEST					
	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	ROUGE↑	DTW-P↓	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	ROUGE↑	DTW-P↓
GEN-OBT w/o [*GLO*]	18.42	10.39	7.14	5.41	18.04	10.87	17.66	10.63	7.48	5.77	18.63	10.74
Before Pose Emb.	20.22	11.78	8.03	6.04	19.94	10.59	20.05	12.31	8.62	6.59	20.05	10.38
Between Pose Emb.&Dec.	**24.92**	**15.72**	**11.20**	**8.68**	**25.21**	**10.37**	**23.08**	**14.91**	**10.48**	**8.01**	23.49	**10.07**
After Pose Dec.	22.45	14.24	10.02	7.60	24.69	10.43	21.48	14.02	9.98	7.68	**24.03**	10.33

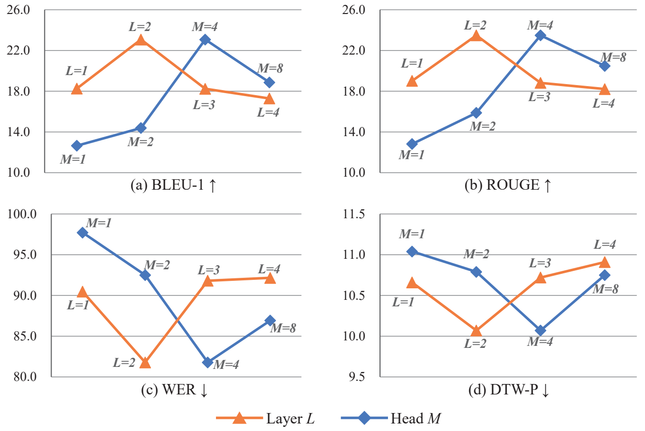
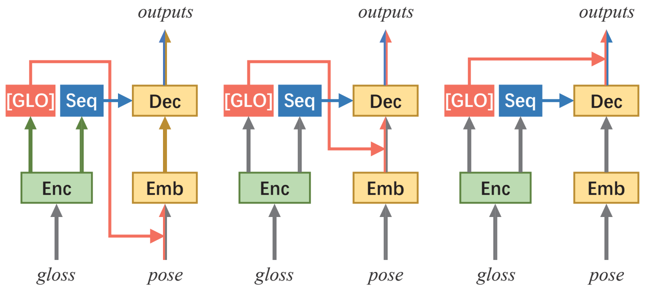


Figure 3: Ablation studies of head number M and layer number L. In the practice, M and L are fixed with optimal values $M = 4$ and $L = 2$.

head number M reflects the interaction diversity of the attention mechanism. $M = 4$ is a optimal setup. In the following experiments, we set $L = 2$ and $M = 4$.

Two essential factors in GEN-OBT. Here, we test the two essential factors in GEN-OBT: gloss token [GLO] and online back-translation. As shown in Table 1, **GEN-OBT w/o Gloss-Token** refers to the removal of token [GLO] in gloss encoder, which deteriorates 6.5%/5.42% *BLEU-1*, 7.17%/4.86% *ROUGE*, and -0.5/-0.67 *DTW-P* on DEV/TEST compared to **GEN-OBT**, respectively. **GET w/o OBT** removes the reverse gloss decoder in **GEN-OBT**, which leads to obvious performance degradation compared to **GEN-OBT** too (*e.g.*, -6.06%/-4.37% *BLEU-1*, -5.78%/-3.70% *ROUGE* on DEV/TEST). The introduction of gloss tokens and online back-translation significantly improves the quality of the generated sign poses.

Different token aggregation strategies. Here we discuss the usage of gloss semantics [GLO] in the pose decoder. We provide three strategies of token aggregation as depicted in Figure 4. The essence of the token is to learn the global semantics of glosses and then guide the pose generation. As the experimental results shown in Table 2, the usage of token **Between Pose Emb.&Dec.** in Figure 4 (a) achieves the best performance on most metrics, especially lifting *BLEU-1* by 4.7%/3.03% and 2.47%/1.6%, *BLEU-4* by 2.64%/1.42% and 1.08%/0.33% on DEV/TEST compared to the others (**Before Pose Emb.** and **After Pose Dec.** in Figure 4 (b)and (c)), respectively. We observe that even simply adding the token up to the pose representation **After Pose Dec.**, the influence of gloss is unavoidable. Considering that **Between Pose Emb.&Dec.** has comprehensive advantages in most metrics, we take it as our choice - the optimal token aggregation setup. Anyway, either usage of token aggregation in Figure 4 outperforms **GEN-OBT w/o** [GLO]. This further shows the effectiveness of gloss token.



(a) Before Pose Emb. (b) Between Pose Emb.&Dec. (c) After Pose Dec.

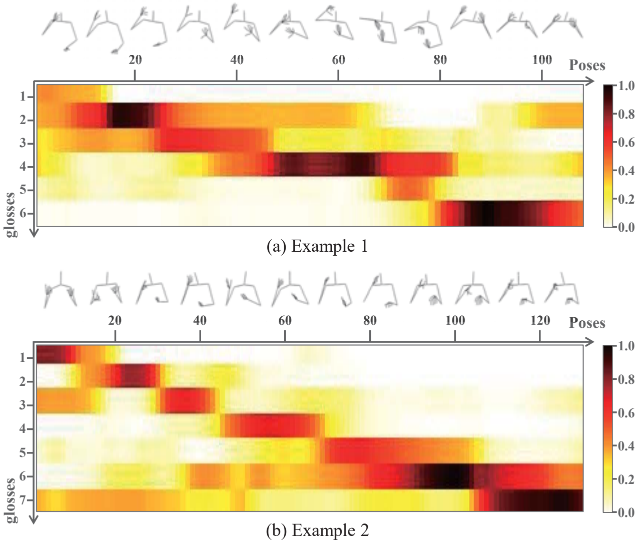
Figure 4: Illustration of different gloss token aggregation strategies. We provide three aggregation positions: (a) aggregating token [$\widehat{GLO}$] before pose embedding (at Step 1 in Sec.3.3), (b) aggregating [$\widehat{GLO}$] between pose embedding and pose transformer (at Step 2 in Sec.3.3), and (c) aggregating [$\widehat{GLO}$] after pose transformer (at Step 3 in Sec.3.3).

4.3 Comparison with State-of-the-Arts

We compare our GEN-OBT with state-of-the-art methods as follows. **PT-base** [26] is a native transformer model for SLP, which addresses T2G and G2P procedures simultaneously. **PT-FP&GN** [26] is an extension of PT-base, which introduces Gaussian noise onto original poses for data augmentation and predicts a pose segment with a 10-frame sliding window at each time. In other words, **PT-FP&GN** predicts a pose segment at once, whereas we produce a solo pose frame. **NAT-AT** [11] is a graph-based model, which first predicts

Table 3: Quantitative results on PHOENIX14T dataset. '†' indicates the reconstructed results.

Methods	DEV							TEST						
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	WER↓	DTW-P↓	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	WER↓	DTW-P↓
Ground Truth	29.77	20.21	15.16	12.13	29.60	74.17	0.00	29.76	20.12	14.93	11.93	28.98	71.94	0.00
PT-base† [26]	9.53	3.45	1.62	0.72	8.61	98.53	29.33	9.47	3.37	1.47	0.59	8.88	98.36	28.48
PT-FP&GN† [26]	12.51	6.50	4.76	3.88	11.87	96.85	11.75	13.35	7.29	5.33	4.31	13.17	96.50	11.54
NAT-AT [11]	–	–	–	–	–	–	–	14.26	9.93	7.11	5.53	18.72	88.15	–
NAT-EA [11]	–	–	–	–	–	–	–	15.12	10.45	7.99	6.66	19.43	82.01	–
DET [39]	17.25	10.17	7.04	5.32	17.85	–	–	17.18	10.39	7.39	5.76	17.64	–	–
GEN-OBT (Ours)	**24.92**	**15.72**	**11.20**	**8.68**	**25.21**	**82.36**	**10.37**	**23.08**	**14.91**	**10.48**	**8.01**	**23.49**	**81.78**	**10.07**



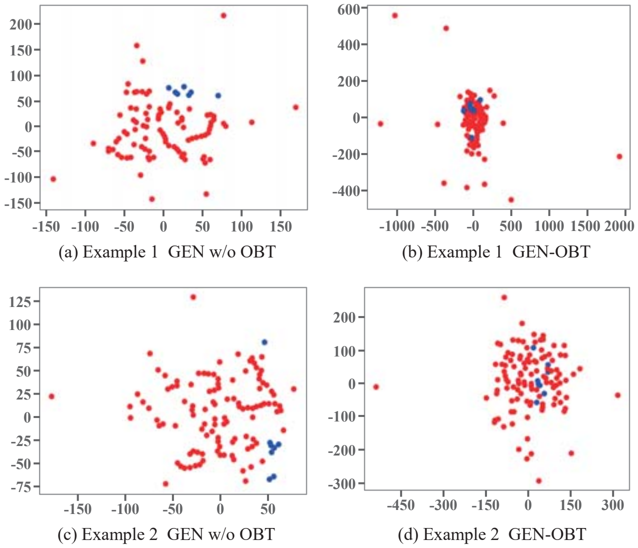
(a) Example 1

(b) Example 2

Figure 5: Interactive cross-modal attention of poses and glosses in the pose transformer. (a) Example 1 - Glosses: FREITAG SONNE WOLKE SUED ANFANG REGEN (Friday sun cloud south beginning rain). (b) Example 2 - Glosses: JETZT WIE-AUSSEHEN WETTER MORGEN DONNERSTAG NEUNZEHN NOVEMBER (now look-like weather morning Thursday nineteen November).

the duration of poses and then utilizes a spatial-temporal graph convolutional generator to produce a pose sequence. Compared with **NAT-AT**, **NAT-EA** [11] further explores a semantic constraint with Gaussian distribution. **DET** [39] introduces extra facial information - facial landmarks and facial action units - to generate sign poses.

As shown in Table 3, **GEN-OBT (Ours)** performs prominent superiority over the others. Compared with transformer-based baseline **PT-base**, all the evaluation metrics of GEN-OPT (Ours) are improved significantly, especially for *WER*↓ and *DTW-P*↓; the performance reductions exceed 16.17%/16.58% and 18.96/18.41 on DEV/TEST, respectively. And compared with the extended transformer method **PT-FP&GN**, our method is still superior in all metrics, such as *BLEU-1* (increasing 12.41%/9.73%), *ROUGE* (increasing by 13.34%/10.32%), and *WER*↓ (reducing by 14.49%/14.72%). For the graph-based generation models, **NAT-AT** and **NAT-EA** have just reported the experimental results on the TEST set. For example, compared with **NAT-EA** (the advanced version in them), **GEN-OBT** has obvious advantages (*e.g.*, 23.08% vs. 15.12% on *BLEU-1* and 23.49% vs. 19.43% on *ROUGE*). Furthermore, compared with **DET**, our approach achieves better performance too (*e.g.*, lifting *ROUGE* by 7.36%/5.85% on DEV/TEST) without additional supervision.



(a) Example 1 GEN w/o OBT

(b) Example 1 GEN-OBT

(c) Example 2 GEN w/o OBT

(d) Example 2 GEN-OBT

Figure 6: Visualization of feature distributions of Examples 1 and 2 (in Figure 5) using t-SNE. Red and blue points mark the generated poses and original input glosses, respectively.

4.4 Qualitative Results

Instantiation of interactive cross-modal attention. Figure 5 shows two examples of cross-modal interaction between the gloss and the pose sequences in the pose transformer. As shown in Figure 5, the highly responsive attention regions are distributed diagonally in the attention map. In other words, the semantic distributions of pose and gloss in our model are consistent along the temporal dimension. At each time stamp, our model generates corresponding poses according to the sequential glosses. In addition, in the task, the short gloss sequence is required to be transformed into the long pose sequence, where $N \ll T$. As shown in Figure 5, each gloss is responsive to a sub-sequence (a 'fragment', not a frame) of pose obviously. And the boundary of produced poses according

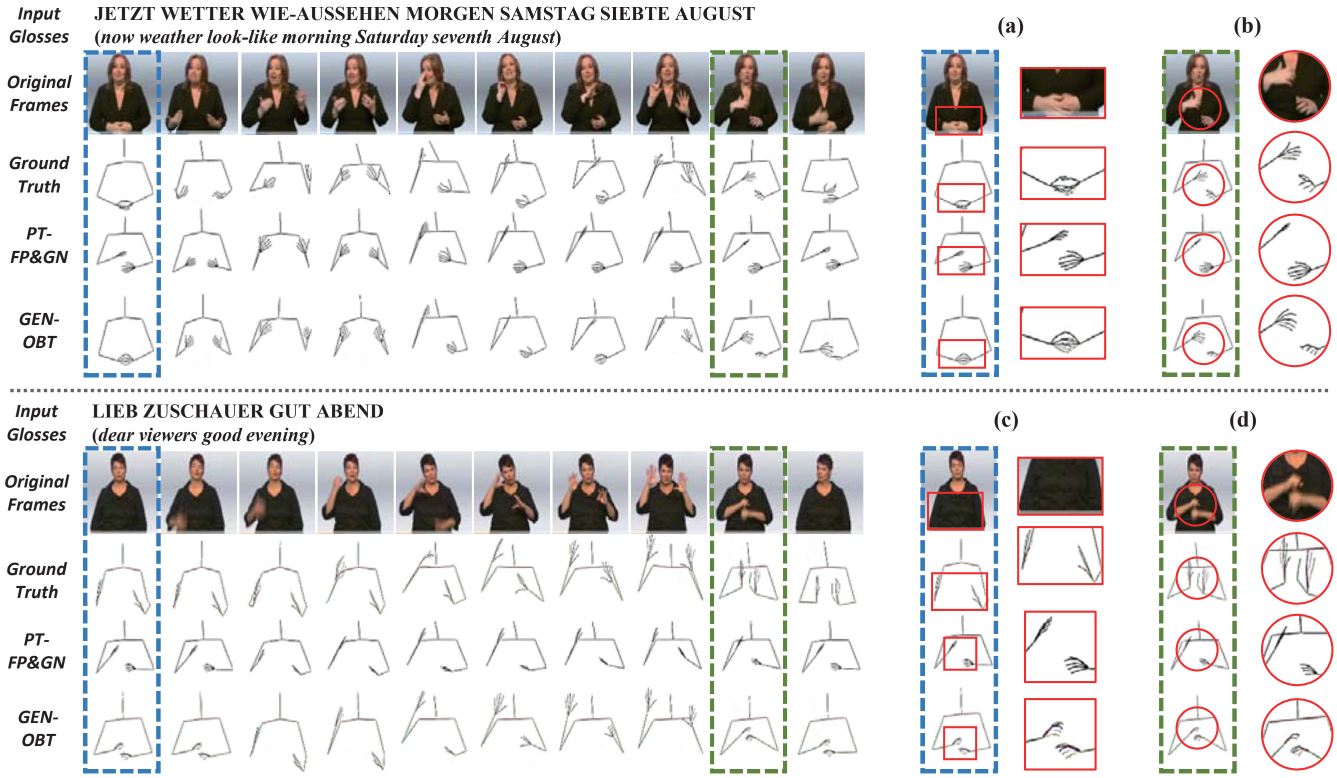


Figure 7: Visualization examples of produced pose sequence. We compare our GEN-OBT with PT-FP&GN [26] and the ground-truth. (a)~(d) display the enlarged region of hand posture. In the upper example, GEN-OBT fits the ground-truth much better. In the lower example, even with wrong pose labels, GEN-OBT generates close-to-natural arm poses. From these visualization examples, our method can tackle the cases of undetected arms or quick motion afterimage.

to two adjacent glosses is obvious too. These experimental conclusions are reasonable. This indicates that the decoded poses and the sequentially input gloss are in a clear many-to-one correspondence.

Effectiveness of online back-translation (OBT). Furthermore, we analyze the feature distributions of the samples in Figure 5 and use t-SNE [35] to visualize the feature distributions of original glosses (blue points) and generated poses (red points) as shown in Figure 6. Thereinto, Figure 6 (a) and (c) are obtained from the comparative GEN w/o OBT, while Figure 6 (b) and (d) are obtained from the proposed GEN-OBT. For Example 1, the spatial distribution of gloss features and pose features are consistent under the constraint of OBT. In the case of removing OBT, the gloss features are distributed at the edge of the pose features, and the two distributions are independent. The feature distributions in Example 2 also exhibit similar characteristics, which illustrates that OBT helps constrain the association between input glosses and output poses.

Visualization examples of generated pose sequences. As shown in Figure 7, another two examples of pose production are displayed to demonstrate the superiority of the proposed GEN-OBT. We compare the generation results of GEN- and PT-FP&GN [26]. For the upper example shown in Figures 7 with normal posture labels, our GEN-OBT produces more realistic poses than PT-FP&GN. For the bottom example with noise labels, our method generates close-to-natural poses, namely performing temporal smoothing and continuity along the generate pose sequence. As shown in Figures 7 (a)~(d), sometimes the ground truth fails to capture posture details due to motion afterimages or the undetected joints, our method still has a good robustness.

5 CONCLUSIONS

In this paper, we propose a Gloss semantic-Enhanced Network with Online Back-Translation (GEN-OBT) for SLP. We develop an encoder with a gloss token to learn the global semantics of glosses. The token is taken as a gloss guidance term, which is aggregated onto the pose sequence and then interacted with the gloss sequence to progressively predict the next pose. In our work, the pose decoder is a recurrent transformer. After the complete collection of the pose sequence, an CTC-based reverse decoder is proposed to produce the poses back into glosses. The CTC optimization guarantees semantic preservation in terms of both pose and gloss. Extensive experiments validate the effectiveness of these techniques.

ACKNOWLEDGMENTS

This work is supported in part by the National Natural Science Foundation of China (NSFC) under grants 61876058, U20A20183, 61932009, 61725203, 62020106007.

REFERENCES

[1] Qian Bao, Wu Liu, Jun Hong, Lingyu Duan, and Tao Mei. 2020. Pose-native Network Architecture Search for Multi-person Human Pose Estimation. In *ACM International Conference on Multimedia*. 592–600.

[2] Donald J Bemdt and James Clifford. 1994. Using Dynamic Time Warping to Find Patterns in Time Series. In *AAAI Workshop on Knowledge Discovery in Databases*. 359–370.

[3] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural Sign Language Translation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7784–7793.

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7291–7299.

[5] Runpeng Cui, Zhong Cao, Weishen Pan, Changshui Zhang, and Jianqiang Wang. 2019. Deep Gesture Video Generation with Learning on Regions of Interest. *IEEE Transactions on Multimedia* 22, 10 (2019), 2551–2563.

[6] Mathieu De Coster, Mieke Van Herreweghe, and Joni Dambre. 2021. Isolated Sign Recognition from RGB Video Using Pose Flow and Self-attention. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3441–3450.

[7] JRW Glauert, Ralph Elliott, SJ Cox, Judy Tryggvason, and Mary Sheard. 2006. VANESSA – A System for Communication Between Deaf and Hearing People. *Technology and Disability* 18, 4 (2006), 207–216.

[8] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *International Conference on Machine Learning*. 369–376.

[9] Dan Guo, Shengeng Tang, and Meng Wang. 2019. Connectionist temporal modeling of video and language: a joint model for translation and sign labeling. In *International Joint Conference on Artificial Intelligence*. 751–757.

[10] Dan Guo, Wengang Zhou, Houqiang Li, and Meng Wang. 2017. Online Early-late Fusion Based on Adaptive HMM for Sign Language Recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications* 14, 1 (2017), 1–18.

[11] Wencan Huang, Wenwen Pan, Zhou Zhao, and Qi Tian. 2021. Towards Fast and High-Quality Sign Language Production. In *ACM International Conference on Multimedia*. 3172–3181.

[12] Euijun Hwang, Jung-Ho Kim, and Jong-Cheol Park. 2021. Non-Autoregressive Sign Language Production with Gaussian Space. In *British Machine Vision Conference*.

[13] Kostas Karpouzis, George Caridakis, S-E Fotinea, and Eleni Efthimiou. 2007. Educational Resources and Implementation of a Greek Sign Language Synthesis Architecture. *Computers & Education* 49, 1 (2007), 54–74.

[14] Dilek Kayahan and Tunga Güngör. 2019. A Hybrid Translation System from Turkish Spoken Language to Turkish Sign Language. In *International Symposium on INnovations in Intelligent SysTems and Applications*. 1–6.

[15] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.

[16] Dimitris Kouremenos, Klimis S Ntalianis, Giorgos Siolas, and Andreas Stafylopatis. 2018. Statistical Machine Translation for Greek to Greek Sign Language Using Parallel Corpora Produced via Rule-Based Machine Translation. In *International Conference on Tools with Artificial Intelligence*. 28–42.

[17] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. 2019. Pifpaf: Composite Fields for Human Pose Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 11977–11986.

[18] Shyam Krishna and Janmesh Ukey. 2021. GAN Based Indian Sign Language Synthesis. In *Indian Conference on Vision, Graphics and Image Processing*. 1–8.

[19] Taro Miyazaki, Yusuke Morita, and Masanori Sano. 2020. Machine Translation from Spoken Language to Sign Language Using Pre-trained Language Model As Encoder. In *Workshop on the Representation and Processing of Sign Languages*. 139–144.

[20] Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. Data Augmentation for Sign Language Gloss Translation. In *International Workshop on Automatic Translation for Signed and Spoken Languages*. 1–11.

[21] B Natarajan and R Elakkiya. 2022. Dynamic GAN for High-Quality Sign Language Video Generation from Skeletal Poses Using Generative Adversarial Networks. *Soft Computing* (2022).

[22] Achraf Othman and Mohamed Jemni. 2011. Statistical Sign Language Machine Translation: from English written text to American Sign Language Gloss. *International Journal of Computer Science Issues* 8, 5 (2011), 65–73.

[23] Gabriele Prato, Ella Charlaix, and Mehdi Rezagholizadeh. 2020. Fully Quantized Transformer for Machine Translation. In *Conference on Empirical Methods in Natural Language Processing*. 1–14.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[25] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Adversarial Training for Multi-Channel Sign Language Production. In *British Machine Vision Conference*.

[26] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Progressive Transformers for End-to-end Sign Language Production. In *European Conference on Computer Vision*. 687–705.

[27] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. AnonySign: Novel Human Appearance Synthesis for Sign Language Video Anonymisation. In *IEEE International Conference on Automatic Face and Gesture Recognition*. 1–8.

[28] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. Continuous 3d Multi-channel Sign Language Production Via Progressive Transformers and Mixture Density Networks. *International Journal of Computer Vision* 129, 7 (2021), 2113–2135.

[29] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. Mixed SIGNals: Sign Language Production via a Mixture of Motion Primitives. In *IEEE International Conference on Computer Vision*. 1919–1929.

[30] Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2020. Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks. *International Journal of Computer Vision* 128, 4 (2020), 891–908.

[31] Stephanie Stoll, Simon Hadfield, and Richard Bowden. 2020. SignSynth: Data-Driven Sign Language Video Generation. In *European Conference on Computer Vision*. 353–370.

[32] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *International Conference on Learning Representations*.

[33] Shengeng Tang, Dan Guo, Richang Hong, and Meng Wang. 2021. Graph-Based Multimodal Sequential Embedding for Sign Language Translation. *IEEE Transactions on Multimedia* (2021).

[34] Inigo Jauregi Unanue, Jacob Parnell, and Massimo Piccardi. 2021. BERTTune: Fine-Tuning Neural Machine Translation with BERTScore. In *Annual Meeting of the Association for Computational Linguistics*. 915–924.

[35] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data Using t-SNE. *Journal of Machine Learning Research* 9 (2008), 2579–2605.

[36] Neel Vasani, Pratik Autee, Samip Kalyani, and Ruhina Karani. 2020. Generation of Indian Sign Language by Sentence Processing and Generative Adversarial Networks. In *International Conference on Information Systems Security*. 1250–1255.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Annual Conference on Neural Information Processing Systems*.

[38] Lucas Ventura, Amanda Duarte, and Xavier Giró-i Nieto. 2020. Can Everybody Sign Now? Exploring Sign Language Video Generation from 2d Poses. In *Sign Language Recognition, Translation & Production*.

[39] Carla Viegas, Mert Inan, Lorna Quandt, and Malihe Alikhani. 2022. Including Facial Expressions in Contextual Embeddings for Sign Language Generation. *arXiv preprint arXiv:2202.05383* (2022).

[40] Hanjie Wang, Xiujuan Chai, and Xilin Chen. 2019. A Novel Sign Language Recognition Framework Using Hierarchical Grassmann Covariance Matrix. *IEEE Transactions on Multimedia* 21, 11 (2019), 2806–2814.

[41] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. 2020. G3AN: Disentangling Appearance and Motion for Video Generation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5264–5273.

[42] Qinkun Xiao, Minying Qin, and Yuting Yin. 2020. Skeleton-based Chinese Sign Language Recognition and Generation for Bidirectional Communication Between Deaf and Hearing People. *Neural Networks* 125 (2020), 41–55.

[43] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. 2018. Pose Guided Human Video Generation. In *European Conference on Computer Vision*. 201–216.

[44] Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Weinan Zhang, Yong Yu, and Lei Li. 2020. Towards Making the Most of Bert in Neural Machine Translation. In *AAAI Conference on Artificial Intelligence*, Vol. 34. 9378–9385.

[45] Jan Zelinka and Jakub Kanis. 2020. Neural Sign Language Synthesis: Words Are Our Glosses. In *International Workshop on Applications of Computer Vision*. 3395–3403.

[46] Jan Zelinka, Jakub Kanis, and Petr Salajka. 2019. NN-based Czech Sign Language Synthesis. In *International Conference on Speech and Computer*. 559–568.

[47] Jiali Zeng, Shuangzhi Wu, Yongjing Yin, Yufan Jiang, and Mu Li. 2021. Recurrent Attention for Neural Machine Translation. In *Conference on Empirical Methods in Natural Language Processing*. 3216–3225.

[48] Ni Zeng, Yiqiang Chen, Yang Gu, Dongdong Liu, and Yunbing Xing. 2020. Highly Fluent Sign Language Synthesis Based on Variable Motion Frame Interpolation. In *IEEE International Conference on Systems, Man, and Cybernetics*. 1772–1777.

[49] Tianfu Zhang, He-Yan Huang, Chong Feng, and Longbing Cao. 2021. Enlivening Redundant Heads in Multi-head Self-attention for Machine Translation. In *Conference on Empirical Methods in Natural Language Processing*. 3238–3248.