Gloss Prior Guided Visual Feature Learning for Continuous Sign Language Recognition

Leming Guo, Wanli Xue, *Member, IEEE*, Bo Liu, Kaihua Zhang, *Member, IEEE*, Tiantian Yuan, and Dimitris Metaxas, *Fellow, IEEE*

***Abstract*— Continuous sign language recognition (CSLR) is to recognize the glosses in a sign language video. Enhancing the generalization ability of CSLR's visual feature extractor is a worthy area of investigation. In this paper, we model glosses as priors that help to learn more generalizable visual features. Specifically, the signer-invariant gloss feature is extracted by a pre-trained gloss BERT model. Then we design a gloss prior guidance network (GPGN). It contains a novel parallel densely-connected temporal feature extraction (PDC-TFE) module for multi-resolution visual feature extraction. The PDC-TFE captures the complex temporal patterns of the glosses. The pre-trained gloss feature guides the visual feature learning through a cross-modality matching loss. We propose to formulate the cross-modality feature matching into a regularized optimal transport problem, it can be efficiently solved by a variant of the Sinkhorn algorithm. The GPGN parameters are learned by optimizing a weighted sum of the cross-modality matching loss and CTC loss. The experiment results on German and Chinese sign language benchmarks demonstrate that the proposed GPGN achieves competitive performance. The ablation study verifies the effectiveness of several critical components of the GPGN. Furthermore, the proposed pre-trained gloss BERT model and cross-modality matching can be seamlessly integrated into other RGB-cue-based CSLR methods as plug-and-play formulations to enhance the generalization ability of the visual feature extractor.**

***Index Terms*— Continuous sign language recognition, cross-modality feature matching, parallel densely-connected temporal feature, optimal transport problem.**

I. Introduction

SIGN languages are used by hearing impaired people for daily communication. The information in sign languages is conveyed by complex joint patterns of manual channels including hand shape and trajectory of hand movements, and non-manual channels such as facial expressions and head gestures [1]. Continuous sign language recognition (CSLR) has received increasing attention in the computer vision community. The task is to recognize the glosses (the text vocabularies corresponding to the sign language actions) in a sign language video, which is fundamental for other applications *e.g.*, sign language translation (SLT) [2].

Compared with the traditional methods that extract hand-craft features for CSLR, deep learning models have achieved remarkable performance improvement. Most deep learning models adopt the paradigm that the spatial-temporal visual feature is extracted, followed by a cross-modality alignment. Various visual feature extraction modules have been developed such as CNN-LSTM [5], 3D-CNN [6], fully convolutional network [7] and transformer [8]. In addition, the cross-modality alignment module such as CTC adopted in [6] look for the correspondence between video segments and glosses, which leverage the sentence-level gloss annotation as the ground truth for the alignment. This annotation is a form of weak supervision signal, as only the order but not the timing of gloss annotations in the video is known.

To learn effective visual features in CSLR, the sub-tasks *i.e.*, video representation learning, gloss representation learning, and cross-modality alignment are coupled together, which often results in model overfitting under the limited size of available benchmarks [9]. Previous methods have employed iterative fine-tuning strategies to mitigate the overfitting problem to enhance the generalization ability of the visual feature extractor [10], updating either the visual feature extractor or sequence module while keeping the other module constant. Additionally, Niu et al. proposed stochastic frame-dropping and stochastic gradient-stopping strategies to alleviate model overfitting [3]. In [4], a visual alignment constraint was devised for the same purpose.

Nevertheless, these approaches have ignored the presence of different signers executing the same gloss introducing significant visual discrepancies. These factors pose obstacles to achieving effective visual feature learning, as depicted in Figure 1. In order to address the aforementioned limitations, we present a novel approach called the gloss prior guidance network (GPGN) as a solution. GPGN aims to alleviate visual feature divergence by aligning the signer-variant visual features with corresponding signer-invariant gloss features in an end-to-end manner. This is accomplished through the use of the proposed cross-modality matching loss. Instead of utilizing one-hot vectors as gloss representations, we leverage language

Manuscript received 14 November 2023; accepted 8 May 2024. Date of publication 30 May 2024; date of current version 4 June 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62376197, Grant 62020106004, Grant 92048301, and Grant 62276141; in part by Tianjin Research Innovation Project for Postgraduate Students under Grant 2021YJSB244; and in part by the Tianjin Science and Technology Program under Grant 23JCYBJC00360. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sebastian Bosse. *(Corresponding author: Wanli Xue.)*

Leming Guo, Wanli Xue, and Tiantian Yuan are with the School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China (e-mail: xuewanli@email.tjut.edu.cn).

Bo Liu is with Walmart Global Tech, Sunnyvale, CA 94086 USA.

Kaihua Zhang is with the School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China.

Dimitris Metaxas is with the Department of Computer Science, Rutgers University, Piscataway, NJ 08854 USA.

Digital Object Identifier 10.1109/TIP.2024.3404869

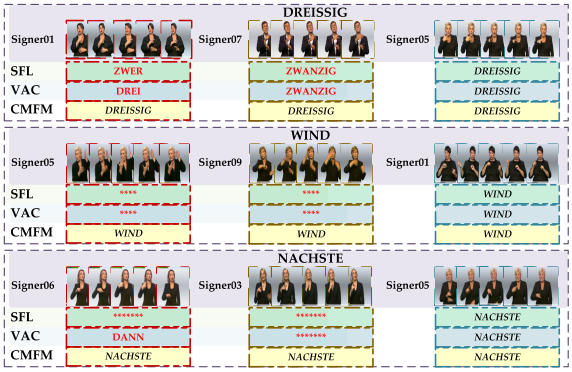

Fig. 1. Illustration of the impact of visual feature generalization on testing recognition performance, where incorrect predictions are highlighted in red. Although SFL [3] and VAC [4] can accurately recognize glosses such as "DREISSIG," "WIND," and "NACHSTE" when performed by the specific signer, they struggle to achieve accurate recognition when these glosses are performed by other signers. In contrast, the GPGN achieves correct recognition of these glosses across different signers. The GPGN overcomes this limitation by incorporating gloss as a prior, thereby learning visual features that are more generalizable.

model pre-training techniques to obtain vectorized gloss features that capture the rich sign contextual and grammatical information from training sentences. These gloss features remain fixed throughout the learning process of the CSLR model, making them invariant to different signers. To train the video feature extractor, we propose to minimize the cross-modality matching loss and the CTC loss. By means of the matching, the pre-trained gloss feature is used as prior semantic information to help learn a more generalizable video feature extractor. The proposed GPGN is distinguished by the following technical novelties:

- We adopt a pre-training approach to learn fixed gloss feature representations for enhancing visual feature extractor learning. Specifically, we fine-tune a variant of the BERT model [11] by utilizing the mask language model scheme.
- A parallel densely-connected temporal feature extraction (PDC-TFE) module is designed to extract multi-resolution visual features.
- We formulate visual and glossy feature matching into an optimal transport problem, and consequently design a cross-modality matching loss that is optimized in an end-to-end manner to facilitate the learning of more generalizable visual features.

Experimental results on several public benchmarks [2], [6], [12], [13] demonstrate the effectiveness of our method.

II. Related Work

A. Continuous Sign Language Recognition

Continuous sign language recognition (CSLR) has been studied in the vision community for decades. To learn robust visual features under sentence-level annotations, some methods have adopted connectionist temporal classication (CTC) [14] constraint with an iterative fine-tuning strategy to enhance the visual feature extractor. However, this strategy proves to be time-consuming [4]. In recent years, several works have focused on relieving the model overfitting problem in an end-to-end manner. For example, FCN [7] has designed a Gloss Feature Enhancement (GFE) module to refine the visual feature. SFL [3] has proposed a reinforcement learning operation to predict the gloss-wise labels. VAC [4] and SMKD [9] have improved the visual model generalization by optimizing the learning of the visual feature extractor and the sequential module. C^2SLR [15] have employed the keypoints heatmaps and proposed spatial and temporal constraints to concentrate on informative regions of signs. TwoStream-SLR [16] proposed a two-stream network to incorporate the videos and keypoint sequences information. TLP [17], SEN [18], and CorrNet [19] focused on squeezing temporal features. Besides, CVT-SLR [20] and CTCA [1] designed a V-L mapper, a contrastive learning based loss and a knowledge distillation based loss, respectively, to transfer gloss knowledge from a pre-trained language model to the visual model. In contrast to these above approaches, the proposed GPGN leverages the gloss as valuable prior to facilitating the learning of more generalizable visual features via optimizing a regularized optimal transport objective.

B. Visual Feature Extraction

Spatio-Temporal visual feature modeling is a fundamental aspect of research in video analysis. In TDN [21], the temporal difference between consecutive frames and segments is computed to extract spatio-temporal features. To address the issue of local semantic consistency and semantic ambiguity, a temporal semantic pyramid network is proposed in [22]. It incorporates inter-scale attention for local semantic consistency and intra-scale attention to resolve semantic ambiguity. MS-TCN++ [23] introduces a stacked single-stage temporal convolution module. This module primarily consists of a dual dilated layer (DDL) with two dilated convolutions having complementary receptive fields. In the case of DenseTCN [24], a dense connection is employed in the stacked TC layers to hierarchically capture signs (means sign language gestures in sign language video). Each TC layer learns temporal resolution information, and a fully-connection layer produces predictions, which are fused using an "argmax" operation. Unlike the above methods, we propose a parallel densely-connected temporal feature extraction (PDC-TFE) to capture the multiscale temporal patterns of glosses.

C. Vision-Text Modeling

Matching visual signals with textual semantics forms the widely used method of computer vision tasks. To address multi-label image and video classification, a method introduced in [25] constructs a label affinity graph and employs graph embedding to extract label semantic information. Furthermore, referring expression comprehension also referred to as image text matching, constitutes a prototypical problem, aiming to localize the object instance described by natural language [26]. Moreover, related studies have extensively emerged in domains such as image captioning and text-based image retrieval. Leveraging large-scale pre-training offers considerable advantages in developing generic vision-and-language models [27]. Subsequent to appropriate fine-tuning,

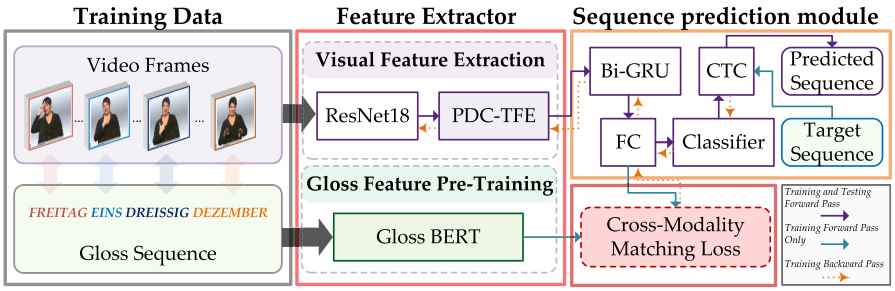

Fig. 2. The pipeline of GPGN. We first fine-tune the pre-trained BERT using sign language glosses, and then extract semantic gloss features (§Sec.III-A). Next, the spatial-temporal visual feature is extracted using the ResNet18 and the proposed PDC-TFE module (§Sec.III-B). Subsequently, the sequence information is encoded by Bi-GRU. Moreover, a linear layer is employed to map the gloss and sequence features into a common feature space to enable the cross-modality matching (§Sec.III-C). Finally, the mapped sequence features are fed into a classifier for sequence prediction.

these pre-trained models can be tailored to various downstream tasks. As a theoretically sound data matching algorithm, optimal transport has found utility in diverse applications, and has recently gained attention in computer vision research [28]. In this work, we utilize optimal transport to align sign language visual features with signer-invariant gloss features.

III. Gloss Prior Guidance Network

The framework of the proposed GPGN model is illustrated in Figure 2. In this section we elaborate in detail on the three model components: a) pre-trained gloss feature repository; b) visual feature extraction module; c) cross-modality matching loss, CTC loss, model training, and model inference.

A. Gloss Feature Pre-Training

Sign language glosses serve as a written form of sign language, they aid in how sign language works by converting video form into text form. It is worth noting that signers may exhibit significant variations when signing the same sign, similar to how speakers of a spoken language may differ in tone or pronunciation. Therefore, the gloss feature representation is ideally supposed to be signer-invariant. To address this, we propose a pre-trained gloss feature repository. Our approach involves training a gloss BERT model for each type of considered sign language *e.g.*, German sign language and Chinese sign language. This enables us to preserve the discriminative qualities of each gloss while also capturing the contextual semantic information between glosses through the masked gloss token.

Existing literature highlights the significant correlation between sign language glosses and natural language vocabulary [13]. Therefore, extracting gloss features by the natural language BERT model can ensure the discrimination between different glosses [11]. Nevertheless, the limited availability of sign language data makes it challenging to learn the contextual nuances of sign language glosses. Nevertheless, if the extracted gloss features can capture certain sign language characteristics, this would greatly benefit sign language recognition models. Therefore, we employ a fine-tuning strategy for each type of sign language under consideration (*e.g.*, German sign language or Chinese sign language) using a Masked Language Modeling (MLM) scheme [11]. This MLM approach enables our models to learn bidirectional representations of sentences, which is particularly well-suited for tasks such as CSLR that necessitate comprehensive contextual understanding.

The gloss BERT is a multi-layer bidirectional Transformer model, with the tokenized gloss sequences and the position information of each gloss as input. We denote a gloss token sequence containing L glosses by $\mathrm{y} = \{y_i\}_{i=1}^{L} \in |\mathcal{C}|$, $|\mathcal{C}|$ indicates the gloss corpus. For each token y_i, an embedding layer (Emb_T) is utilized to map gloss tokens into a feature space, then token position embedding (Emb_P) is added:

$$\mathrm{Emb}(y_i) = \mathrm{Emb}_T(y_i) + \mathrm{Emb}_P(i).$$

The embedded gloss features $f_{G_0} = \{\mathrm{Emb}(y_i)\}_{i=1}^{L}$ are fed into a 12-layer stacked Transformer Blocks $\{\mathrm{Trans}_i\}_{i=1}^{12}$ [29], that is

$$f_{G_i} = \mathrm{Trans}_i(f_{G_{i-1}}), \quad i = 1, \ldots, 12.$$

In our experimental study, the German and Chinese sign language benchmarks are initialized using pre-trained German and Chinese BERT models [30], respectively. Subsequent to initialization, the models are trained through a masked gloss token prediction task. Specifically, M ratio of tokens in a given sentence are randomly masked, and the models are tasked with predicting these tokens. Let d represent the dimension of the gloss feature. After the MLM training process, the model will output the pre-trained gloss feature repository $f_G \in R^{L\times d}$, according to the given gloss sequence with L glosses.

B. Visual Feature Extraction

The visual feature extraction of our proposed GPGN is composed of three parts: a backbone per-frame feature extraction module, a parallel densely-connected temporal feature extraction module, and a sequential predictor.

We adopt ResNet18 as the backbone per-frame feature extractor, denoted as $\mathcal{G}_{pf}$. This choice aligns with recent CSLR works [3], [9]. We denote a T-frame sign language video by $I = \{I_i\}_{i=1}^{T}$, where $I_i \in R^{3\times h\times w}$ indicates the RGB channels of the i-th frame image. Here, h and w indicate the height and width of the frame image, respectively. The per-frame feature, denoted as f_{F_i}, is obtained by $f_{F_i} = \mathcal{G}_{pf}(I_i; \theta_{pf})$, where $f_{F_i} \in R^c$ presents the output of the ResNet18 global average pooling layer; θ_{pf} is the module parameter. By aggregating all $\{f_{F_i}\}_{i=1}^{T}$, the $R^{c\times T}$ spatial feature matrix is obtained and it will be utilized for subsequent parallel densely-connected temporal feature extraction module.

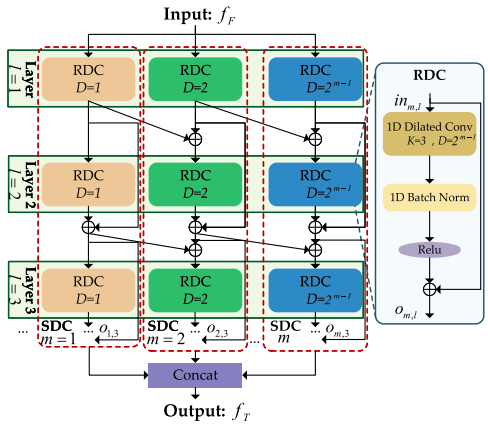

Fig. 3. Illustration of the PDC-TFE module and the RDC layer. The PDC-TFE module utilizes m parallel densely-connected SDC blocks, each with l RDC layers. The RDC layer is a residual convolution layer with the kernel size of K and dilated factor of $D = 2^{m-1}$.

1) Parallel Densely-Connected Temporal Feature Extraction Module: Sign language glosses exhibit significant variations in terms of their temporal length, as observed in benchmarks such as RWTH-2014 [12], where glosses can range from 5 to 16 frames. To effectively capture the temporal patterns inherent in these glosses, we propose a parallel densely-connected temporal feature extraction (PDC-TFE) module $\mathcal{G}_{pdc}(f_F; \theta_{pdc})$, parameterized by θ_{pdc}. Its input is the spatial feature f_F. As illustrated in Figure 3, the PDC-TFE module comprises m parallel stacked densely-connected (SDC) blocks. Each SDC block consists of l layers employing a residual dilated convolution (RDC) architecture, as depicted in Figure 3 as well. The RDC architecture involves a 1D dilated temporal convolution layer, followed by batch normalization (BN) and a ReLU activation function. Take the RDC in the m-th block and l-th layer as an example, we denote the input and output as $in_{m,l}$, $o_{m,l}$ and the operation is:

$$o_{m,l} = in_{m,l} + \text{Relu}(\text{BN}(in_{m,l} * W_{m,l} + b_{m,l})), \tag{1}$$

where $*$ denotes the dilated convolution imposed on the temporal dimension, with the dilated factor of 2^{m-1}. The convolutional kernel parameter is represented by $W_{m,l} \in R^{1\times 3}$, and the bias is denoted as $b_{m,l} \in R^1$.

As depicted in Figure 3, these m SDC blocks and l RDC layers of each block are designed to have dense connections. For example, the input of the l-th layer in the m-th block is the sum of the $l-1$ layers' output of both the m-th block and the $(m-1)$-th block, that is:

$$in_{m,l} = \sum_{i=1}^{l-1} (o_{m,i} + o_{m-1,i}).$$

To facilitate effective information flow, we introduce dense connections within the m SDC blocks and l RDC layers. The dense connection strategy has proven to be effective for image feature extraction [31], and we adopt this strategy to learn the high temporal resolution patterns of glosses by reusing the low. Finally, the output of all SDC blocks $\{o_{1,l}, o_{2,l}, \ldots, o_{m,l}\}$ is aggregated as a spatial-temporal feature $f_T \in R^{m\times c\times T}$, where the aggregation is the concatenation operation on the channel dimension.

2) Sequential Predictor: The sequential predictor $\mathcal{G}_{sp}(f_T; \theta_{sp})$ consists of two-layer Bidirectional Gated Recurrent Unit (Bi-GRU), a fully-connected layer, and a classifier. The spatial-temporal feature f_T is passed through the Bi-GRU and then the fully-connected layer to generate d-dimension vectors $f_V \in R^{d\times T}$. f_V also serves as the visual feature input for computing the cross-modality matching loss. For each $f_{V_i} \in f_V$, the classifier produces frame-level gloss prediction $p_i \in R^{|\mathcal{C}|}$ as probabilistic outputs.

C. Objective Function

The training objective function contains the CTC loss and the proposed cross-modality matching loss (CM loss).

1) Connectionist Temporal Classification (CTC): To handle the alignment between the video sequence and the target gloss sequence, we employ the CTC algorithm [14]. Given the video's per-frame categorical probability prediction $\text{p}_V = \{p_i\}_{i=1}^T$, $p_i \in R^{|\mathcal{C}|}$ and the gloss target sequence $\text{y}_V = \{y_j\}_{j=1}^L$, CTC aims to find an alignment path $\pi = \{\pi_i\}_{i=1}^T$. Where $\pi_i \in |\mathcal{C}| \cup \{blank\}$, $|\mathcal{C}|$ is the gloss corpus and $blank$ is an additional label used to indicate categories outside the vocabulary and to separate consecutive repeating glosses. The latent variables $\{\pi_i\}_{i=1}^T$ capture the alignment between the frames $\{I_i\}_{i=1}^T$ and their associated glosses. The conditional probability of an alignment path $\pi = \{\pi_i\}_{i=1}^T$ is defined as the follows:

$$\text{pr}(\boldsymbol{\pi}|\text{p}_V) = \prod_{i=1}^{T} \text{pr}(\pi_i|p_i). \tag{2}$$

The conditional probability $\text{pr}(\text{y}_V|\text{p}_V)$ is calculated by summing up the probabilities of all alignment paths π:

$$\text{pr}(\text{y}_V|\text{p}_V) = \sum_{\boldsymbol{\pi}\in\mathcal{B}^{-1}(\text{y}_V)} \text{pr}(\boldsymbol{\pi}|\text{p}_V), \tag{3}$$

where $\mathcal{B} : (|\mathcal{C}| \cup \{blank\})^T \rightarrow |\mathcal{C}|^L$; $\mathcal{B}^{-1}$ is the inverse mapping function of $\mathcal{B}$; $\mathcal{B}$ maps the alignment path $\boldsymbol{\pi}$ to its corresponding gloss sequence by removing blanks and repeated glosses. Finally, the CTC loss is defined as

$$\mathcal{L}_{ctc} = -\log p(\text{y}_V|\text{p}_V). \tag{4}$$

2) Cross-Modality Matching Loss: Visual feature learning for CSLR poses challenges due to its weakly supervised nature. To address this, we design a cross-modality matching loss (CM loss) to facilitate effective visual feature learning.

For the CM loss, we first model the gloss features $f_G = \{f_{G_i}\}_{i=1}^L$, $f_{G_i} \in R^d$ and the visual features $f_V = \{f_{V_j}\}_{j=1}^T$, $f_{V_j} \in R^d$ as two sets in a shared feature space. Subsequently, given a cost matrix $\text{C} \in R^{L\times T}$, where $\text{C}_{i,j} = 1 - \frac{f_{G_i}^\top f_{V_j}}{\|f_{G_i}\|\|f_{V_j}\|}$, the goal is to maximize the correlation between the gloss features and their corresponding visual features using a variant of the optimal transport distance $\mathcal{D}_{ot}$, that is:

$$\mathcal{D}_{ot} = \min_{\text{M}\in R_+^{L\times T}} \sum_{i=1}^{L}\sum_{j=1}^{T} \text{M}_{i,j}\text{C}_{i,j},$$
$$\text{s.t.} \quad \text{M}\mathbf{1}_T = \mathbf{1}_L, \quad \text{M}^\top\mathbf{1}_L = \mathbf{1}_T. \tag{5}$$

where $\mathbf{1}_T$ and $\mathbf{1}_L$ are T-dimension and L-dimension vectors with all entries equal to one, respectively. M can be interpreted as the transport matrix and the optimal solution of $\mathrm{M}_{i,j}$ is the optimal amount of mass required to move f_{G_i} to f_{V_j} with a minimum overall cost.

Solving (5) has $O(n^3 \log(n))$ worst case computational complexity, where n is proportional to L and T. To reduce this complexity, we finally adopt the relaxation proposed in [32], which adds an entropy-based regularization term to (5) to optimize the transport matrix by using Sinkhorn iteration [32]. As a result, it achieves an approximate solution to the original optimal transportation problem to reduce the computational complexity. The cross-modality matching loss is defined as:

$$\begin{aligned} \mathcal{L}_{cm} = & \min_{\mathrm{M} \in R_{+}^{L \times T}} \sum_{i=1}^{L} \sum_{j=1}^{T} \mathrm{M}_{i,j} \mathrm{C}_{i,j} - \beta \mathrm{H}(\mathrm{M}), \\ & \text{s.t.} \quad \mathrm{M}\mathbf{1}_T = \mathbf{1}_L, \quad \mathrm{M}^{\top}\mathbf{1}_L = \mathbf{1}_T. \end{aligned} \tag{6}$$

where $\mathrm{H}(\mathrm{M}) = \sum_{i,j} \mathrm{M}_{ij} \log \mathrm{M}_{ij}$ and β is the regularization parameter controlling the importance of the entropy term. The larger the β, the denser the final transport matrix will be. The optimal solution of M can be obtained by a variant of the Sinkhorn algorithm proposed in [32].

D. Model Training and Inference

During the training stage, the parameters of ResNet18 θ_{pf}, PDC-TFE module θ_{pdc} and Sequential predictor θ_{sp} are optimized by:

$$\arg \min_{\theta_{pf}, \theta_{pdc}, \theta_{sp}} L_{ctc} + \lambda L_{cm}, \tag{7}$$

where λ is the hyperparameter to balance the contribution of L_{ctc} and L_{cm}.

During the inference stage, given a testing sign language video, the spatial-temporal visual feature is extracted by $\mathcal{G}_{pf}$ and $\mathcal{G}_{pdc}$ modules. Subsequently, the sequence predictor $\mathcal{G}_{sp}$ generates per-frame category probability predictions. To obtain the gloss sequence prediction, a CTC beam search algorithm is employed, with the beam width set to 10. This approach efficiently explores multiple potential gloss sequences and selects the most likely prediction based on the category probabilities.

IV. Experiments

A. Benchmarks and Evaluation Metric

1) RWTH-2014 [12]: It is a German CSLR benchmark divided into a training set with 5,672 videos, a development set with 540 videos, and a test set with 629 videos. Across all videos, a total of 1,295 words are signed by 9 different signers.

2) RWTH-2014T [2]: It encompasses a vocabulary of 1,085 words, and comprises 7,096, 519, and 642 videos in the training set, development set, and test set, respectively.

3) CSL-500 [6]: It is a Chinese CSLR benchmark containing 100 sentences and a vocabulary of 178 words. It is split into a training set with 4,700 videos and a test set with 300 videos.

4) CSL-Daily [13]: It is a large-scale Chinese sign language benchmark. It is composed of 18,401 videos for the training set, 1,077 videos for the development set, and 1,176 videos for the test set. And it has 2,000 words for the CSLR task.

5) Evaluation Metric: We adopt the word error rate (WER) metric that is widely used for CSLR evaluation [1]. The WER measures the minimum number of substitutions, deletions, and insertions needed to convert one predicted sentence to an associated reference sentence:

$$WER = \frac{N_I + N_D + N_S}{L}, \tag{8}$$

where N_I, N_D, N_S are the number of operations for insertions, deletions, and substitutions, respectively.

B. Experimental Setup

1) Gloss BERT: For all benchmarks, we employ the last hidden state of pre-trained BERT as the gloss feature f_G, dimension d is 768, mask ratio M of the MLM scheme is set to 0.5.

2) Visual Feature Extraction: In the PDC-TFE module, we set $m = 3$ (except $m = 2$ for the CSL-500), $l = 3$, and $c = 512$. For the sequence predictor, the dimension of Bi-GRU is 1024, and the dimension of f_V matches that of f_G.

3) Data Augmentation: For all benchmarks, half of all video frames are randomly discarded like SFL [3]. The remaining frames are resized to 256×256 and then randomly cropped to 224×224 with random horizontal flipping employed (center cropping 224×224 during the inference stage).

4) Training: GPGN was trained using the Adam optimizer for 80 epochs, with weight decay and learning rate set to $1e$-4 for two RWTH benchmarks and the CSL-500 benchmark, while they were set to $1e$-6 and $5e$-5 for the CSL-Daily benchmark, respectively. Through experiments, λ and β are set to 2 and 0.3, respectively for all benchmarks. The experiments are implemented using PyTorch and conducted on an A100 GPU.

C. Ablation Study

The GPGN consists of two key components: the cross-modality matching (CM) strategy and the PDC-TFE module. We define the combination of the pre-trained gloss BERT model and the CM loss as the CM strategy.

"Baseline" refers to ResNet18 is equipped with a three-layer 1D-temporal convolution network. This baseline is similar to state-of-the-art baseline [4], and it is optimized by CTC loss. As shown in Table I, baseline achieves performance of 24.0% and 24.5% on both the dev and test sets.

1) Effect of PDC-TFE: "R+PDC-TFE" indicates ResNet18 is equipped with PDC-TFE module. In Table I, since the PDC-TFE effectively captures the rich temporal information inherent in glosses of various lengths, it achieves improvements of 2.3% and 2.9% compared to the baseline.

2) Effect of Different Values of m and l of PDC-TFE Module: Large m and l represent large temporal resolutions, in addition, large m also means more diverse temporal resolutions. In Table II, experiments are based on "R+PDC-TFE".

TABLE I
ABLATION EXPERIMENTS (%) OF THE GPGN ON THE RWTH-2014, "B" AND "R" INDICATES THE BASELINE AND RESNET18, RESPECTIVELY

Variants	CM	PDC-TFE	Dev%	Test%
Baseline	-	-	24.0	24.5
R+PDC-TFE	-	✓	21.7 (↓ 2.3)	21.6 (↓ 2.9)
B+CM	✓	-	22.8 (↓ 1.2)	23.0 (↓ 1.5)
GPGN	✓	✓	**19.9** (↓ 4.1)	**20.4** (↓ 4.1)

TABLE II
EFFECT (%) OF DIFFERENT VALUES OF m AND l OF THE PDC-TFE MODULE. EXPERIMENTS (%) ARE BASED ON THE RWTH-2014 TEST SET

Modules	l=1	l=2	l=3	l=4
PDC-TFE (m=1)	23.4	23.0	23.1	22.8
PDC-TFE (m=2)	22.8	22.7	22.7	22.6
PDC-TFE (m=3)	22.6	21.9	**21.6**	22.0
PDC-TFE (m=4)	21.9	22.4	22.7	22.8

We observe that setting $m = 3$, $l = 3$ is optimal, but larger m and l lead to performance decrease. We consider that a large value of m or l brings many parameters, leading to overfitting. Additionally, a too-large temporal resolution may capture collections of several short signs, resulting in temporal boundary ambiguity.

3) Effect of CM Strategy: "B+CM" denotes that the cross-modality matching (CM) strategy is applied to the baseline. Table I illustrates "B+CM" outperforming the baseline. Further, we plug the CM strategy into other RGB-cue-based state-of-the-art methods, such as VAC, TLP, SEN, CorrNet, and CTCA. As shown in Table III, all methods gain significant improvements through the use of the CM strategy, which convincingly shows its effectiveness and transferability. For the above results, we consider that gloss features extracted by the fine-tuned gloss BERT model exhibit promising signer-invariant characteristics and carry rich gloss correlation semantics. Furthermore, with the effect of the cross-modality matching loss, the gloss features serve as a reference to promote semantic correlation learning among visual features. As a result, visual features will be clustered with their corresponding gloss features. This clustering approach can effectively reduce the intra-class divergence of visual features.

4) Effect of PDC-TFE Module & CM Strategy: As shown in Table I, GPGN combining the PDC-TFE module and CM strategy in the baseline leads to remarkable performance gains, surpassing "Baseline", "B+CM" and "R+PDC-TFE" on both the dev and test sets. These improvements can be attributed to the utilization of PDC-TFE in capturing fine differences among various temporal glosses within a video, which further enhances the performance of the CM strategy.

5) Effect of Different Values of λ and β: We have done extensive experiments and concluded that GPGN performs optimally when λ=2 and β=0.3. If the CM loss is assigned a high λ, the resulting gradient feedback may weaken the CTC loss and drive the model optimization in an undesirable direction. A larger value of β generates a denser transport matrix in (6). In Figure 4, we can observe that the GPGN performance is less dependent on the choice of hyperparameters, thus indicating the stability of the CM loss.

TABLE III
EFFECT (%) OF THE CROSS-MODALITY MATCHING STRATEGY ("CM") PLUGGED INTO RECENT STATE-OF-THE-ART RGB-CUE-BASED METHODS ON THE RWTH-2014. "*" DENOTES EXPERIMENT RESULTS ACCOMPLISHED BY OUR IMPLEMENTATION

Methods	WER	
	Dev%	Test%
VAC [4] (ICCV 2021)	21.2	22.3
VAC+CM*	20.3 (↓ 0.9)	21.1 (↓ 1.2)
TLP [17] (ECCV 2022)	19.7	20.8
TLP+CM*	19.3 (↓ 0.4)	20.2 (↓ 0.6)
SEN [18] (AAAI 2023)	19.5	21.0
SEN+CM*	19.2 (↓ 0.3)	20.2 (↓ 0.8)
CorrNet* [19] (CVPR 2023)	19.0	19.7
CorrNet*+CM	18.7 (↓ 0.3)	19.4 (↓ 0.3)
CTCA [1] (CVPR 2023)	19.5	20.1
CTCA+CM*	19.3 (↓ 0.2)	19.8 (↓ 0.3)

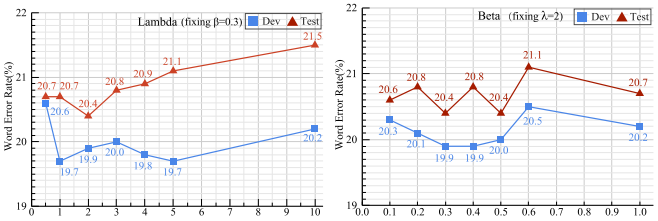

Fig. 4. Experimental results of different λ and β on the RWTH-2014. When changing λ, β=0.3 will be fixed and vice versa λ=2 will be fixed.

D. Comparison With Baseline Methods

1) Evaluation on RWTH-2014: Table IV presents a comparison between our GPGN and several baseline methods on the RWTH-2014. Only employing the RGB cue, our GPGN achieves the competitive performance (19.9% and 20.4%) on both the dev and test sets. FCN, VAC, and SMKD neglect the gloss prior guidance and the capturing of glosses with various temporal patterns, as a result, their performances are significantly inferior to our GPGN. In addition, performances of TLP, SEN, and CorrNet demonstrate the importance of capturing glosses with various temporal patterns, and CTCA and CVT-SLR have validated employing the gloss prior guidance plays significant enhancement roles for the CSLR model.

2) Evaluation on RWTH-2014T: Table V shows the comparison between our GPGN and baseline methods on the RWTH-2014T. V-L Mapper lacks the ability to capture glosses with complex temporal patterns, C^2SLR, and TLP neglects the gloss prior guidance. Our GPGN achieves remarkable performance, outperforming the V-L Mapper and TLP by 2.0% and 0.7%, on the test set, respectively. Furthermore, on the dev set, GPGN outperforms C^2SLR by 0.9%.

TABLE IV
COMPARISON (%) WITH BASELINE METHODS ON THE RWTH-2014. THE ENTRIES DENOTED BY "∗" USED EXTRA CUES (KEYPOINTS OR REGIONS OF HAND AND FACE)

Methods	Dev%		Test%	
	del/ins	WER	del/ins	WER
FCN (2020) [7]	-	23.7	-	23.9
DNF+SBD-RL (2020) [33]	10.5/3.3	23.4	10.6/2.8	23.5
VAC (2021) [4]	7.9/2.5	21.2	8.4/2.6	22.3
SMKD (2020) [9]	6.8/2.5	20.8	6.3/2.3	21.0
SEN (2023) [18]	5.8/2.6	19.5	7.3/4.0	21.0
TLP (2022) [17]	6.3/2.8	19.7	6.1/2.9	20.8
STMC* (2021) [34]	7.7/3.4	21.1	7.4/2.6	20.7
C^2SLR* (2022) [15]	-	20.5	-	20.4
CVT-SLR (2023) [20]	6.4/2.6	19.8	6.1/2.3	20.1
CTCA (2023) [1]	6.2/2.9	19.5	6.1/2.6	20.1
CorrNet (2023) [19]	5.6/2.8	18.8	5.7/2.3	19.4
TwoStream-SLR* (2022) [16]	-	**18.4**	-	**18.8**
GPGN (2022) (Ours)	5.8/3.6	19.9	6.3/2.8	20.4

TABLE V
COMPARISON (%) WITH BASELINE METHODS ON THE RWTH-2014T. THE ENTRIES DENOTED BY "∗" USED EXTRA CUES (KEYPOINTS)

Methods	WER	
	Dev%	Test%
SFL (2020) [3]	25.1	26.1
SLT (Gloss+Text) (2020) [8]	24.6	24.5
BN-TIN+Transf (2021) [13]	22.7	23.9
V-L Mapper (2022) [35]	21.9	22.5
SMKD [9] (2021)	20.8	22.4
TLP (2022) [17]	19.4	21.2
SEN (2023) [18]	19.3	20.7
CorrNet (2023) [19]	18.9	20.5
C^2SLR* (2022) [15]	20.2	20.4
CVT-SLR (2023) [20]	19.4	20.3
CTCA (2023) [1]	19.3	20.3
TwoStream-SLR* (2022) [16]	**17.7**	**19.3**
GPGN (2022) (Ours)	19.3	20.5

3) Evaluation on CSL-500: Table VI delivers the comparison between our GPGN and baseline methods on the CSL-500. S2VT, LS-HAN, LS-HAN, GEU, and SBD-RL lack capturing glosses with various temporal patterns. CTM, DenseTCN, and SBD-RL focus on gloss temporal boundary detection, they have no gloss prior guidance. Although GEU is SOTA in Split I, our GPGN brings obvious gains by GEU of 21.5% in the more challenging Split II. Furthermore, although SBD-RL surpasses GPGN by 1.6% on Split II, its enhanced variant's performance [33] on RWTH-2014 (23.4%/23.5%) is remarkably inferior to our GPGN by (3.5%/3.1%).

4) Evaluation on CSL-Daily: Table VII illustrates the results of GPGN and other baseline methods. Both BN-TIN+Transf, TIN-Iterative, and FCN utilize a single-temporal resolution structure to capture signs, they are inferior to our GPGN, which achieves effective WERs of 31.1% and 30.0% on both the dev set and test sets.

TABLE VI
COMPARISON (%) WITH BASELINE METHODS ON THE CSL-500 SPLIT I AND SPLIT II. THE ENTRIES DENOTED BY "∗" USED EXTRA CUES (KEYPOINTS OR REGIONS OF HAND AND FACE OR DEPTH INFORMATION)

Methods	WER	
	Split I	Split II
S2VT (2015) [36]	25.5	67.0
LS-HAN (2018) [6]	17.3	-
HLSTM-atten (2018) [37]	10.2	64.1
CTM (2019) [38]	-	61.9
DenseTCN (2019) [24]	14.3	44.7
STMC* (2020) [39]	2.1	-
FCN (2020) [7]	3.0	-
VAC (2021) [4]	1.6	-
C^2SLR* (2022) [15]	0.9	-
SEN (2023) [18]	0.8	-
CorrNet (2023) [19]	0.8	-
GEU* (2021) [40]	**0.6**	49.9
SBD-RL (2020) [33]	-	**26.8**
GPGN (2022) (Ours)	0.9	28.4

TABLE VII
COMPARISON (%) WITH BASELINE METHODS ON THE CSL-DAILY. THE ENTRIES DENOTED BY "∗" USED EXTRA CUES (KEYPOINTS)

Methods	Dev%		Test%	
	del/ins	WER	del/ins	WER
BN-TIN+Transf (2021) [13]	13.9/3.4	33.6	13.5/3.0	33.1
TIN-Iterative (2019) [41]	12.8/3.3	32.8	12.5/2.7	32.4
SLT(Gloss+Text) (2020) [8]	10.3/4.4	33.1	9.6/4.1	32.0
SEN (2023) [18]	-	31.1	-	30.7
CorrNet (2023) [19]	-	30.6	-	30.1
CTCA (2023) [1]	9.2/2.5	31.3	8.1/2.3	29.4
TwoStream-SLR* (2022) [16]	-	**25.4**	-	**25.3**
GPGN (Ours)	9.6/2.5	31.1	9.6/2.1	30.0

TABLE VIII
EVALUATION (%) OF DISTINCT BERT MODEL STRATEGIES ON THE RWTH-2014 AND CSL-DAILY

Methods	RWTH-2014		CSL-Daily	
	Dev%	Test%	Dev%	Test%
Embedding	21.5	21.5	-	-
BERT	20.9	21.2	32.1	30.8
BERT fine-tuning	**19.9**	**20.4**	**31.1**	**30.0**

E. Module Analysis

1) Comparison With Different Gloss Feature Extractors: As shown in Table VIII, "Embedding" presents an embedding method consisting of an embedding layer, an LSTM layer, and an MLP layer. "BERT" indicates the BERT model as the initialization. While "BERT fine-tuning" denotes that the BERT model has been fine-tuned on sign language benchmarks as the initialization. The performance of "BERT" outperforms

TABLE IX
EVALUATION (%) OF DIFFERENT MASK RATIOS M FOR THE MLM SCHEME OF THE FINE-TUNED BERT ON THE RWTH-2014

M	0.15	0.3	0.5	0.6	0.8	0.9
Dev/Test	20.4/20.7	20.1/20.4	19.9/20.4	20.3/20.7	20.7/20.8	21.2/21.4

TABLE X
COMPARISON (%) WITH DIFFERENT TEMPORAL FEATURE EXTRACTORS AND CM LOSSES ON THE RWTH-2014. * DENOTES THAT EXPERIMENTS ARE ACCOMPLISHED BY OUR IMPLEMENTATION

Methods	Dev%		Test%	
	del/ins	WER	del/ins	WER
FCN [7]	-	-	-	26.0
Dilated* [42]	7.6/3.0	23.6	7.4/3.2	24.8
MS-TCN* [43]	7.8/2.7	24.4	7.5/2.6	24.6
DDL* [23]	8.6/3.2	23.3	8.1/2.6	23.2
HRDC	7.7/3.4	21.9	7.5/2.8	22.3
DenseTCN [24]	10.7/5.1	35.9	10.5/5.5	36.5
3D+PDC-TFE*	12.9/4.2	33.3	12.6/4.1	32.9
R+PDC-TFE	7.3/3.4	21.7	6.7/2.8	21.6
GPGN (soft-DTW)	7.1/2.5	21.3	7.2/2.5	21.8
GPGN (CM)	5.8/3.6	**19.9**	6.3/2.8	**20.4**

the embedding method. Because the BERT model can extract more discriminative gloss features. Furthermore, employing the fine-tuned BERT model achieves the best performance on both RWTH-2014 and CSL-Daily. We attribute this to the MLM scheme, which allows the BERT model to capture contextual semantics and subtle sign grammatical information from sign language sentences, although it may be overfitting.

2) Comparison With Different Mask Ratios M for the Fine-Tuned BERT: As shown in Table IX, setting the mask ratio M to 0.5 yields optimal performance. when M is set to 0.9, the performance of GPGN is near the "Embedding" in Table VIII. This result demonstrates that a suit larger mask rate will force BERT to provide enough contexts to learn good representations, but the BERT is unable to learn contextual knowledge among glosses under a very large mask ratio.

3) Comparison With Different Temporal Feature Extractors: In Table X, we replaced the backbone of Dilated, MS-TCN, and DDL with ResNet18. We also replaced the temporal convolution layers with the PDC-TFE module in DenseTCN to construct the "3D+PDC-TFE". "R+PDC-TFE" outperforms FCN, Dilated, MS-TCN, and DDL, because the first three methods capture single-temporal resolution, and DDL captures two-scale temporal resolution. Additionally, "3D+PDC-TFE" achieves better results than the multi-temporal resolution extractor, DenseTCN. We attribute this result to the multi-temporal resolution parallel learning of PDC-TFE. To verify the dense-connection strategy's effectiveness, we remove the dense-connection from the PDC-TFE module that we refer to as "HRDC", and its performance decreases.

4) Comparison With Different Cross-Modality Matching Methods: In Table X, "GPGN (soft-DTW)" indicates the replacement of CM loss with soft Dynamic Time Warping [44]. The performance of CM loss outperforms the soft-DTW. This can be attributed to the limitations of soft-DTW, such as erroneous alignments for periodic sequences with different starting points [45] and the tendency to match noisy frames due to its strict order constraint. In contrast, the CM loss can avoid these cases due to the optimal transport, achieving more flexibility with increased interpretability.

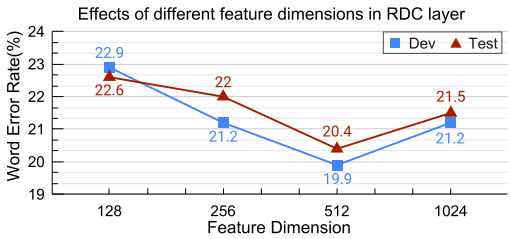

Fig. 5. Effects (%) of different feature dimension c in the RDC layer on the RWTH-2014.

Fig. 6. Qualitative results of GPGN and its variants. An example from the test set of the RWTH-2014 is presented. The wrongly recognized glosses are marked in red/cyan/orange words that denote the deletion(D)/substitution(S)/insertion(I) errors, respectively. "*" denotes the model cannot identify the gloss.

5) Comparison With Different Feature Dimensions of RDC Layer: In Figure 5, experiments based on GPGN, setting $c = 512$ achieves the optimal results. When c exceeds 512, overfitting occurs, causing a performance decline, but it still performs better than cases with smaller c. This observation indicates that balanced high-dimensional features are more effective in capturing temporal information.

F. Qualitative Results

To qualitatively evaluate the GPGN, Figure 6 visualizes an example with video frames with their predicted glosses. We can observe that when the baseline incorporates the PDC-TFE module, the gloss "GRAUPEL" can be correctly recognized. Furthermore, with the help of the CM strategy, the model corrects the recognition errors for "ABEND" and "GRAUPEL" made by the baseline. Remarkably, when employing GPGN, only a single substitution recognition error (ANFANG) occurs. These results highlight that the GPGN performs favorably, particularly in handling challenging signs.

To assess the generalizability of visual features, Figure 7 visualizes the class activation maps for the 7×7 feature maps from the ResNet18. Comparing GPGN with VAC and SFL, it is evident that GPGN effectively focuses on learning crucial sign-related information, such as the hand, arm, and head (mouth shape). This result demonstrates the sign sensitivity of GPGN and its ability to capture nuanced distinguishing signs.

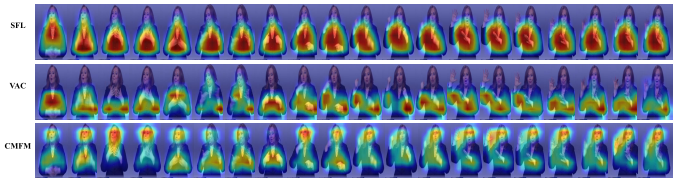

Fig. 7. Visualizing the CAMs results generated by the GPGN and its counterparts SFL and VAC with an example from the test set of the RWTH-2014. The colors of images change from blue to yellow and to red, meaning the model pays more attention to the positions (e.g. head, hand, and arm).

V. Conclusion and Discussion

In this paper, we present a novel gloss prior guidance network (GPGN) aimed at enhancing the generalizability of the visual feature extractor in CSLR model learning. Our approach capitalizes on the gloss prior information to improve the generalizability of the visual features. To achieve this, we design a pre-trained gloss BERT model to extract the gloss feature that is signer-invariant and encompasses rich contextual gloss sequences and sign grammatical information. Furthermore, we devise a PDC-TFE module that collaborates with a ResNet18 backbone for multi-resolution spatial-temporal visual feature extraction. Moreover, we propose a cross-modality matching loss that formulates the alignment between gloss and visual features as a regularized optimal transport problem. We refer to the combination of the gloss BERT and this matching loss as the cross-modality matching strategy. Extensive experiments on RWTH-2014, RWTH-2014T, CSL-500, and CSL-Daily have demonstrated the competitive performance of our GPGN and the transferability of the cross-modality matching strategy.

References

[1] L. Guo et al., "Distilling cross-temporal contexts for continuous sign language recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10771–10780.

[2] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7784–7793.

[3] Z. Niu and B. Mak, "Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition," in *Proc. ECCV*, 2020, pp. 1–19.

[4] Y. Min, A. Hao, X. Chai, and X. Chen, "Visual alignment constraint for continuous sign language recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11522–11531.

[5] O. Koller, N. C. Camgoz, H. Ney, and R. Bowden, "Weakly supervised learning with multi-stream CNN-LSTM-HMMS to discover sequential parallelism in sign language videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2306–2320, Apr. 2019.

[6] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in *Proc. AAAI*, 2018, pp. 1–11.

[7] K. L. Cheng, Z. Yang, Q. Chen, and Y.-W. Tai, "Fully convolutional networks for continuous sign language recognition," in *Proc. ECCV*, 2020, pp. 1–19.

[8] N. Cihan Camgöz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10020–10030.

[9] A. Hao, Y. Min, and X. Chen, "Self-mutual distillation learning for continuous sign language recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11283–11292.

[10] J. Pu, W. Zhou, and H. Li, "Iterative alignment network for continuous sign language recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4160–4169.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019, *arXiv:1810.04805*.

[12] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Comput. Vis. Image Understand.*, vol. 141, pp. 108–125, Dec. 2015.

[13] H. Zhou, W. Zhou, W. Qi, J. Pu, and H. Li, "Improving sign language translation with monolingual data by sign back-translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1316–1325.

[14] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.

[15] R. Zuo and B. Mak, "C2SLR: Consistency-enhanced continuous sign language recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5121–5130.

[16] Y. Chen, R. Zuo, F. Wei, Y. Wu, S. Liu, and B. Mak, "Two-stream network for sign language recognition and translation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 1–6.

[17] Z. L. L. Hu, L. Gao, and W. Feng, "Temporal lift pooling for continuous sign language recognition," in *Proc. ECCV*, 2022, pp. 1–14.

[18] L. Hu, L. Gao, Z. Liu, and W. Feng, "Self-emphasizing network for continuous sign language recognition," in *Proc. AAAI*, 2023, pp. 1–13.

[19] L. Hu, L. Gao, Z. Liu, and W. Feng, "Continuous sign language recognition with correlation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1–18.

[20] J. Zheng et al., "CVT-SLR: Contrastive visual-textual transformation for sign language recognition with variational alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23141–23150.

[21] L. Wang, Z. Tong, B. Ji, and G. Wu, "TDN: Temporal difference networks for efficient action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1895–1904.

[22] D. Li et al., "TSPNet: Hierarchical feature learning via temporal semantic pyramid for sign language translation," in *Proc. NIPS*, 2020, pp. 1–19.

[23] S.-J. Li, Y. A. Farha, Y. Liu, M.-M. Cheng, and J. Gall, "MS-TCN: Multi-stage temporal convolutional network for action segmentation," in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 3570–3579, Jun. 2019.

[24] D. Guo, S. Wang, Q. Tian, and M. Wang, "Dense temporal convolution network for sign language translation," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 1–15.

[25] R. You, Z. Guo, L. Cui, X. Long, Y. Bao, and S. Wen, "Cross-modality attention with semantic graph embedding for multi-label classification," in *Proc. AAAI*, 2020, pp. 1–11.

[26] Y. Liao et al., "A real-time cross-modality correlation filtering method for referring expression comprehension," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10877–10886.

[27] Y.-C. Chen et al., "Uniter: Universal image-text representation learning," in *Proc. ECCV*, 2020, pp. 1–17.

[28] S. Yuan et al., "Advancing weakly supervised cross-domain alignment with optimal transport," in *Proc. BMVC*, 2020, pp. 1–9.

[29] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.

[30] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, "Revisiting pre-trained models for Chinese natural language processing," in *Proc. Findings Assoc. Comput.*, 2020, pp. 1–19.

[31] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[32] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1–26.

[33] C. Wei, J. Zhao, W. Zhou, and H. Li, "Semantic boundary detection with reinforcement learning for continuous sign language recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 1138–1149, Mar. 2021.

[34] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial-temporal multi-cue network for sign language recognition and translation," *IEEE Trans. Multimedia*, vol. 24, pp. 768–779, 2022.

[35] Y. Chen, F. Wei, X. Sun, Z. Wu, and S. Lin, "A simple multi-modality transfer learning baseline for sign language translation," 2022, *arXiv:2203.04287*.

[36] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence - video to text," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4534–4542.
[37] D. Guo, W. G. Zhou, H. Li, and M. Wang, "Hierarchical lstm for sign language translation," in *Proc. AAAI*, 2018, pp. 1–18.
[38] D. Guo, S. Tang, and M. Wang, "Connectionist temporal modeling of video and language: A joint model for translation and sign labeling," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 1–17.
[39] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial-temporal multi-cue network for continuous sign language recognition," in *Proc. AAAI*, 2020, pp. 1–13.
[40] S. Tang, D. Guo, R. Hong, and M. Wang, "Graph-based multimodal sequential embedding for sign language translation," *IEEE Trans. Multimedia*, vol. 24, pp. 4433–4445, 2022.
[41] R. Cui, H. Liu, and C. Zhang, "A deep neural framework for continuous sign language recognition by iterative training," *IEEE Trans. Multimedia*, vol. 21, no. 7, pp. 1880–1891, Jul. 2019.
[42] J. Pu, W. Zhou, and H. Li, "Dilated convolutional network with iterative optimization for continuous sign language recognition," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 12–19.
[43] Y. A. Farha and J. Gall, "MS-TCN: Multi-stage temporal convolutional network for action segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3575–3584.
[44] M. Cuturi and M. Blondel, "Soft-DTW: A differentiable loss function for time-series," in *Proc. ICML*, 2017, pp. 1–14.
[45] B. Su and G. Hua, "Order-preserving optimal transport for distances between sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 2961–2974, Dec. 2019.

Leming Guo is currently pursuing the Ph.D. degree in computer science and technology with Tianjin University of Technology. His research interests include computer vision and multimedia analysis.

Wanli Xue (Member, IEEE) received the Ph.D. degree in the technology of computer application from Tianjin University in 2019. He is currently an Associate Professor with the School of Computer Science and Engineering, Tianjin University of Technology. His research interests include visual tracking, image stitching, and sign language recognition.

Bo Liu received the Ph.D. degree in computer science from Rutgers, The State University of New Jersey, in 2018. He currently holds the position of Staff Data Scientist at Walmart Global Tech. Prior to this role, he served as a Senior Research Scientist at JD.com Silicon Valley Research Center. His areas of expertise and research interests encompass machine learning, computer vision, and data science.

Kaihua Zhang (Member, IEEE) received the Ph.D. degree from the Department of Computing, The Hong Kong Polytechnic University, in 2013. From 2009 to 2010, he was a Research Assistant with the Department of Computing, The Hong Kong Polytechnic University. He is currently a Professor with the School of Computer and Software, Nanjing University of Information Science and Technology. His research interests include image segmentation, level sets, and visual tracking.

Tiantian Yuan received the B.S. degree in education technology from Tianjin Normal University in 2006 and the Ph.D. degree in technology of computer application from Nankai University in 2012. She is currently a Professor with the Technical College for the Deaf, Tianjin University of Technology, China. Her research interests include computer vision, computer education for the deaf, and the IoT.

Dimitris Metaxas (Fellow, IEEE) received the B.E. degree from the National Technical University of Athens, Greece, in 1986, the M.S. degree from the University of Maryland in 1988, and the Ph.D. degree from the University of Toronto in 1992. He is currently a Professor with the Computer Science Department, Rutgers University. He is also directing the Computational Biomedicine Imaging and Modeling Center (CBIM). He has been conducting research toward the development of formal methods upon, which computer vision, computer graphics, and medical imaging can advance synergistically.