# Reconstructing Signing Avatars From Video Using Linguistic Priors
## ***Supplementary Material***

Maria-Paola Forte    Peter Kulits    Chun-Hao Huang    Vasileios Choutas    Dimitrios Tzionas
Katherine J. Kuchenbecker    Michael J. Black
Max Planck Institute for Intelligent Systems, Stuttgart and Tübingen, Germany
{forte,kjk}@is.mpg.de    {kulits,chuang2,vchoutas,dtzionas,black}@tue.mpg.de

This document is a companion to our main paper, providing additional details and results. In addition, please see the supplemental video at sgnify.is.tue.mpg.de, which provides video results that illustrate the performance of our method.

## S.1. Examples of Sign Classes

Table S.1 provides representative images of our eight sign classes to supplement Tab. 1 in the main paper. The videos of these signs appear in the supplemental video.

## S.2. SGNify Objective

The full objective function of SGNify is:

$$\begin{aligned} E(\theta,\psi,\beta) = \; & \lambda_{\theta_b} E_{\theta_b} + \lambda_{m_h} E_{m_h} + \\ & E_J + \lambda_\alpha E_\alpha + E_O + \\ & \lambda_P E_P + \lambda_A E_A + \\ & L_s + \sum_{h \in \{r,l\}} L_i^h + \\ & \lambda_t L_t + \lambda_{st} L_{st}, \end{aligned} \tag{S.1}$$

where $\theta$ is the full set of optimizable pose parameters, and $\theta_b$ and $m_h$ are the pose vectors for the body and the two hands. The body pose is modeled by a VAE (called Vposer) that transforms the body pose, $\theta_b$, into a latent vector $Z$. We enforce an L2 prior in this space, *i.e.*, $E_{\theta_b}(\theta_b) = \|Z\|^2$. For the hands, SMPL-X uses a low-dimensional PCA pose space such that $\theta_h = \sum_{n=1}^{|m_h|} m_{h_n} \mathcal{M}$, where $\mathcal{M}$ are principal components capturing the finger pose variations and $m_{h_n}$ are the corresponding PCA coefficients. Thus, $E_{m_h}(m_h)$ is an L2 prior on the coefficients $m_h$. $E_J$ represents the joint re-projection loss, and $E_\alpha(\theta_b)$ is a prior penalizing extreme bending only for elbows and knees. For more details on these terms, please refer to the original paper of SMPLify-X [9]. $E_O$ is a bone-orientation term, which factors out the residual of the parent joint from the residual of the child joint. For more details about this term, please refer to the original paper of RICH [6]. $E_P$ and $E_A$ are used to prevent self-interpenetration. When self-contact occurs, the $E_P$ term pushes vertices that are inside the mesh to the surface, and $E_A$ aligns the surface normals of the vertices in contact. For more details, please refer to the original paper of TUCH [7].

We added $L_s$ and $L_i^h$ to enforce our linguistic constraints: $L_s$ represents the symmetry constraints, and $L_i^h$ the hand-pose invariance of the right ($r$) and left ($l$) hands, as described in Sec. 3.2 in the paper. We also added a temporal loss $L_t$ on the body- and hand-pose vectors and a standing loss $L_{st}$ to penalize deviations from a standing pose when none of the feet keypoints are detected; specifically, this penalization is applied to the joints below the pelvis and to the spine.

Finally, each $\lambda$ denotes the influence weight of each loss term. For more details on the exact $\lambda$ values and insights on the full SGNify objective, please see the code, which can be reached from the project URL.

We optimize our objective function using the trust-region Newton conjugate gradient method [8]. Note that we do not optimize for the shape $\beta$ and the facial expressions $\psi$, as explained in the main paper.

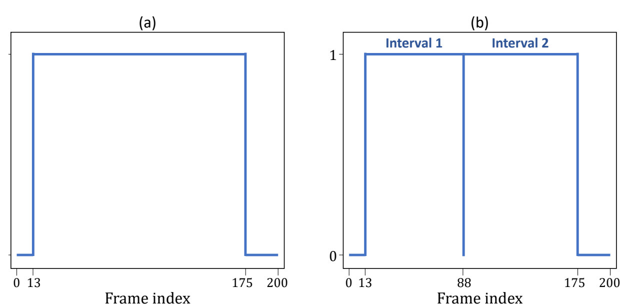



Figure S.1. We consider an example sequence of 200 frames. (a) Static hand: Frames whose value on the y-axis is 1 are candidates for identifying $\theta^h_{ref}$. (b) Transitioning hand and input features for the sign-group classifier: The first interval shows candidates for $\theta^h_{ref,i}$, and the second one for $\theta^h_{ref,f}$.

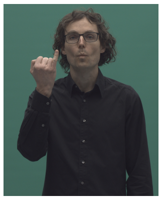
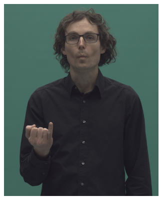
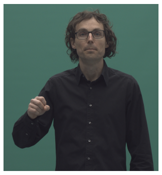
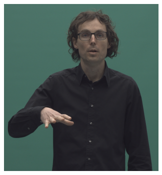
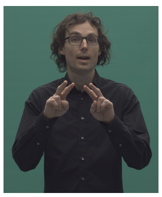
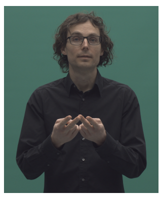
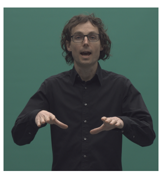
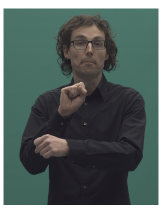
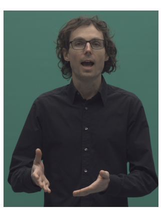
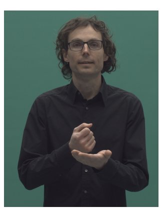
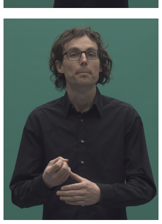
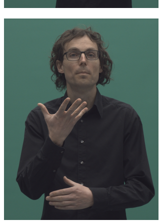
| Initial Hand Pose | Final Hand Pose | Class | Hand-Pose Symmetry | Hand-Pose Invariance Dominant | Hand-Pose Invariance Non-dominant |
|---|---|---|---|---|---|
|  |  | 0a | ✗ | static | ✗ |
|  |  | 0b | ✗ | transitioning | ✗ |
|  |  | 1a | ✓ | static | static |
|  |  | 1b | ✓ | transitioning | transitioning |
|  |  | 2a | ✓ | static | static |
|  |  | 2b | ✗ | transitioning | static |
|  |  | 3a | ✗ | static | static |
|  |  | 3b | ✗ | transitioning | static |

Table S.1. Linguistic constraints defining the eight sign classes. See supplemental video.

⟨*hns*⟩ ::= [SYMMETRY] ⟨*block*⟩

⟨*block*⟩ ::= [⟨*handshape_block*⟩ | ⟨*non_handshape_block*⟩]*

⟨*handshape_block*⟩ ::= HANDSHAPE [HANDSHAPE_MODIFIER | HANDSHAPE_FINGER_LOCATION]*

⟨*non_handshape_block*⟩ ::= ⟨*par*⟩ | ⟨*seq*⟩ | ⟨*fusion*⟩ | EXTENDED_FINGER_LOCATION | PALM_ORIENTATION | MOVEMENT | MOVEMENT_MODIFIER | LOCATION | LOCATION_MODIFIER | OTHER_SYMBOL__NO_GROUP

⟨*par*⟩ : HAMPARBEGIN ⟨*block*⟩ [HAMPLUS ⟨*block*⟩] HAMPAREND

⟨*seq*⟩ : HAMSEQBEGIN ⟨*block*⟩ HAMSEQEND

⟨*fusion*⟩ : HAMFUSIONBEGIN ⟨*block*⟩ HAMFUSIONEND

Figure S.2. Constructed HamNoSys EBNF grammar.

## S.3. Intervals for Selecting the Candidate Frames for the Reference Hand Poses ($\theta^h_{\textit{ref}}$, $\theta^h_{\textit{ref},i}$, and $\theta^h_{\textit{ref},f}$)

When articulating an isolated sign, signers start and end in a rest pose. SGNify identifies the beginning and end of the sequence based on when the hands begin to move. After automatic trimming, the initial and final frames of the sequence show the transition from the rest pose to the pose(s) characteristic of the sign. We observe that the transition from the rest pose to the core part of the sign usually happens around $t = 0.5 * T/8$, and the transition from the sign to the rest pose typically occurs around $t = 7 * T/8$, where $T$ is the number of frames in the motion sequence. As a result, we assume the core part of a sign to happen between $0.5*T/8 < t < 7*T/8$. Figure S.1a shows the frames during which the transitions from/to the rest pose happen (indicated with 0) and the frames during which the sign is articulated (indicated with 1) for a sample trimmed recording containing 200 frames. To identify the two key poses representing the initial and final hand poses ($\theta^r_{\textit{ref},i}$ and $\theta^r_{\textit{ref},f}$), we consider two different intervals; we expect to see the first hand pose at the beginning of the sequence (first interval shown in Fig. S.1b) and the second hand pose at the end (second interval shown in Fig. S.1b).

## S.4. HamNoSys Parsing

We construct an Extended Backus–Naur form (EBNF) grammar (see Fig. S.2) to parse HamNoSys [5] annotations to a form where we can extract labels to train our sign group classifier. HamNoSys is a universal sign-language phonetic transcription system that can be used to represent all hand poses and movements that constitute a sign; *i.e.,* someone reading a HamNoSys annotation would be able to fully reproduce the sign it represents. We parse these transcriptions on the annotated Corpus-Based Dictionary of Polish Sign Language (CDPSL) [1], and we assign our classes to the clips as follows:

**Class 0a:** There is one *handshape_block* nonterminal and no SYMMETRY terminal is present.

**Class 0b:** There are two *handshape_block* nonterminals, the two *handshape_block* nonterminals are not equal, a HAMREPLACE terminal is present, and no SYMMETRY or REPEAT terminals are present.

**Class 1a:** There is one *handshape_block* nonterminal and a SYMMETRY terminal is present.

**Class 1b:** There are two *handshape_block* nonterminals, they are not equal, a HAMREPLACE terminal is present, and no SYMMETRY or REPEAT terminals are present.

**Class 2a:** There are two *handshape_block* nonterminals, they are equal, they fall within a *par* nonterminal, and no SYMMETRY terminal is present.

**Class 2b:** There are three *handshape_block* nonterminals, the first two are equal, a HAMREPLACE terminal is present, and no SYMMETRY or REPEAT terminals are present.

**Class 3a:** There are two *handshape_block* nonterminals, they are not equal, they fall within a *par* nonterminal, and no SYMMETRY terminal is present.

**Class 3b:** There are three *handshape_block* nonterminals, the first is not equal to the second, a HAMREPLACE terminal is present, and no SYMMETRY or REPEAT terminals are present.

Note that the SYMMETRY parameter from HamNoSys

Figure S.3. The multi-view setup comprises 12 synchronized RGB cameras. A close-up frontal camera is zoomed in to focus on the hands and face. Another frontal camera captures the entire front of the body. Two top-lateral cameras acquire images with a top-down view. Four lateral cameras are placed at hip level and capture the whole body; two are slightly behind the signer, and the other two are slightly in front. Two frontal-lateral cameras also have a full-body view, looking slightly down. Finally, two other frontal cameras, one with a bottom-up view and one with a top-down view, are focused on the hands. The participant stands on a $1.5$ m $\times$ $1.5$ m platform of adjustable height located in front of a green screen.


Figure S.4. Sample frames and reconstructions from segments of the German sentence: *Der* ***Vater*** *muss für die Reparatur seines* ***Autos*** *viel Geld* ***ausgeben***.

refers to Battison's symmetry condition [3], which also includes the signer's arm movement and not only the hand pose; in contrast, our symmetry constraint applies only to hand pose.

## S.5. SGNify Extensions

### S.5.1. Multi-view

If multi-view video is available, SGNify is easily extended to this case. We used 12 synchronized RGB cameras (see Fig. S.3) at 90 fps to capture the same participant used in the quantitative evaluation plus two additional signers, a native signer and an interpreter with 17 years of experience. Each participant articulated all signs in our German Sign Language (DGS) corpus (see Sec. 4 in the paper). A close-up frontal camera is zoomed in to focus on the hands and face of the signer and has a view similar to existing sign-language videos. Another frontal camera captures the whole front body of the participant. Two top-lateral cameras acquire images with a top-down view. Four lateral cameras are placed at hip level and capture the whole body; two are slightly behind the signer, and the other two are slightly in front. Two frontal-lateral cameras also have a full-body view, looking slightly down. Finally, two other frontal cameras, one with a bottom-up view and one with a top-down view, are focused on the hands. The participant stands on a 1.5 m $\times$ 1.5 m platform of adjustable height located in front of a green screen. Then, multi-view SGNify is used to fit SMPL-X. We follow Huang *et al.* [6] to combine the keypoint predictions of different cameras. A person-specific $\beta$ is obtained with a 3D scanner. Sample multi-view results are shown in the supplemental video.

### S.5.2. Continuous Sign Language Capture (CSLC)

SGNify can also be used for CSLC. Besides isolated signs, our corpus contains ten sentences articulated by the three interpreters during the four sessions; one session with a 54-camera Vicon mocap system at 120 fps synchronized with a frontal $4112 \times 3008$ RGB camera at 60 fps, framing an upper-body view as typically found in SL video (Sec. 4 in the paper) and three sessions with the multi-view setup (see Sec. S.5.1). Depending on the interpreter, different DGS versions of the same German sentences were proposed.

We conduct an exploratory quantitative study with twelve sentences (ten main sentences and two variations) collected as in Sec. 4 in the paper and analyzed as in Sec. 5.1 in the paper. Tab. S.2 shows the mean TR-V2V error across the twelve sentences for four methods and three body regions. This experiment compares SGNify with FrankMocap [11], PIXIE [4], PyMAF-X [13], and our baseline SMPLify-SL. SGNify achieves the lowest error for the upper body and both hands, beating the state-of-the-art methods. It is interesting to notice that while FrankMocap has a hand-pose error lower than PyMAF-X in our previous quantitative experiment (see Sec. 4 in the paper), this is not true in this second experiment. This inconsistency further emphasizes the limitations of a per-frame metric for sign language. In the future, a perceptual study should be conducted to evaluate the recognition of the reconstructed sentences with proficient signers. Such an experiment will give more insights about the next crucial steps for CSLC. Fig. S.4 shows sample frames and SGNify's reconstructions from a sentence of this exploratory study.

| Method | Upper Body | Left Hand | Right Hand |
|---|---|---|---|
| FrankMocap [11] | 74.93 | 23.70 | 19.57 |
| PIXIE [4] | 59.09 | 24.79 | 20.19 |
| PyMAF-X [13] | 68.30 | 22.51 | 18.49 |
| SMPLify-SL | 55.71 | 21.14 | 18.60 |
| SGNify | **54.72** | **20.28** | **17.44** |

Table S.2. Mean TR-V2V error (mm) on fluid sentences.

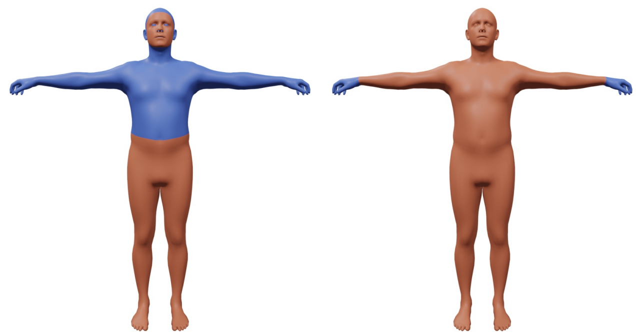


Figure S.5. Blue vertices are used to calculate vertex error metrics, while red vertices are ignored. The left image shows the vertices used for the column of quantitative results labeled "Upper Body", *i.e.*, upper-body vertices. The right image shows the vertex subsets for the left and right hands. Best viewed in color.

## S.6. Vertices for Quantitative Analysis

Figure S.5 illustrates the subsets of vertices selected for the quantitative evaluation.

## S.7. Second Perceptual Study

Fig. S.6 shows a sample frame represented with each of the four methods used in the second perceptual study: real video, the solid purple avatar from the first study, the same avatar wearing a black long-sleeved t-shirt, and a fully textured human character adapted from Meshcapade [2].

## S.8. Additional Examples

Figure S.7 shows additional examples from the Real SASL [10] and CDPSL [1] datasets. Figure S.8 shows addi-

Figure S.6. Sample frames from the four methods presented in the second perceptual study: real video, the solid purple avatar from the first study, the same avatar wearing a black long-sleeved t-shirt, and a fully textured human character.

tional examples from The American Sign Language Handshape Dictionary [12] and our collected DGS dataset (see Sec. 4 in the paper).


Figure S.7. Additional examples on the Real SASL and CDPSL sign-language dictionaries.

Figure S.8. Additional examples on The American Sign Language Handshape Dictionary and our captured dataset.

## References


[1] CDPSL: Corpus-based Dictionary of Polish Sign Language. https://www.slownikpjm.uw.edu.pl/. 3, 5

[2] Meshcapade GmbH, Tübingen, Germany. https://meshcapade.com, 2022. 5

[3] Robbin Battison. *Lexical Borrowing in American Sign Language*. Education Resources Information Center (ERIC) , 1978. 5

[4] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, pages 792–804, 2021. 5

[5] Thomas Hanke. HamNoSys – representing sign language data in language resources and language processing contexts. In *International Conference on Language Resources and Evaluation (LREC)*, volume 4, pages 1–6, 2004. 3

[6] Chun-Hao Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human–scene contact. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13274–13285, 2022. 1, 5

[7] Lea Müller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, and Michael J. Black. On self-contact and human pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9990–9999, 2021. 1

[8] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, NY, USA, second edition, 2006. 1

[9] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 1

[10] Real SASL: Real South African Sign Language. https://www.realsasl.com/. 5

[11] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. FrankMocap: A monocular 3D whole-body pose estimation system via regression and integration. In *International Conference on Computer Vision Workshops (ICCVw)*, pages 1749–1759, 2021. 5

[12] Richard A. Tennant, Marianne Gluszak, and Marianne Gluszak Brown. *The American Sign Language Handshape Dictionary*. Gallaudet University Press, 2010. 6

[13] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. PyMAF-X: Towards well-aligned full-body model regression from monocular images. *arXiv preprint arXiv:2207.06400*, 2022. 5