Supplementary Materials

Isolated Sign Language Recognition in Low-Resolution Videos via Privileged Knowledge Distillation

Zeynep Gökce-Aker[1*], Yunus Can Bilge[2], Nazli Ikizler-Cinbis[3], Pinar Duygulu[3]

[1*]Department of Computer Engineering, Hacettepe University, Ankara, 06800, Turkey.
[2]Department of Information Systems, Hacettepe University, Ankara, 06800, Turkey.
[3]Department of Artificial Intelligence Engineering, Hacettepe University, Ankara, 06800, Turkey.

*Corresponding author(s). E-mail(s): zeynep.gokce@hacettepe.edu.tr;
Contributing authors: yunuscan.bilge@hacettepe.edu.tr;
nazli@cs.hacettepe.edu.tr; pinar@cs.hacettepe.edu.tr;

1 Introduction

In this document, we provide supplementary details to accompany our main paper, ensuring clarity and facilitating reproducibility. Specifically, this document covers:

- Qualitative comparison of input enhancement (super-resolution) strategies (Section 2)
- Complete training and evaluation settings for PKD-ISLR (Section 3)
- Architectural details of all baseline models (Section 4)

2 Qualitative Analysis of Input Enhancement Methods

The visual differences among image enhancement approaches are illustrated in Figure 1. Bilinear upsampling enlarges spatial resolution without introducing additional detail, producing blurry outputs that preserve the original low-frequency content. FlashVSR [1] generates visually plausible reconstructions at a global level; however, fine-grained sign-relevant regions, particularly finger articulation and facial expressions, may not be faithfully recovered, as SR models optimized for perceptual quality are not specifically trained on sign language content and may produce outputs inconsistent with the original gesture.

PKD-ISLR does not rely on HR reconstruction at inference time. The student is trained directly on 64×64 inputs, and bilinear upsampling is applied only to meet the input size requirement of the VideoMAE backbone. LR robustness is instead learned through knowledge distillation from a pose-aware HR teacher during training. This reduces exposure to the reconstruction artifacts and information loss associated with SR-based bridging.

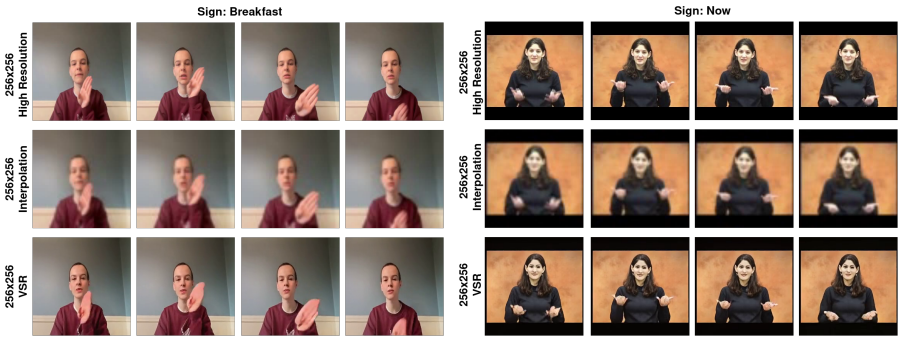

Fig. 1 Qualitative comparison of input enhancement strategies on samples from (left) the ASLCitizen test set (sign: "Breakfast") and (right) the WLASL test set (sign: "Now"). Each row in order shows: (a) native 256×256 high-resolution reference, (b) bilinear interpolation upsampling to 256×256, and (c) FlashVSR-based super-resolution output. While bilinear interpolation does not recover missing details, FlashVSR can introduce visual artifacts and may fail to faithfully restore fine-grained hand details and facial details.

3 Model Settings

Table 1 consolidates the full training and evaluation protocol used for PKD-ISLR. The settings cover the input pipeline (HR 256×256 teacher and LR 64×64 student streams, 16 uniformly sampled frames per clip), the data-augmentation pipeline (color jitter, horizontal flip, mixup, CutMix, label smoothing, and repeated augmentation), the ViT-Small encoder (12 transformer blocks, 384-dim, ∼21.9 M parameters) together with the ST-GCN pose encoder used by the teacher, and the shared AdamW optimization recipe (learning rate 10^{-4}, layer-wise LR decay 0.7, cosine schedule, weight decay

5×10^{-2}). The lower half of the table specifies the response-based knowledge-distillation loss used to transfer the teacher's predictions to the LR student, the dataset-specific schedule (60 epochs for ASLCitizen, 40 for WLASL), and the evaluation protocol (official train/val/test splits, single-crop top-1 accuracy).

4 Baseline Model Details

Table 2 summarizes the baseline models used to evaluate PKD-ISLR in isolated sign language recognition.

Swin [2] uses a hierarchical shifted-window attention design, Uniformer-S [3] combines 3D convolutions with self-attention, and VideoMAE-V2-S [4] is a Vision Transformer pretrained using masked video autoencoding on Kinetics-710 before fine-tuning on the target dataset. Together, these models represent three widely used transformer architectures for video action recognition. All transformer-based baselines are implemented using their officially released small variants.

VideoMAE-V2-S [4], SPOTER [5], SSL-STC [6] and Uni-Sign [7] utilize the pose information. VideoMAE-V2-S [4] and SPOTER [5] are trained from scratch on MediaPipe Holistic [8] landmarks and represent the classical GCN-based and Transformer-based baselines, respectively. SSL-STC [6] replaces MediaPipe with the more accurate HRNet-W48 wholebody estimator and benefits from self-supervised pretraining on American Sign Language (ASL) data. Uni-Sign [7] is structurally distinct from the rest: its backbone is the multilingual language model mT5-base [9], conditioned on RTMPose-m [10] keypoints, making it the only baseline that leverages a large-scale multilingual sign-language pretraining corpus.

Finally, our proposed **PKD-ISLR** adopts the VideoMAE-V2 (ViT-S) backbone but distills pose-feature knowledge from MediaPipe Holistic during training, combining the representational strength of RGB transformers with the sign-specific priors of pose-based models without incurring pose inference at test time.

Table 1 Complete training and evaluation settings for PKD-ISLR on WLASL100 and ASLCitizen100 where settings differ between the two datasets, both values are shown.

Input and Data	
Input resolution (HR teacher)	256×256 pixels
Input resolution (LR student)	64×64 pixels
Temporal length	16 frames (uniform sampling)
Tubelet size	16x16x2 (Width x Height x Time)
Data Augmentation	
Center crop	224×224 from 256×256
Horizontal flip	$p = 0.5$ (`flip_horizontal=1`; *disabled for student on WLASL100*)
Color jitter	$p = 0.5$
Label smoothing	0.1
Mixup	$\alpha = 0.8$ (*disabled for student on WLASL100 subset.*)
CutMix	$\lambda \sim \text{Beta}(1.0, 1.0)$ (*disabled for student on WLASL subsets*)
Repeated augmentation	2 augmented clips per video
Model Architecture	
Visual backbone	ViT-Small encoder from VideoMAE-V2, output feat dim=384
Pose encoder (teacher only)	ST-GCN on MediaPipe Holistic (T×27×2); output feat dim=256
Pose fusion module	Concatenation / Adaptive weighted average / Cross attention
Classification head	Linear ($384 \to C$); C is the number of gloss to predict.
Optimization (teacher and student share these)	
Optimizer	AdamW (β_1=0.9, β_2=0.999, ϵ=10^{-8})
Base learning rate	1×10^{-4}
Layer-wise LR decay	0.7
Weight decay	5×10^{-2}
LR schedule	Cosine annealing to `min_lr`= 10^{-6}
Warmup	5 epochs in teachers, 10 epochs in student, linear from 10^{-8}
Effective batch size	8
Drop-path	0.05 (student, teacher on ASLCitizen100) / 0.1 (student, teacher on WLASL100)
Loss (teacher)	Cross-entropy with label smoothing
Schedule	
Teacher epochs	60 (ASLCitizen100) / 40 (WLASL100)
Student total epochs	60 (ASLCitizen100) / 40 (WLASL100)
Optional two-stage student	Stage 1: 10 epochs encoder frozen; Stage 2: 25 epochs full fine-tune
Student Distillation	
Distillation type	Response-based (Hinton soft-target)
Soft-target loss	$-\sum \text{softmax}(z_T/\tau) \cdot \log \text{softmax}(z_S/\tau)\, \tau^2$
Temperature	$\tau = 3$
Combined loss	$\mathcal{L} = 0.75\, \mathcal{L}_{CE} + 0.25\, \mathcal{L}_{KD}$
Evaluation	
Splits	Official train/val/test for each dataset
Metric	Top-1 accuracy (%)
Checkpoint selection	Best validation Top-1

Table 2 Detailed architectural specifications of baseline models. For each method we report the backbone architecture, the pose-estimation model used to produce keypoint inputs (RGB-only models do not require this), and the pretraining dataset.

Method	Backbone	Pose Estimation Model	Pre-trained
Swin [2]	3D Swin Transformer	–	Kinetics-400
UniFormer [3]	3D Conv + MHRA Transformer	–	Kinetics-400
VideoMAE-V2-S [4]	ViT-S	–	Kinetics-710
ST-GCN [11]†	GCN	MediaPipe Holistic	None
SPOTER [5]	Transformer	MediaPipe Holistic	None
SSL-STC [6]	GCN+Transformer	HRNet-W48	ASL
Uni-Sign [7]	mT5-base (LLM)	RTMPose-m	Multilingual SL
PKD-ISLR	ViT-S	MediaPipe Holistic	Kinetics-710

† ST-GCN reimplemented using the proposed implementation in [11].

References

[1] Zhuang, J., Guo, S., Cai, X., Li, X., Liu, Y., Yuan, C., Xue, T.: Flashvsr: Towards real-time diffusion-based streaming video super resolution. In: CVPR, pp. 43482–43493 (2026)

[2] Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: CVPR, pp. 3202–3211 (2022)

[3] Li, K., Wang, Y., Gao, P., Song, G., Liu, Y., Li, H., Qiao, Y.: Uniformer: Unified transformer for efficient spatiotemporal representation learning. ICLR (2022)

[4] Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: Videomae v2: Scaling video masked autoencoders with dual masking. In: CVPR, pp. 14549–14560 (2023)

[5] Boháček, M., Hrúz, M.: Sign pose-based transformer for word-level sign language recognition. In: WACV, pp. 182–191 (2022)

[6] Zhao, W., Zhou, W., Hu, H., Wang, M., Li, H.: Self-supervised representation learning with spatial-temporal consistency for sign language recognition. IEEE Transactions on Image Processing (2024)

[7] Li, Z., Zhou, W., Zhao, W., Wu, K., Hu, H., Li, H.: Uni-sign: Toward unified sign language understanding at scale. ICLR (2025)

[8] Grishchenko, I., Bazarevsky, V.: Mediapipe holistic—simultaneous face, hand and pose prediction, on device. Google AI Blog. Dec (2020)

[9] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mT5: A massively multilingual pre-trained text-to-text transformer (2021). https://arxiv.org/abs/2010.11934

[10] Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., Chen, K.: Rtmpose: Real-time multi-person pose estimation based on mmpose. arXiv preprint arXiv:2303.07399 (2023)

[11] Desai, A., Berger, L., Minakov, F., Milano, N., Singh, C., Pumphrey, K., Ladner, R., Daumé III, H., Lu, A.X., Caselli, N., et al.: Asl citizen: a community-sourced dataset for advancing isolated sign language recognition. Advances in Neural Information Processing Systems **36** (2024)