

HySUP: Seam-Aware Hand-Body Fusion for Accurate Whole-Body 3-D Reconstruction from a Single Image

Le Cong Thuong*, Thu Phuong Le^{†‡}, Minh Hang Dang*, Son Tung Ha*, Thanh Ha Le^{*§}

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

[†]Manning College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, United States

Abstract—Expressive human pose and shape estimation (EHPS)—which jointly captures body, hand, and facial motion—is vital for emerging AR/VR and telepresence applications. State-of-the-art *holistic* regressors accurately recover overall body pose yet still yield fused or imprecise fingers, whereas *hand-specific* networks resolve fine articulation but sacrifice global coherence. We benchmark two leading holistic models (SMPLest-X, SMPLer-X) and two hand-specialized models (HaMeR, WiLoR) on the EHF dataset and confirm this trade-off. Leveraging their complementary strengths, we introduce HySUP—a lightweight pipeline that replaces a SMPLest-X hand region with HaMeR output through a two-stage seam-aware alignment: (i) a PCA-guided Procrustes transform for scale and pose normalization, followed by (ii) a rigid-as-possible refinement that enforces vertex- and normal-level consistency along the wrist seam. Without retraining and with only 150 ms of overhead, HySUP reduces hand PA-MPVPE and PA-MPJPE on EHF by 3.30 mm (27.11 %) and 3.55 mm (29.48 %), respectively, while preserving full-body accuracy. These results highlight a simple path forward: “plug-and-play” fusion of specialist parts can outperform ever-larger holistic models at a fraction of the computational cost. Code and more results can be found on our [project page](#).

Index Terms—Expressive human pose estimation, SMPL-X, 3-D reconstruction, optimization, hybrid models

I. INTRODUCTION

Expressive human pose and shape estimation (EHPS) reconstructs a complete 3-D body—including hands and face—from a single RGB frame, enabling immersive telepresence, XR content creation, and ergonomics-aware human-computer interaction. Holistic foundation models such as SMPLerX [1] and SMPLest-X [2] now predict torso and limb geometry with impressive fidelity, yet their hand outputs remain coarse: fingers are frequently fused or mis-articulated (Fig. 1). The opposite trade-off appears in hand-focused networks like HaMeR and WiLoR [3], [4], which deliver millimetre-level finger detail but lose alignment with the global body frame. This persistent mismatch prompts our central question: *Can we combine the global coherence of holistic models with the fine-*

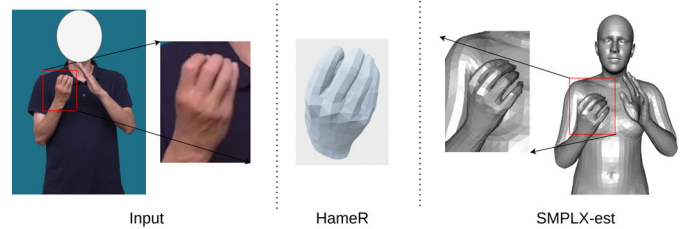


Fig. 1. Illustration of the limitations in current whole-body human models, especially fused or imprecise fingers.

grained accuracy of hand specialists—without retraining and with minimal latency?

To address this, we first systematically benchmark leading holistic and specialized models on the EHF dataset [5]. Our analysis reveals that holistic networks accurately predict torso and limb joints, including surface normals at key landmarks such as wrists, but consistently underperform in finger articulations. Conversely, specialized regressors exhibit the opposite behavior, with HaMeR demonstrating superior hand pose reconstruction. Guided by these insights, we select SMPLest-X as our body expert and HaMeR as our hand specialist.

Building on these models, we introduce **HySUP** (Hybrid Sub-module Unification Pipeline), a lightweight optimization approach that effectively fuses individual part meshes. HySUP aligns wrist seam via a vertex-and-normal consistency objective. Specifically, HySUP employs a two-step optimization: it first uses Procrustes analysis to automatically rescale and align individual part meshes, followed by a surface-aware rigid-as-possible optimization. This second stage refines the mesh integration by enforcing vertex-and-normal consistency at anatomically critical landmarks, including wrist keypoints, to achieve seamless transitions between body parts.

On the EHF benchmark, HySUP achieves an error reduction in hand MPVPE by 3.30 mm (27.11 %) and MPJPE by (29.48 %) compared to the best holistic baseline with just approximately 150 ms of overhead. This heuristic-driven fusion strategy successfully combines the global coherence provided by foundation models with the fine-grained precision of specialized regressors, delivering state-of-the-art whole-

[‡] Work done during internship at Human-Machine Interaction Laboratory, University of Engineering and Technology, Vietnam National University, Hanoi.

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

body accuracy with minimal computational overhead.

II. RELATED WORK

Expressive 3D Human Reconstruction. SMPL-X [6] significantly advanced human modeling by extending the popular SMPL model [5] to incorporate fully articulated hands and expressive facial features. Mathematically, SMPL-X parameterizes the human body through a combination of pose parameters θ (joint rotations), shape parameters β (body proportions), and facial expression parameters ψ . This comprehensive modeling enables expressive representations crucial for nuanced applications such as animation, virtual reality, and affective computing. Building on this, SMPLer-X [1] scales expressive human pose estimation by training Vision Transformer (ViT) backbones on a diverse set of 32 datasets totaling 4.5 million samples, introducing dedicated modules for improved face and hand localization. Subsequently, SMPLest-X [2] further amplifies performance by leveraging minimalist ViT-Huge backbones trained on an unprecedented scale of 10 million instances, demonstrating state-of-the-art results on benchmarks including AGORA [7], UBody [8], and EgoBody [9]. However, despite these advancements, fine-grained reconstruction of hand articulations remains notably coarse.

Hand Mesh Recovery. Hand reconstruction from single RGB images has made substantial strides with the advent of specialized models. MANO [10] has become the standard parametric hand model, parameterizing hand pose and shape succinctly for downstream regression tasks. HaMeR [3] significantly enhances reconstruction precision through a fully transformer-based architecture, employing iterative refinement strategies with a ViT-H backbone trained on extensive hand-specific datasets like FreiHAND [11] and InterHand2.6M [12]. Concurrently, WiLoR [4] integrates localization and reconstruction into a unified pipeline, effectively handling complex scenarios involving multiple hands in uncontrolled environments. Its transformer-based regressor demonstrates superior robustness and real-time performance by training on diverse, large-scale in-the-wild datasets. These specialized hand models outperform holistic approaches in local hand detail reconstruction but still face challenges in ensuring global coherence with full-body meshes.

III. BENCHMARK EVALUATION

A. Setup

a) Dataset: All evaluations are conducted on the *Expressive Hands and Faces* (EHF) dataset [5]. EHF comprises high-resolution RGB frames showing a single subject performing diverse body poses, finger gestures and facial expressions. Ground-truth SMPL-X meshes are obtained by carefully fitting the model to multi-view scans, yielding millimetre-level vertex accuracy for the body, hands and face. Because the dataset is purely for benchmarking, no training images from EHF are used at any stage.

TABLE I
HAND ACCURACY ON EHF (LOWER IS BETTER; BEST IN **BOLD**).

Model	PA-MPVPE ↓			PA-MPJPE ↓		
	Left	Right	Mean	Left	Right	Mean
<i>Holistic models</i>						
SMPLer-X	14.73	14.63	14.68	14.81	14.74	14.77
SMPLest-X	12.42	11.91	12.17	12.32	11.76	12.04
<i>Specialized regressors</i>						
WiLoR	11.44	12.42	11.95	11.17	12.56	11.88
HaMeR	8.92	8.81	8.87	8.43	8.55	8.49

TABLE II
WHOLE-BODY ACCURACY ON EHF (LOWER IS BETTER; BEST IN **BOLD**).

Model	PA-MPVPE ↓	PA-MPJPE ↓
SMPLer-X	38.79	42.79
SMPLest-X	36.33	39.89

b) Metrics: Following recent work [3], [4], we report *Procrustes-aligned mean per-vertex position error* (**PA-MPVPE**) and *Procrustes-aligned mean per-joint position error* (**PA-MPJPE**). Both metrics measure the Euclidean distance (in millimetres) between predicted and ground-truth vertices or joints after a similarity alignment that removes global translation, rotation and uniform scale. Lower values indicate better accuracy.

c) Hardware: All experiments ran on a workstation powered by an Intel Core i9-10900 CPU (2.80 GHz) and an NVIDIA GeForce RTX 3090 GPU.

B. EHF benchmark evaluation

We evaluate our candidate models on the EHF dataset. For hands we test the state-of-the-art regressors **HaMeR** [3] and **WiLoR** [10]; for full-body reconstruction we compare the recent foundation models **SMPLer-X** [1] and **SMPLest-X** [2]. The quantitative results are summarised in Tables I and II.

a) Hand-specific findings: HaMeR is the clear winner, with a mean PA-MPVPE of 8.87 mm—about 3.08 mm ($\approx 26\%$) lower than the next specialized regressor, WiLoR (11.95 mm). Even this weaker hand expert still surpasses the strongest holistic model, SMPLest-X (12.17 mm), confirming that dedicated hand networks consistently outperform whole-body approaches on fine finger geometry.

b) Whole-body performance: SMPLest-X improves on SMPLer-X by 2.46 mm PA-MPVPE ($\approx 6.3\%$ drop) and 2.9 mm PA-MPJPE ($\approx 6.8\%$ drop), delivering the strongest global accuracy among the holistic baselines.

IV. HYSUP: HYBRID SUB-MODULE UNIFICATION PIPELINE

Holistic regressors such as SMPLest-X accurately recover global body poses but struggle with detailed finger articulations, resulting in fused or imprecise hands. In contrast, specialized networks like HaMeR deliver high-fidelity finger

TABLE III
INFLUENCE OF VERTEX-NORMAL TRADE-OFF λ ON OPTIMIZATION
CONVERGENCE. LOWER IS BETTER.

λ	Right hand $\downarrow$			Left hand $\downarrow$		
	Steps	Vert. $\times 10^{-3}$	Norm.	Steps	Vert. $\times 10^{-3}$	Norm.
0 (verts)	86.47	3.73	20.52	106.13	30.98	54.98
10^6 (norms)	36.40	0.30	0.25	28.63	0.71	0.54
10^{-2}	87.51	0.29	0.27	82.40	0.59	0.55
10^{-1}	81.33	0.29	0.26	72.00	0.68	0.54
1	83.24	0.29	0.25	62.78	0.70	0.54
10	80.90	0.29	0.25	54.32	0.70	0.54

meshes but fail to maintain coherence within the full-body topology. Motivated by this insight, we introduce HySUP, a lightweight, seam-aware, two-stage fitting pipeline that fuses SMPLest-X’s coarse hand output with HaMeR’s detailed finger meshes. By aligning vertex- and normal-level consistency along the wrist seam, HySUP achieves both global coherence and fine local detail in just approximately 150 ms per frame. Qualitative results are shown in Figure 2.

A. Algorithm

— **Stage 1 (S1): PCA-guided Procrustes.** For each hand mesh $\mathcal{M}_h$ and its corresponding body mesh $\mathcal{M}_b$, we first identify the wrist-anchor vertices $\mathbf{V}^A$ on the template mesh (see Figure 3; details are provided in Appendix A). From these anchors, we calculate initial estimates of translation $\mathbf{t}_0$ and scale s_0 using their centroids and maximal extents. Principal component analysis (PCA) applied to these anchors yields orthonormal bases $\mathbf{U}_h, \mathbf{U}_b \in \text{SO}(3)$. The rotation $\mathbf{R}_0 = \mathbf{U}_b \mathbf{U}_h^T$ then aligns the principal axes, giving us the initial similarity transform $(s_0, \mathbf{R}_0, \mathbf{t}_0)$.

Stage 2 (S2): Seam Optimization. Starting from the initial transform parameters $(s, \mathbf{R}, \mathbf{t}) = (s_0, \mathbf{R}_0, \mathbf{t}_0)$, we minimize the following loss function:

$$\mathcal{L}(s, \mathbf{R}, \mathbf{t}) = \sum_i \|s\mathbf{R}\mathbf{V}_h^{(i)} + \mathbf{t} - \mathbf{V}_b^{(i)}\|^2 + \lambda \sum_i \left(1 - \langle \mathbf{R}\mathbf{N}_h^{(i)}, \mathbf{N}_b^{(i)} \rangle\right), \quad (1)$$

where $\mathbf{V}_h^{(i)}, \mathbf{V}_b^{(i)}$ and $\mathbf{N}_h^{(i)}, \mathbf{N}_b^{(i)}$ are the corresponding seam vertices and unit normals on HaMeR and SMPLest-X, respectively.

We used the L-BFGS-B optimizer [13], as our formulation involves only soft, smooth constraints and a simple linear bound on the scale parameter ($s > 0$), with convergence declared when the change in the objective between successive iterations fell below 10^{-4} or when 5,000 iterations were reached.

B. Ablation experiments and Analysis

a) *Role of Vertex and Normal Loss:* Table III shows that relying solely on vertex loss ($\lambda = 0$) prioritizes vertex accuracy but results in substantial normal errors (20.52° right hand, 54.98° left hand), highlighting severe surface orientation

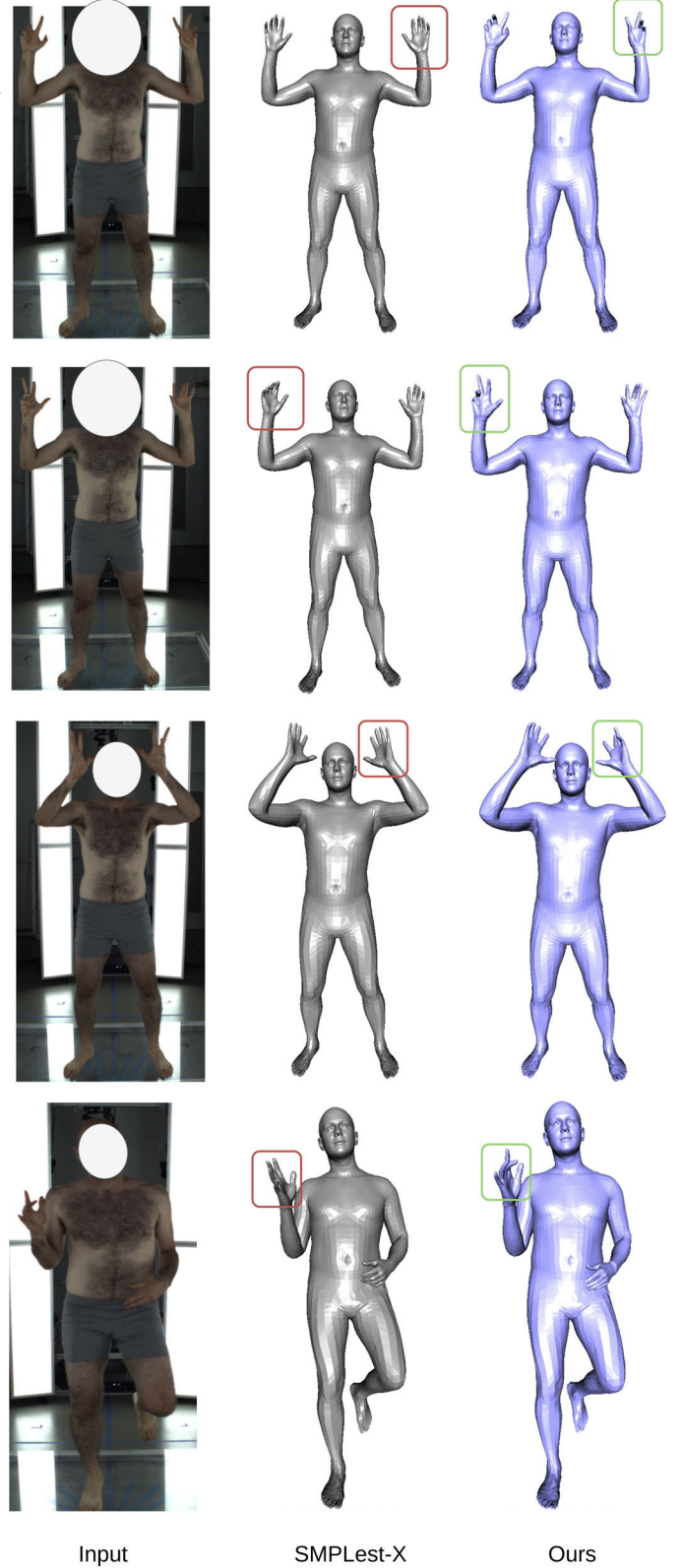


Fig. 2. Qualitative comparison of SMPLest-X results before and after applying HySUP. As can be seen, the hand articulation—especially the fingers—is noticeably improved.

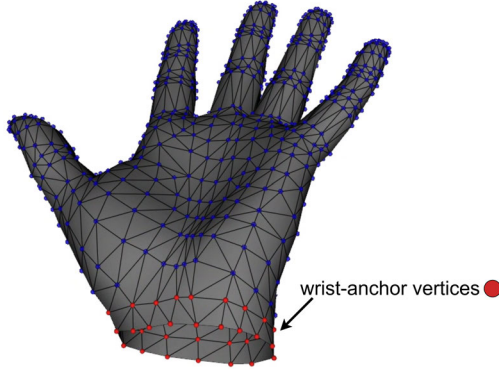


Fig. 3. Position of wrist-anchor vertices

TABLE IV
EFFECT OF ADDING STAGE 1 ON OPTIMIZATION METRICS. LOWER IS BETTER.

Config.	Right hand ↓			Left hand ↓		
	Steps	Vert. $\times 10^{-3}$	Norm.	Steps	Vert. $\times 10^{-3}$	Norm.
Only S2	134.81	0.29	0.26	109.30	0.67	0.54
S1 + S2	31.68	0.29	0.25	26.44	0.67	0.54

issues (Illustrated in Figure 4). Even a minimal inclusion of normal loss ($\lambda = 10^{-2}$) drastically reduces these errors to roughly 0.26° without significantly affecting vertex precision. Moreover, the robustness of the vertex-normal trade-off is evident: parameter choices between 10^{-2} and 10 yield consistently stable results, greatly simplifying tuning. Notably, incorporating normal loss also significantly accelerates optimization convergence; a purely normal-oriented loss ($\lambda = 10^6$) reduces iteration counts by nearly half. Nevertheless, using exclusively vertex or normal loss presents limitations: pure-vertex optimization leads to orientation distortion, while pure-normal optimization accelerates convergence at the cost of vertex accuracy. Thus, combining vertex and normal losses with a balanced weighting ($\lambda \approx 1$) is optimal, ensuring high geometric fidelity, accurate surface normals, and efficient convergence.

b) Role of Optimization Stages: Table IV demonstrates the effect of introducing PCA-guided Procrustes (Stage 1). Although this initial alignment stage does not significantly

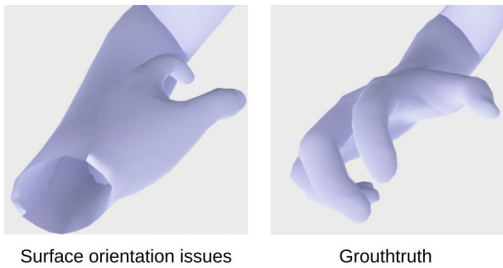


Fig. 4. Illustration of optimization without normal loss, causing surface orientation issues.

improve final vertex or normal accuracy, it dramatically reduces computational cost. Specifically, adding Stage 1 leads to roughly a fourfold decrease in iteration counts and corresponding computational overhead, cutting optimization time substantially (from about 134 iterations to approximately 32 for the right hand and from 109 to about 26 for the left hand). This stage introduces negligible computational overhead (less than 1 ms), thereby significantly boosting overall pipeline efficiency.

V. CONCLUSION AND LIMITATION

We introduced **HySUP**, a light-weight yet effective pipeline that fuses a holistic foundation model (SMPLest-X) with a specialist hand regressor (HaMeR) through a two-stage, seam-aware similarity optimization. On the EHF benchmark HySUP lowers hand PA-MPVPE/MPJPE by 3.3 mm (27.11 % relative) and 3.55 mm (29.48 % relative) over the strongest holistic baseline, while adding only 150 ms of computation. More broadly, the “plug-and-play part fusion” paradigm can easily generalise to other domains such as faces, footwear, or garments as stronger specialized models emerge. As accuracy improves through this method, previously challenging applications—such as sign language understanding—become more approachable.

Limitations. HySUP’s gain is bounded by the quality of the underlying part networks; if a future holistic model fully solves hand articulation our relative advantage will shrink.

REFERENCES

- [1] Z. Cai, W. Yin, A. Zeng, C. Wei, Q. Sun, W. Yanjun, H. E. Pang, H. Mei, M. Zhang, L. Zhang, et al., Smpler-x: Scaling up expressive human pose and shape estimation, *Advances in Neural Information Processing Systems* 36 (2023) 11454–11468.
- [2] W. Yin, Z. Cai, R. Wang, A. Zeng, C. Wei, Q. Sun, H. Mei, Y. Wang, H. E. Pang, M. Zhang, et al., Smplest-x: Ultimate scaling for expressive human pose and shape estimation, *arXiv preprint arXiv:2501.09782* (2025).
- [3] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, J. Malik, Reconstructing hands in 3d with transformers, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9826–9836.
- [4] R. A. Potamias, J. Zhang, J. Deng, S. Zafeiriou, Wilor: End-to-end 3d hand localization and reconstruction in-the-wild, *arXiv preprint arXiv:2409.12259* (2024).
- [5] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, M. J. Black, Smpl: A skinned multi-person linear model, in: *Seminal Graphics Papers: Pushing the Boundaries*, Volume 2, 2023, pp. 851–866.
- [6] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, M. J. Black, Expressive body capture: 3d hands, face, and body from a single image, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10975–10985.
- [7] P. Patel, C.-H. P. Huang, J. Tesch, D. T. Hoffmann, S. Tripathi, M. J. Black, Agora: Avatars in geography optimized for regression analysis, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13468–13478.
- [8] Z. Yang, A. Zeng, C. Yuan, Y. Li, Effective whole-body pose estimation with two-stages distillation, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4210–4220.
- [9] S. Zhang, Q. Ma, Y. Zhang, Z. Qian, T. Kwon, M. Pollefeys, F. Bogo, S. Tang, Egobody: Human body shape and motion of interacting people from head-mounted devices, in: *European conference on computer vision*, Springer, 2022, pp. 180–200.

- [10] J. Romero, D. Tzionas, M. J. Black, [Embodied hands: modeling and capturing hands and bodies together](#), ACM Trans. Graph. 36 (6) (Nov. 2017). doi:10.1145/3130800.3130883.
URL <https://doi.org/10.1145/3130800.3130883>
- [11] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, T. Brox, *Freihand: A dataset for markerless capture of hand pose and shape from single rgb images*, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 813–822.
- [12] G. Moon, S.-I. Yu, H. Wen, T. Shiratori, K. M. Lee, *Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image*, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16, Springer, 2020, pp. 548–564.
- [13] C. Zhu, R. H. Byrd, P. Lu, J. Nocedal, *Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization*, ACM Transactions on mathematical software (TOMS) 23 (4) (1997) 550–560.

APPENDIX A

WRIST–ANCHOR VERTEX INDICES

Table V lists the MANO¹ vertex IDs that define the wrist seam used by HySUP in both Stage 1 (PCA–guided Procrustes) and Stage 2 (Seam Optimization). These anchors were selected by inspecting the MANO template; they form a closed loop that is one-to-one with the corresponding SMPL-X wrist ring.

TABLE V
MANO WRIST–ANCHOR VERTEX IDs USED IN HYSUP.

33	34	37	38	39	44	52	53	78
79	85	91	92	107	108	116		
117	118	119	120	121	122	190	191	210
211	214	215	234	239	279			

¹hand topology, mesh resolution = 778 vertices.