

HySUP-V: Quality-Aware 3D Vietnamese Sign Language Reconstruction with Anchor-Guided Kinematic Fusion

Quang Nguyen Viet*, Diep Nguyen Hoang*, Hai-Anh Bui-Vu*, Le Cong Thuong*, Thanh Ha Le*[§]

*Faculty of Information Technology, University of Engineering and Technology, Hanoi, Vietnam

Abstract—Reconstructing expressive 3D signing motion from in-the-wild monocular video remains difficult due to motion blur, compression artifacts, frequent hand and face occlusions. These degradations make standard pipelines temporally unstable and prone to losing fine articulator detail. We introduce HySUP-V, a reliability-aware framework that assesses and manages visual evidence quality across sequences. In Stage I (EQA), we score part-level image quality to select high-fidelity anchor frames as references. In Stage II (AKF), we enforce shape consistency by fusing specialized hand (MANO) and face (FLAME) estimates into a unified SMPL-X model and use geodesic pose transport to bridge degraded frames. Applying HySUP-V, we curate VSL-3D, a public dataset of 1,540 Vietnamese Sign Language clips with expressive SMPL-X parameters. Human evaluations yield a Mean Opinion Score of 3.47/5, plausible reconstructions with strong semantic preservation and hand accuracy, while highlighting remaining challenges in facial detail and motion naturalness.

Index Terms—sign language, 3D human reconstruction, SMPL-X, FLAME, hands, IQA, interpolation, temporal smoothing

Our project page: <https://hysup-video-page.github.io>

I. INTRODUCTION

Reconstructing accurate 3D human motion from monocular 2D video is a standing challenge in computer vision, particularly for in-the-wild footage, which is subject to motion blur and compression artifacts. These difficulties are especially pronounced in the analysis of sign languages, which depend on rapid, precise movements of the hands, face, and body to convey meaning. Our work focuses on Vietnamese Sign Language (VSL), the primary sign language of the Deaf community in Vietnam. While linguistic research has identified its major varieties and phonological structure [1], [2], [3], and 2D video is abundant, the community lacks a public 3D motion-capture corpus. This data gap for VSL stands in contrast to resources available for other languages [4], motivating methods to produce high-quality 3D reconstructions from existing 2D video.

[§] Corresponding author: ltha@vnu.edu.vn

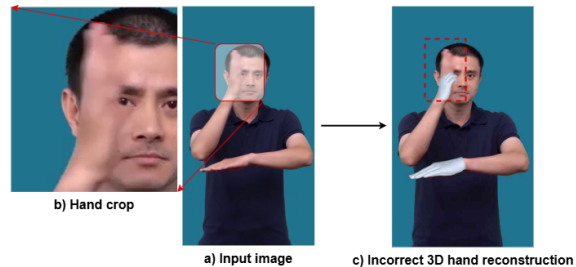


Figure 1: Blur compromises 3D reconstruction. (a) Input frame from a VSL video. (b) A crop of the hand reveals severe motion blur. (c) Standard reconstruction models can produce unstable hand poses from such input, motivating our approach of reliability-aware anchoring and guided motion refinement.

This reliance on 2D video, however, reveals limitations in current 3D estimation pipelines. As shown in Fig. 1, fast signing induces severe blur and self-occlusions [5], yielding temporally jittery or implausible poses in standard 2D-to-3D models [6], [7]. Holistic SMPL-X-based approaches (e.g., SMPLest-X [8], SMPLer-X [9]) offer unified body-hand-face parameterization but often prioritize larger body regions during optimization, neglecting fine finger and facial details. In contrast, specialized models like MANO for hands [10] and FLAME for faces [11] provide superior fidelity but require careful fusion to maintain kinematic consistency [12], [13].

To address these issues, we introduce **HySUP-V**, a two-stage framework for robust 3D reconstruction from challenging video. The core principle of our method is to dynamically manage the variable quality of visual evidence. First, an Evidence Quality Assessment (EQA) stage identifies high-quality anchor frames that serve as trustworthy references. Second, an Anchor-Guided Kinematic Fusion (AKF) stage leverages these anchors to establish a consistent body shape and fuse high-fidelity hand and face models into a unified SMPL-X representation. For degraded segments between anchors, our framework preserves

motion coherence by using geodesic interpolation rather than relying on compromised visual data.

To advance research in expressive 3D human motion and sign language analysis, we present two primary contributions. We introduce **HySUP-V**, a two-stage pipeline engineered to overcome common failures in 3D reconstruction from difficult video footage. Applying this method, we deliver VSL-3D, a new benchmark dataset of **1,540** 3D-reconstructed Vietnamese Sign Language clips with rich SMPL-X parameters. A human evaluation study confirms our reconstructions achieve high semantic fidelity and plausible hand gestures (**3.47/5** Likert score), while also identifying current limitations in capturing nuanced facial expressions and overall naturalness.

II. RELATED WORK

A. Parametric Human Models and Mesh Recovery

Parametric models are foundational for monocular 3D human reconstruction. SMPL [14] provides shape and pose parameters for the body, while SMPL-X [8] integrates articulated hands (MANO [10]) and expressive faces (FLAME [11]). Single-image methods include direct regression [15] and optimization-based fitting like SMPLify-X [16].

Video extensions enhance temporal coherence: VIBE uses GRUs for feature aggregation [17], TCMR applies causal convolutions for refinement [18], and PyMAF-X leverages attention on mesh-aligned features [19]. However, these holistic approaches often falter on sign language footage, where hands and faces key for linguistic cues occupy minimal pixels, and uniform priors suppress rapid non-manual markers. Our pipeline counters this by quality-gated inference and fusion of specialized components.

B. Specialized Hand and Face Reconstruction

Dedicated hand models like HaMeR [20] excel over holistic counterparts, trained on datasets such as InterHand2.6M [21]. For faces, DECA [12] and EMOCA [22] yield superior expression and identity detail. Integrating these into SMPL-X poses challenges, including kinematic mismatches (e.g., wrist dislocations or chirality errors). Our fusion protocol ensures consistency by respecting the model’s chain and geometry, with targeted head-joint refinement.

C. Sign Language Datasets and 3D Reconstruction

Most sign language datasets are 2D video corpora with linguistic annotations, e.g., WLASL [23], MS-ASL [24], RWTH-PHOENIX-Weather 2014T [25], and How2Sign [26]—none focused on Vietnamese Sign Language. This 2D-to-3D disparity drives methods to uplift existing videos. Our approach offers

a robust, scalable pipeline tailored to sign language challenges.

III. METHOD

We introduce **HySUP-V**, a framework for robust 3D human reconstruction from monocular videos. The core challenge in this setting is the variable quality of per-frame visual evidence. Our method is designed to explicitly address this by first identifying trustworthy temporal anchors and then leveraging them to guide a temporally coherent and kinematically consistent reconstruction. The framework comprises two main stages: Stage I, *Evidence Quality Assessment (EQA)*, which analyzes per-frame data quality, and Stage II, *Anchor-Guided Kinematic Fusion (AKF)*, which uses this quality information to regularize and refine the final 3D output. Fig. 2 illustrates the HySUP-V pipeline.

Our goal is to recover the parameters of an expressive SMPL-X body model [16] from a sequence of images $\{I_t\}_{t=1}^T$. This includes a single, sequence-level shape vector β to represent identity, and per-frame parameters for pose θ_t , expression ψ_t , and translation t_t . We leverage specialized part models—MANO for hands and FLAME for faces [10], [11].

A. Stage I: Evidence Quality Assessment (EQA)

The **EQA** stage quantifies the reliability of visual evidence for the face, left hand, and right hand on a per-frame basis. For each frame t and body part ℓ , we isolate the region using segmentation-guided cropping and evaluate it with a no-reference Image Quality Assessment (IQA) model to produce a quality score q_t^ℓ . Based on these scores, we partition the sequence into a high-quality **anchor set** $\mathcal{A}^\ell$ (containing the most reliable frames, including the top-scoring one t_ℓ^*) and a **degraded set** $\mathcal{D}^\ell$ for the remainder. Critically, identity and shape parameters are estimated exclusively using the anchor set.

B. Stage II: Anchor-Guided Kinematic Fusion (AKF)

The AKF stage uses the EQA anchors (and optional reliability masks) to enforce temporal consistency and to fuse the part-specific estimates into a unified, kinematically consistent SMPL-X body model.

a) Sequence-Level Shape Consistency: To suppress shape (identity) flicker, we estimate a subject-specific shape once and keep it fixed over the sequence. Concretely, we extract the SMPL-X shape β , MANO hand shape β_{MANO} , and FLAME face shape β_{FLAME} from their anchor frames (t_{hand}^* , t_{face}^*) and hold these vectors constant for all t . This follows standard video practice that treats identity shape as time-invariant, while per-frame pose θ_t and expressions ψ_t

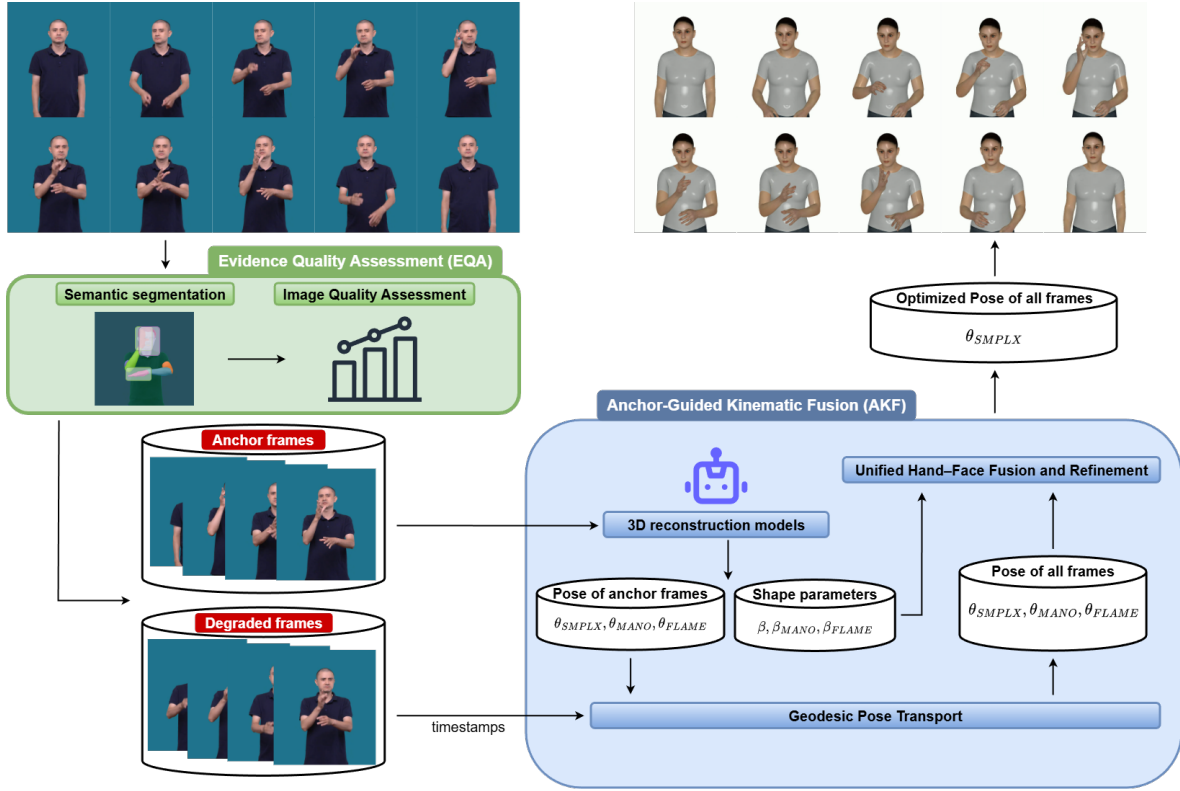


Figure 2: Overview of the HySUP-V pipeline: Stage I-EQA segments and scores frames to choose anchors, then Stage II-AKF reconstructs shapes/poses with SMPL-X+MANO+FLAME and uses geodesic transport to fill degraded intervals, yielding optimized poses for all frames.

remain dynamic, improving temporal stability without sacrificing motion/detail.

b) Geodesic Pose Transport: To ensure temporally coherent motion across low-quality segments for the hands and face, we employ **Geodesic Pose Transport**. For any degraded frames situated between two consecutive reliable anchors, t_a and t_b , we infer the pose by interpolating the joint rotations along the shortest geodesic path on the $SO(3)$ manifold.

Specifically, for each relevant joint j (including finger, jaw, and eye joints), the rotation $R_j^\ell(t)$ at an intermediate frame t is computed via Spherical Linear Interpolation (SLERP):

$$R_j^\ell(t) = \text{SLERP}(R_j^\ell(t_a), R_j^\ell(t_b); s), \quad (1)$$

where $s = (t - t_a)/(t_b - t_a)$ is the normalized temporal weight. This is implemented in unit-quaternion space, with standard sign correction to guarantee the shortest arc. For segments at the beginning or end of a sequence bounded by only a single anchor, we extrapolate by holding the nearest reliable pose (zero-order hold).

c) Unified Hand-Face Fusion and Refinement: Hand specialist regressors (e.g., HaMeR) output *global*

wrist/finger rotations in their own canonical hand frame, whereas SMPL-X expects *local* rotations relative to the parent (forearm) in the SMPL-X kinematic tree. We introduce process addresses this mismatch. First, we align the coordinate systems using a fixed transformation $C_h \in SO(3)$ (where $C_h = I$ if the frames are already aligned). Then, we convert the specialist's global rotation for a joint, R_j^{spec} , into a local rotation relative to its SMPL-X parent. This local rotation for a joint j on hand h is computed as:

$$R_{j,h}^{\text{local}} = (R_{\text{parent}(j)}^{\text{glob}})^{-1} (C_h R_j^{\text{spec}} C_h^\top),$$

where $R_{\text{parent}(j)}^{\text{glob}}$ is the global rotation of the parent joint in the SMPL-X hierarchy. Finally, the resulting local rotation matrices are converted to the axis-angle representation and assigned to the corresponding entries of the SMPL-X pose parameter vector θ .

For the face, we refine SMPL-X facial parameters by aligning SMPL-X head-relative joints to FLAME targets while regularizing expressions and small-angle rotations. All landmarks are expressed in the SMPL-X *head frame*. Denote SMPL-X joints by $\mathbf{j}_\bullet^{\text{sx}}$ and

FLAME targets by $\mathbf{j}_\bullet^{\text{fl}}$. Define

$$\Delta_{\text{jaw}} = \mathbf{j}_{\text{jaw}}^{\text{sx}} - \mathbf{j}_{\text{jaw}}^{\text{fl}}, \quad \Delta_{\text{eyek}} = \mathbf{j}_{\text{eyek}}^{\text{sx}} - \mathbf{j}_{\text{eyek}}^{\text{fl}} \quad (k \in \{L, R\}),$$

and let $d_{\text{sx}} = \|\mathbf{j}_{\text{eyeR}}^{\text{sx}} - \mathbf{j}_{\text{eyeL}}^{\text{sx}}\|_2$ with d_{fl} the FLAME inter-ocular distance. Using axis-angle vectors $\mathbf{r}_{\text{jaw}}, \mathbf{r}_{\text{eyeL}}, \mathbf{r}_{\text{eyeR}}, \mathbf{r}_{\text{head}}$ and expression coefficients ψ , we minimize

$$\begin{aligned} \mathcal{L}_{\text{face}} = & \lambda_j \rho(\Delta_{\text{jaw}}) + \lambda_e \sum_{k \in \{L, R\}} \rho(\Delta_{\text{eyek}}) \\ & + \lambda_d (d_{\text{sx}} - d_{\text{fl}})^2 \\ & + \lambda_\psi \|\psi\|_2^2 + \lambda_{p, \text{jaw}} \|\mathbf{r}_{\text{jaw}}\|_2^2 \\ & + \lambda_{p, \text{eye}} (\|\mathbf{r}_{\text{eyeL}}\|_2^2 + \|\mathbf{r}_{\text{eyeR}}\|_2^2), \end{aligned} \quad (2)$$

where ρ is the Huber penalty (weights listed in Table I).

C. Implementation Details

We instantiate the HySUP-V framework using a selection of state-of-the-art components and empirically determined hyperparameters, summarized in Table I.

a) Component Architecture: For the Evidence Quality Assessment (EQA) stage, we use Sapiens [27] for its robust face and hand segmentation, which is essential for accurate part localization. The quality of each extracted patch is then quantified by MUSIQ [28], a no-reference IQA metric adept at capturing perceptual artifacts like motion blur, providing the core signal for our reliability gates and anchor selection. For the Anchor-Guided Kinematic Fusion (AKF) stage, we adopt the methodology established in [29]. Following this approach, we employ SMPLest-X [8] as the core backbone to construct a unified SMPL-X representation. This framework integrates high-fidelity part-level estimates: detailed MANO-based hand poses from HaMeR [6] and nuanced facial geometry from the FLAME-based Pixel3DMM [30].

Instead of fixed per-part thresholds, we form a shared reliability gate by dropping the lowest 10% of frames per sequence using $s_t = \min_\ell q_t^\ell$: $m_t = \mathbf{1}[s_t \geq \text{Perc}_{10}(\{s_u\})]$. Anchors are then selected over the reliable set $\mathcal{R} = \{t \mid m_t = 1\}$.

IV. DATASET AND EVALUATION

A. Dataset

We apply HySUP-V to Vietnamese Sign Language (VSL) dictionary videos curated under the QIPEDC initiative.¹ We process **1,540** videos into full 3D motion sequences with per-frame SMPL-X parameters $(\beta; \{\psi_t, \theta_t, \mathbf{t}_t\}_{t=1}^T)$. Examples of the resulting 3D reconstructions are shown in Fig. 3.

¹ <https://qipcdc.moet.gov.vn/dictionary>

Table I: HySUP-V configuration: components and chosen values. Notation follows the method: quality traces q_t^ℓ , shared gate m_t , anchors $t_{\text{face}}^*, t_{\text{hand}}^*$. Face-refinement loss weights from Eq. 2.

Component	Implementation choice
Evidence Quality Assessment (EQA)	
Part quality	$q_t^\ell \in [0, 1], \ell \in \{\text{face}, \text{hand}_L, \text{hand}_R\}$.
Sequence score	$s_t = \min_\ell q_t^\ell$ (conservative across parts).
Shared gate	$m_t = \mathbf{1}[s_t \geq \text{Perc}_{10}(\{s_u\})]$ (drop lowest 10%).
Anchor set	$\mathcal{R} = \{t \mid m_t = 1\}$.
Face anchor	$t_{\text{face}}^* = \arg \max_{t \in \mathcal{R}} q_t^{\text{face}}$.
Hand anchor	$t_{\text{hand}}^* = \arg \max_{t \in \mathcal{R}} \min(q_t^{\text{hand}_L}, q_t^{\text{hand}_R})$.
Semantic segmentation	Sapiens [27].
Image quality (IQA)	MUSIQ [28].
Hand model estimator	HaMeR (MANO) [6].
Hand pose interpolation	Geodesic SLERP on SO(3) [31].
Face model estimator	FLAME-based (Pixel3DMM) [30].
Body / fusion backbone	SMPL-X-based (SMPLest-X) [8].
Seq-level shapes	MANO β from t_{hand}^* ; FLAME β from t_{face}^* .
Face-refinement loss weights (Eq. 2)	
Jaw landmarks	$\lambda_j = 2 \times 10^3$.
Eye landmarks	$\lambda_e = 4 \times 10^3$.
Inter-ocular distance	$\lambda_d = 5 \times 10^1$.
Expression prior	$\lambda_\psi = 1 \times 10^{-2}$.
Rotation priors	$\lambda_{p, \text{jaw}} = 1 \times 10^{-2}, \lambda_{p, \text{eye}} = 1 \times 10^{-2}$.

B. Subjective Evaluation

a) Design & protocol: We conducted a subjective evaluation with four raters with basic familiarity with sign language, who assessed 25 reconstructed VSL clips. Using a 5-point Likert scale, they scored each clip on four criteria: *comprehensibility*, *hand-gesture accuracy*, *naturalness of motion*, and *facial expressions*.

b) Results and Analysis: The evaluation results, summarized in Table II, indicate that our method successfully preserves semantic content, with **comprehensibility** receiving the highest average score (3.94/5). However, a performance gradient is evident across other criteria: **hand-gesture accuracy** (3.39) and **naturalness** (3.34) were rated moderately, while **facial expressions** (3.20) were identified as the primary limitation.

Follow [32], lexical formation in Vietnamese Sign Language can be organized into six mechanisms: (i) *Direct pointing*, (ii) *Simulation*, (iii) *Reflection and Analysis of characteristics*, (iv) *Derivatives*, (v) *Borrow*, (vi) *Finger letters and numbers*. An analysis by sign type, following the VSL formation mechanisms in [32] (Table III), clarifies these trends. Reconstructions of simple manual shapes like **finger letters and numbers** (Type vi) achieved near-ceiling scores

Table II: Average scores per Criterion (1–5 Likert scale)

Criterion	Mean $\pm$ Std
Comprehensibility	3.94 $\pm$ 0.71
Hand gesture accuracy	3.39 $\pm$ 0.71
Naturalness of motion	3.34 $\pm$ 0.78
Facial expression	3.20 $\pm$ 0.76
Overall	3.47 $\pm$ 0.74

Table III: Average scores per Sign type (1–5 Likert scale)

Type	Comprehensibility	Hand gesture accuracy	Naturalness of motion	Facial expressions
(i)	3.50 $\pm$ 0.77	2.81 $\pm$ 0.68	2.75 $\pm$ 0.80	2.50 $\pm$ 0.82
(ii)	3.80 $\pm$ 0.83	3.60 $\pm$ 0.88	3.45 $\pm$ 0.96	2.15 $\pm$ 1.08
(iii)	3.67 $\pm$ 0.77	2.75 $\pm$ 0.60	3.00 $\pm$ 0.52	3.67 $\pm$ 1.35
(iv)	3.50 $\pm$ 1.17	2.88 $\pm$ 1.24	2.81 $\pm$ 1.15	1.94 $\pm$ 0.27
(v)	3.88 $\pm$ 0.85	2.88 $\pm$ 0.84	2.81 $\pm$ 0.98	4.06 $\pm$ 0.94
(vi)	5.00 $\pm$ 0.00	4.85 $\pm$ 0.10	4.75 $\pm$ 0.26	4.85 $\pm$ 0.30

(**4.86/5**), demonstrating the method’s robustness on high-contrast, distinct gestures. Conversely, categories involving more complex or rapid motions, such as **direct pointing** (Type i) and **derivatives** (Type iv), received the lowest ratings, which we attribute to challenges like motion blur and occlusions.

V. CONCLUSION

We introduced **HySUP-V**, a reliability-aware framework for 3D signing motion reconstruction from in-the-wild video. HySUP-V scores framewise visual evidence to select high-fidelity anchors, then kinematically fuses MANO (hands) and FLAME (face) into a unified SMPL-X model, with geodesic interpolation and quaternion-based smoothing. This yields higher articulator fidelity and better temporal coherence than holistic baselines. On Vietnamese Sign Language footage—including a human study with a mean score of **3.47/5**—HySUP-V proves effective.

To accelerate progress, we release **VSL-3D**, a large public corpus of 3D VSL sequences with expressive parameters and per-frame quality labels. Remaining challenges under extreme blur and low light motivate future work on blur-adaptive probabilistic modeling and leveraging linguistic priors.

ACKNOWLEDGE

We appreciate the Sign by Sign Club’s participation in the subjective evaluation at the University of Languages and International Studies (ULIS), Vietnam National University, Hanoi. We especially appreciate the contributions of Nguyen Bao Ngoc, Nguyen Khanh Linh, To Dieu Linh, and Tran Thi Nhat Linh for their direct participation and constructive feedback.

REFERENCES

- [1] J. Woodward, “Sign languages and sign language families in thailand and viet nam,” in *The signs of language revisited*. Psychology Press, 2013, pp. 30–52.
- [2] J. B. Jepsen, G. De Clerck, S. Lutalo-Kiingi, and W. B. McGregor, *Sign languages of the world: A comparative handbook*. Walter de Gruyter GmbH & Co KG, 2015.
- [3] C. T. X. Mỹ, “Các yếu tố cấu thành kí hiệu trong ngôn ngữ kí hiệu của người khiếm thính việt nam,” *Tạp chí Khoa học ĐHSP TPHCM*, no. 37, pp. 11–14, 2012.
- [4] Z. Yu, S. Huang, Y. Cheng, and T. Birdal, “Signavatars: A large-scale 3d sign language holistic motion dataset and benchmark,” in *European Conference on Computer Vision*. Springer, 2024, pp. 1–19.
- [5] Y. Zhao, D. Rozumnyi, J. Song, O. Hilliges, M. Pollefeys, and M. R. Oswald, “Human from blur: Human pose tracking from blurry images,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 14 905–14 915.
- [6] Y. Hasson, M. Habermann, W. Xu, C. Theobalt, M. J. Black, J. Romero, D. Tzionas *et al.*, “Hamer: Revisiting 3d hand pose and mesh estimation,” *arXiv:2312.03991*, 2023.
- [7] R. A. Potamias, J. Zhang, J. Deng, and S. Zafeiriou, “Wilor: End-to-end 3d hand localization and reconstruction in-the-wild,” *arXiv preprint arXiv:2409.12259*, 2024.
- [8] W. Yin, Z. Cai, R. Wang, A. Zeng, C. Wei, Q. Sun, H. Mei, Y. Wang, H. E. Pang, M. Zhang *et al.*, “Smplst-x: Ultimate scaling for expressive human pose and shape estimation,” *arXiv preprint arXiv:2501.09782*, 2025.
- [9] Z. Cai, W. Yin, A. Zeng, C. Wei, Q. Sun, W. Yanjun, H. E. Pang, H. Mei, M. Zhang, L. Zhang *et al.*, “Smplst-x: Scaling up expressive human pose and shape estimation,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 11 454–11 468, 2023.
- [10] J. Romero, D. Tzionas, and M. J. Black, “Embodied hands: Modeling and capturing hands and bodies together,” *ACM Trans. Graphics (TOG)*, 2017.
- [11] T. Li, T. Bolkart, M. J. Black, S. Romdhani, and A. Tewari, “Flame: Learning a model of facial shape and expression from 4d scans,” in *SIGGRAPH Asia*, 2017.
- [12] Y. Feng *et al.*, “Deca: Learning an animatable detailed 3d face model from in-the-wild images,” *ACM Trans. Graphics (TOG)*, 2021.
- [13] P. Filntisis, G. Retsinas, R. Danecek, V. F. Abrevaya, A. Roussos, T. Bolkart, and P. Maragos, “Smirk: 3d facial expressions through analysis-by-neural-synthesis,” in *Greeks in AI Symposium 2025*, 2024.
- [14] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, “Smpl: a skinned multi-person linear model,” *ACM Trans. Graph.*, vol. 34, no. 6, Oct. 2015. [Online]. Available: <https://doi.org/10.1145/2816795.2818013>
- [15] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, “End-to-end recovery of human shape and pose,” in *CVPR*, 2018.
- [16] G. Pavlakos, V. Choutas *et al.*, “Expressive body capture: 3d hands, face, and body from a single image,” in *CVPR*, 2019.
- [17] M. Kocabas, N. Athanasiou, and M. J. Black, “Vibe: Video inference for human body pose and shape estimation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5253–5263.
- [18] H. Choi *et al.*, “Tcmr: Temporal convolutional mesh regression for human mesh recovery,” in *ICCV*, 2021.
- [19] Y. Zhang *et al.*, “PyMAF-X: Towards well-localized 3d human mesh estimation under occlusion,” <https://github.com/Human-Pose3D/pymaf-x>, 2021.
- [20] Y. Hasson *et al.*, “Hamer: 3d hand mesh regression from a single rgb image,” *arXiv:2306.09319*, 2023.
- [21] G. Moon *et al.*, “Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation,” in *ECCV*, 2020.

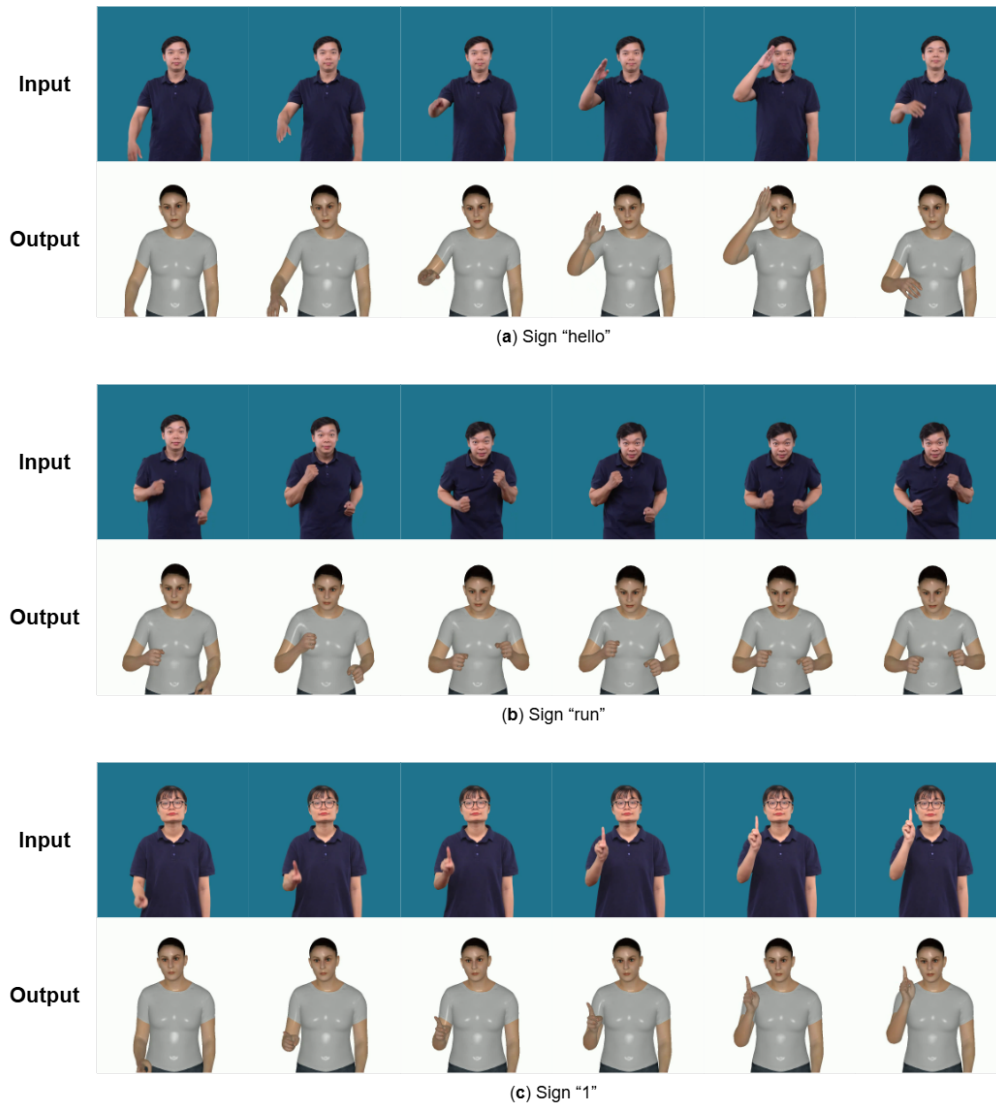


Figure 3: Examples of 3D video reconstruction from 2D Vietnamese Sign Language videos. Each subfigure shows input frames (top row) and reconstructed 3D avatars (bottom row).

- [22] R. Daněček, T. Bolkart, and M. J. Black, "Emoca: Emotion driven monocular face capture and animation," in *CVPR*, 2022.
- [23] D. Li *et al.*, "Word-level deep sign language recognition from video: A new large-scale dataset and methods," in *WACV*, 2020.
- [24] H. R. Vaezi Joze and O. Koller, "MS-ASL: A large-scale data set and benchmark for understanding american sign language," *BMVC*, 2019.
- [25] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems," *CVIU*, 2015.
- [26] A. Duarte *et al.*, "How2sign: A large-scale multimodal dataset for continuous ASL," in *CVPR*, 2021.
- [27] R. Khrodar, T. Bagautdinov, J. Martinez, S. Zhaoen, A. James, P. Selednik, S. Anderson, and S. Saito, "Sapiens: Foundation for human vision models," in *European Conference on Computer Vision*. Springer, 2024, pp. 206–228.
- [28] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "Musiq: Multi-scale image quality transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5148–5157.
- [29] L. Thuong, T. Le, M. Dang, S. Ha, and H. Le, "Hysup: Seam-aware hand-body fusion for accurate whole-body 3-d reconstruction from a single image," 08 2025, pp. 1–5.
- [30] S. Giebenhain, T. Kirschstein, M. Rünz, L. Agapito, and M. Nießner, "Pixel3dmm: Versatile screen-space priors for single-image 3d face reconstruction," *arXiv preprint arXiv:2505.00615*, 2025.
- [31] K. Shoemake, "Animating rotation with quaternion curves," in *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, 1985, pp. 245–254.
- [32] L. T. T. Uyên, "Một số phương thức hình thành kí hiệu của người điếc việt nam," *Tạp chí Khoa học Giáo dục Việt Nam (Vietnam Journal of Educational Sciences)*, September 2020, received 06/04/2020; revised 05/06/2020; accepted 15/09/2020. [Online]. Available: https://vjes.vnies.edu.vn/sites/default/files/baivietso6_0.pdf