This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Distilling Cross-Temporal Contexts for Continuous Sign Language Recognition

Leming Guo[1] Wanli Xue[1*] Qing Guo[2*] Bo Liu[3] Kaihua Zhang[4]
Tiantian Yuan[5] Shengyong Chen[1]
[1] School of Computer Science and Engineering, Tianjin University of Technology,
[2] Centre for Frontier AI Research (CFAR), A*STAR, Singapore, [3] Walmart Global Tech, Sunnyvale, CA, USA,
[4] School of Computer and Software, Nanjing University of Information Science and Technology,
[5] Technical College for the Deaf, Tianjin University of Technology

Abstract

Continuous sign language recognition (CSLR) aims to recognize glosses in a sign language video. State-of-the-art methods typically have two modules, a spatial perception module and a temporal aggregation module, which are jointly learned end-to-end. Existing results in [9, 20, 25, 36] have indicated that, as the frontal component of the overall model, the spatial perception module used for spatial feature extraction tends to be insufficiently trained. In this paper, we first conduct empirical studies and show that a shallow temporal aggregation module allows more thorough training of the spatial perception module. However, a shallow temporal aggregation module cannot well capture both local and global temporal context information in sign language. To address this dilemma, we propose a cross-temporal context aggregation (CTCA) *model. Specifically, we build a dual-path network that contains two branches for perceptions of local temporal context and global temporal context. We further design a cross-context knowledge distillation learning objective to aggregate the two types of context and the linguistic prior. The knowledge distillation enables the resultant one-branch temporal aggregation module to perceive local-global temporal and semantic context. This shallow temporal perception module structure facilitates spatial perception module learning. Extensive experiments on challenging CSLR benchmarks demonstrate that our method outperforms all state-of-the-art methods.*

1. Introduction

Sign language is a visual language for deaf and hearing-impaired people for ease of communication. Because sign language has a different grammatical structure and expression from natural spoken language, deaf and hearing-impaired people hardly communicate normally with hearing people in daily life. To eliminate this communication gap, continuous sign language recognition (CSLR) enforces to recognition of various glosses from a sign language video. Due to the data collection and annotation being labor-intensive, CSLR benchmarks adopt a sentence-annotation manner for all sign language videos [1, 15, 33].

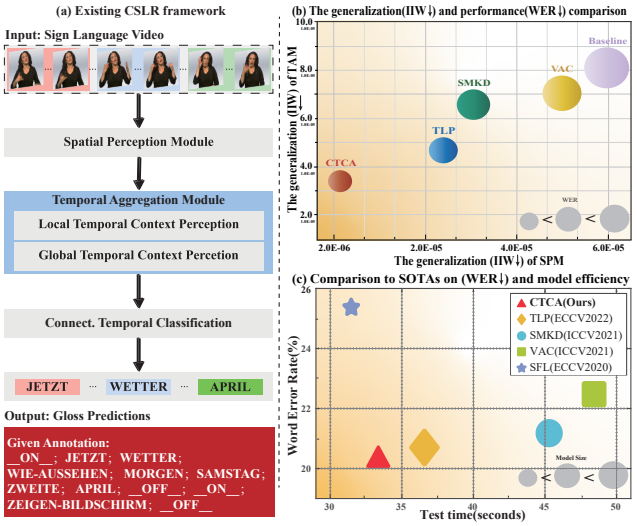

Figure 1. (a) is the common CSLR framework. (b) presents the generalization capability (*i.e.*, information stored in weights (IIW) [31]) of TAMs and SPMs of baseline, state-of-the-art methods and the proposed method. (c) is the performance of the state-of-the-art methods including word error rate, test time, and model size.

In recent years, there is a consensus among state-of-the-art methods on a baseline framework (See Fig. 1 (a)). It is made up of a spatial perception module (SPM), and a temporal aggregation module (TAM) including two components for local and global temporal perception module, and the connectionist temporal classification (CTC) loss [8] for training. At present, these methods [9, 20, 25, 36] have perceived one limitation of this framework that the temporal aggregation module can lead to insufficiently trained spatial

*Wanli Xue and Qing Guo are corresponding authors (xuewanli@email.tjut.edu.cn and tsingqguo@ieee.org).

perception module and affect the final accuracy. We use a recent interpretation method (*i.e*., the compression of information stored in weights (IIW) [31]) to measure the generalization capability of different neural modules. Fig. 1 (b) shows the IIWs of TAMs and SPMs of baseline and state-of-the-art works and infers their positive relation, *i.e*., low-generalization TAM (*i.e*., high IIW) usually leads to low-generalization SPM. We provide more studies in Sec. 3.

Y. Min *et al*. [20] measure the difference of correctly and incorrectly recognized results between auxiliary and primary classifiers to evaluate model overfitting, and A. Hao *et al*. [9] visualized heatmaps of TAM's self-similarity matrices to show what the local and global temporal perception learning. However, there are no straightforward quantitative studies to discuss the effects of TAM on SPM, and we do not know how significant the effects could be and have no idea about the desired temporal aggregation. In this paper, we extensively study the limitation and desirable properties of the temporal aggregation module in the CSLR framework via constructing a baseline framework and extensive empirical studies. We insight that a desired temporal aggregation module should be a shallow architecture to allow more effective training of spatial perception module but also should be a deep one for a high temporal aggregation capability. Whereas, it is quite challenging for the temporal aggregation module to achieve these properties simultaneously.

To overcome this challenge, we propose the cross-temporal context aggregation (CTCA) that a shallow temporal aggregation module has capable of incorporating local-global temporal contexts and the linguistic prior. Specifically, we construct a dual-path network, which decouples the local and global perception modules and imposes a linguistic module in parallel. This architecture ensures the local context perception, global context perception, and linguistic prior extraction. Furthermore, we propose a cross-context knowledge distillation loss function to transfer the local temporal context and the linguistic prior to the global perception module. Notice that the spatial perception module can facilitate itself by receiving cross-context knowledge as supervision during distillation. Fig. 1(b) shows that both SPM and TAM in CTCA achieve higher generalization than the ones of baselines. Consequently, Fig. 1(c) delivers CTCA's superiority and it outperforms the state-of-the-art methods on WER, test time, and model size.

2. Related Work

Continuous sign language recognition. To learn stronger representation under the sentence-annotated benchmarks, current deep learning-based approaches exploit the Connectionist Temporal Classification (CTC) [8], which provides a many-to-one mapping between frames and glosses. To mitigate the insufficient training problem [9, 20, 25, 36], a time-consuming iterative fine-tuning strategy [6, 25] is utilized to boost the model generalization. In contrast, the end-to-end methods proposed, such as VAC [20] and SMKD [9] proposed an end-to-end alignment constraint to make the learning of local and global temporal context to be consistent. C^2SLR [36] designed two constraints for building spatial consistency and temporal perception consistency to enhance the representation power of the model. Y. Min *et al*. [21] proposed an optimized CTC loss, which can constrain features on a hypersphere and control the peak behavior of CTC loss to enhance the feature generalization. In this paper, we achieve a shallower temporal perception module that not only promotes more thorough training of spatial perception model but also aggregates local-global temporal context via a cross-temporal context distillation.

Knowledge distillation. Knowledge distillation (KD) is a model compression technique that facilitates student models to achieve strong performance by excavating knowledge from large teacher models [11] or itself [16]. To eliminates huge training cost needs the ONE [16] proposes a multi-branch network to conduct an ensemble on-the-fly. VAC [20] conducts its local and global temporal context module as an implicit teacher to promote each one to gain advanced generalization. Wang *et al*. [30] propose that consistent knowledge of multi-modal contains in each channel, and can achieve multi-modal fusion by exchanging their channels. POS-SCAN [35], LGD [32], and ELG [27] assume the linguistic information shares similar semantics with the visual. They transfer the image-text alignment from teacher to student. In this paper, we distill the local-global temporal information and linguistic context to a single one-branch temporal aggregation module.

3. Preliminaries and Analysis

3.1. General CSLR Framework

Given a sign language video $\mathcal{X} = \{\mathbf{X}_t\}_{t=1}^{T}$ having T frames, a continuous sign language recognition method (CSLR) denoted as $\phi(\cdot)$ contributes to predicting L glosses (*i.e*., $\mathcal{Y} = \{y_i\}_{i=1}^{L}$) contained in the video where y_i is the i-th gloss, and we have $\mathcal{Y} = \phi(\mathcal{X})$. As shown in Fig. 1(a), we show three clips in the input video, which are highlighted with pink, blue, and green colors. A CSLR method is desired to predict the corresponding three glosses, *i.e*., 'JETZT', 'WETTER', and 'APRIL'. The existing training dataset only provides sentence annotation (See the annotation in Fig. 1) instead of the ground truth of each frame and we even do not know the exact number of glosses in the input video. Hence, the task is a weakly-supervised learning problem and significantly challenging.

The dominant CSLR framework [9, 20, 36] mainly involves three modules: spatial perception module (SPM) to extract feature representation of each frame independently, temporal aggregation module (TAM) to sequentially con-

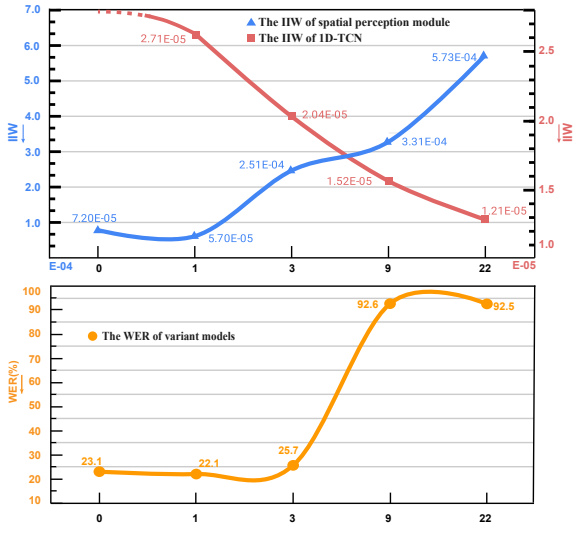

Figure 2. Experiments about chain depth variants of TAM in the general CSLR framework. The IIW [31] is exploited to measure the generalization gap (the larger gap, the larger the IIW) of SPM and TAM. The WER value is utilized to estimate the performance.

duct local and global aggregation across frames' features, and the classification module for the final prediction. We denote the three modules as $\phi_{\text{spm}}(\cdot)$, $\phi_{\text{tam}}(\cdot)$, and $\phi_{\text{cls}}(\cdot)$, respectively, and can reformulate the whole process as $\mathcal{Y} = \phi(\mathcal{X}) = \phi_{\text{cls}}(\phi_{\text{tam}}(\phi_{\text{spm}}(\mathcal{X})))$ (See Fig. 1(a)). Specifically, $\phi_{\text{spm}}(\cdot)$ first extracts features of all frames via a convolution neural network like ResNet18 [10], which are denoted as $\mathcal{V} = \{\mathbf{V}_t\}_{t=1}^{T}$. Then, $\phi_{\text{tam}}(\cdot)$ uses a local temporal context perception component like temporal convolution network (*i.e.*, 1D-TCN) [5, 9, 20, 36] to capture the correlation across adjacent frames and map the representations $\mathcal{V}$ to a new one, *i.e.*, $\mathcal{V}^{\text{loc}} = \{\mathbf{V}_t^{\text{loc}}\}_{t=1}^{T}$. $\phi_{\text{tam}}(\cdot)$ further feeds the locally-aggregated representations to the global perception component like a two-layer BLSTM [9, 20, 34, 36] to capture the global temporal patterns across all frames and get $\mathcal{V}^{\text{glo}} = \{\mathbf{V}_t^{\text{glo}}\}_{t=1}^{T}$. Then, the $\mathcal{V}^{\text{glo}}$ is further passed to ϕ_{cls}, which is a fully-connected layer to predict logits $\mathcal{Z}^{\text{glo}} = \{\mathbf{Z}_t^{\text{glo}}\}_{t=1}^{T}$. Finally, the $\mathcal{Z}^{\text{glo}}$ is fed into the CTC to align the prediction and ground truth and calculate the loss.

3.2. Empirical Studies

Following the fact about chain rules of back-propagation, we study the effects of the chain depth of the temporal aggregation module to the spatial perception module. Intuitively, the temporal aggregation module with fewer layers (*i.e.*, a shallow TAM) has fewer effects on the back-propagation of gradients, and a shallower TAM could lead to a more powerful SPM. To conduct a solid analysis, we have the following setups:

- **Baseline CSLR.** We follow the general CSLR framework in Sec. 3.1 and construct a CSLR method including a cascaded ResNet18 as the SPM, K-layer 1D-TCNs and a two-layer BLSTM as the TAM, and the classifier and the CTC as the classification module. The K-layer 1D-TCNs and BLSTM are used for local and global temporal contexts perception, respectively.
- **TAM with different chain depths.** We set five variants of TAM by stacking K 1D-TCNs where we set $K \in \{0, 1, 3, 9, 22\}$ *. We do not tune the depth of the BLSTM since the BLSTM is prone to over-fitting [9, 20], we contribute to exploring the influence of the depth between SPM and BLSTM.
- **Evaluation of SPM and CSLR.** Following existing works, we use the word error rate (WER) to evaluate the performance of the whole CSLR method. In contrast, for TAM and SPM, we choose another metric, *i.e.*, the compression of information stored in weights (IIW) [31], which is designed to understand the behavior of neural networks and can be an indicator of generalization gap (*i.e.*, the accuracy difference on the test and training datasets) of neural networks [31] . Intuitively, with the same training dataset, a lower IIW of a module means that the module contributes more to the accuracy of the test dataset.

With the above setups, we get five CSLR variants with five TAMs having different chain depths. Then, we train and test all variants on the RWTH-2014 dataset [15]. Specifically, we can calculate the word error rate (WER) of the five variants for the prediction accuracy evaluation. We further count the information stored in weights (IIW) of the SPMs and TAMs of the five variants, respectively. As the results are shown in Fig. 2, we have the following observations: ❶ When we consider the chain depth (*i.e.*, K) from 1 to 22, the IIW of SPM gradually increases and reaches the maximum at $K = 22$ while the IIW of TAM gradually decreases, which means the effects of chain depth to the capability of SPM and TAM have completely opposite trends. A powerful SPM desires a shallow TAM while the TAM itself requires a deeper architecture. ❷ Considering the changes of WER of the CSLR along different chain depths, we see that the trend of WER is consistent with the IIW's variation trend, which means the SPM has higher effects than the TAM on the final prediction accuracy.

3.3. Motivations

According to the above studies, a desired temporal aggregation module should be a shallow architecture to allow more effective training of the spatial perception module but also should be a deep one for a high temporal aggregation capability. *Such a contradiction makes the designing of a suitable TAM challenging.* Because SPM is directly related to the final accuracy, we tend to select a shallow TAM.

*0 illustrates the TAM drop the 1D-TCN. We adjust channels of 1D-TCN to make sure different variants have the same model size. Such as $1, 3, 9, 22$ denote their channels are 1024,512,256 and 128, respectively.

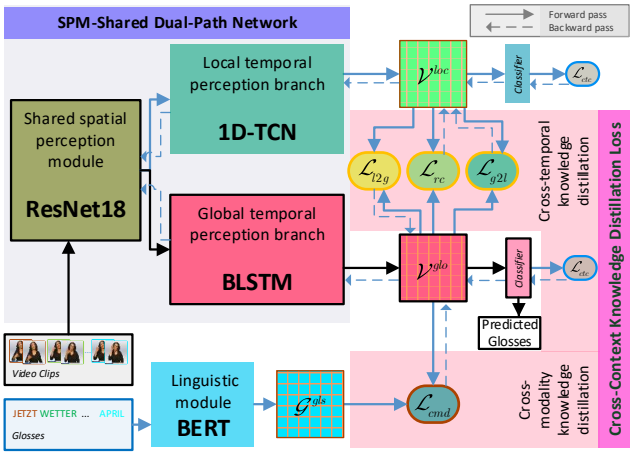

Figure 3. Pipeline of the CTCA. In particular, the steps with black borders or black arrows belong to the CTCA test process.

Then, the key problem becomes how to make TAM shallower and contribute more to the final prediction accuracy.

Existing works have noticed that both local and global temporal perception are critical to prediction accuracy. We also conduct a simple study and observe that removing any temporal perception components leads to a significant accuracy drop (See the supplementary material). All the above observations motivate us to develop a shallower but more powerful temporal aggregation module that has the capability of perceiving local and global context perception.

4. Cross-Temporal Context Aggregation

An intuitive idea is to utilize knowledge distillation to integrate local and global contexts into a shallow architecture. As a result, the two opposite requirements for TAM in Sec. 3.3 could be handled simultaneously. To this end, we first provide a vanilla distillation solution in Sec. 4.1, which has several inherent limitations. Then, we impose a more advanced distillation method, *i.e.*, cross-temporal context aggregation (CTCA), in Sec. 4.2 and Sec. 4.3.

4.1. A Vanilla Distillation Solution

We follow the baseline CSLR in Sec. 3.2 and construct two independent CSLRs that have the same spatial perception module (SPM) and classification module but a different temporal aggregation module (TAM). The first one retains the global perception component (*i.e.*, a BLSTM) and removes the local component (*i.e.*, 1D-TCN). Such a CSLR has the shallowest TAM (*i.e.*, $K = 0$ in Sec. 3.2) and we denote it as CSLR-GloTAM. The second one only contains the local perception component and the BLSTM layer is removed and we denote it as CSLR-LocTAM. The main goal is to train a CSLR-GloTAM to let its TAM have the local temporal contexts. To this end, we use the CSLR-LocTAM to teach the CSLR-GloTAM for acquiring local information and only use the CSLR-GloTAM for inference. The teaching manner follows the vanilla distillation [11].

However, such a vanilla distillation approach has some limitations: ❶ In terms of CSLR-GloTAM, when BLSTM over-fits the training data, the loss can be hardly propagated to the SPM. The distillation based on two separated CSLRs cannot address this issue properly. ❷ By adapting existing distillation methods directly, we need to design two exclusive optimization schemes for the two independent CSLRs, respectively, which inevitably leads to high time and computing costs. ❸ The vanilla distillation leverages a fully-supervised distillation loss, which cannot be adapted to the weakly-supervised optimization like CSLR directly.

4.2. SPM-Shared Dual-Path Network

Enforcing to resolve the first and second limitations in Sec. 4.1, we propose the SPM-Shared Dual-Path Network (SDPN). Inspired by [16], we adopt a multi-branch architecture with a shared SPM to merge the training processes of teacher and student, which distills from each other in a peer-teaching manner. As a result, this structure avoids the requirement of two optimization schemes trade-offs. It is easy to train with only one optimization scheme. As such, different from the vanilla distillation approach has two separate networks, the SDPN is composed of a shared SPM and followed a dual-path (See Fig. 3): local temporal perception branch (*i.e.*, 1D-TCN branch), global temporal perception branch (*i.e.*, BLSTM branch). Specifically, each branch itself of SDPN stands for the specific knowledge concurrently, which conducts a native ensemble model on-the-fly. With the above architecture, even if the BLSTM branch is over-fitted, other branches like the 1D-TCN branch can still back-propagate the loss to the shared SPM straightforwardly. More importantly, the SDPN is easy to adopt distillation losses (See Sec. 4.3) to distill the local temporal context and the linguistic prior to reinforcing BLSTM. The two branches of SDPN have different objectives and architectures as done in [9, 17, 20]. In particular, we only use the BLSTM branch during inference, rather than both branches. We detail the architectural designs in the following:

Shared spatial perception module. We use the ResNet18 for the spatial perception module (See Sec. 3.1) where the fully-connection layer is removed, which takes the video clips as inputs and outputs the spatial representation $\mathcal{V}$.

Local & global temporal perception branches. The local branch consists of a 1D-TCN and a local classifier $\mathcal{F}^{\text{loc}}$. The 1D-TCN is able to provide local temporal representations $\mathcal{V}^{\text{loc}}$ maintained sign-wise information. And the local classifier outputs the local prediction $\mathcal{Z}^{\text{loc}}$. The global branch contains a two-layer BLSTM and a global classifier $\mathcal{F}^{\text{glo}}$. The BLSTM outputs global temporal representations $\mathcal{V}^{\text{glo}}$, which contains the order and correlation of co-occurring signs. The global classifier outputs the global prediction

$\mathcal{Z}^{\text{glo}}$. Note that, both $\mathcal{Z}^{\text{glo}}$ and $\mathcal{Z}^{\text{loc}}$ are optimized by CTC loss to achieve a knowledge complementary to the spatial perception module reinforcement, and the 1D-TCN is removed during inference.

4.3. Cross-Context Knowledge Distillation Loss

To make the shallow BLSTM branch learn both local & global perceptions and achieve the desired property against the third limitation in Sec. 4.1, we propose the Cross-Context Knowledge Distillation Loss (CCKD). CCKD composes a cross-temporal knowledge distillation (CTD) and a cross-modality knowledge distillation (CMD). The CTD builds a mutual temporal context distillation that enables the BLSTM and 1D-TCN branches to incorporate complementary knowledge to refine themselves and the SPM. The CMD distills linguistic prior *i.e.*, inter-gloss semantic correlation from glosses sequence to BLSTM branch to reinforce its temporal aggregation. In contrast to the vanilla distillation that enhances the student's ability to achieve consistent prediction scores with the teacher, the CCKD focuses on distilling channel-wise knowledge from intermediate feature hints [28]. By doing so, CCKD facilitates students to learn how to learn, rather than how to get prediction scores, which is more applicable to CSLR.

Cross-temporal knowledge distillation. ❶ Local temporal context guidance loss ($\mathcal{L}_{l2g}$). Given the local and global temporal representation $\mathcal{V}^{\text{loc}}$ and $\mathcal{V}^{\text{glo}}$, the $\mathcal{L}_{l2g}$ encourages the $\mathcal{V}^{\text{glo}}$ to learn sign-wise context maintained in $\mathcal{V}^{\text{loc}}$ to remain local-global temporal contexts. Due to the knowledge of sign-wise context being distinct from the global context, knowledge confusion may occur during the $\mathcal{L}_{l2g}$ procedure. To relieve the confusion, we weight the $\mathcal{V}^{\text{loc}}$ by the intersection information $I(\mathcal{V}^{\text{loc}}, \mathcal{V}^{\text{glo}})$ among $\mathcal{V}^{\text{loc}}$ and $\mathcal{V}^{\text{glo}}$ to suppress signs information of $\mathcal{V}^{\text{loc}}$ who differs with $\mathcal{V}^{\text{glo}}$.

$$\mathcal{L}_{l2g}(\mathcal{V}^{\text{glo}}, \mathcal{V}^{\text{loc}}) = \varphi(\mathcal{V}^{\text{glo}}, \mathcal{V}^{\text{loc}} \cdot I(\mathcal{V}^{\text{loc}}, \mathcal{V}^{\text{glo}})), \quad (1)$$

$$\varphi(\mathcal{V}^{\text{glo}}, \mathcal{V}^{\text{loc}}) = \frac{\tau^2}{C} \sum_{c=1}^{C} \sum_{i=1}^{T} KL((\mathbf{V}_{c,i}^{\text{glo}}), (\mathbf{V}_{c,i}^{\text{loc}})), \quad (2)$$

$$I(\mathcal{V}^{\text{loc}}, \mathcal{V}^{\text{glo}}) = \frac{1}{\frac{\sigma(\mathbf{V}_{c,i}^{\text{loc}})}{\sigma(\mathbf{V}_{c,i}^{\text{glo}})} + 1}, \quad (3)$$

where if $\frac{\sigma(\mathbf{V}_{c,i}^{\text{loc}})}{\sigma(\mathbf{V}_{c,i}^{\text{glo}})}$ is large and thus $I(\mathcal{V}^{\text{loc}}, \mathcal{V}^{\text{glo}})$ is small, meaning the large different between $\mathcal{V}^{\text{loc}}$ and $\mathcal{V}^{\text{glo}}$. The KL denotes the KL divergence, τ is the temperature factor for smooth, σ indicates the softmax function and C corresponds to the feature channels. As indicated in [28], the KL divergence can be used between intermediate feature hints.

❷ Global temporal context guidance loss ($\mathcal{L}_{g2l}$). In addition, the $\mathcal{L}_{g2l}$ loss evolves distilling global contexts *i.e.*, signs order and correlation among co-occurring signs to the 1D-TCN branch. As a result, remaining the local-global context, the 1D-TCN is applicable to learning sign-wise contexts by taking into account global contexts. The $\mathcal{L}_{g2l}$ loss is also formulated as a similarly form with $\mathcal{L}_{l2g}$:

$$\mathcal{L}_{g2l}(\mathcal{V}^{\text{loc}}, \mathcal{V}^{\text{glo}}) = \varphi(\mathcal{V}^{\text{loc}}, \mathcal{V}^{\text{glo}}). \quad (4)$$

❸ Reconstruction loss ($\mathcal{L}_{rc}$). The intuition behind this loss is that the distillation between 1D-TCN and BLSTM should be further enhanced when the local representation is well-learned. Therefore, we impose the $\mathcal{L}_{rc}$ as the reconstruction function for the local temporal representation $\mathcal{V}^{\text{loc}}$ to reinforce cross-temporal context distillation. Notably, we experimentally follow the entropy principle [13, 18] to measure knowledge uncertainty by $w(\mathbf{V}_i^{\text{loc}})$. If the estimated $\mathcal{V}^{\text{loc}}$ is well-learned and has a high certainty, $H(\mathbf{V}_i^{\text{loc}})$ becomes small, leading to a large weight $(1 + w(\mathbf{V}_i^{\text{loc}}))$. Then, the consistency between local & global representations is further enhanced via the $\mathcal{L}_{rc}$. We also adopt a residual connection for stable weighting.

$$\mathcal{L}_{rc} = \sqrt{\sum_{i=1}^{T} (\mathbf{V}_i^{\text{glo}} - \mathbf{V}_i^{\text{loc}})^2} \cdot \sum_{i=1}^{T} (1 + w(\mathbf{V}_i^{\text{loc}})), \quad (5)$$

$$w(\mathbf{V}_i^{\text{loc}}) = \exp^{-H(\mathbf{V}_i^{\text{loc}})}, \quad (6)$$

As mentioned above, the CTD loss is formulated as:

$$\mathcal{L}_{ctd} = \alpha_1 \mathcal{L}_{l2g} + \alpha_2 \mathcal{L}_{g2l} + \alpha_3 \mathcal{L}_{rc}, \quad (7)$$

Where the α_1, α_2, α_3 achieve trade-offs among losses in the CTD loss. Furthermore, the CTD loss ensures the spatial perception module to be propagated the local-global information during the whole training time, which avoids the propagation effect of BSLTM overfitting and enhances SPM generalization ability, implicitly.

Cross-modality knowledge distillation. The CTD loss only restricts the cross-temporal context statistics yet has weak semantic context supervision *i.e.*, the linguistic prior. Motivated by CLIP [26] and ELG [27], the cross-modality knowledge distillation (CMD) is proposed to distill inter-gloss semantic correlation from glosses sequence to BLSTM branch. In practice, given the gloss sequence $\mathcal{Y} = \{y_i\}_{i=1}^{L}$ corresponding to $\mathcal{X}$ with L glosses, it will be embedded by pre-trained BERT to generate high-dimensional semantic features $\mathcal{G}^{\text{gls}} = \{\mathbf{G}_i^{\text{gls}}\}_{i=1}^{L} \in \mathbb{R}^{Y \times L}$. Further, the $\mathcal{G}^{\text{gls}}$ and global temporal representation $\mathcal{V}^{\text{glo}}$ will be computed by cross-attention [29] to generate the linguistic-visual relation $\mathcal{G}^{\widehat{t}} \in \mathbb{R}^{Y \times L}$.

$$\mathcal{G}^{\widehat{t}} = softmax(\frac{f_{\mathcal{Q}}(\mathcal{G}^{\text{gls}}) f_{\mathcal{K}}(\mathcal{V}^{\text{glo}})^{\text{T}}}{\sqrt{C}}) f_{\mathcal{V}}(\mathcal{V}^{\text{glo}}), \quad (8)$$

Moreover, the self-correlation matrix $cor^{\widehat{t}}$ will be obtained by the matrix product of $\mathcal{G}^{\widehat{t}}$ and $(\mathcal{G}^{\widehat{t}})^{\text{T}}$. The cor^{t} is a

self-correlation matrix of gloss features $\mathcal{G}^{\text{gls}}$, *i.e.*, $cor^t = \left\|\mathcal{G}^{\text{gls}}\right\|_2 \left(\left\|\mathcal{G}^{\text{gls}}\right\|_2\right)^{\mathrm{T}}$. Finally, we distill the inter-gloss discrimination from the cor^t to the $cor^{\hat{t}}$ via the cross-modality distillation loss $\mathcal{L}_{cmd}$. The t and $\hat{t}$ represent the gloss modality and visual modality, respectively.

$$\mathcal{L}_{cmd} = \frac{1}{L}\sum_{i=1}^{L} \sigma(cor_i^{\hat{t}}) \cdot \log\left[\frac{\sigma(cor_i^{\hat{t}})}{\sigma(cor_i^{t})}\right], \tag{9}$$

where σ indicates the softmax function. We involve $cor^{\hat{t}}$ to have the same knowledge with cor^t to enhance the linguistic-visual relation, which encourages $\mathcal{V}^{\text{glo}}$ to have the inter-gloss discrimination indirectly.

Combining all loss functions, the CCKD is denoted as:

$$\mathcal{L}_{CCKD} = \alpha\mathcal{L}_{ctd} + \beta\mathcal{L}_{cmd} + \gamma\mathcal{L}_{ctc}, \tag{10}$$

where the α, β, and γ control the trade-off among losses.

4.4. Implementation Details

Network details. We set $K = 3$ and kernel size is 3 for 1D-TCN, each layer equips with a 1D Batch Norm layer and a ReLU activation. The output channel of each temporal layer in 1D-TCN is set to 1024 and BLSTM is set to 2048.
Hyperparameters setting. The τ in Eq. (2) is set to 4, the α_1, α_2, α_3 in Eq. (7) are set to 0.2, 1, 1 respectively. And α, β, and γ in Eq. (10) are set to 3, 130, 1 respectively. We find that the α_1 is chosen in $[0.1, 1)$ and the β needs to be set to among $[50, 150]$ to obtain good performance.
Training and test strategy. During both the training and test stages for all benchmarks, we randomly discard half of video frames as inputs following SFL [22]. We then resize the frame size to 256×256 and perform randomly cropping frames to 224×224, as well as random horizontal flip (50%) in the training stage (only center crop frames to 224×224 in the test stage). The BERT in the CMD loss is frozen and only employed in the training stage. In the test, only the spatial perception module and the global temporal perception branch are exploited to generate predictions. The CTCA is implemented in PyTorch with one A100 GPU.

5. Experiments

5.1. Dataset and Evaluation

RWTH-2014 [15] is a German sign language dataset, it consists of a total of 6,841 sentences with a vocabulary of 1,295 glosses, signed by 9 different singers.
RWTH-2014T [1] can be used for both CSLR and sign language translation tasks [1]. It involves a vocabulary of 1,085 glosses, split into 7,096, 519, and 642 samples in the train set, dev set, and test set, respectively.
CSL-Daily [33] is a large Chinese sign language dataset about people's daily life. It can be employed in both CSLR and sign language translation. There are 18,401, 1,077, and 1,176 videos for the training set, dev set, and test set, respectively. It has a vocabulary of 2,000 glosses for CSLR.
Evaluation metric. In this paper, we adopt the word error rate (WER) metrics for CSLR methods evaluation. It measures the minimum number of substitutions, deletions, and insertions that need to convert one predicted glosses sequence to a given reference sequence [15].

Table 1. Comparison with state-of-the-art methods on the RWTH-2014 dataset. (WER(%) the lower is the better).

Methods	Dev		Test	
	del/ins	WER	del/ins	WER
DNF [6]	7.8/3.5	23.8	7.8/3.4	24.4
FCN [5]	-	23.7	-	23.9
VAC [20]	7.9/2.5	21.2	8.4/2.6	22.3
CMA [23]	7.3/2.7	21.3	7.3/2.4	21.9
SMKD [9]	6.8/2.5	20.8	6.3/2.3	21.0
C^2SLR [36]	-	20.5	-	20.4
TLP [17]	6.3/2.8	19.7	6.1/2.9	20.8
RadialCTC [21]	6.5/2.7	**19.4**	6.1/2.6	20.2
CTCA(Ours)	6.2/2.9	19.5	6.1/2.6	**20.1**

Table 2. Comparison with state-of-the-art methods on the RWTH-2014T dataset. (WER(%) the lower is the better).

Methods	WER	
	Dev	Test
SLT [2]	24.6	24.5
CNN+LSTM+HMM [14]	22.1	24.1
BN-TIN+Transf [33]	22.7	23.9
V-L Mapper [4]	21.9	22.5
SMKD [9]	20.8	22.4
C^2SLR [36]	20.2	20.4
TLP [9]	19.4	21.2
CTCA(Ours)	**19.3**	**20.3**

5.2. Comparison with State-of-the-arts

Evaluation on RWTH-2014. The objective of the VAC and SMKD is to enhance the discriminative capacity of the visual module (*i.e.*, SPM+1D-TCN). Meanwhile, our CTCA focuses on investigating and designing the desired TAM for CSLR. As demonstrated in Tab. 1, the proposed CTCA solely utilizes the RGB information of sign language videos, yet it achieves competitive performance in terms of the WER (19.5% and 20.1%). Notably, it surpasses other RGB cue-based approaches, *i.e.*, SMKD, C2SLR, and TLP, thereby validating the effectiveness of CTCA.
Evaluation on RWTH-2014T. In Tab. 2, it is evident that the CTCA gains the best performance (19.3% and 20.3%) compared with other state-of-the-art approaches.
Evaluation on CSL-Daily. To further assess the CTCA's generalization capacity, we evaluate it on the CSL-Daily benchmark. As presented in Tab. 3, the CTCA outperforms

Table 3. Comparison with state-of-the-art methods on the CSL-Daily. (WER(%) the lower is the better).

Methods	Dev		Test	
	del/ins	WER	del/ins	WER
LS-HAN [12]	14.6/5.7	39.0	14.8/5.0	39.4
SLT(Gloss+Text) [2]	10.3/4.4	33.1	9.6/4.1	32.0
FCN [5]	12.8/4.0	33.2	12.6/3.7	32.5
BN-TIN+Transf [33]	13.9/3.4	33.6	13.5/3.0	33.1
TIN-Iterative [6]	12.8/3.3	32.8	12.5/2.7	32.4
CTCA(Ours)	9.2/2.5	**31.3**	8.1/2.3	**29.4**

Table 4. Effect of the SDPN architecture on the RWTH-2014.

Methods	IIW	Dev		Test	
		del/ins	WER	del/ins	WER
Baseline	6.7E-5	7.0/3.0	21.8	6.7/2.7	22.1
Baseline+SDPN	5.6E-5	7.3/3.1	21.7	7.4/2.4	21.8
VAC(VE) [20]	5.4E-5	-	23.3	-	23.8
VAC(VE)+SDPN	4.2E-5	7.6/3.1	22.0	7.6/3.0	22.6

all other methods and achieves a 3% WER advantage over the second-best performer, the TIN-Iterative.

5.3. Ablation Study

Benefits of SPM-Shared Dual-Path Network. Tab. 4 illustrates that when both baseline and VAC are constructed to SDPN architecture, achieving lower generalization gap (lower IIW) and improved performance (lower WER). It is noteworthy that during the testing, only the spatial perception module and the BLSTM branch of SDPN are utilized. These results suggest that the SDPN architecture can facilitate thorough training for the spatial perception module, leading to improved performance.

Benefits of components of Cross-Context Knowledge Distillation loss. ❶ **Effect of vanilla distillation solution.** We compare the vanilla distillation approach, as discussed in Sec. 4.1 with the baseline. As presented in Tab. 5, the vanilla distillation yields an improvement over the baseline, yet it is still inferior to each component of the CCKD.

❷ **Effect of $\mathcal{L}_{l2g}$ loss and $\mathcal{L}_{g2l}$ loss.** We introduced two distillation methods, SDPN A and SDPN B based on the SDPN model. SDPN A is the 1D-TCN distilling sign-wise knowledge to the BLSTM via the $\mathcal{L}_{l2g}$. This method improves performance on the test set by 1.0% compared to the baseline. We also removes the intersection information weighted term $I(.;)$ from SDPN A, which results in a 0.4% drop in performance on the test set, implying its effectiveness. SDPN B, on the other hand, is an SDPN model optimized by the $\mathcal{L}_{g2l}$, which yields a 1.4% WER improvement over the baseline on the test set. It is noteworthy that SDPN A's performance is inferior to SDPN B. This can be explained that BLSTM tends to overfit on the sequential order of signs, which is easy to learn and leads to fast convergence

Table 5. Ablation study on cross-context knowledge distillation loss on the RWTH-2014.

Method	$\mathcal{L}_{ctd}$			$\mathcal{L}_{cmd}$	Dev	Test
	$\mathcal{L}_{l2g}$	$\mathcal{L}_{g2l}$	$\mathcal{L}_{rc}$			
Baseline	-	-	-	-	21.8	22.1
Vanilla	-	-	-	-	21.7	21.9
SDPN A	✓	-	-	-	21.0	21.1
SDPN A$-I(.;)$	✓	-	-	-	21.3	21.5
SDPN B	-	✓	-	-	20.8	20.7
SDPN C	✓	✓	-	-	20.4	20.6
SDPN D	✓	✓	✓	-	20.0	20.4
SDPN D$-w(.;)$	✓	✓	✓	-	20.2	20.6
SDPN E	-	-	-	✓	21.3	21.0
CTCA	✓	✓	✓	✓	**19.5**	**20.1**

Table 6. Comparison of different knowledge fusion schemes on the RWTH-2014. "Wasserstein" is the Wasserstein distance.

Method	Knowledge fusion	Dev	Test
SDPN	-	21.7	21.8
	Vanilla distillation [11]	21.6	21.6
	Wasserstein [3]	21.6	21.5
	JMMD [19]	21.3	21.3
	CKD [28]	21.3	21.5
	CTCA($\mathcal{L}_{l2g}$)	21.0	21.1
	concatenation	22.7	23.6
	point-wise addition	21.2	22.3
	attention	22.2	22.6

optimized by CTC [9, 25]. Although the BSLTM in SDPN A receives local context knowledge, it still propagates limited local-global information to SPM, resulting in limited gain for SPM when it is overfitting. Furthermore, the 1D-TCN is less prone to overfitting [5] and in SDPN B facilitated by the $\mathcal{L}_{g2l}$ the 1D-TCN can learn sign-wise context by taking into account global contexts. Then it can provide richer sign-wise supervision to the SPM for locating signs, which can eliminate the limited back-propagation. As a result, more robust features SPM can generate, and they are fed into BLSTM, which outperforms SDPN A. This finding is consistent with the observation ❷ in Sec. 3.2 that the SPM has a high effect on the final prediction performance. The results of SDPN C show that the mutual distillation approach gains 20.6% on the test set. This technique can enhance both the $\mathcal{L}_{l2g}$ and $\mathcal{L}_{g2l}$, resulting in improved 1D-TCN and BLSTM. This success is also verified by [9].

❸ **Effect of $\mathcal{L}_{rc}$ loss and $\mathcal{L}_{cmd}$ loss.** Seen from Tab. 5, based on the SDPN C model the SDPN D combines the $\mathcal{L}_{rc}$, which is effective in enhancing performance. Specifically, when SDPN D drops the $w(.;)$, it gives a worse result, it shows the effect of entropy measurement. Moreover, SDPN E, achieves relatively better performance, which validates the need for linguistic prior supervision.

Table 7. Performance comparison of local temporal perception module with distinct temporal window widths on the RWTH-2014. Ft and $Ft(d)$ correspond to the 1D temporal convolution layer with kernel of t and dilation of d, respectively.

Method	variants	windows	Dev	Test
1D-TCN	$F3$-$F3$-$F3$	7*2	19.5	20.1
	$F3(1)$-$F3(2)$	7*2	19.8	20.3
	$F5$-$F5$	9*2	20.6	20.6
	$F5$-$F5$-$F5$	13*2	19.9	20.6
	$F7$-$F7$	13*2	20.1	20.3

Table 8. Comparison of CTCA with distinct global temporal perception modules (GTPM) on the RWTH-2014.

Method	variants	Dev	Test
GTPM-branch	BLSTM	19.5	20.1
	Dilated blocks	22.2	22.6
	Transformer	28.7	28.9
	Transformer+BLSTM	24.4	24.1

5.4. Module Analysis

Other choices of knowledge fusion schemes. In Tab. 6, we compare distinct knowledge fusion schemes based on our SDPN, including the knowledge distillation strategies (upper part of Tab. 6) and feature fusion strategies (bottom of Tab. 6). Specifically, all schemes use only SPM and BLSTM in testing and all knowledge distillation schemes distill knowledge from 1D-TCN to BLSTM. We observe that feature fusion strategies have limited promotion and even worse results than the SDPN, while all knowledge distillation schemes gained an improvement by the SDPN. This validates the effectiveness of knowledge distillation and shows that simply concatenating or summing features lacks the constraint for 1D-TCN and BLSTM to guide them to possess context knowledge of the other. And the attention mechanism only learns the correlation between 1D-TCN and BLSTM, neglecting context knowledge transferring.

Impact of different distillation objects. The Tab. 6 also shows the performance of using intermediate feature hints versus prediction scores as distillation objects. The Wasserstein distance [3, 7] and JMMD [19] enforce to align the intermediate feature distributions. And the CKD [28] distills significant activation values in each channel of intermediate feature hints. Further, based on the CKD our $\mathcal{L}_{l2g}$ weights the distillation by the intersection information of intermediate feature hints. On the other hand, the vanilla distillation [11] is designed for aligning glosses prediction scores, gives the worst performance. Overall, results suggest that adopting intermediate feature hints as distillation objects is more effective than employing prediction scores to teach how to generate the desired output.

Impact of different temporal window widths of 1D-TCN. Tab. 7 delivers that the window width designed to approximate the average length of isolated sign [6] performs the best. (Since half of the frames are selected for training, *i.e.*, SFL [22], the actual window width needs to be multiplied by 2). Whereas, variants of 1D-TCN with substantially different window widths (ranging from 14 to 26) do not show a significant performance gap. We explain that variants of 1D-TCN can learn sign-wise knowledge through global context guidance optimized by the $\mathcal{L}_{g2l}$, which relieves the effect of receptive field discrepancy.

Other choices of global temporal perception modules. Tab. 8 ablates the performance of distinct global temporal perception modules, such as BLSTM, Dilated blocks [24], Transformer [2, 22], and Transformer+BLSTM. The BLSTM achieves the best performance. The plausible reason is that sign language videos have a semantic sequence with grammatical rules, which means both forward and backward frames should be taken into consideration [25], which BLSTM achieves natively. However, Transformer is unable to do it to learn the semantic sequence well.

6. Conclusion

In this work, we have extensively studied the limitation and desired properties of the temporal aggregation module (TAM) in the continuous sign language recognition framework. In particular, with an advanced analysis tool (*i.e.*, information stored in weights [31]), we showed that the desired TAM should be a shallow architecture to allow more effective training of spatial perception module but also should be a deep one for a high temporal aggregation capability. To achieve this property, we proposed cross-temporal context aggregation (CTCA) with a dual-path network. Moreover, we proposed a cross-context knowledge distillation loss function to transfer the local temporal context and the linguistic prior to the global perception module. Extensive experimental results demonstrated that our CTCA effectively enhances the generalization of the spatial perception module while achieving the leading accuracy with fewer parameters and higher efficiency.

Limitations. In the redesign of the temporal aggregation module, the improvement of global temporal context perception still follows the current scheme, which still tends to cause overfitting. Therefore finding a more suitable global temporal context perception for the weakly supervised CSLR task is a challenge to be solved. Meanwhile, the CTCA is not tested on real-world with complex scenes, which deserves further study.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (61906135, 62020106004, 92048301, 62276141), and the Tianjin Research Innovation Project for Postgraduate Students (2021YJSB244). It was also supported by A*STAR Centre for Frontier AI Research.

References

[1] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *CVPR*, 2018. 1, 6

[2] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *CVPR*, 2020. 6, 7, 8

[3] Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. Graph optimal transport for cross-domain alignment. In *ICML*, 2020. 7, 8

[4] Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. A simple multi-modality transfer learning baseline for sign language translation. In *CVPR*, 2022. 6

[5] Ka Leong Cheng, Zhaoyang Yang, Qifeng Chen, and Yu-Wing Tai. Fully convolutional networks for continuous sign language recognition. In *ECCV*, 2020. 3, 6, 7

[6] Runpeng Cui, Hu Liu, and Changshui Zhang. A deep neural framework for continuous sign language recognition by iterative training. *TMM*, 21(7), 2019. 2, 6, 7, 8

[7] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *NIPS*, 2013. 8

[8] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, 2006. 1, 2

[9] Aiming Hao, Yuecong Min, and Xilin Chen. Self-mutual distillation learning for continuous sign language recognition. In *ICCV*, 2021. 1, 2, 3, 4, 6, 7

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[11] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 2, 4, 7, 8

[12] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. Video-based sign language recognition without temporal segmentation. In *AAAI*, 2018. 7

[13] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4), 1957. 5

[14] Oscar Koller, Necati Cihan Camgoz, Hermann Ney, and Richard Bowden. Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. *PAMI*, 42(9), 2019. 6

[15] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141, 2015. 1, 3, 6

[16] Xu Lan, Xiatian Zhu, and Shaogang Gong. Knowledge distillation by on-the-fly native ensemble. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *NIPS*, 2018. 2, 4

[17] Zekang Liu Lianyu Hu, Liqing Gao and Wei Feng. Temporal lift pooling for continuous sign language recognition. In *ECCV*. Springer, 2022. 4, 6

[18] Hu Liu, Sheng Jin, and Changshui Zhang. Connectionist temporal classification with maximum entropy regularization. *NIPS*, 31, 2018. 5

[19] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *ICML*. PMLR, 2017. 7, 8

[20] Yuecong Min, Aiming Hao, Xiujuan Chai, and Xilin Chen. Visual alignment constraint for continuous sign language recognition. In *ICCV*, 2021. 1, 2, 3, 4, 6, 7

[21] Yuecong Min, Peiqi Jiao, Yanan Li, Xiaotao Wang, Lei Lei, Xiujuan Chai, and Xilin Chen. Deep radial embedding for visual sequence learning. In *ECCV*, 2022. 2, 6

[22] Zhe Niu and Brian Mak. Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition. In *ECCV*, 2020. 6, 8

[23] Junfu Pu, Wengang Zhou, Hezhen Hu, and Houqiang Li. Boosting continuous sign language recognition via cross modality augmentation. In *ACMMM*, 2020. 6

[24] Junfu Pu, Wengang Zhou, and Houqiang Li. Dilated convolutional network with iterative optimization for continuous sign language recognition. In *IJCAI*, 2018. 8

[25] Junfu Pu, Wengang Zhou, and Houqiang Li. Iterative alignment network for continuous sign language recognition. In *CVPR*, 2019. 1, 2, 7, 8

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021. 5

[27] Karsten Roth, Oriol Vinyals, and Zeynep Akata. Integrating language guidance into vision-based deep metric learning. In *CVPR*, 2022. 2, 5

[28] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *ICCV*, 2021. 5, 7, 8

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 5

[30] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. *NIPS*, 33, 2020. 2

[31] Zifeng Wang, Shao-Lun Huang, Ercan Engin Kuruoglu, Jimeng Sun, Xi Chen, and Yefeng Zheng. Pac-bayes information bottleneck. In *ICLR*, 2022. 1, 2, 3, 8

[32] Peizhen Zhang, Zijian Kang, Tong Yang, Xiangyu Zhang, Nanning Zheng, and Jian Sun. Lgd: label-guided self-distillation for object detection. In *AAAI*, volume 36, 2022. 2

[33] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *CVPR*, 2021. 1, 6, 7

[34] Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. Spatial-temporal multi-cue network for continuous sign language recognition. In *AAAI*, volume 34, 2020. 3

[35] Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. More grounded image captioning by distilling image-text matching model. In *CVPR*, 2020. 2

[36] Ronglai Zuo and Brian Mak. C2slr: Consistency-enhanced continuous sign language recognition. In *CVPR*, 2022. 1, 2, 3, 6