Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Zeynep Gokce Aker

zeynep.gokce@hacettepe.edu.tr

Hacettepe University

Yunus Can Bilge

Hacettepe University

Nazli Ikizler-Cinbis

Hacettepe University

Pinar Duygulu

Hacettepe University

Research Article

Keywords: Sign language recognition, Knowledge distillation, Low resolution sign language recognition, Isolated sign language recognition

Posted Date: September 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10320065/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

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;

Abstract

We study isolated sign language recognition under low-resolution settings and propose a Privileged Knowledge Distillation framework for Isolated Sign Language Recognition (PKD-ISLR), that transfers discriminative supervision from a multimodal high-resolution teacher to a low-resolution student model. The teacher integrates high-resolution RGB appearance and pose information during training, while the student operates only on low-resolution RGB at inference. Extensive experiments on WLASL and ASLCitizen benchmarks show that the proposed PKD-ISLR framework outperforms strong baselines, including input-enhancement methods. These findings demonstrate that low-resolution sign language recognition is better addressed as a supervision transfer problem, highlighting the effectiveness of privileged multimodal distillation under visual degradation.

Keywords: Sign language recognition, Knowledge distillation, Low resolution sign language recognition, Isolated sign language recognition

1 Introduction

Isolated sign language recognition (ISLR) refers to the recognition of individual signs performed separately, one at a time. ISLR requires modeling fine-grained spatiotemporal patterns, including hand shapes, finger articulations, motion trajectories, body posture, and facial expressions. These subtle visual cues are linguistically critical in distinguishing visually similar signs. Recent progress in transformer-based architectures [1–3] together with large-scale pretraining strategies has significantly advanced ISLR performance. In addition, publicly available datasets such as WLASL [4] and ASLCitizen [5] have enabled more realistic benchmarking at larger vocabulary scales.

Despite these advances, deploying ISLR systems in real-world environments remains a challenge due to the low-resolution settings captured under bandwidth limitations, low-cost acquisition devices, and compression. On the contrary, most existing ISLR methods implicitly assume sufficiently high-resolution video inputs where fine-grained motion details remain clearly observable. Unlike generic action recognition, which can rely on coarse motion patterns, ISLR requires precise

observation of hand shapes, finger configurations, facial expressions, and subtle motion trajectories. All these discriminative cues which are essential for disambiguating visually similar signs become imperceptible in low resolution videos, making even linguistically distinct signs ambiguous to recognition systems. As illustrated in Figure 1, hand shapes, finger positions, and facial expressions, which are clear in the HR become indistinguishable in LR. Many visually similar signs differ only in subtle hand configurations or facial details, which vanish under spatial degradation in low resolution.

The loss of fine-grained visual information in low resolution directly affects sign language recognition and thus necessitates low resolution-specific modeling strategies tailored to the unique demands of sign language. Recently, methods dedicated to low-resolution action recognition have been proposed [6–10]. Nevertheless, low-resolution ISLR remains largely underexplored.

In this work, we address low-resolution isolated sign language recognition through privileged knowledge distillation (PKD). Specifically, we propose PKD-ISLR (Privileged Knowledge Distillation framework for Isolated Sign Language Recognition), a framework grounded in the Learning Using Privileged Information (LUPI) [11]. Our approach employs a multimodal teacher architecture that jointly models high-resolution RGB appearance and pose information during training, and transfers this knowledge to a low-resolution RGB-only student through knowledge distillation. In this way, the student benefits from pose-guided supervision during training while operating solely on low-resolution RGB videos at inference time.

We evaluate the proposed framework on the WLASL and ASLCitizen benchmarks [4, 5]. Experimental results demonstrate that our method consistently improves ISLR performance compared to baselines. Our contributions are summarized as follows:

- We propose PKD-ISLR, a privileged knowledge distillation framework which leverages pose information during training while maintaining RGB-only inference.
- We provide detailed analysis of teacher design and distillation objectives, offering insights into effective privileged supervision for low-resolution ISLR.

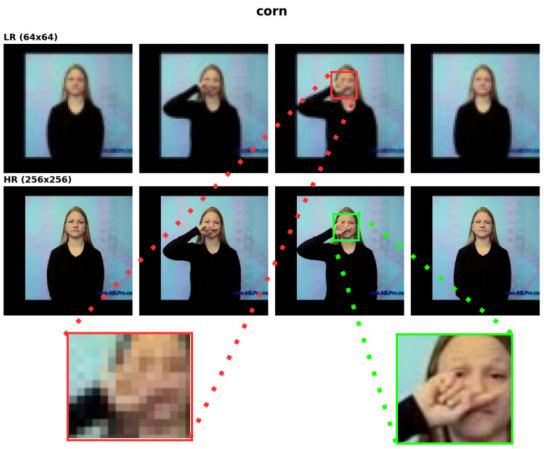


Fig. 1 Resolution degradation and loss of discriminative sign cues (sign: CORN). Top: downsampled low-resolution frames. Bottom: the corresponding native high-resolution frames. Although hand configurations, finger positions and articulations, and facial expressions are clearly visible in high-resolution, these fine-grained visual details critical for distinguishing signs become imperceptible at low resolution. This spatial degradation is the primary challenge addressed in this work.

- We present a systematic evaluation on two widely used benchmark datasets for low resolution setting.
- We conduct comprehensive ablation studies that demonstrate the importance of pose extraction quality, multimodal fusion strategies, and distillation objectives for low-resolution sign language recognition.

2 Related Work

Recent advances in SLR have been largely driven by deep learning and transformer-based architectures [12–14]. Transformer architectures have achieved strong performance by modeling long-range spatiotemporal dependencies in sign videos [1–3, 15]. Large-scale datasets WLASL [4], MS-ASL [16], and ASLCitizen [5] have enabled more realistic evaluation for large-vocabulary SLR.

Beyond RGB-based modeling, pose information has become an important complementary modality. Human pose estimation frameworks such as OpenPose [17], MediaPipe [18], and MMPose [19] are commonly used to extract skeletal representations, which are then processed with graph-based models such as ST-GCN [20] and GCN-BERT [21]. Several multimodal approaches

have demonstrated that combining RGB and pose improves recognition performance by jointly modeling appearance and structural motion cues [22–25]. Despite these advances, most existing SLR methods assume sufficiently HR video inputs where hand and finger details remain visible.

Low-resolution action recognition has been extensively studied where human actions often appear with severely degraded spatial details, such as surveillance, privacy-preserving vision systems, and long-distance video capture [10, 26]. Early studies addressed this problem through cross-resolution learning strategies. [6] introduced inverse super-resolution, where multiple transformed LR samples are generated from HR videos to improve robustness against resolution degradation. [7] and [8] further proposed coupled HR-LR architectures that jointly learn representations across resolutions by sharing parameters or feature spaces. Multi-Siamese embedding learning [26] aligned multiple degraded views of the same action within a shared embedding space.

Another line of work integrates super-resolution reconstruction into the recognition pipeline. [9] jointly optimize video super-resolution and action classification, while TinyVIRAT [10] evaluates recognition under realistic surveillance style low-resolution settings.

Recently, knowledge distillation has emerged as an effective solution for low-resolution video recognition. Cross-resolution distillation methods transfer knowledge from high-resolution teachers to low-resolution students, allowing low-resolution models to benefit from missing spatial details during training [27, 28]. Although these approaches demonstrate that cross-resolution supervision is highly effective for low-resolution video understanding, they mainly focus on generic human actions rather than the fine-grained motion patterns required for sign language recognition.

To the best of our knowledge, low-resolution ISLR has not been explicitly targeted in the literature, with the exception of TinySign [29]. TinySign combines super-resolution reconstruction with pose-based modeling and introduces a dedicated benchmark derived from ASLLVD [30] at 28×28 resolution. However, it depends on an explicit super-resolution stage and on accurate pose estimation at inference, which are unreliable on severely degraded inputs.

In contrast, our work investigates privileged pose supervision for ISLR, where pose information is available only during training while the student model operates solely on low-resolution RGB videos at inference, requiring neither super-resolution nor pose estimation at test time. Our proposed model can also be exploited on large-scale isolated sign language recognition datasets.

3 Methodology

Our framework, referred to as Privileged Knowledge Distillation framework for Isolated Sign Language Recognition (PKD-ISLR), consists of two main components: (i) a multi-modal teacher provides the spatio-temporal representation of the video that integrates appearance-based and pose-based information and (ii) a student model distilled from the multimodal teacher. Figure 2 depicts the overview of our model.

3.1 Multimodal Teacher

The teacher is a dual-stream architecture that processes high-resolution RGB video in the appearance stream and pose information in the pose stream in parallel and fuses their embeddings before classification. The appearance stream encodes a high-resolution video clip $X^{HR} \in \mathbb{R}^{T\times H'\times W'\times 3}$, where T denotes the number of frames, and H' and W' denote the frame height and width, respectively. A pretrained spatio-temporal encoder (ViT) extracts a clip-level appearance embedding $F_a \in \mathbb{R}^{d_a}$ where subscript a denotes the appearance stream and d_a is the dimensionality of the appearance feature representation.

The pose stream extracts J body landmarks (in 2D) per frame from the same high-resolution RGB video, yielding pose sequences $P \in \mathbb{R}^{T\times J\times 2}$. We use an off-the-shelf 2D pose extractor for this purpose, specifically MediaPipe Holistic [31]. These landmark sequences are encoded by a spatio-temporal graph convolutional network. The pose stream produces a clip-level pose embedding $F_p \in \mathbb{R}^{d_p}$ where subscript p denotes the pose stream and d_p is the dimensionality of the pose feature representation.

As a result, the appearance stream produces a clip-level appearance embedding F_a, while the

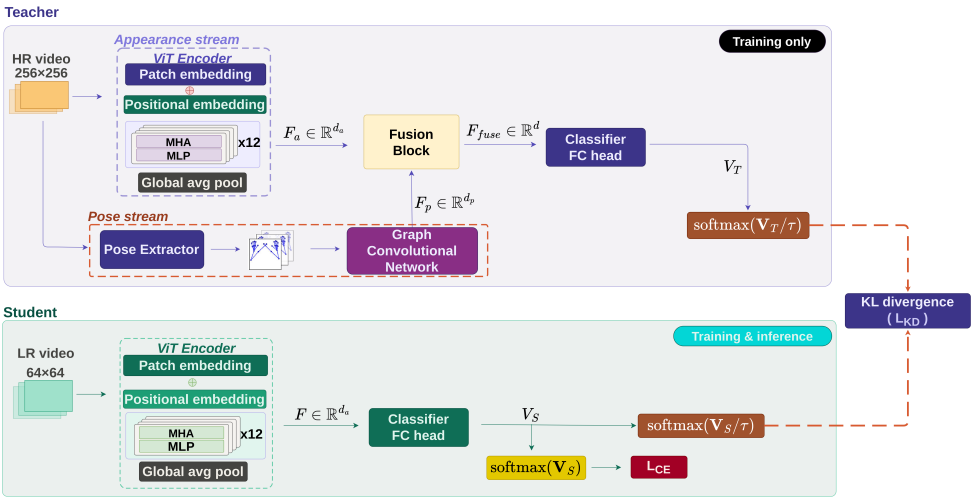


Fig. 2 The proposed PKD-ISLR framework for low-resolution ISLR. During training, a dual-stream teacher encodes high-resolution RGB appearance and pose, fuses them, and transfers its softened predictions to the student through KL-based response distillation (L_{KD}), combined with cross-entropy on ground-truth glosses (L_{CE}). The student is a single-stream RGB encoder operating on LR inputs and is the only component used at inference.

pose stream produces a pose embedding F_p in different dimensionalities. To integrate appearance and structural information, we explore three multimodal fusion strategies: concatenation, adaptive weighted average fusion, and cross-attention (analyzed in Section 4.3). The resulting fused embedding, F_{fuse}, is passed through a fully connected classification head to produce the teacher logits $\mathbf{V}_T \in \mathbb{R}^C$, where C is the number of glosses.

Concatenation fusion.

The appearance feature $F_a \in \mathbb{R}^{d_a}$ and the pose feature $F_p \in \mathbb{R}^{d_p}$ reside in spaces of different dimensionality. Each modality is therefore first projected into a common d-dimensional (d=384) embedding space via a linear layer followed by a ReLU non-linearity and obtained $\tilde{F}_a$ and $\tilde{F}_p$ for appearance and pose features, respectively. These are then fused as following:

$$F_{\text{fuse}} = \text{ReLU}\big(W_{\text{fuse}}\,[\tilde{F}_a; \tilde{F}_p]\big) \in \mathbb{R}^d, \qquad (1)$$

where $W_{\text{fuse}} \in \mathbb{R}^{d\times 2d}$ is a learnable projection matrix mapping the concatenated representation back to the d-dimensional embedding space. The fused representation (F_{fuse}) is then passed through the classifier.

Adaptive weighted average fusion.

As a second fusion method, we investigate to fuse the appearance feature F_a and the pose feature F_p using an adaptive weighted fusion approach inspired by the GMU [32]. Both embeddings are first projected into a shared d-dimensional latent space via tanh-activated linear layers:

$$\hat{f}_a = \tanh(F_a W_a), \qquad \hat{f}_p = \tanh(F_p W_p), \qquad (2)$$

where $W_a \in \mathbb{R}^{d_a\times d}$ and $W_p \in \mathbb{R}^{d_p\times d}$ are learnable projection matrices. A scalar importance weight is then estimated for each modality by passing its projected feature through a dedicated fully-connected layer followed by a sigmoid activation:

$$\alpha_a = \sigma(\hat{f}_a\, w_a), \qquad \alpha_p = \sigma(\hat{f}_p\, w_p), \qquad (3)$$

where $w_a, w_p \in \mathbb{R}^{d\times 1}$ are learnable weight vectors. The two scalars are normalized as follows:

$$\tilde{\alpha}_a = \frac{\alpha_a}{\alpha_a + \alpha_p}, \qquad \tilde{\alpha}_p = \frac{\alpha_p}{\alpha_a + \alpha_p}, \qquad (4)$$

ensuring $\tilde{\alpha}_a + \tilde{\alpha}_p = 1$ by construction. The fused representation is then computed as the adapted weighted average of the two projected features:

$$F_{\text{fused}} = \tilde{\alpha}_a\, \hat{f}_a + \tilde{\alpha}_p\, \hat{f}_p \in \mathbb{R}^d. \qquad (5)$$

This mechanism allows the model to dynamically adjust the relative contribution of the appearance and pose modalities on a per-sample basis, without imposing a fixed modality preference.

Cross-attention fusion.

In the last fusion method, we combine the appearance and pose embeddings using a symmetric cross-attention block. Each modality is first linearly projected into a shared d-dimension of query (Q), key (K) and value (V) representations:

$$\begin{aligned} Q_a &= F_a W_a^Q, \quad K_a = F_a W_a^K, \quad V_a = F_a W_a^V, \\ Q_p &= F_p W_p^Q, \quad K_p = F_p W_p^K, \quad V_p = F_p W_p^V, \end{aligned} \tag{6}$$

where $W_a^{\{Q,K,V\}} \in \mathbb{R}^{d_a \times d}$ and $W_p^{\{Q,K,V\}} \in \mathbb{R}^{d_p \times d}$ are learnable projection matrices where $d = 256$. Cross-attention is then computed in both directions:

$$\tilde{F}_a^{\text{att}} = \sigma\big(Q_a K_p^\top\big)\, V_p, \quad \tilde{F}_p^{\text{att}} = \sigma\big(Q_p K_a^\top\big)\, V_a \ \in \mathbb{R}^d \tag{7}$$

Here, $\tilde{F}_a^{\text{att}}$ is a pose-aware appearance representation, and $\tilde{F}_p^{\text{att}}$ is an appearance-aware pose representation. These representations are re-projected original dimensions before fusion operation by a modality-specific output projection $W_{a,p}^O \in \mathbb{R}^{d_{a,p} \times h}$:

$$F_a^{\text{att}} = W_a^O\, \tilde{F}_a^{\text{att}} \in \mathbb{R}^{d_a}, \qquad F_p^{\text{att}} = W_p^O\, \tilde{F}_p^{\text{att}} \in \mathbb{R}^{d_p}. \tag{8}$$

These cross-attended embeddings are concatenated and given to a single d_a-dimensional representation through a final linear layer:

$$F_{\text{fuse}} = W_f\, [\, F_a^{\text{att}} ;\, F_p^{\text{att}} \,] + b_f \ \in \ \mathbb{R}^{d_a}, \tag{9}$$

with $W_f \in \mathbb{R}^{d_a \times (d_a + d_p)}$ and the final fused embedding that passed through the classification head.

3.2 Privileged Knowledge Distillation

The student is a single-stream RGB encoder that observes only low-resolution video $X^{LR} \in \mathbb{R}^{T \times H \times W \times 3}$ processed by transformer-based spatio-temporal network (ViT). The student outputs class logits $\mathbf{V}_S \in \mathbb{R}^C$.

The asymmetry between teacher and student inputs is what makes this a *privileged* setting: during training, the teacher observes the richer input pair (X^{HR}, P) that includes both high-resolution video and pose, whereas the student observes only X^{LR}, the reduced resolution version of X^{HR}.

Response-based distillation.

We transfer knowledge from the teacher to the student by aligning their softened class distributions, following the idea of [33]. Using a temperature parameter τ, the softened predictions are computed as:

$$p_T = \text{softmax}(\mathbf{V}_T/\tau)\,, \quad p_S = \text{softmax}(\mathbf{V}_S/\tau)\,, \tag{10}$$

The distillation loss L_{KD} is the Kullback–Leibler (KL) divergence [33] between these temperature-scaled outputs:

$$L_{\text{KD}} = \tau^2 \cdot \text{KL}(p_T \,\|\, p_S)\,. \tag{11}$$

The final training objective combines the distillation loss with the standard cross-entropy loss:

$$L = (1-\alpha)\, L_{CE} + \alpha\, L_{\text{KD}}, \tag{12}$$

where L_{CE} is the cross-entropy loss computed over the ground-truth gloss labels, and α controls the balance between teacher supervision and ground-truth supervision.

4 Experiments

Datasets: We evaluate our method on two large-scale isolated sign language recognition benchmarks:WLASL [4] and ASLCitizen [5]. Since no standard low-resolution benchmark exists in this domain, we construct low-resolution (LR, 64 × 64) versions of the datasets by downsampling all videos via bilinear interpolation, and refer to the native videos (256 × 256) as high-resolution (HR). We report results on WLASL100 and ASLCitizen100 for small-vocabulary evaluation, and WLASL2000 and ASLCitizen2731 for large-vocabulary evaluation. Dataset statistics are given in Table 1.

Implementation Details: We use the VideoMAEV2 encoder [34] small variant (ViT small) as the backbone for both the teacher

Table 1 Summary statistics of our proposed low-resolution benchmarks employed in our experiments.

Dataset	Classes	Train	Val	Test	Video
WLASL100	100	1442	338	258	2038
WLASL2000	2000	14296	3920	2879	21095
ASLCitizen100	100	1707	393	1444	3544
ASLCitizen2731	2731	40154	10304	32941	83399

appearance stream and the student network. We initialize this backbone with officially released pretrained weights [1]. This initialization provides strong prior spatiotemporal representations learned from large-scale video data. The architecture consists of 12 transformer encoder blocks, each with 6 attention heads. The patch embedding dimension is 384, computed from 16×16 spatial patches.

The teacher's appearance stream processes high-resolution RGB videos $X_{HR} \in \mathbb{R}^{T\times H'\times W'\times 3}$ with $H' = W' = 256$ and temporal length $T = 16$. The student network receives low-resolution RGB videos $X_{LR} \in \mathbb{R}^{T\times H\times W\times 3}$ with $H = W = 64$. To satisfy VideoMAE-V2's input requirements, low-resolution videos are upsampled to HR via bilinear interpolation, preserving low-resolution content statistics while conforming to the backbone's input format. This upsampling does not introduce additional spatial information recovery. The VideoMAE-V2 encoder produces a global clip-level appearance embedding $F_a \in \mathbb{R}^{d_a=384}$.

Teacher pose stream. Pose keypoints are extracted using MediaPipe Holistic [31] that provides 27 pose landmarks including hand, face, and upper-body joints. The resulting pose sequences $P \in \mathbb{R}^{T\times 27\times 2}$ (where $J = 27$ landmarks and each landmark has (x, y) coordinates) are processed using Spatial-Temporal Graph Convolutional Networks (ST-GCN) model following the configuration of [5]. The pose stream produces clip-level pose embeddings $F_p \in \mathbb{R}^{d_p=256}$.

Training and evaluation. The multimodal teacher is trained in two stages. First, the appearance encoder is fine-tuned on high-resolution RGB videos. Then, the pose stream is introduced and fused with the appearance features, while the pose encoder remains frozen.

[1]The VideoMAE-V2 (distilled ViT-S variant) weights are available in the official repository.

The student shares the same backbone architecture and is trained using privileged knowledge distillation using with temperature $\tau = 3$ and weight $\alpha = 0.25$. The teacher is frozen during student training. At inference, only the student is used, taking LR RGB video clips as input. We report Top-1 accuracy on the test split of each dataset for recognition performance evaluation.

4.1 Comparisons with SOTA

Two groups of baseline methods are evaluated on low-resolution WLASL100 and ASLCitizen100 datasets: (1) three well-known transformer-based action recognition models (Swin [35], UniFormer [36], VideoMAE-V2 [34] with small versions) originally designed for general video understanding, (2) four sign-language-specific models (ST-GCN [20], SPOTER [37], SSL-STC [38], Uni-Sign [3]), all of which rely on pose keypoint extracted from the input video. All models are finetuned using their official pre-trained weights and evaluated on low-resolution inputs; results are reported in Table 2.

In these experiments, action recognition based models receive only RGB video as input and rely on appearance information. At low resolutions, the fine-grained spatial details essential for sign language recognition, such as finger configuration and handshape, are largely lost, limiting the effectiveness of this modality when used alone. Pose-based models, by contrast, take skeletal keypoints as input. However, since pose estimation itself degrades at low resolutions, the keypoints extracted are often inaccurate.

Our proposed PKD-ISLR addresses both of these limitations. During inference, the student model takes only low-resolution RGB as input. During training, however, it is guided by a multimodal teacher that has access to both high-resolution RGB and accurate pose information, allowing the student to learn richer representations despite the degraded input. PKD-ISLR achieves the best performance on both datasets, outperforming VideoMAE-V2 [34], the strongest RGB-based baseline, and Uni-Sign [3], the strongest pose-based baseline (based on multilingual language model), on both WLASL100 and ASLCitizen100. Specifically, it achieves accuracy improvements of 7.34% and 6.72% over

Table 2 Comparison of isolated sign language recognition performance (Top-1 accuracy (%)).

Model	Accuracy	
	WLASL100	ASLCitizen100
Swin-S [35]	49.61	72.71
UniFormer-S [36]	48.45	72.58
VideoMAE-V2-S [34]	69.37	75.69
ST-GCN [20]	27.13	40.30
SPOTER [37]	27.52	58.38
SSL-STC [38]	46.90	52.91
Uni-Sign[2] [3]	68.99	74.03
PKD-ISLR (Ours)	**76.74**	**82.41**

Table 3 Computational efficiency and model size comparison.

Model	Params	Size	GFLOPs	Latency
Swin-S [35]	49.59	189.16	83.1	3641.61
UniFormer-S [36]	21.25	81.05	41.8	2425.79
VideoMAE-V2-S [34]	21.92	83.61	57.0	436.63
ST-GCN [20]	8.07	30.79	19.54	329.92
SPOTER [37]	231.87	884.53	0.73	31.23
SSL-STC [38]	112.92	430.74	22.62	591.01
Uni-Sign [3]	605.7	2310.57	7.53	177.37
PKD-ISLR (Ours)	21.92	83.61	56.95	538.51

VideoMAE-V2-S on WLASL100 and ASLCitizen100, respectively.

Computational Efficiency. Table 3 compares the computational efficiencies. All measurements are performed per clip with $T = 16$ on CPU. For pose-based methods, pose extraction cost is included in the results. PKD-ISLR achieves improvement over the RGB-based VideoMAE-V2-S baseline without introducing additional inference cost.

PKD-ISLR and VideoMAE-V2-S share the same student backbone and therefore have identical inference complexity ($\sim$21.92M parameters and $\sim$57G FLOPs). However, PKD-ISLR achieves significantly higher recognition accuracy through privileged knowledge distillation during training, while requiring only low-resolution RGB input during inference. Compared with most baseline methods, PKD-ISLR provides a favorable trade-off between recognition performance and computational cost. Although several pose-based methods achieve lower inference latency, they rely on additional pose estimation pipelines and generally obtain lower recognition accuracy under low-resolution conditions.

4.2 Comparison to input enhancement methods

A straightforward strategy for handling low-resolution (LR) sign language videos is to directly reuse existing high-resolution (HR) action recognition models without modifying their architectures or training procedures. In this setting, LR inputs are simply restored or upsampled at test time to match the spatial resolution expected by pre-trained HR models. To investigate whether this strategy alone is sufficient to address the low resolution sign language recognition problem, we compare PKD-ISLR against two such configurations on WLASL100 and ASLCitizen100.

We consider three conditions in this experimental setup. The first is a native HR reference, where models are trained and tested on HR inputs. The second is *bilinear bridging* (LR+UPS), where LR inputs are upsampled to HR using bilinear interpolation before inference. The third is *Video Super Resolution (VSR) bridging* (LR+VSR), where LR inputs are first enhanced using FlashVSR [39] before being passed to an HR-trained model.

As shown in Table 4, bilinear bridging yields limited improvements because it enlarges the spatial size without recovering missing visual information. Video Super Resolution (VSR) bridging performs better than naive interpolation, indicating that learned super-resolution can partially recover useful spatial structure. However, a substantial performance gap still remains relative to native HR evaluation. This suggests that LR degradation is not fully eliminated even after super resolution reconstruction.

PKD-ISLR performs competitively against both bridging strategies. Combined with the results in Table 2, this suggests that test-time input restoration alone is not sufficient, and that building LR robustness into the training procedure offers a complementary advantage.

Table 4 Effect of test-time input enhancement strategies on low-resolution isolated sign language recognition performance (Top-1 %) on WLASL100 and ASLCitizen100. PKD-ISLR (HR) results corresponds to teacher accuracy while PKD-ISLR (LR+UPS) corresponds to our proposed knowledge distillation based student model results.

	WLASL100			ASLCitizen100		
Model	**HR**	**LR+UPS**	**LR+VSR**	**HR**	**LR+UPS**	**LR+VSR**
Swin-S [35]	58.91	31.78	54.65	77.35	50.21	72.58
Uniformer-S [36]	60.08	32.95	53.10	76.25	53.95	73.55
VideoMAE-V2-S [34]	76.36	52.71	70.54	83.38	67.93	80.40
ST-GCN [5]	64.34	38.37	48.06	65.58	35.04	61.22
SSL-STC [38]	77.52	21.71	61.24	81.37	45.01	78.88
SPOTER [37]	77.52	36.43	52.71	70.84	38.23	62.60
Uni-Sign [3]	82.17	71.32	**74.81**	85.04	80.47	**81.16**
PKD-ISLR (Ours)	**83.72**	**76.74**	—	**89.47**	**82.41**	—

Table 5 Effect of pose extraction model on WLASL100 (Top-1 %). All experiments use HR inputs and the same global graph representation.

Pose Extraction Model	**Top-1 (%)**
MediaPipe Holistic [31]	**74.80**
RTMPose-m [40]	56.58

4.3 Ablation Studies

Impact of pose extraction quality on pose-based supervision. The pose stream serves as the privileged stream through which structural information is transferred to the teacher model. Therefore, the quality of the pose extractor directly influences the supervision signal later distilled into the LR student. Since inaccurate or unstable pose estimation can limit the quality of the distilled representation, we analyze the impact of different pose extraction models on WLASL100 using HR inputs.

We compare two recent pose estimation models: MediaPipe Holistic [31] and RTMPose-m [40]. To isolate the effect of pose extraction quality, the experiments use global graph representation containing the detected joints. An ST-GCN classifier is trained on each pose configuration, and Top-1 accuracy is reported.

Table 5 shows that MediaPipe Holistic substantially outperforms RTMPose-m, achieving 74.80% Top-1 accuracy compared to 56.58%. These results suggest that robust and temporally consistent landmark extraction is critical for constructing effective pose-based supervision in isolated sign language recognition. Thus, MediaPipe Holistic is adopted as the pose extractor in all subsequent experiments.

Impact of fusion strategy in the multimodal teacher model. After determining the pose extraction pipeline, the next question is how structural pose information should be integrated with appearance features in the teacher model. Since the teacher is responsible for generating the privileged supervision later distilled into the LR student, the fusion strategy directly affects the quality of the transferred representation.

We evaluate three multimodal fusion strategies for combining pose and RGB appearance embeddings. *Concatenation* directly stacks the appearance and pose embeddings before classification, *adapted weighted average* combines the two modalities using learnable scalar coefficients, and *cross-attention* corresponds to the bidirectional attention-based fusion module, where each modality dynamically attends to the other to model cross-modal interactions. These modules are described in detail in Section 3.1.

Table 6 reports performances of all teacher configurations across WLASL and ASLCitizen subsets. Two main observations emerge from the results. First, all multimodal teacher variants outperform both the pose-only baseline (*Pose*) and the RGB-only teacher baseline (*TBase*), demonstrating that pose and appearance information provide complementary cues for sign recognition. This confirms the importance of integrating structural and visual information within the privileged teacher.

Second, no single fusion strategy consistently dominates across all datasets. Concatenation achieves the best performance on WLASL100, fusion of adapted weighted average performs best on ASLCitizen100, and cross-attention achieves the highest accuracy on WLASL2000. We

attribute this behavior to the differing characteristics of the datasets, particularly the balance between vocabulary size, intra-class variation, and the number of training samples per class.

Impact of distillation objective. Our PKD-ISLR model uses response-based knowledge distillation as its default training objective (Section 3.2). We compare it with a feature-based alternative to evaluate how the distillation target affects low-resolution student learning.

The features are extracted from fully connected layer after the final transformer block of the teacher and student models, respectively. In the feature-based setting, the response-based loss L_{KD} is replaced with a cosine similarity-based feature alignment objective which directly encourages the student and teacher features to be directionally aligned in the representation space.

Table 7 reports results under different teachers and datasets. Response-based distillation consistently achieves better and more stable performance than feature-based distillation across all settings. In particular, it reaches 75.58% on WLASL100 and 82.41% on ASLCitizen100.

Feature-based distillation shows more variation across teacher architectures and datasets. In some cases it performs competitively, but overall it is less stable. We attribute this to the difficulty of aligning high-dimensional feature spaces under strong low-resolution degradation. Overall, response-based distillation provides a simpler and more robust training signal for low-resolution sign language recognition.

5 Conclusion

We propose PKD-ISLR, a privileged knowledge distillation framework that transfers structural knowledge from a multimodal HR teacher to an efficient LR appearance-only student. Through extensive experiments, we demonstrate that existing HR-oriented action recognition models cannot fully overcome the LR sign recognition problem through direct LR training, naive upsampling, or super-resolution-based input restoration. Although video super-resolution improves performance compared to bilinear interpolation, a substantial gap relative to native HR performance remains. These findings indicate that LR degradation persists even after input reconstruction and motivate the need for LR-specific learning strategies.

Our analysis further shows that the quality of privileged supervision strongly depends on teacher design. Robust pose extraction significantly improves structural supervision quality, multimodal fusion consistently outperforms unimodal teachers. Among different distillation objectives, response-based distillation provides the most stable and effective supervision under severe spatial degradation, outperforming feature-based alternatives across multiple datasets and teacher configurations.

Overall, the proposed PKD-ISLR framework consistently improves LR isolated sign language recognition while maintaining a practical inference-time pipeline that relies only on LR RGB input. By leveraging pose information exclusively during training, the proposed approach avoids the computational overhead and reliability issues associated with pose estimation or super-resolution during deployment. Beyond the proposed framework itself, this work provides empirical insights into how privileged multimodal supervision can be effectively utilized for recognition tasks operating under severe visual degradation and resource-constrained conditions.

Supplementary information. The supplementary material available at *supplementary.pdf*

Author Contributions. Zeynep Gokce-Aker: Conceptualization, Methodology, Software, Writing - original draft. Yunus Can Bilge: Supervision, Methodology, Writing - review & editing. Nazli Ikizler-Cinbis: Supervision, Methodology, Writing - review & editing. Pinar Duygulu: Supervision, Methodology, Writing - review & editing.

Funding . No funding was received for conducting this study.

Acknowledgements. The numerical calculations reported in this paper were partially performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

References

[1] Hu, H., Zhao, W., Zhou, W., Wang, Y.,

Table 6 Performance comparison of teacher models with different multimodal fusion strategies across WLASL and ASLCitizen subsets (Top-1 %). HR RGB inputs and pose features are jointly used in multimodal teacher configurations.

Model	Fusion Strategy	WLASL100	WLASL2000	ASLCitizen100	ASLCitizen2731
Pose	—	74.80	45.33	77.66	56.95
TBase	—	77.13	48.00	83.93	74.73
$TPose_{\text{Concat}}$	Concatenation	**83.72**	51.58	89.47	76.81
$TPose_{\text{Average}}$	Adaptive weighted average	81.01	49.46	**92.11**	73.19
$TPose_{\text{Attention}}$	Cross-Attention	82.17	**51.82**	90.93	**76.85**

Table 7 Comparison of response-based and feature-based distillation objectives under LR input across WLASL and ASLCitizen subsets. Response-based distillation ($\text{S-PKD}_{\text{response}}$) consistently provides stronger and more stable performance compared to feature-based distillation ($\text{S-PKD}_{\text{feat}}$)across different teacher configurations.

Model	Teacher	WLASL100	WLASL2000	ASLCitizen100	ASLCitizen2731
SBase	—	69.37	40.95	75.69	69.71
$\text{S-PKD}_{\text{response}}$	TBase	73.64	43.45	81.99	75.81
	$\text{TPose}_{\text{Concat}}$	**76.74**	44.95	**82.41**	**77.42**
	$\text{TPose}_{\text{Average}}$	72.48	44.49	82.06	74.33
	$\text{TPose}_{\text{Attention}}$	75.58	45.09	**82.41**	76.96
$\text{S-PKD}_{\text{feat}}$	TBase	70.54	43.52	76.87	71.08
	$\text{TPose}_{\text{Concat}}$	75.19	45.05	72.16	72.43
	$\text{TPose}_{\text{Average}}$	74.41	**45.47**	78.81	73.88
	$\text{TPose}_{\text{Attention}}$	74.41	44.84	79.36	74.01

Li, H.: Signbert: pre-training of hand-model-aware representation for sign language recognition. In: ICCV, pp. 11087–11096 (2021)

[2] Zhao, W., Hu, H., Zhou, W., Mao, Y., Wang, M., Li, H.: Masa: Motion-aware masked autoencoder with semantic alignment for sign language recognition. IEEE Transactions on Circuits and Systems for Video Technology **34**(11), 10793–10804 (2024)

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

[4] Li, D., Rodriguez, C., Yu, X., Li, H.: Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In: WACV, pp. 1459–1469 (2020)

[5] 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)

[6] Ryoo, M., Rothrock, B., Fleming, C., Yang, H.J.: Privacy-preserving human activity recognition from extreme low resolution. In: AAAI, vol. 31 (2017)

[7] Chen, J., Wu, J., Konrad, J., Ishwar, P.: Semi-coupled two-stream fusion convnets for action recognition at extremely low resolutions. In: WACV, pp. 139–147 (2017). IEEE

[8] Xu, M., Sharghi, A., Chen, X., Crandall, D.: Fully-coupled two-stream spatiotemporal networks for extremely low resolution action recognition. arXiv preprint (2018)

[9] Zhang, H., Liu, D., Xiong, Z.: Two-stream action recognition-oriented video super-resolution. In: ICCV, pp. 8799–8808 (2019)

[10] Demir, U., Rawat, Y.S., Shah, M.: Tinyvirat: Low-resolution video action recognition. In: ICPR, pp. 7387–7394 (2021). IEEE

[11] Vapnik, V., Izmailov, R.: Learning using privileged information: similarity control and knowledge transfer. The Journal of Machine Learning Research **16**(1), 2023–2049 (2015)

[12] Koller, O., Ney, H., Bowden, R.: Deep sign: Hybrid cnn-hmm for continuous sign language recognition. In: BMVC (2016)

[13] Camgoz, N.C., Hadfield, S., Koller, O., Bowden, R.: Subunets: End-to-end hand shape and continuous sign language recognition. In: ICCV, pp. 3056–3065 (2017)

[14] Cui, R., Liu, H., Zhang, C.: Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In: CVPR, pp. 1610–1618 (2017)

[15] Camgoz, N.C., Saunders, B., Athar, A., Bowden, R.: Sign language transformers: Joint end-to-end sign language recognition and translation. In: CVPR, pp. 10023–10033 (2020)

[16] Joze, H.R.V., Koller, O.: Ms-asl: A large-scale data set and benchmark for understanding american sign language. In: BMVC (2019)

[17] Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., Sheikh, Y.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE transactions on pattern analysis and machine intelligence **43**(1), 172–186 (2019)

[18] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M.G., Lee, J., et al.: Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172 (2019)

[19] OpenMMLab Pose Estimation Toolbox and Benchmark. https://github.com/open-mmlab/mmpose (2020)

[20] Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: AAAI, vol. 32 (2018)

[21] Tunga, A., Nuthalapati, S.V., Wachs, J.: Pose-based sign language recognition using gcn and bert. In: WACV, pp. 31–40 (2021)

[22] Gruber, I., Krnoul, Z., Hrúz, M., Kanis, J., Bohacek, M.: Mutual support of data modalities in the task of sign language recognition. In: CVPR, pp. 3424–3433 (2021)

[23] Zhou, H., Zhou, W., Zhou, Y., Li, H.: Spatial-temporal multi-cue network for sign language recognition and translation. IEEE Transactions on Multimedia **24**, 768–779 (2021)

[24] Ren, Y., Li, H., Li, Y., Pu, J., Pu, X., Jing, S., Jin, P., He, L.: Multi-modal isolated sign language recognition based on self-paced learning. Expert Systems with Applications **291**, 128340 (2025)

[25] Chen, Y., Zuo, R., Wei, F., Wu, Y., Liu, S., Mak, B.: Two-stream network for sign language recognition and translation. NeurIPS **35**, 17043–17056 (2022)

[26] Ryoo, M., Kim, K., Yang, H.: Extreme low resolution activity recognition with multi-siamese embedding learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)

[27] Chen, B., Qiao, Y., Wang, Y.: Low-resolution action recognition for tiny actions challenge. arXiv preprint arXiv:2209.14711 (2022)

[28] Bai, Y., Zou, Q., Chen, X., Li, L., Ding, Z., Chen, L.: Extreme low-resolution action recognition with confident spatial-temporal attention transfer. International Journal of Computer Vision **131**(6), 1550–1565 (2023)

[29] Hüseyinoğlu, A., Bilge, F.A., Bilge, Y.C., Ikizler-Cinbis, N.: Tinysign: sign language recognition in low resolution settings. SIVP **18**(10), 6881–6890 (2024)

[30] Neidle, C., Thangali, A., Sclaroff, S.: Challenges in development of the american sign language lexicon video dataset (asllvd) corpus. In: LREC (2012)

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

[32] Arevalo, J., Solorio, T., Montes-y-Gómez,

M., González, F.A.: Gated multimodal units for information fusion. arXiv preprint arXiv:1702.01992 (2017)

[33] Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)

[34] 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)

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

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

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

[38] 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)

[39] 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)

[40] 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)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary.pdf