

Visual Alignment Constraint for Continuous Sign Language Recognition

Supplementary Material

Yuecong Min^{1,2}, Aiming Hao^{1,2}, Xiujuan Chai³, Xilin Chen^{1,2}

¹Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China

²University of Chinese Academy of Sciences, Beijing, 100049, China

³Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing, 100081, China

{yuecong.min, aiming.hao}@vipl.ict.ac.cn, chaixiujuan@caas.cn, xlchen@ict.ac.cn

This supplementary material provides details that are not shown in the main paper. We first present the training process of the proposed VAC (§ A.1) and ablations on dataset size (§ A.2), temperature (§ A.3), loss weight (§ A.4) and augmentation (§ A.5). Then we present the details of the temporal convolution designs (§ B.1), the proposed metrics (§ B.2) and the performance gap (§ B.3). Finally, we visualize the spatial activations (§ C.1), magnitudes (§ C.2) and more qualitative results (§ C.3).

A. Additional Results

A.1. Training process of VAC

We compare the curves with different constraints in Fig. 1. Adopting VAC can significantly accelerate the training process, which achieves better performance than baseline after the first learning rate decay. The \mathcal{L}_{VE} can immediately accelerate the training process at the beginning and the \mathcal{L}_{VA} takes effect when the alignment model begins to converge, which happens after the first learning rate decay.

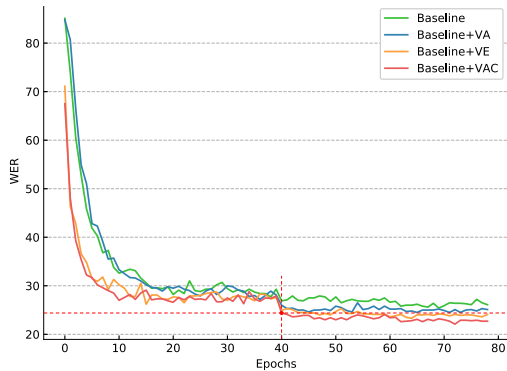


Figure 1. Learning curves of WER(%) on PHOENIX14 with different settings. The learning rate is decayed at 40 and 60 epochs.

A.2. Ablation on Dataset Size

We visualize the recognition results with different sizes of training data in Fig. 2 below. It can be seen that VAC can steadily improve performance as the training data size increases, while the visual extractor of the baseline (WER_a) shows a saturation trend, which implies the available training data is **NOT** sufficient for the visual extractor.

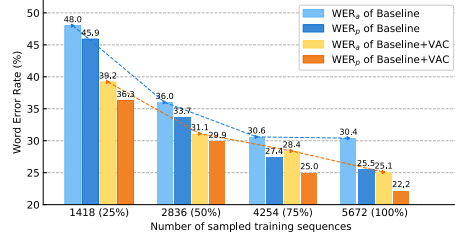


Figure 2. Results on PHOENIX14 with sampled training set.

A.3. Ablation on Temperature τ

To determine the temperature τ in Equ. 5 of the main paper, we evaluate its effect in Table 1. Low temperatures leads to spike responses and high temperatures will produce noisy supervision. According to ablation results, $\tau=8$ is a proper choice.

Table 1. Ablation results (WER, %) of temperature τ .

τ	1	4	8	12	16
Dev	22.1	22.0	21.2	21.7	21.6
Test	22.8	22.9	22.3	22.9	22.7

A.4. Ablation on Loss Weight α

Another hyperparameter need to be carefully tuned is the loss weight α in Equ. 6 of the main paper. We conduct ablation study on it and present results in Table 2. As weight of distillation increases, the performance first increases and then decreases after certain value. The optimal weight for distillation loss is 25.

Table 2. Ablation results (WER, %) of loss weight α .

α	10	15	20	25	30	35
Dev	22.1	21.9	21.5	21.2	21.5	22.0
Test	23.0	22.4	22.1	22.3	22.6	23.2

Table 3. Ablation results (WER, %) of augmentation.

Crop	Flip	Temporal Scaling	Dev	Test
			28.1	28.4
✓			23.8	24.6
	✓		26.1	26.4
		✓	27.4	27.3
✓	✓		23.2	23.8
✓	✓	✓	22.1	23.0

A.5. Ablation on Data Augmentation

As mentioned in Sect. 5.1, we adopt three kinds of data augmentation strategies (random crop, horizontal flip and random temporal scaling) during training, which is the same as previous work [2]. In Table 3, we evaluate the effect of data augmentation. We can observe that adopting data augmentation can significantly improve the performance, especially with random crop. We assume that the network has a tendency to use shortcuts, such as the absolute position of hands in video, and adopting random crop can enforce the network to learn more high-level features and mitigate these shortcuts. It is interesting to see that the horizontal flip can improve the results although all signers in PHOENIX14 use their right hand as the dominant hand when signing, which brings about 0.6% performance gain.

B. Additional Implementation Details

B.1. Details on Temporal Layer Designs

As mentioned in Sect. 5.2, we evaluate three kinds of basic temporal convolution layers and present the details in Table 4. The output dimension C of the ResNet18 [1] is 512, and the output dimension C' of the temporal layer is 1024. Conv1x α (1x α Convolution-BN-ReLU) and Max-pooling 1x β are used to extract different levels of features. The lengths T' of output sequences of (Frame-wise Raw, Frame-wise Conv1x3, Subgloss-wise, Gloss-wise) are $(T, T-2, T/2-2, T/4-3)$. The alignment model contains a two-layer BiLSTM (512 hidden states for each direction) and a fully-connected layer with N output units is adopted to make the final prediction.

B.2. Details on Proposed Metrics

In Sect. 4.2, we propose two metrics, Word Deterioration Rate (WDR) and Word Amelioration Rate (WAR), to evaluate the performance of the recognition results. To calculate WDR and WAR, we need to align the reference sentence and the recognized sentences from the auxiliary classifier

		WER _p = WER _a = 2/9 ≈ 22.2%	
REF _p :	ON HEUTE NACHT MEHR SCHNEE NORD **** SUEDEST **** ABER KALT		
HYP _p :	ON HEUTE NACHT MEHR SCHNEE NORD SUEDEST SUEDEST ABER KALT		
REF _a :	ON HEUTE NACHT MEHR SCHNEE NORD SUEDEST ABER KALT		
HYP _a :	ON HEUTE NACHT **** SCHNEE NORD SUEDEST ABER ****		
REF*:	ON HEUTE NACHT MEHR SCHNEE NORD **** SUEDEST **** ABER KALT		
HYP*:	ON HEUTE NACHT **** SCHNEE NORD **** SUEDEST **** ABER ****		
HYP _p :	ON HEUTE NACHT MEHR SCHNEE NORD SUEDEST SUEDEST ABER KALT		
		WAR = 2/9 ≈ 22.2%	WDR = 2/9 ≈ 22.2%

(a) Alignment process of three sentences.

		WER _p = 4/8 = 50.0%	WER _a = 3/8 ≈ 37.5%
REF _p :	ON KUEHL KOMMEN **** TEMPERATUR SAMSTAG SONNTAG IX GLEICH		
HYP _p :	ON KUEHL WEHEN TEMPERATUR SAMSTAG SONNTAG ** WIE		
REF _a :	ON KUEHL KOMMEN **** TEMPERATUR SAMSTAG SONNTAG IX GLEICH		
HYP _a :	ON KUEHL KOMMEN WEHEN TEMPERATUR SAMSTAG SONNTAG ** DASSELBE		
REF*:	ON KUEHL KOMMEN **** TEMPERATUR SAMSTAG SONNTAG IX GLEICH		
HYP*:	ON KUEHL **** WEHEN TEMPERATUR SAMSTAG SONNTAG ** WIE		
		WER _p = 5/8 = 62.5%	WER _a = 3/8 ≈ 37.5%

(b) An example of performance deterioration of the primary classifier.

		WER _p = 2/8 = 25.0%	WER _a = 2/8 = 25.0%
REF _p :	MORGEN BESONDERS NORD REGION UND **** WEST **** REGEN KOENNEN		
HYP _p :	MORGEN BESONDERS NORD REGION UND WEST WEST DANN REGEN KOENNEN		
REF _a :	MORGEN BESONDERS NORD REGION UND WEST REGEN KOENNEN		
HYP _a :	MORGEN BESONDERS NORD NACHT WEST WEST REGEN KOENNEN		
REF*:	MORGEN BESONDERS NORD REGION UND **** WEST **** REGEN KOENNEN		
HYP*:	MORGEN BESONDERS NORD **** NACHT WEST WEST REGEN KOENNEN		
		WER _p = 2/8 = 25.0%	WER _a = 4/8 = 50.0%

(c) An example of performance amelioration of the primary classifier.

Figure 3. Alignment results of the proposed alignment method. We highlight **wrong recognition glosses** and the alignment results of **the auxiliary classifier**, **the primary classifier**.

and the primary classifier first. As shown in Fig. 3(a), we first align the reference and the recognized sentences and refer the alignment results as to (REF_p, HYP_p) and (REF_a, HYP_a) for the primary classifier and the auxiliary classifier, respectively. Then we align REF_a and REF_p to obtain the aligned reference REF*. The final alignment results (REF*, HYP*, HYP_a, HYP_p) are presented in the last row of Fig. 3(a) by aligning (REF*, HYP_a) and (REF*, HYP_p), respectively.

With the help of alignment results, we can compare the performance of the two classifiers. As shown in Fig. 3(a), both of the auxiliary and the primary classifiers have the same WER 22.22% (HYP_p has two insertion errors, and REF_a has two deletion errors). The primary classifier corrects the misrecognized results of the auxiliary classifier but makes new mistakes, which can not be measured by WER. WDR measures the ratio that is correctly recognized by the auxiliary classifier but misrecognized by the primary classifier (two ‘SUED’ in HYP_p), and WAR does in the opposite direction (‘MEHR’ and ‘KALT’ in HYP_p). Based on the proposed metrics, we can calculate that both WAR and WDR are 22.22% and better understand the recognition results: the introduction of the alignment model brings 22.22% gains and extra 22.22% errors, so the total WER

Table 4. More details about the temporal layer design. Conv $1 \times \alpha$ ($1 \times \alpha$ Convolution-BN-ReLU) and Max-pooling $1 \times \beta$ are used to extract different levels of features.

Backbone		Layer		Output Size
		ResNet18		$(B, C, 1, T)$
Temporal Layer	Frame wise		Subgloss wise	Gloss wise
	Conv 1×1	Conv 1×3	Conv 1×5 Max-pooling 1×2	Conv 1×5 Max-pooling 1×2 Conv 1×5 Max-pooling 1×2
				$(B, C', 1, T')$
Alignment model		BiLSTM($C', 512, 2$) Linear($1024, N$)		(B, T', N)

Table 5. Train/Dev/Test performance comparison (%) with different evaluate metrics on PHOENIX14. WER_p^* and WER_a^* correspond to the WER^* results of primary classifier and auxiliary classifier, respectively, and $\Delta WER^* = WER_a^* - WER_p^* = WAR - WDR$.

	WER_a^*	WER_p^*	WAR	WDR	ΔWER^*
Baseline	12.9 / 30.4 / 29.4	2.5 / 25.5 / 26.9	11.5 / 11.3 / 10.0	1.2 / 6.5 / 7.4	10.4 / 4.9 / 2.5
Baseline + iteration	7.9 / 27.5 / 27.0	1.9 / 25.1 / 26.3	7.0 / 9.8 / 8.8	0.9 / 7.3 / 8.2	6.0 / 2.4 / 0.7
Baseline + VE	3.8 / 26.2 / 26.3	2.5 / 23.4 / 24.0	2.1 / 6.2 / 5.8	0.8 / 3.4 / 3.4	1.3 / 2.8 / 2.3
Baseline + VA	13.4 / 26.8 / 26.9	2.0 / 24.7 / 25.2	12.0 / 7.7 / 7.8	0.6 / 5.7 / 6.2	11.4 / 2.1 / 1.7
Baseline + VAC	4.5 / 24.0 / 24.6	1.0 / 21.3 / 22.4	3.7 / 5.7 / 5.4	0.2 / 3.0 / 3.2	3.6 / 2.7 / 2.2

remains unchanged.

Due to the alignment process and different weights of operations, the proposed three-sentence alignment strategy leads to a little performance degradation than the general WER, as discussed in Sect. 5.2. Fig. 3(b) and Fig. 3(c) show some examples. Aligning REF_a and REF_p changes the alignment results, which often breaks substitution errors to more deletion and insertion errors. However, only a small ratio of sequences has such a problem, and we believe this problem is acceptable for results analysis.

B.3. Details on the Performance Gap

Figure 6 in Sect. 5.2 visualizes the performance gap with different settings, and we present the detailed results in Table 5. The conclusions in Sect. 5.2 are consistent on both dev and test sets.

Ref : UND TAG BLEIBEN KUEHL SECHS GRAD BAYERN IX REGION AUCH 5+H IX AUCH ABER
Baseline : ABER TAG BLEIBEN SECHS GRAD FLUSS IX REGION SCHLESWIG FREITAG AUCH ABER
VAC : DANN TAG BLEIBEN SECHS GRAD BAYERN IX REGION AUCH 5 H SOLL ABER

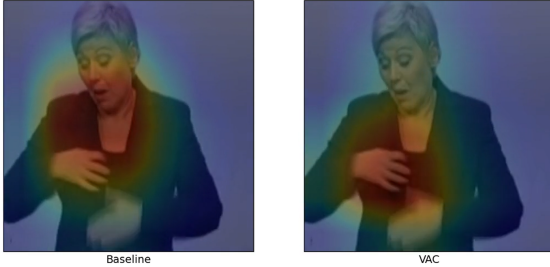


Figure 4. Interface of the recognition result animation. We highlight the wrong recognition glosses.

C. Qualitative Results

C.1. Visualization of Spatial Activations

We visualize some recognition results in the animation folder. As shown in Fig. 4, the reference and the predictions of baseline and Visual Alignment Constraint (VAC) are presented above the videos. The bottom videos visualize the activation changes during the signing. The activation maps are obtained by calculating the l_2 norm of the 7×7 ResNet18 feature maps. From the animation, we can observe that the baseline mainly focuses on the central area of frames, and the proposed method can dynamically focus on hands and facial expressions, which extracts more discriminative visual features.

C.2. Visualizing of Magnitudes

In Sect. 3.3, we propose a magnitude hypothesis that the l_2 norms of features reflect the importance of frames. Besides, experimental results in Sect. 5.2 verify that the proposed VAC is more compatible with the spiky activations. Fig. 5 presents the gate values, the l_2 norms of features, and the final predictions on dev and training sets. The baseline shows different behavior on training and dev sets: the norms of gloss and sequence features have consistent tendencies on the training set but the correlations become weakened on the dev set. Baseline+VAC shows consistent behavior on both sets, which indicates the effectiveness of the proposed VAC.

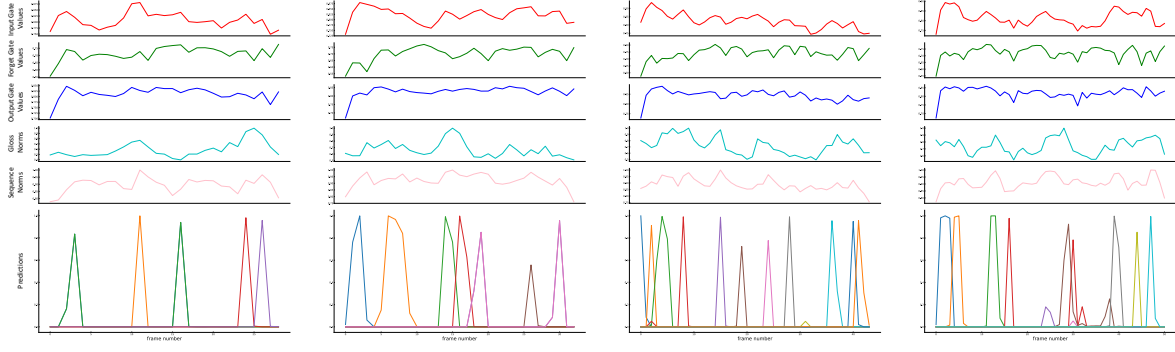
C.3. More Qualitative Recognition Results

We visualize more sequences in Fig. 6, and we can notice that the prediction results of two classifiers are not always

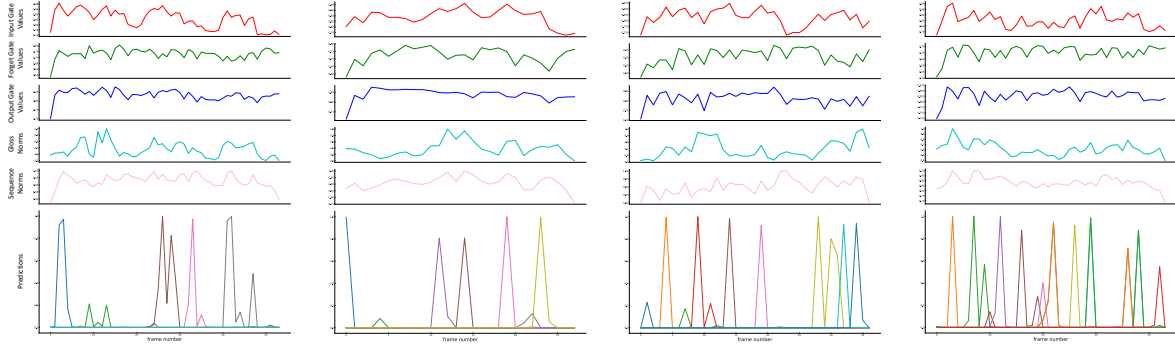
consistent. As shown in Fig. 6(a), the primary classifier can provide better results by incorporating more context information. However, the primary classifier may neglect visual information or predict wrong glosses, which gives worse results in some cases, as shown in Fig. 6(b). The proposed VAC attempts to make better use of visual and context information.

References

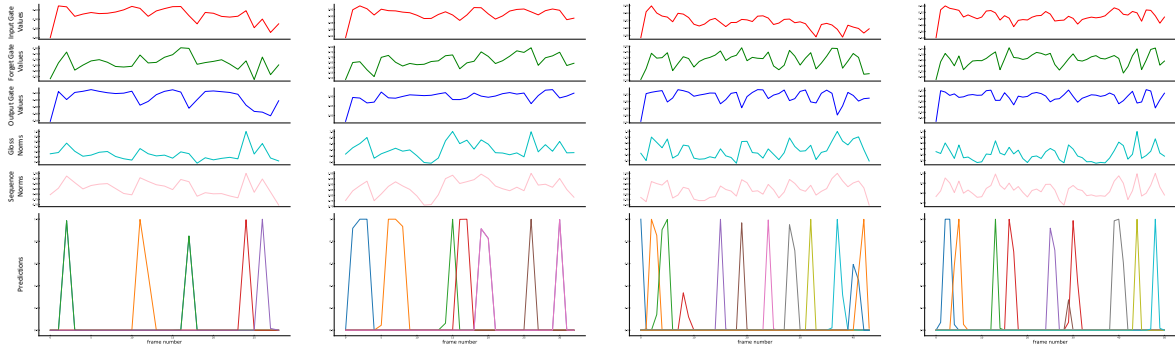
- [1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2
- [2] Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. Spatial-temporal multi-cue network for continuous sign language recognition. In *Proceedings of the Association for the Advancement of Artificial Intelligence*, pages 13009–13016, 2020. 2



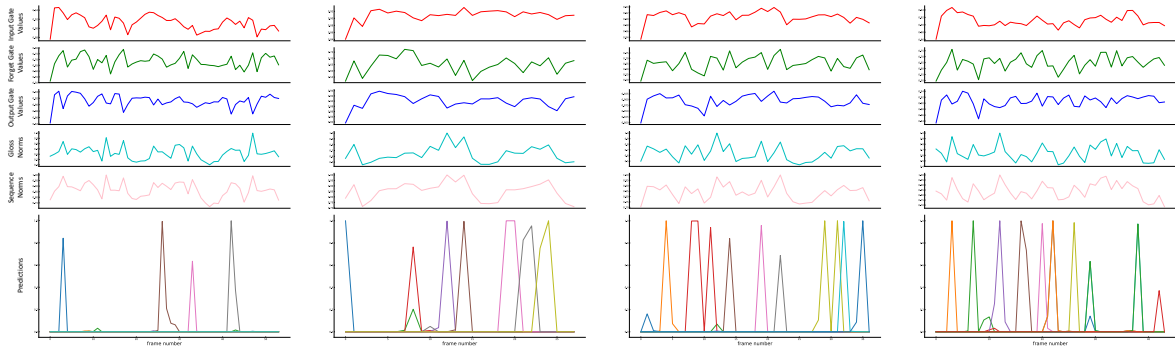
(a) Baseline on training set



(b) Baseline on dev set

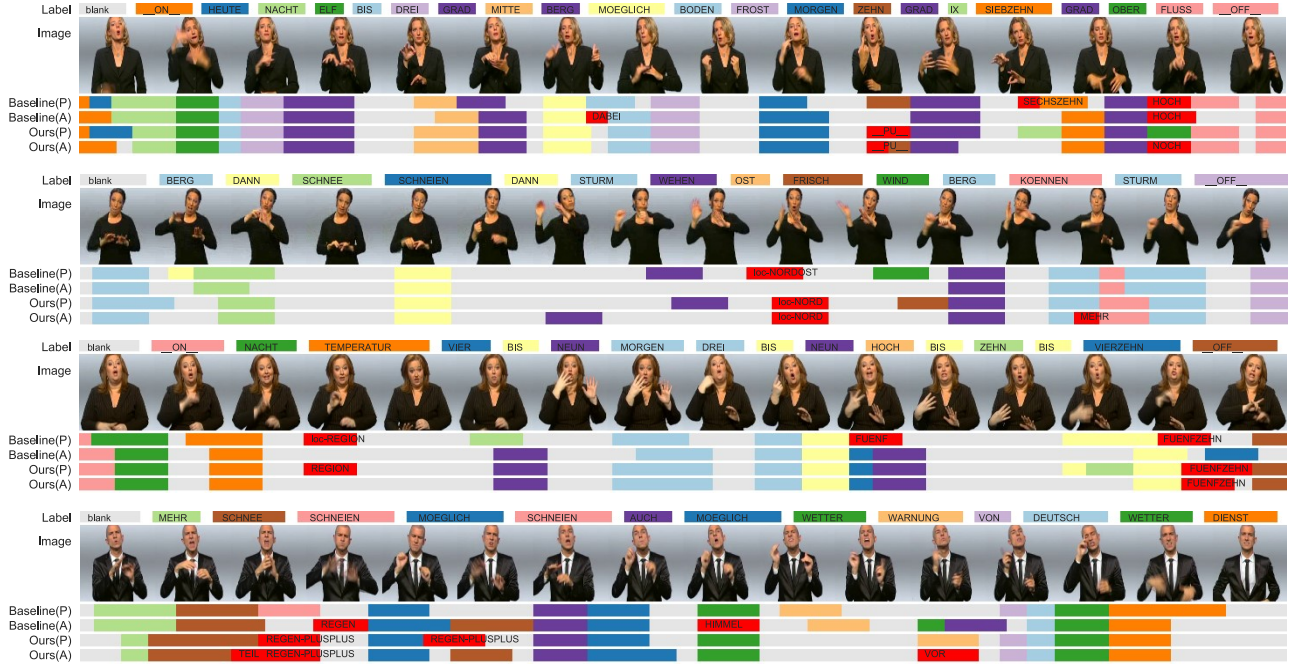


(c) Baseline+VAC on training set

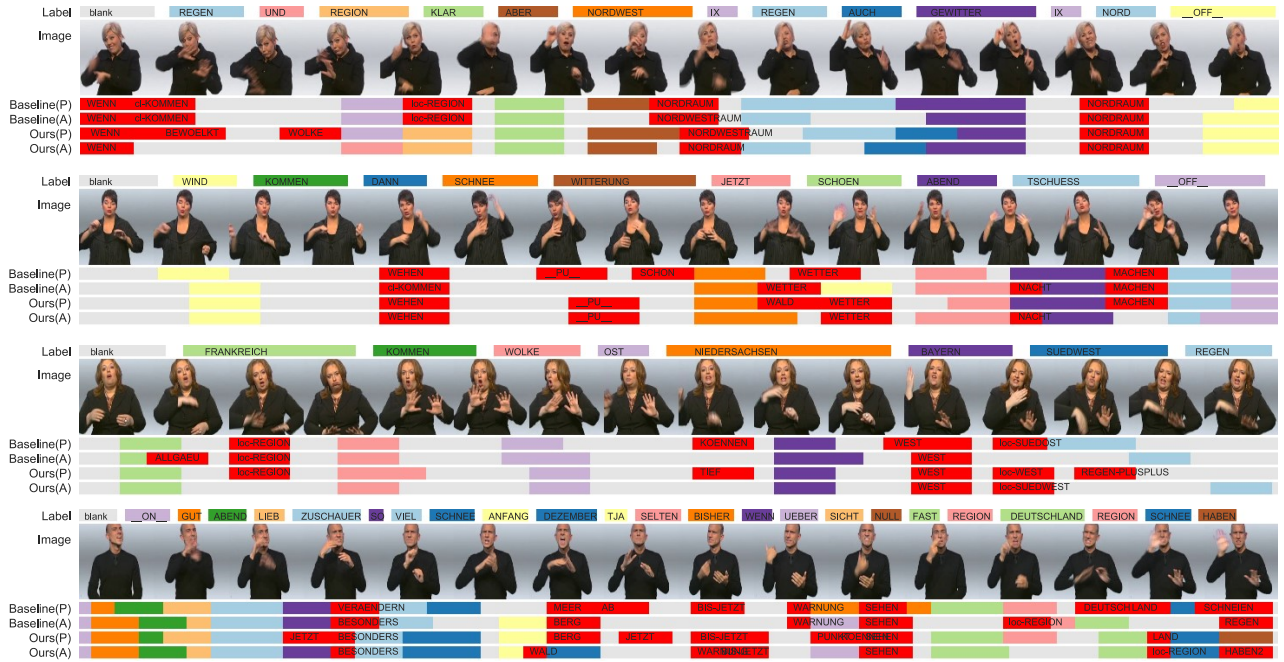


(d) Baseline+VAC on dev set

Figure 5. Visualization of the gate values, the l_2 norm of features and the final prediction on PHOENIX14.



(a) The primary classifier provides better results than the auxiliary.



(b) The auxiliary classifier provides better results than the primary.

Figure 6. Qualitative comparison among different network settings with examples from Dev set on PHOENIX14. Wrong recognition results (except deletion operations) are marked in red. The primary classifier and auxiliary classifier outputs are marked as (P) and (A).