Seeing in 2D, Thinking in 3D: 3D Hand Mesh-Guided Feature Learning for Continuous Fingerspelling

K. Papadimitriou[1,3], P. Filntisis[3], G. Retsinas[3], G. Potamianos[1,3], P. Maragos[2,3,4]
[1]Department of Electrical & Computer Engineering, University of Thessaly, 38334 Volos, Greece
[2]School of Electrical & Computer Engineering, National Technical University of Athens, Greece
[3]Robotics Institute, Athena Research Center, 15125 Maroussi, Greece
[4]HERON - Hellenic Robotics Center of Excellence, Athens, Greece
{k.papadimitriou, pfilntisis, george.retsinas, gpotam, petros.maragos}@athenarc.gr

Abstract

Recognizing continuous fingerspelling from monocular RGB video is a highly challenging task due to complex hand articulation, coarticulation effects, and significant inter-signer variability. Prior methods use either raw visual features, which lack structural awareness of fine-grained finger dynamics, or parallel RGB–pose streams from explicit pose estimation, which add substantial inference-time overhead. In this work, we propose a novel knowledge distillation framework that transfers rich hand articulation knowledge from HAMER, a foundation model for 3D hand mesh/pose reconstruction, into a lightweight, RGB-only fingerspelling recognizer. We extract high-level pose embeddings from HAMER's Transformer head, which encode detailed hand structure, and distill them into a ResNet34-based appearance encoder via a dedicated training objective. Subsequently, the learned pose-aware features are fed into a 1D-CNN and BiGRU for temporal modeling, with the full system trained using both connectionist temporal classification (CTC) and a knowledge distillation loss. Notably, our approach does not rely on the teacher model (HAMER) at inference time, thus enabling real-time performance. We evaluate our method on two American sign language (ASL) benchmark fingerspelling datasets, as well as a studio-quality Greek fingerspelling corpus. Our model achieves state-of-the-art accuracy with over 3× lower inference time than prior methods, offering an effective trade-off between accuracy and efficiency for real-time deployment.

1. Introduction

While the majority of prior work in sign language recognition (SLR) has focused on isolated signs or the more complex task of continuous SLR, the recognition of fingerspelling remains relatively understudied. Fingerspelling constitutes a fundamental component of sign language communication, as it enables deaf and hard-of-hearing individuals to express names, technical terms, and foreign words that lack dedicated signs. Automatically recognizing and transcribing a fingerspelling video into a sequence of letters has the potential to support accessible, real-time communication in scenarios where verbal interaction is not feasible. However, continuous fingerspelling recognition remains an inherently complex task due to the fast and subtle finger movements, with signed letters often suffering from high visual similarity. These challenges are further exacerbated in real-world settings, where inter-signer variability, self-occlusion, and degraded video quality introduce additional ambiguity. Further, unlike gloss-level signs, which typically exhibit distinctive manual and non-manual movements, continuous fingerspelling offers limited motion contrast, relies solely on manual-only articulation, and lacks explicit temporal segmentation, making its recognition particularly challenging. As a result, conventional vision models often struggle to capture the structural nuances required for accurate fingerspelling recognition, especially when relying solely on appearance-based cues. To address these limitations, recent works have explored explicit hand modeling techniques that capture the underlying kinematics and articulatory patterns of the hand.

Several state-of-the-art systems adopt skeletal or mesh-based hand models to enhance fingerspelling performance. For example, Fingerspelling PoseNet [5] uses a Transformer-based encoder-decoder trained over 3D skeletal trajectories with a joint CTC/attention loss, while the method in [26] combines 3D-CNNs over RGB input with graph convolutional networks that operate on hand pose-rotation parameters. Although these systems achieve notable gains, especially under signer-independent evaluation, they require dedicated pose estimation modules during inference, typically as parallel RGB–pose streams. This de-

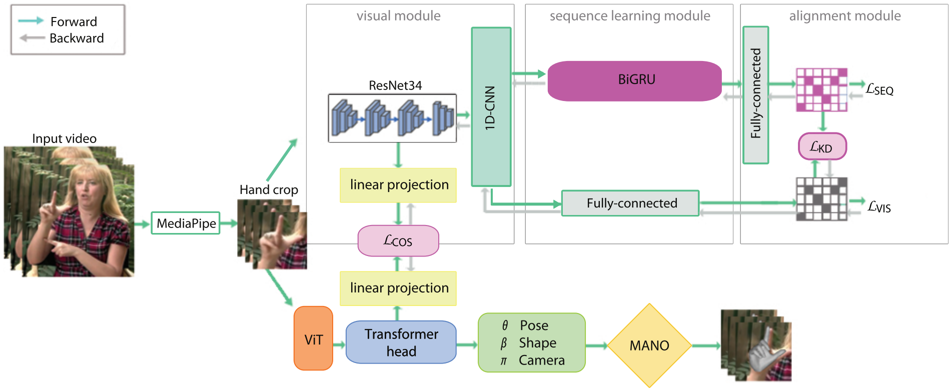

Figure 1. Architecture of the proposed fingerspelling recognition system. Given an input video, the signing hand is cropped and passed through a visual encoder, while a secondary stream extracts high-level structural features for guidance during training. The visual features are then processed by temporal modeling modules, with supervision applied at both intermediate and final prediction stages. During inference, only the main visual stream is used, enabling efficient and real-time recognition.

sign increases computational overhead and limits suitability for real-time deployment. Moreover, pose information is often treated as an auxiliary input stream rather than being fully integrated into the feature learning process.

In this work, we propose a new perspective on continuous fingerspelling recognition by leveraging the structural representation power of a large 3D hand reconstruction model without incurring its computational cost at inference time (see also Fig. 1). In particular, instead of integrating an additional pose/skeleton modality into the recognition pipeline, we distill pose-aware supervision from HAMER [31], a powerful off-the-shelf 3D hand reconstruction model. Given an RGB image of a hand, HAMER predicts the pose, shape, and camera parameters of a 3D morphable model (MANO [33]), which explains the hand image. These rich pose representations offer valuable structural priors for fingerspelling recognition. HAMER can be regarded as a foundation model, trained on large-scale, diverse hand instances, capable of generalizing across multiple downstream tasks (e.g., hand tracking, mesh recovery, gesture understanding) and producing semantically rich pose representations with strong transferability. During training, we guide a ResNet34-based [10] visual encoder using high-level pose embeddings from HAMER's Transformer head, aligning the two modalities in a shared latent space via a cosine similarity objective. This training scheme teaches the ResNet34 encoder to learn both fine-grained appearance and hand pose-aware representations.

To model the temporal structure of fingerspelled sequences, the resulting representations are fed into a 1D-CNN to capture short-term motion patterns between consecutive frames, followed by a bidirectional GRU (BiGRU) [40] sequential model to encode long-range dependencies and contextual information. To guide temporal learning and reinforce alignment at multiple stages, we apply CTC losses after both the 1D-CNN and the final BiGRU layer, encouraging linguistically meaningful representations at both the frame-level and sequence-level. Additionally, we employ a KL-divergence loss between the soft predictions of the two stages to promote temporal consistency across the model's intermediate and final outputs. At inference, the model relies *solely on RGB input* and operates in real-time, as the visual encoder has effectively internalized the pose information during training.

To summarize, our key contributions are:

- We propose a novel knowledge distillation framework that transfers hand articulation features from a state-of-the-art 3D hand reconstruction model (HAMER) into a lightweight RGB-only fingerspelling recognizer. This cross-modal transfer enables the recognizer to internalize rich 3D hand articulation cues, as evidenced by our ablation studies that demonstrate significant performance degradation when pose supervision is removed.
- To the best of our knowledge, this is the first approach to leverage HAMER's high-level pose embeddings for fingerspelling recognition. We qualitatively and quantitatively evaluate the impact of pose supervision from HAMER compared to other 3D hand models and demonstrate that HAMER provides significantly richer guidance.
- The proposed approach constitutes an efficient real-time recognition system that operates solely on RGB input, removing the need for pose estimation during inference.

We evaluate our proposed approach on two benchmark ASL fingerspelling in-the-wild datasets, namely ChicagoFSWild [35] and ChicagoFSWild+ [36], as well as a smaller, studio-quality Greek fingerspelling dataset [27]. The proposed method achieves state-of-the-art performance across all datasets, demonstrating the effectiveness of distilling HAMER pose representations into an efficient RGB-only recognition model.

2. Related Work

Accurate representation of hand articulation is one of the core challenges in fingerspelling recognition, as fine-grained finger movements are critical for conveying letter-level distinctions. Early methods primarily focused on appearance-based models, extracting features from raw RGB videos using CNNs [1, 16, 32, 35, 36], or capturing motion using optical flow [23, 25]. However, these approaches are highly susceptible to varying lighting, occlusions, and camera-signer relative positioning.

To overcome such limitations, researchers introduced skeletal-based representations derived from 2D human pose estimators like OpenPose [37] and HRNet [38], extracting keypoints for the hands and body [15, 21, 23]. Although more robust to visual noise, 2D keypoints cannot capture the full spatial structure of the hand, especially its depth and articulation. To address these shortcomings, 3D skeletal representations were adopted [5], often generated by monocular RGB-based systems such as MediaPipe [20] or learned projection models [28]. While such methods capture richer spatial motion, these representations are limited to joint coordinates and typically ignore detailed articulation, such as joint rotations. In response, the community has turned to parametric 3D hand models [26], which represent hand articulation using pose rotation parameterization. These models offer a more expressive and anatomically accurate description of finger configurations, which is especially beneficial for SLR tasks and fingerspelling in particular that relies on manual articulation only [17, 24, 26, 29]. However, a key limitation is that these models are trained on general-purpose hand datasets and are not tailored to the unique articulation patterns found in signing.

While accurate hand representation is essential, fingerspelling recognition also demands robust sequence modeling to capture the dynamic nature of letter sequences. Early work in continuous fingerspelling recognition addressed these challenges using hybrid approaches that combined frame-wise visual features with probabilistic sequence models. The work in [13] introduced a segmental conditional random field model paired with CNN features, enabling lexicon-free recognition of letter sequences in studio-quality data. Further, the introduction of more challenging in-the-wild datasets shifted the focus toward real-time recognition through powerful sequence learners. In response, the work in [36] proposed an end-to-end architecture based on recurrent neural networks (RNNs) and attention mechanisms for predicting fingerspelled letters in unconstrained conditions. Their iterative visual attention model, in particular, demonstrated strong performance by dynamically refining attention over the signing hand during decoding. In addition, Transformer-based models have more recently emerged as the state-of-the-art for fingerspelling recognition, offering greater capacity for long-range dependencies and improved temporal reasoning. In particular, the work in [5] proposed a Transformer-based encoder-decoder trained on 3D hand keypoint trajectories with a joint CTC/attention loss. While achieving strong results under signer-independent settings, this architecture introduces significant computational overhead and requires pose inference at test time, limiting real-time applicability. Similarly, the work in [26] proposed a multimodal framework that combines 3D-CNN visual features with pose-rotation parameters derived from the PIXIE 3D hand reconstruction model [6]. This design achieves high accuracy but again relies on parallel streams, increasing inference time.

These approaches underscore the value of pose information for fingerspelling recognition, yet highlight a key limitation in that most methods treat pose as an additional stream rather than integrating its structure directly into the visual recognition process.

3. Methodology

Our framework addresses continuous fingerspelling recognition by training an RGB-based model under pose-aware supervision, as illustrated in Fig. 1. The core idea is to use the HAMER 3D hand reconstruction model during training to guide the learning of structural hand representations while keeping the final model lightweight and RGB-only for inference.

The overall pipeline involves the following stages: (i) each video frame is processed by the MediaPipe framework [20] to detect and crop the signing hand, reducing background interference and isolating the region-of-interest; (ii) the resulting hand crops are first encoded by a ResNet34 and then passed through a 1D-CNN and a BiGRU to capture both local and global motion patterns; and (iii) the same hand crops are fed into a frozen HAMER model, extracting high-level pose embeddings from its Transformer head. These features are used during training to supervise the visual encoder via a distillation objective.

Notably, the HAMER model is used only during training, while it is entirely discarded at inference time. As a result, the system operates solely on RGB input during testing, enabling efficient and real-time recognition. The following subsections describe in detail each component of the proposed approach.

3.1. Preprocessing

To enable effective continuous fingerspelling recognition, our preprocessing pipeline focuses on precise localization and cropping of the signing hand region. For this purpose, we employ the MediaPipe hand landmark framework [20], which estimates 21 3D skeletal joint coordinates per detected hand. Although the MediaPipe hand model provides a handedness label for each detected hand, this classification is not stable across frames, as it is not anatomically grounded and may be unreliable in multi-hand scenarios. To address this, we group detected hands across consecutive frames based on their spatial proximity, forming landmark trajectories for each hand candidate. These trajectories enable the estimation of per-hand motion statistics, such as temporal variance and cumulative joint displacement, which are subsequently used to distinguish between signing and non-signing hands. The trajectory exhibiting the highest aggregate motion is designated as the dominant (signing) hand.

Once the signing hand is determined, we crop a padded bounding box around its landmarks, estimated using the minimum and maximum x and y coordinates, which are scaled to the image dimensions. This ensures that the cropped region fully encompasses the signing hand while minimizing background noise. The resulting cropped hand region is then propagated to the recognition model.

3.2. Visual Recognition Baseline

As illustrated in Fig. 1, the backbone of the fingerspelling recognition system consists of three core components: (i) a visual module; (ii) a sequence learning module; and (iii) an alignment module. Given an input signing video that consists of T RGB frames of $H \times W$-pixel size, denoted as $\mathbf{x} = \{\mathbf{x}_t\}_{t=1}^{T} \in \mathbb{R}^{T\times 3\times H\times W}$, the visual module (ResNet34 and 1D-CNN) extracts discriminative spatio-temporal features $\mathbf{v} = \{\mathbf{v}_{t'}\}_{t'=1}^{T'} \in \mathbb{R}^{T'\times d}$. These features are then fed into the sequence learning module, which models global temporal dependencies, generating sequential embeddings $\mathbf{s} = \{\mathbf{s}_{t'}\}_{t'=1}^{T'} \in \mathbb{R}^{T'\times d}$. Subsequently, the embeddings are projected through a fully-connected layer followed by a softmax activation to produce posterior probabilities. During training, CTC losses and an auxiliary loss function are deployed, collectively forming the alignment module.

3.2.1. Visual Module

The visual module extracts per-frame spatial features from the input signing video. It is based on a ResNet34 [10] backbone pretrained on ImageNet [3] and operates on cropped hand regions resized to 256×256 pixels. Each frame output feature map is passed through a global average pooling layer, yielding a 512-dimensional feature vector. To handle variable-length input sequences and preserve temporal alignment across the batch, we apply masked batch normalization before convolution. This allows normalization to operate only on valid (i.e., non-padded) frames, avoiding distortions caused by padding.

To capture short-range temporal dependencies across neighboring frames, the sequence of spatial features extracted by the visual encoder is processed by a 1D-CNN module. This temporal convolutional block combines 1D convolutions and max-pooling layers with varying kernel sizes, designed to capture local motion patterns in the signing hand. The resulting spatio-temporal features are then passed to the sequence learning module for higher-level modeling.

3.2.2. Sequence Learning Module

To capture long-range temporal dependencies across the signing sequence, we employ a multi-layer BiGRU [40] network. The BiGRU receives the spatio-temporal feature sequence produced by the visual module and models both forward and backward temporal context. To handle variable-length input sequences, we use sequence packing and unpacking operations [8], which allow efficient processing without being affected by padding tokens. Our temporal model consists of four stacked BiGRU layers with hidden dimensionality of 512. This recurrent structure enables the network to learn both short- and long-range motion patterns across time, producing sequence-aware representations for each frame in the video.

The output of the final BiGRU layer is passed through a fully-connected classifier, which maps the per-frame representations into a set of class logits over the target vocabulary. These logits are subsequently aligned with the ground-truth letter sequence through the alignment module described next.

3.2.3. Alignment Module

In continuous fingerspelling recognition, the absence of frame-level annotations introduces ambiguity in aligning input frames with target letters. To address this, we employ the CTC loss [9], which enables training without explicit frame-to-letter alignment. Our baseline architecture produces predictions from two different processing stages: one from the visual module and another from the sequence modeling module. Particularly, in addition to the sequence-level predictions from the BiGRU (described in Sec. 3.2.2), we apply a separate classifier to the visual features after the 1D-CNN to obtain an auxiliary prediction path, encouraging the visual module to directly learn features that align with the target sequences. We supervise both with separate CTC losses on the corresponding logits, i.e., $\mathcal{L}_{\text{SEQ}}$ and $\mathcal{L}_{\text{VIS}}$, supporting both modules in learning meaningful temporal alignments to the target letters.

To further ensure consistency between the two prediction paths, we introduce an auxiliary KL-divergence loss [11], denoted as $\mathcal{L}_{\text{KD}}$. This loss internally applies a softmax operation with temperature equal to 2 to the raw logits, producing smoothed distributions suitable for distillation - this acts exactly like soft-label distillation. This loss encourages the features extracted from the visual module to approximate the higher-level temporal dynamics captured by the sequence model, which serves as a fixed teacher during training.

The total training objective is a weighted sum of these three losses:

$$\mathcal{L}_{\text{M}} = \lambda_{\text{SEQ}}\mathcal{L}_{\text{SEQ}} + \lambda_{\text{VIS}}\mathcal{L}_{\text{VIS}} + \lambda_{\text{KD}}\mathcal{L}_{\text{KD}}.$$

In our experiments, we set the loss weights to $\lambda_{\mathrm{SEQ}} = 1.0$, $\lambda_{\mathrm{VIS}} = 1.0$, and $\lambda_{\mathrm{KD}} = 5.0$. The weights are empirically determined based on the validation performance.

3.3. Pose-Guided Supervision via HAMER

Continuous fingerspelling recognition requires fine-grained modeling of subtle and dynamic hand articulations in 3D space. To effectively capture hand motion dynamics, we utilize HAMER [31], a cutting-edge model for 3D hand pose and shape estimation. HAMER directly infers hand articulation parameters from monocular RGB images without relying on intermediate skeletal representations. Specifically, it utilizes a Vision Transformer (ViT) [4] feature learner to capture global spatial context from hand-centric RGB inputs. These features are then processed by a Transformer decoder [39], which maps the visual tokens to pose (θ), shape (β), and camera parameters (π) via the MANO [33] parametric 3D hand model, which represents hand geometry and articulation. The hand configuration is parameterized using 10 shape coefficients and 48 pose parameters, represented as 6D joint rotations, while camera parameters ensure accurate 2D-to-3D alignment (3 parameters). The model outputs a 3D hand mesh with 778 vertices and 21 joint locations, effectively encoding both global configuration and fine-grained finger articulation.

To inject structural pose knowledge into the learning process, in our framework, we extract intermediate latent embeddings from the Tranformer head just before the MANO regression module. We could also use lower-level features directly from the ViT encoder, however we choose the output of the Transformer head, as it provides more refined and semantically rich representations. These pose-aware embeddings, with a dimensionality of 1024, encode detailed 3D spatial structure and serve as a source of structural supervision during training. Specifically, they guide the RGB recognition stream toward learning articulation-sensitive representations. Note that HAMER is used only during training and remains entirely frozen, while at inference time it is altogether dropped.

3.4. Training Strategy

Our model is trained end-to-end using a multi-objective strategy that jointly supervises sequence-level recognition and preserves structural consistency with respect to fine-grained 3D hand articulation. The training pipeline integrates two complementary information streams: (i) an RGB-based visual encoder and (ii) a frozen 3D knowledge distillation stream driven by HAMER. While each stream maintains modality-specific representations, their features are projected into a shared latent space to facilitate cross-modal alignment. Directly enforcing supervision in the original feature spaces may introduce noise or unintended biases into the visual representation. To address this, we align the RGB and pose features in a shared latent space, allowing the model to extract only the structurally relevant information. This setup encourages partial alignment between the two modalities, enabling the visual encoder to benefit from pose-derived structure while preserving the flexibility to learn additional modality-specific cues.

In particular, the visual features derived from the ResNet34 visual encoder are compressed into 256-dimensional embeddings via a single-layer linear projection module. Simultaneously, pose-aware representations extracted from HAMER's Transformer decoder are linearly projected into a shared 256-dimensional latent space. We enforce a cosine similarity loss on these feature vectors, effectively guiding the RGB encoder to internalize pose-relevant information through a teacher-student distillation mechanism. Note, however, that this scheme is susceptible to collapsing toward a trivial zero solution. To mitigate this, we introduce an additional cycle-consistency regularization term that encourages the projected features to retain meaningful information. Specifically, we require that the representations in the shared 256-dimensional space be able to reconstruct the original features via an auxiliary linear projection. It should be noted that the learned projection layers are auxiliary modules that are dropped during inference.

In total, the training objective consists of six losses:

- CTC losses ($\mathcal{L}_{\mathrm{VIS}}$ and $\mathcal{L}_{\mathrm{SEQ}}$), applied to both the 1D-CNN and the BiGRU outputs, supervising frame-to-letter alignments without requiring explicit annotations.
- Self-distillation loss ($\mathcal{L}_{\mathrm{KD}}$) that aligns the soft predictions from the classifiers attached to the 1D-CNN and BiGRU outputs using temperature scaling.
- Cosine similarity loss ($\mathcal{L}_{\mathrm{COS}}$), which enforces consistency between the projected RGB embeddings and the pose embeddings from HAMER.
- Cycle-consistency regularization loss ($\mathcal{L}_{\mathrm{REG}}$), a mean squared error between the original and reconstructed features of both RGB ($\mathcal{L}_{\mathrm{REG}}^{\mathrm{RGB}}$) and pose ($\mathcal{L}_{\mathrm{REG}}^{\mathrm{POS}}$) modalities. This regularizer prevents the learned embeddings from collapsing to a trivial zero solution.

The contribution of each term is controlled by empirically tuned weights based on validation performance. The overall training objective is given by:

$$\mathcal{L}_{total} = \mathcal{L}_{\mathrm{M}} + 0.1\,\mathcal{L}_{\mathrm{COS}} + 0.01\,\mathcal{L}_{\mathrm{REG}}^{\mathrm{RGB}} + 0.01\,\mathcal{L}_{\mathrm{REG}}^{\mathrm{POS}}.$$

3.5. Inference

At inference time, the model operates solely on RGB inputs, without any reliance on 3D pose embeddings. More precisely, given a video sequence, the frames are passed through the ResNet34 backbone and the temporal convolutional encoder to extract spatio-temporal representations. These are then processed by the sequence modeling module and a fully-connected classifier, producing frame-wise logits over the target vocabulary. A softmax is applied over the logits to obtain frame-level probability distributions. Final predictions are generated via beam-search decoding with beam width 5 over these probabilities. To enhance linguistic coherence, we integrate an RNN-based character-level language model (LM) into the decoding process.

4. Experimental Framework

4.1. Datasets

We evaluate our method on three publicly available continuous fingerspelling datasets. The first two, i.e., Chicago Fingerspelling in the Wild (Chicago-FS-Wild) [35] and its extended version Chicago-FS-Wild+ [36], are well-established benchmarks for continuous fingerspelling in ASL. Both datasets provide official signer-independent (SI) splits, ensuring that signers in the training, validation, and test sets do not overlap. In particular, Chicago-FS-Wild includes 5,455 training video sequences from 87 signers, 981 validation sequences from 37 signers, and 868 test sequences from 36 signers. Chicago-FS-Wild+ offers a larger and more diverse set, including 50,402 training sequences from 216 signers, 3,115 validation sequences from 22 signers, and 1,715 test sequences from a separate group of 22 signers. Additionally, we utilize the Greek fingerspelling (FGSL) dataset [27], which contains a total of 1,554 video samples from 21 signers (recorded under studio-like visual conditions) and follows a 7-fold SI cross-validation protocol. Each fold uses data from 18 signers for training and validation (80/20 split), while the remaining 3 signers are used exclusively for testing. Note that in the ASL datasets, the average letters per sequence is ≈ 5 (maximum 45 letters), whereas in FGSL the average is ≈ 3.5 (maximum 6 letters).

4.2. Implementation Details

We train our method for 40 epochs with a batch size of 2. We use the Adam optimizer [14] with an initial learning rate of 10^{-4}, which is reduced by a factor of 0.1 at epochs 20 and 35. To improve robustness and generalization, we perform data augmentation via random cropping and horizontal flipping during the training phase. The system is implemented in PyTorch [30], and the experiments are carried out on an Nvidia RTX 3090 GPU.

5. Experimental Results

In this section, we present the experimental evaluation of our method on the continuous fingerspelling datasets of Sec. 4.1. Our analysis addresses three main axes: (1) the benefit of pose supervision via HAMER embeddings, (2) the impact of architectural and training design choices, and (3) comparison with state-of-the-art models under SI settings. The system's performance is assessed in terms of letter accuracy (LAcc, %), taking into account letter insertions, deletions, and substitutions when comparing recognized (system output) against ground-truth letter strings.

3D hand reconstruction method	LAcc (%) ↑
EXPOSE [2]	89.11
FrankMocap [34]	90.97
PIXIE [6]	94.55
HAMER	**96.27**

Table 1. Letter accuracy (LAcc, %) of our system evaluated on the FGSL dataset, when replacing the visual encoder with pose embeddings extracted from different 3D hand reconstruction models.

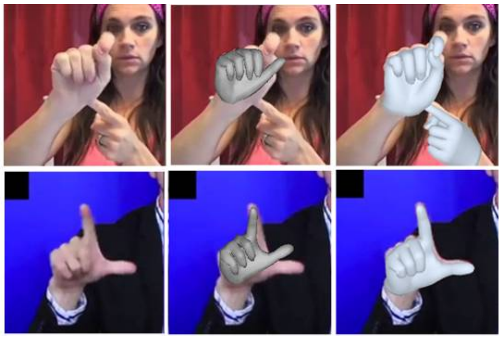

Figure 2. Visualization of 3D hand reconstructions across samples from the ChicagoFSWild (upper row) and ChicagoFSWild+ (lower row) datasets. The first column presents the cropped hand region from the input RGB frames. The second column showcases the reconstructed 3D mesh using the PIXIE model, while the third column displays the hand articulation as estimated by the HAMER model, demonstrating its superiority.

Pose supervision via HAMER embeddings: To justify the selection of HAMER as the supervision backbone in our framework, we conduct a comparative evaluation against alternative 3D hand reconstruction models, namely PIXIE [6], FrankMocap [34], and EXPOSE [2]. To ensure a fair comparison, we extract features consistently from the internal representations of each model before their final parameter regression layer. In this analysis, we replace the visual encoder (ResNet34) output with the extracted hand embeddings and feed them directly into the temporal convolutional block (1D-CNN). This design allows us to assess the discriminative power of each model's pose representation in isolation. All models are evaluated under the same training and testing protocol on the FGSL dataset. As shown in Table 1, HAMER-based features yield the highest accuracy of 96.27%, significantly outperforming the others. While models like PIXIE and FrankMocap are primarily designed for full-body or face-body modeling, and EXPOSE introduces body-hand integration, they fall short in capturing fine-grained finger articulation. In contrast, HAMER produces rich and stable representations tailored specifically for hand pose, which translate into superior performance when used for recognition. These results empir-

Model		LAcc (%)
Unimodal	RGB-only	63.50
	Pose-only (ViT)	60.50
	Pose-only (Trans.)	62.00
Fusion	Early - RGB & Pose (Trans.)	64.30
	Late - RGB & Pose (Trans.)	66.32
Distillation	Pose (ViT) - MSE	64.15
	Pose (ViT) - Cosine	63.70
	Pose (Trans.) - MSE	66.24
	Pose (Trans.) - Cosine	65.60
	Proj. Pose (Trans.) - MSE	66.95
	Proj. Pose (Trans.) - Cosine (**Ours**)	**67.10**

Table 2. Ablation study examining the impact of architectural and training design choices, evaluated on the ChicagoFSWild dataset in LAcc (%). All models share the same recognition backbone.

ically validate our decision to adopt HAMER as the pose-aware teacher in our proposed method. To further support this observation, Fig. 2 presents qualitative comparisons between HAMER and PIXIE, the two best-performing models in our evaluation. The figure visualizes the reconstructed 3D hand meshes from both models, revealing that HAMER produces more detailed and anatomically consistent finger articulations. This visual evidence reinforces the quantitative results, showcasing HAMER's superior ability to capture fine-grained hand structure.

Architectural and training design: Table 2 provides an ablation study of our model on the ChicagoFSWild dataset, focusing on the impact of architectural and training design choices. All models share the same visual recognition backbone of Sec. 3.2 and differ only in how pose information is used. In particular, we examine single-modality baselines, fusion-based models, and various distillation approaches that use pose information as an auxiliary training signal. This analysis helps isolate the contribution of each design choice to the final recognition performance. We begin with the unimodal baselines to establish reference points for RGB and pose modalities. The RGB-only model achieves 63.50% LAcc, indicating that strong visual cues alone provide reasonable performance. For pose-only models, we evaluate two variants using features extracted from HAMER. Specifically, we use the ViT encoder features and the Transformer head output as input to the recognition model. Since the ViT encoder outputs patch-level tokens for each frame, we apply max-pooling to aggregate spatial information and obtain a fixed-length feature vector per frame. In both cases, the standard ResNet34 visual encoder is removed from the pipeline. The ViT-based model reaches 60.50%, while the variant using Transformer output features improves this to 62.00%.

Model	Streams	LAcc (%) ↑
R-CNN-Att [36]	FF	45.10
FG-Transformer [7]	FF	48.36
CNN-Att [28]	H/M & SK	47.93
Siam-LSTM [22]	FF	48.00
Iterative-LM [19]	FF	49.60
F-ResNet-BiLSTM [12]	FF	57.84
3D-CNN-BiGRU [26]	H	64.85
F-PoseNet [5]	H	66.30
Ours	H	**67.10**

Table 3. LAcc, % comparison of state-of-the-art on the Chicago-FSWild dataset. Notation: full frame (FF), hand (H), mouth (M), and skeleton (SK).

Model	Feature streams	LAcc (%) ↑
R-CNN-Att [36]	FF	46.70
RNN-Att [18]	FF	66.20
F-PoseNet [5]	H	71.10
3D-CNN-BiGRU [26]	H	73.57
Ours	H	**75.38**

Table 4. LAcc, % comparison of state-of-the-art on the Chicago-FSWild+ dataset. Notation as in Table 3.

We next examine fusion-based models that jointly leverage RGB and pose inputs. We compare two standard fusion strategies: early fusion, where features from both modalities are concatenated prior to the 1D-CNN temporal encoder and jointly processed; and late fusion, where each modality is trained separately and their predictions are averaged over logits. Early fusion yields a modest improvement over the RGB-only baseline (64.30% vs. 63.50%), suggesting that the model benefits from complementary pose information. However, late fusion performs substantially better, achieving 66.32% LAcc.

We also investigate feature-level distillation, comparing several variants of this approach. In all cases, pose and RGB embeddings are linearly projected before computing the loss, either into the space of one of the two modalities or into a shared latent space. We compare the effect of the feature source (ViT vs. Transformer head), the loss function (MSE vs. cosine similarity), and the projection strategy. Distillation based on Transformer head features consistently outperforms ViT-based distillation (66.24% vs. 64.15% with MSE). Regarding losses, MSE seems to perform slightly better than cosine similarity. However, once both pose and RGB embeddings are projected into a shared latent space, cosine similarity becomes more effective, achieving our best overall performance (67.10%). Note also that unlike early and late fusion strategies that require both RGB and pose streams at inference, introducing significant runtime overhead, our distillation strategy achieves superior performance while relying solely on RGB inputs. This leads to a four-fold improvement in efficiency (14ms vs. 60ms for late fusion) without sacrificing recognition accuracy.

Comparison with state-of-the-art models: We further as-

Model	Feature streams	LAcc (%) ↑
R-CNN-Att [36]	FF	65.62
TDCNN [23]	FF & SK	84.98
RNN-Att [18]	FF	85.12
3D-CNN-BiGRU [27]	H	92.49
Ours	H	**96.70**

Table 5. LAcc, % comparison of our model against literature on the FGSL dataset under the SI setting. Notation as in Table 3.

Model	LAcc (%)	# Params (M)	Inference time (ms)
3D-CNN-BiGRU [26]	64.85	190	55
F-PoseNet [5]	66.30	80	43
Ours	**67.10**	65	14

Table 6. Comparison with state-of-the-art methods on the ChicagoFSWild dataset, reporting LAcc, number of parameters, and inference time per frame.

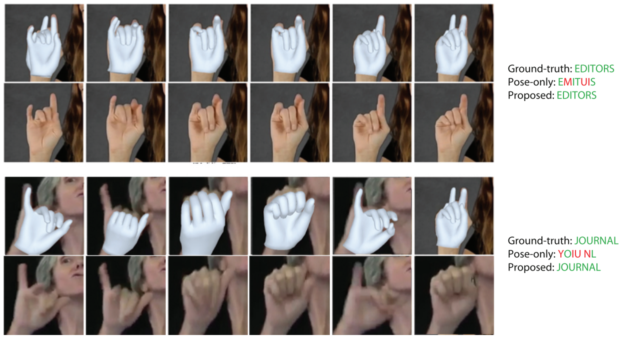

Figure 3. Qualitative comparison of letter sequence predictions from the pose-only model (HAMER) and our proposed RGB-based model. For each example, the top row shows rendered HAMER frames and the corresponding RGB frames below. On the right, we provide the ground-truth label, the prediction from the pose-only model, and the one from our proposed model.

sess the performance of our proposed methodology through comparative evaluations against state-of-the-art methods across the considered datasets. Specifically, Tables 3, 4, and 5 summarize the LAcc (%) results on the ChicagoFSWild, ChicagoFSWild+, and FGSL datasets, respectively. As it can be observed, on the ChicagoFSWild dataset (Table 3), our approach achieves the highest LAcc of 67.12%, surpassing previous top-performing approaches, including the recent fingerspelling F-PoseNet system [26]. Specifically, our method yields an absolute improvement of 0.80%. Similarly, on the larger-scale ChicagoFSWild+ dataset (Table 4), our method achieves state-of-the-art performance with an LAcc of 75.38%, substantially outperforming by 1.81% absolute the previous state-of-the-art, namely the 3D-CNN-BiGRU method [26] that employs RGB and 3D hand joint parameters inferred from PIXIE, training the two streams independently and fusing them during inference. Evaluating our approach on the FGSL corpus (Table 5), we also achieve superior performance, obtaining 96.70% LAcc, significantly exceeding the previously reported best results of the 3D-CNN-BiGRU method [26] (92.49%). These outcomes validate our model's ability to perform well across different sign languages, as it achieves strong performance on ASL and Greek datasets. Note that the different LAcc ranges across datasets are not unexpected. For example, the FGSL corpus is studio-quality with shorter letter sequences compared to the two in-the-wild ASL datasets. In contrast, the ASL datasets (ChicagoFSWild and ChicagoFSWild+) consist of in-the-wild videos with considerable signer, lighting, and viewpoint variability. Among the two, ChicagoFSWild+ provides broader signer diversity and more footage, while ChicagoFSWild includes noisier and more visually ambiguous signing instances.

Efficiency evaluation: Next, Table 6 provides a unified comparison of our model against the best-performing approaches on the ChicagoFSWild dataset. We report LAcc, number of parameters, and inference time per frame to enable an objective evaluation of both recognition performance and computational efficiency. Our model achieves the highest accuracy, while also being smaller and significantly faster than prior work. With an inference time of just 14ms per frame, it is well-suited for real-time deployment scenarios.

Qualitative Ablation: Finally, although the HAMER-based pose stream serves as a strong supervisory signal during training, it can sometimes produce inaccurate or noisy predictions when used alone. As illustrated in Fig. 3, such failure cases do not significantly impact our model's performance. Thanks to its architectural design and multi-objective training strategy, our method avoids over-reliance on the pose modality by aligning features in a shared space without enforcing hard coupling.

6. Conclusion

In this work, we presented a novel framework for continuous fingerspelling recognition that integrates structural supervision from a powerful 3D hand reconstruction model. Our approach leverages latent pose-aware embeddings extracted from the frozen HAMER model, a general-purpose foundation architecture trained for hand mesh recovery, and aligns them with RGB-based representations through a targeted knowledge distillation objective. By guiding the RGB encoder to internalize pose-sensitive information, our method benefits from the structural richness of 3D hand articulation without requiring any annotations at test time. We validated our approach on three datasets, including two in ASL and one in Greek, achieving state-of-the-art results. Overall, our method highlights the potential of repurposing general 3D hand reconstruction models as structural priors for downstream SLR tasks.

7. Acknowledgements

Funded by the European Union

This project was partially funded by the European Union under Horizon Europe (grant No. 101136568 - project HERON).

References

[1] N. C. Camgöz, S. Hadfield, O. Koller, H. Ney, and R. Bowden. Neural sign language translation. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7784–7793, 2018. 3

[2] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, and M. J. Black. Monocular expressive body regression through body-driven attention. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 20–40, 2020. 6

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 4

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2021. 5

[5] P. Fayyazsanavi, N. Nejatishahidin, and J. Košecká. Fingerspelling PoseNet: Enhancing fingerspelling translation with pose-based Transformer models. In *Proc. of the Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 1120–1130, 2024. 1, 3, 7, 8

[6] Y. Feng, V. Choutas, T. Bolkart, D. Tzionas, and M. J. Black. Collaborative regression of expressive bodies using moderation. In *Proc. of the International Conference on 3D Vision (3DV)*, pages 792–804, 2021. 3, 6

[7] K. Gajurel, C. Zhong, and G. Wang. A fine-grained visual attention approach for fingerspelling recognition in the wild. In *Proc. of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3266–3271, 2021. 7

[8] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional LSTM networks. In *Proc. of the IEEE International Joint Conference on Neural Networks (IJCNN)*, pages 2047–2052, 2005. 4

[9] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proc. of the International Conference on Machine Learning (ICML)*, page 369–376, 2006. 4

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2, 4

[11] G. E. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *CoRR*, arXiv:1503.02531, 2015. 4

[12] A. E. Kabade, P. Desai, C. Sujatha, and G. Shankar. American sign language fingerspelling recognition using attention model. In *Proc. of the IEEE International Conference for Convergence in Technology (I2CT)*, pages 1–6, 2023. 7

[13] T. Kim, G. Shakhnarovich, and K. Livescu. Fingerspelling recognition with semi-Markov conditional random fields. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pages 1521–1528, 2013. 3

[14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, arXiv:1412.6980, 2014. 6

[15] S. Ko, J. Son, and H. Jung. Sign language recognition with recurrent neural network using human keypoint detection. In *Proc. of the Conference on Research in Adaptive and Convergent Systems (RACS)*, pages 326–328, 2018. 3

[16] O. Koller, S. Zargaran, and H. Ney. Re-Sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3416–3424, 2017. 3

[17] A. Kratimenos, G. Pavlakos, and P. Maragos. Independent sign language recognition with 3d body, hands, and face reconstruction. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4270–4274, 2021. 3

[18] S. Kruthiventi SS, G. Jose, N. Tandon, R. Biswal, and A. Kumar. Fingerspelling recognition in the wild with fixed-query based visual attention. In *Proc. of the ACM International Conference on Multimedia (ACM MM)*, pages 4362–4370, 2021. 7, 8

[19] W. Kumwilaisak, P. Pannattee, C. Hansakunbuntheung, and N. Thatphithakkul. American sign language fingerspelling recognition in the wild with iterative language model construction. *APSIPA Transactions on Signal and Information Processing*, 11(1):22, 2022. 7

[20] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann. MediaPipe: A framework for perceiving and processing reality. In *Proc. of the Workshop on Computer Vision for AR/VR at IEEE CVPR*, 2019. 3

[21] F. Nugraha and E. C. Djamal. Video recognition of American sign language using two-stream convolution neural networks. In *Proc. of the International Conference on Electrical Engineering and Informatics (ICEEI)*, pages 400–405, 2019. 3

[22] P. Pannattee, W. Kumwilaisak, C. Hansakunbuntheung, and N. Thatphithakkul. Novel American sign language fingerspelling recognition in the wild with weakly supervised learning and feature embedding. In *Proc. of the International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pages 291–294, 2021. 7

[23] K. Papadimitriou and G. Potamianos. Multimodal sign language recognition via temporal deformable convolutional sequence learning. In *Proc. of the Interspeech*, pages 2752–2756, 2020. 3, 8

[24] K. Papadimitriou and G. Potamianos. Sign language recognition via deformable 3D convolutions and modulated graph

convolutional networks. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 3

[25] K. Papadimitriou and G. Potamianos. Multimodal locally enhanced Transformer for continuous sign language recognition. In *Proc. of the Interspeech*, pages 1513–1517, 2023. 3

[26] K. Papadimitriou and G. Potamianos. Multimodal continuous fingerspelling recognition via visual alignment learning. In *Proc. of the Interspeech*, pages 922–926, 2024. 1, 3, 7, 8

[27] K. Papadimitriou, G. Sapountzaki, K. Vasilaki, E. Efthimiou, S.-E. Fotinea, and G. Potamianos. A large corpus for the recognition of Greek sign language gestures. *Computer Vision and Image Understanding*, 249:104212–104232, 2024. 2, 6, 8

[28] M. Parelli, K. Papadimitriou, G. Potamianos, G. Pavlakos, and P. Maragos. Exploiting 3D hand pose estimation in deep learning-based sign language recognition from RGB videos. In *Proc. of the European Conference on Computer Vision Workshop on Sign Language Recognition, Translation and Production (ECCVW-SLRTP)*, pages 249–263, 2020. 3, 7

[29] M. Parelli, K. Papadimitriou, G. Potamianos, G. Pavlakos, and P. Maragos. Spatio-temporal graph convolutional networks for continuous sign language recognition. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8457–8461, 2022. 3

[30] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in PyTorch. In *Proc. of the Conference on Neural Information Processing Systems Workshop on Automatic Differentiation (NIPS-W)*, 2017. 6

[31] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik. Reconstructing hands in 3D with transformers. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9826–9836, 2024. 2, 5

[32] J. Pu, W. Zhou, and H. Li. Dilated convolutional network with iterative optimization for continuous sign language recognition. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 885–891, 2018. 3

[33] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6):245:1–245:17, 2017. 2, 5

[34] Y. Rong, T. Shiratori, and H. Joo. FrankMocap: A monocular 3D whole-body pose estimation system via regression and integration. *Proc. of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1749–1759, 2021. 6

[35] B. Shi, A. Martinez Del Rio, J. Keane, J. Michaux, D. Brentari, G. Shakhnarovich, and K. Livescu. American sign language fingerspelling recognition in the wild. In *Proc. of the IEEE Spoken Language Technology Workshop (SLT)*, pages 145–152, 2018. 2, 3, 6

[36] B. Shi, A. Martinez Del Rio, J. Keane, D. Brentari, G. Shakhnarovich, and K. Livescu. Fingerspelling recognition in the wild with iterative visual attention. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pages 5399–5408, 2019. 2, 3, 6, 7, 8

[37] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4645–4653, 2017. 3

[38] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 5693–5703, 2019. 3

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017. 5

[40] C. Yu, L. Tianrui, J. Zhen, and Y. Chengfeng. BGRU: A new method of Chinese text sentiment analysis. *Journal of Physics: Conference Series*, 13(06):973–981, 2019. 2, 4