Reconstructing Signing Avatars From Video Using Linguistic Priors

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

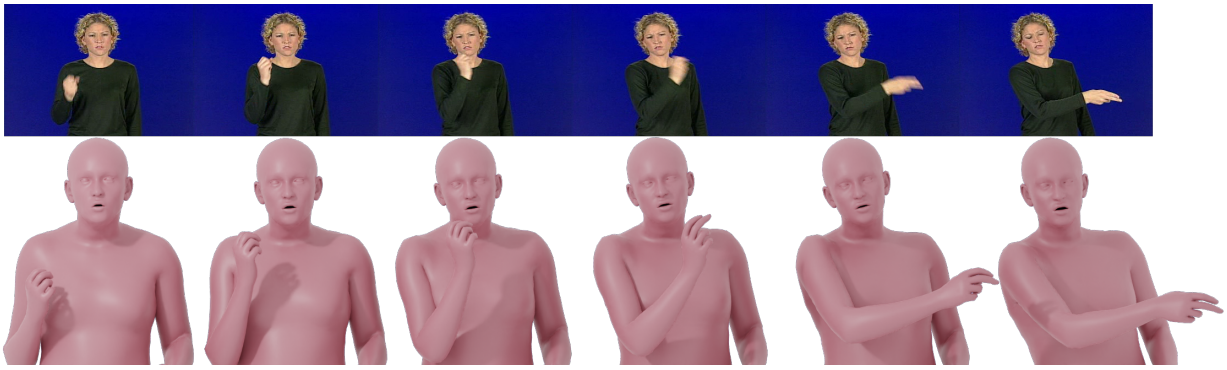

Figure 1. Given a monocular, in-the-wild video of a sign-language sign, SGNify automatically reconstructs a 3D body with accurate hand pose, facial motion, and body pose. Note that motion blur obscures the finger articulations in several video frames; this is a common problem. Our novel linguistic priors enable accurate 3D reconstruction despite such image degradation.

Abstract

Sign language (SL) is the primary method of communication for the 70 million Deaf people around the world. Video dictionaries of isolated signs are a core SL learning tool. Replacing these with 3D avatars can aid learning and enable AR/VR applications, improving access to technology and online media. However, little work has attempted to estimate expressive 3D avatars from SL video; occlusion, noise, and motion blur make this task difficult. We address this by introducing novel linguistic priors that are universally applicable to SL and provide constraints on 3D hand pose that help resolve ambiguities within isolated signs. Our method, SGNify, captures fine-grained hand pose, facial expression, and body movement fully automatically from in-the-wild monocular SL videos. We evaluate SGNify quantitatively by using a commercial motion-capture system to compute 3D avatars synchronized with monocular video. SGNify outperforms state-of-the-art 3D body-pose- and shape-estimation methods on SL videos. A perceptual study shows that SGNify's 3D reconstructions are significantly more comprehensible and natural than those of previous methods and are on par with the source videos. Code and data are available at sgnify.is.tue.mpg.de.

1. Introduction

It is estimated that over 466 million people have disabling hearing loss [13] and more than 70 million people use sign language (SL) as their primary means of communication [52]. Increasing use of digital communication motivates research on capturing, understanding, modeling, and synthesizing expressive 3D SL avatars. Existing datasets and dictionaries used in SL recognition (SLR), translation (SLT), and production (SLP) are primarily limited to 2D video because the technology required to capture 3D movement is prohibitively expensive, requires expertise to operate, and may limit the movements of the signer. Dictionaries of isolated signs are a core SL learning tool, and many SLs have online 2D video dictionaries. The Deaf community is actively seeking 3D dictionaries of isolated signs to aid learning [40]. The current approach to creating such 3D signing dictionaries is fully manual, requiring an artist or a HamNoSys [21] expert, and the resulting avatars often move unnaturally [3]. We aim to automatically reconstruct expressive 3D signing avatars from monocular SL video, which we term *Sign Language Capture (SLC)*. We focus on SLC of isolated signs.

3D reconstruction of human pose and shape has received significant attention, but accurate 3D hand-pose estimation

remains challenging from in-the-wild video. Challenges include the high number of degrees of freedom present in hands [5], frequent occurrence of self-contact and self-occlusions [37, 46], low resolution, and motion blur caused by fast motions [50] that cause hand pose to be unrecognizable in many frames (see Fig. 1). To address these issues, we exploit the linguistic nature of SL itself to develop novel priors that help disambiguate hand poses in SL videos, leading to accurate 3D reconstructions. This is a novel use of linguistic "side information" to improve 3D reconstruction.

Based on hand movements and poses, Battison [4] defines five linguistic classes that contain all SL signs. We build on that work to define eight classes and formalize these as mathematical priors on 3D hand shape. We combine Battison's first two classes and place all one-handed signs in class 0, while two-handed signs are arranged in classes 1, 2, or 3, depending on how the non-dominant hand participates in the articulation of the sign. We then divide each of these four classes into two subclasses depending on whether the pose of the active hand(s) changes during the articulation of the sign. We introduce two class-dependent SL linguistic constraints that capture 1) symmetry and 2) hand-pose invariance. Under Battison's SL symmetry condition [4], when both hands actively move, the articulation of the fingers must be identical; the same is true for one class of two-handed signs in which only the dominant hand moves. We formalize this concept as a regularization term that encourages the pose of the two hands to be similar for such signs. Coupling the hand poses in this way effectively increases the image evidence for a pose, which improves estimates for challenging videos. Our invariance constraint uses the observation that hand pose is either static or transitions smoothly from one pose to another during the articulation of the sign; other significant changes to hand pose are not common in SL. Specifically, we extract a characteristic "reference pose sequence" (RPS) to describe each local hand pose during the sign articulation, and we penalize differences between the RPS and the estimated hand pose in each frame. These two priors of symmetry and hand-pose invariance are universally applicable to all sign languages.

The hands alone, however, are not sufficient to accurately reproduce SL. Information is conveyed holistically in SL through hand gestures, facial expressions, and upper-body movements in 3D space. To combine these, we use a 3D whole-body model, SMPL-X [41], that jointly models this information (see Fig. 1).

Our novel hand-pose constraints are formulated to be incorporated into the loss function for training a neural network regressor or into the objective function of optimization-based methods. In general, optimization-based methods are more computationally intensive but produce more accurate results when limited training data is available, so we take this approach here and build on the SMPLify-X method [41]. To appropriately incorporate our terms into the objective function, we need to know the class of the sign. We train a simple model that extracts features from the raw video and determines the class to which the depicted sign belongs. While SMPLify-X is a good foundation for the hands and body, we find that it does not capture expressive facial motions well. Consequently, we use a more expressive face regressor, SPECTRE [19], to capture the face parameters. We call our method SGNify.

To quantitatively evaluate SGNify, we capture a native German (DGS) signer with a frontal RGB camera synchronized with a 54-camera Vicon motion capture system and recover ground-truth meshes from the Vicon markers [34]. We run SGNify on the RGB video and compute 3D vertex-to-vertex (V2V) error between our resulting avatars and the ground-truth meshes. We find that SGNify reconstructs SMPL-X meshes more accurately than the competition.

We conduct a perceptual evaluation in which we present proficient signers with a video of either an estimated SMPL-X avatar or the real-person source video and task them with identifying the sign being performed. Participants also rate their ease in recognizing the sign and the naturalness of the articulation. Our results show that SGNify reconstructs 3D signs that are as recognizable as the original videos and consistently more recognizable, easier to understand, and more natural than the existing state of the art. We also evaluate SGNify in a multi-view setting and on continuous signing videos. Despite not being designed for the latter, SGNify captures the meaning in continuous SL.

SGNify represents a step towards the automatic reconstruction of natural 3D avatars from sign-language videos. Our key contribution is the introduction of novel linguistic priors that are universal and helpful to constrain the problem of hand-pose estimation from SL video. SGNify is designed to work on video from different SL dictionaries across languages, backgrounds, people, trimming, image resolution, and framing, as visible in Sup. Mat. and in the video on our project page. This capability is critical to capture 3D signing at scale, which will enable progress on learning SL avatars. Our code and motion-capture dataset are available for research purposes at sgnify.is.tue.mpg.de.

2. Related Work

Expressive 3D Humans From RGB Images: Until recently, human-pose estimation has focused on the estimation of 2D [12] or 3D [51] joints of the hands and body, as well as those of facial features [6] from single images. In addition to methods that estimate a sparse set of landmarks, there are multiple methods that estimate the parameters of morphable models for the hand [22, 32, 36, 58], face [11, 16, 18, 20], and body [8, 25, 27–29, 33, 39, 57]. The advent of expressive 3D body models like SMPL-X [41], Adam [26], and GHUM [54] has enabled research on esti-

mating the full 3D body surface [9, 17, 41, 44, 49, 53, 55]. Such body models are ideal for representing the expressiveness of SL but have rarely been applied to this domain [30].

Human Pose for Sign Language: To enable detailed 3D pose estimation from images, How2Sign [14] provides 3D skeleton reconstructions for three hours of data captured in a Panoptic Studio [24]. However, the skeletal representation lacks the richness of a full 3D body model and omits surface details that are important for communication [38]. Kratimenos *et al*. [30] use SMPLify-X to estimate 3D pose and shape on the GSLL sign-language dataset [48]. They compare SL recognition accuracy using features from raw RGB images, OpenPose [7] 2D skeletons, and SMPL-X bodies and observe the best automated recognition results with SMPL-X, illustrating the benefit of using a 3D model. They also highlight the importance of capturing the face and body; in an ablation study, they show that neglecting the face and body harms recognition accuracy [30]. However, their SMPL-X reconstructions use existing off-the-shelf methods and lack visual realism. SMPLify-X [41] and other recent 3D pose-reconstruction methods [17, 44], as well as keypoint detectors, struggle when applied to SL video due to challenging self-occlusion, hand–hand and hand–body interactions [38], motion blur [50], and cropping inherent to SL. SignPose [31] is a 3D-pose-lifting method for SL; it uses manually created synthetic SL animations to infer a textured avatar from single RGB images. SignPose requires all OpenPose keypoints above the pelvis to be detected, which is unrealistic in noisy SL videos. We address these challenges by incorporating sign-language knowledge in the form of linguistic constraints. Since the early 2000s, the integration of linguistic information has been known to be beneficial to both SLR [10] and SLP [35], but this strategy has not previously been applied to SLC.

3. Method

We introduce SGNify, an offline method for reconstructing 3D body shape and pose of SL from monocular RGB video. SGNify centers around a key insight: SL signs follow universal linguistic rules that can be formulated as class-specific priors and used to improve hand-pose estimation. Our full pipeline is shown in Fig. 2.

3.1. SMPLify-SL: Baseline for Sign-Language Video

Our baseline method builds on SMPLify-X [41], which estimates SMPL-X [41] parameters from RGB images. SMPL-X is a 3D body model, representing whole-body pose and shape, including finger articulations and facial expressions. SMPL-X is a function, $M(\theta, \beta, \psi)$, parameterized by body pose θ (including hand pose θ_h), body shape β, and facial expressions ψ, that outputs a 3D body mesh.

To create a strong baseline, we extend SMPLify-X to video by adapting it in the following ways: (1) We cope with the upper-body framing typical of SL videos by changing the heuristic used for camera initialization and the estimation of the out-of-view lower-body joints. (2) Since human motion is locally smooth in time, we initialize $\theta_t \in \mathbb{R}^{|\theta|}$ with θ_{t-1} and include a zero-velocity loss on the hands and body to encourage smooth reconstructions. (3) We estimate shape parameters (β) over multiple frames by taking the median of the parameter estimates and not-optimizing them during the per-frame reconstruction. (4) To better capture the frequent hand–hand and hand–body interactions (mainly with the face and the chest), we employ the more robust self-contact loss of Müller *et al*. [39] instead of the original SMPLify-X interpenetration term. (5) For each frame, we pre-compute the facial expressions (ψ) and jaw poses with SPECTRE [19]. These parameters are substituted into SMPL-X at the end of the optimization. SPECTRE can be swapped for any method whose expression parameters are consistent with those of SMPL-X , *e.g.*, EMOCA [11]. We denote the baseline SMPLify-SL.

3.2. Linguistic Constraints

State-of-the-art optimization- and regression-based human pose estimation methods struggle on SL video, particularly with the estimation of hand pose. We address this challenge by formulating linguistic constraints as additional losses on hand pose and integrating them into the SMPLify-SL objective function. First, we adapt the five sign-classification and morpheme-structure conditions introduced for American Sign Language (ASL) by Battison [4] to divide signs into four primary classes:

Class 0: one-handed signs in which only the dominant hand articulates the sign.

Class 1: two-handed signs in which both hands are active. They share the same poses and perform the same movement in a synchronous or alternating pattern. This class includes all signs that follow Battison's symmetry condition [4].

Class 2: two-handed signs in which the dominant hand is active, the non-dominant hand is passive (its position and pose do not change during the articulation of the sign), and the two hands have the same initial pose.

Class 3: two-handed signs in which the dominant hand is active, the non-dominant hand is passive, and the two hands have different hand poses. All signs in this class follow Battison's dominance condition [4].

We further divide each class into two subclasses: *subclass a* contains signs in which the hand pose of the active hand(s) does not change throughout the articulation of the sign (**static**), and *subclass b* contains all signs in which the hand pose changes (**transitioning**).

Note that the division into these classes is not limited to ASL; Eccarius *et al*. [15] show that the phonological and prosodic properties of ASL can be successfully transferred to other sign-language lexicons.

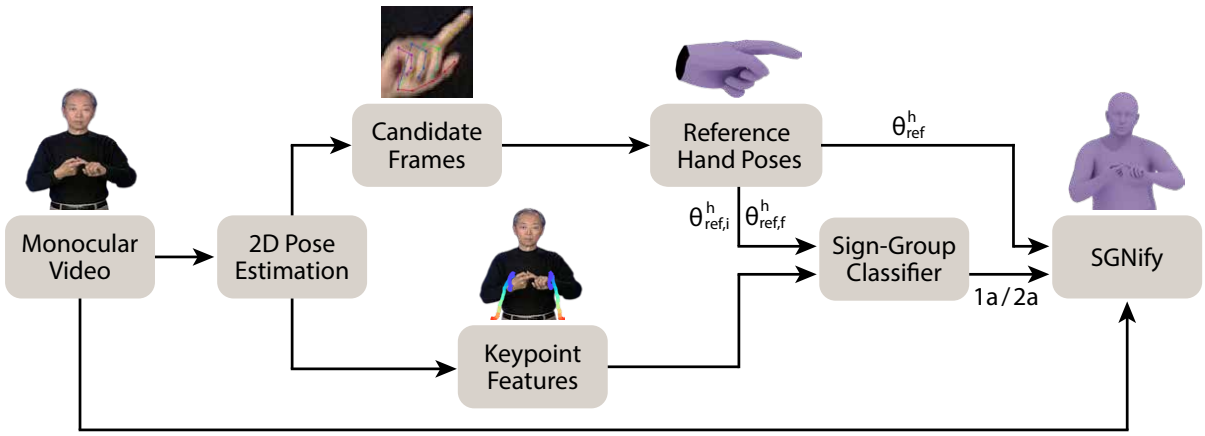

Figure 2. Given a video of a sign-language (SL) sign as input, our method preprocesses the data to first extract 2D keypoints. The hand keypoints are used to select candidate frames for estimating the reference hand poses ($\theta^h_{ref,i}$, $\theta^h_{ref,f}$, and, for static hand poses, also θ^h_{ref}). The initial and final reference hand poses ($\theta^h_{ref,i}$ and $\theta^h_{ref,f}$), together with wrist-keypoint features detected across the sequence, are then fed into our sign-group classifier, which automatically classifies signs in monocular SL video into six groups based on linguistic rules universally applicable to SL [4]. Using the predicted group labels and the relevant reference hand poses, SGNify applies the appropriate linguistic constraints to improve SL 3D hand-pose estimation, especially when the video frame is ambiguous.

Class	Hand-Pose Symmetry	Hand-Pose Invariance	
		Dominant	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 1. Linguistic constraints defining the eight sign classes.

We then convert these linguistic classes into two 3D pose constraints: hand-pose symmetry and hand-pose invariance. Signs in the same class share the same constraints (see Tab. 1 and Tab. S.1 in Sup. Mat.).

Below we describe only the new terms added to the SMPLify-X objective. Please see Sup. Mat. for the full SGNify objective.

3.2.1 Hand-Pose Symmetry

We encourage the left and right hand poses to match for the relevant classes (classes 1a, 1b, and 2a in Tab. 1):

$$L_s = \lambda_s ||\theta^r_t - r(\theta^l_t)||^2_2, \tag{1}$$

where θ^r_t is the finger articulation of the right hand, and $r(\theta^l_t)$ is a reflection function to represent the articulation of the fingers of the left hand as if it were a right hand. This loss penalizes differences in finger poses between the hands.

3.2.2 Hand-Pose Invariance

Each sign has a characteristic reference hand pose sequence (RPS). The RPS defines the hand pose that we expect at each time t during the articulation of the sign. The hand-pose-invariance constraint penalizes differences between the reference hand pose $\theta^h_{ref,t} \in \text{RPS}^h$ and the estimated hand pose θ^h_t:

$$L^h_i = \lambda_i ||\theta^h_{ref,t} - \theta^h_t||^2_2, \tag{2}$$

where h represents either the left or the right hand.

Throughout each sign, the hand pose either stays static or transitions between two poses. When static, only one hand pose, θ^h_{ref}, is representative of the RPS. Signs where the hand pose is transitioning are characterized by two reference hand poses, $\theta^h_{ref,i}$ and $\theta^h_{ref,f}$, corresponding respectively to the initial and final poses. We interpolate $\theta^h_{ref,i}$ and $\theta^h_{ref,f}$ with spherical linear interpolation [45] to obtain intermediate poses. We presently do not consider signs with repeated hand-pose transitions, *e.g.*, STORY in ASL, which occur in a small percentage of signs (∼3%).

3.3. Automatization

To work fully automatically, SGNify must 1) estimate the poses needed to enforce the hand-pose-invariance constraint and 2) classify which sign group is present in a video sequence (see Fig. 2).

To estimate the reference hand poses (θ^h_{ref}, $\theta^h_{ref,i}$, and $\theta^h_{ref,f}$), our method selects candidate frames in the core part of the sign using hand-keypoint detection confidences, and it uses SMPLify-X (adapted to SL cropping) to reconstruct

a preliminary 3D hand pose for each candidate frame. With static hand poses, θ^h_{ref} is obtained by taking the average hand poses of these candidates. With transitioning hand poses, the core part of a sign is divided into two intervals, and $\theta^h_{ref,i}$ and $\theta^h_{ref,f}$ correspond to the average hand poses of the candidate frames in the first and second intervals, respectively (see Sup. Mat. for more details).

The constraints applied to each sign depend on its sign group; we have six sign groups because classes 1a & 2a share the same constraints, as do 2b & 3b (Tab. 1). There is insufficient paired data to train a CNN classifier, so we use an intuitive and interpretable decision tree trained on extracted 2D and 3D pose features. Our features are invariant to the handedness of the signer and include: 1) the minimum of the maximum height differences of each wrist across the sequence: $\min(\{w_r\}_{\max} - \{w_r\}_{\min}, \{w_l\}_{\max} - \{w_l\}_{\min})$, where w_l and w_r are the heights of the left and right wrists, respectively. 2) the cosine distance between the initial poses of each hand: $\text{CosDist}(\theta^r_{ref,i}, \theta^l_{ref,i})$, 3) the maximum of the two cosine distances between each initial and final hand pose: $\max(\text{CosDist}(\theta^r_{ref,i}, \theta^r_{ref,f}), \text{CosDist}(\theta^l_{ref,i}, \theta^l_{ref,f}))$.

We train our sign-group classifier on the over 3,000 videos from the Corpus-based Dictionary of Polish Sign Language (CDPSL) [1]; these are annotated with HamNoSys [21]. We construct a grammar to convert HamNoSys annotations into our group labels (see Sup. Mat.) Row 1 of Figure 3 shows a sample frame from CDPSL. This dataset is not used in our quantitative analysis or perceptual study.

3.4. SGNify Extensions

First, we follow Huang *et al*. [23] to extend SGNify to work on multi-view videos. Second, we propose a baseline method for continuous SLC (CSLC). CSLC introduces additional challenges, such as the segmentation of sentences into signs; this is an active field of research. When a sentence is given as input, we use Renz *et al*. [43] to segment the input video and then process each segment with SGNify. The first frame of each segment is initialized from the last frame of the previous one. These extensions are not our main contribution and are described in Sup. Mat.

4. Dataset

To quantitatively evaluate SGNify as a viable method for SLC, we collected motion-capture data with ground-truth SMPL-X bodies articulating signs. Our dataset represents the first publicly available expressive full-4D capture of isolated SL signs. The experimental procedure was reviewed by the ethics council of the University of Tübingen without objections or remarks (709/2021B02).

In consultation with a Deaf DGS teacher and a DGS interpreter, we defined a German SL corpus consisting of 57 isolated signs. The selected signs cover a wide range of

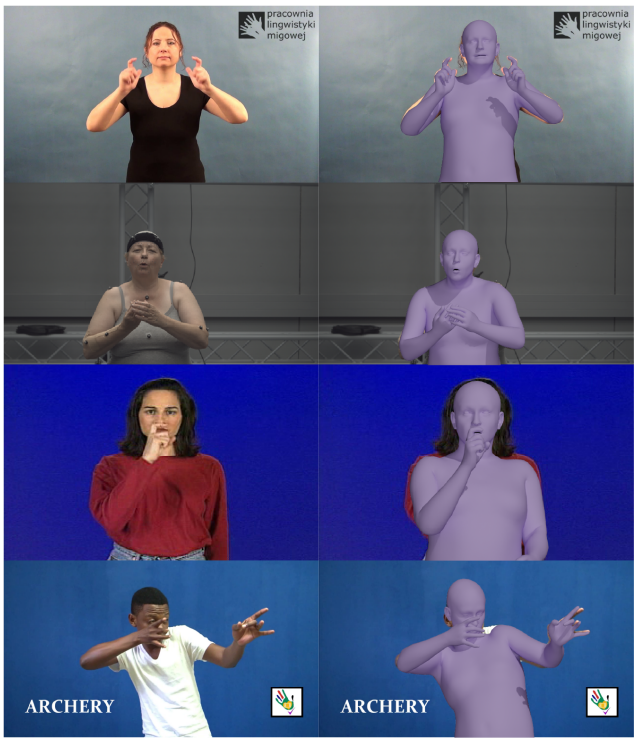

Figure 3. Samples frames reconstructed by SGNify. The input videos are from diverse multilingual datasets. Row 1: Polish SL sign GDAŃSK (GDANSK, class 1a) from the dataset [1] used for training our sign-group classifier. Row 2: DGS sign BLUME (FLOWER, class 3b) from our captured dataset used in the quantitative evaluation. Row 3: ASL sign DOLL (class 0a) from the dataset [47] used in the perceptual study. Row 4: South African SL sign ARCHERY (class 2a) from the in-the-wild dataset [42].

Class	0a	0b	1a	1b	2a	2b	3a	3b
# Signs	12	3	14	3	11	2	10	2

Table 2. Number of signs captured for each class.

challenges for SLC, such as self-contact and self-occlusion. Table 2 summarizes the number of signs collected for each of the eight classes. Signs of *subclass b* are less common, and this is reflected in our corpus.

We captured a native right-handed DGS signer with a Vicon mocap system at 120 fps, synchronized with a frontal 4112×3008 RGB camera at 60 fps, framing an upper-body view as typically found in SL video. The hands start and end at rest at the signer's sides, and each sign lasts between 1.7 and 3.5 seconds after trimming. In total, our dataset comprises 16,608 mocap frames and 8,304 RGB frames. To obtain ground-truth SMPL-X meshes, we scanned the participant in a 4D body scanner in several poses. The SMPL-X mesh was registered to these scans and averaged to obtain a personalized body-shape mesh. MoSh++ was then used

Method	Upper Body	Left Hand	Right Hand
FrankMocap [44]	78.07	20.47	19.62
PIXIE [17]	60.11	25.02	22.42
PyMAF-X [56]	68.61	21.46	19.19
SMPLify-SL	56.07	22.23	18.83
SGNify	**55.63**	**19.22**	**17.50**

Table 3. Evaluation on our ground-truth mocap dataset: mean TR-V2V error (mm) for five methods and three body regions.

Method	Sym	Inv	Left Hand	Right Hand	Both Hands
SMPLify-SL	✗	✗	20.30	18.78	19.54
SGNify	✓	✗	18.44	18.39	18.41
SGNify	✗	✓	19.76	**17.16**	18.46
SGNify	✓	✓	**17.72**	17.29	**17.50**

Table 4. Evaluating how the linguistic constraints of symmetry and invariance affect mean TR-V2V error (mm) in symmetric signs.

Method	Inv	Left Hand	Right Hand	Both Hands
SMPLify-SL	✗	26.09	18.89	20.50
SGNify	✓	**22.22**	**17.70**	**18.60**

Table 5. Evaluating how the linguistic constraint of hand-pose invariance affects mean TR-V2V error (mm) in asymmetric signs.

to fit this mesh to the mocap markers [34]. Marker-based mocap is useful for evaluating ground truth but is not practical for SLC at scale: it is expensive and requires expertise, the reflective markers attached to the signer can influence contact-heavy motions, and processing the resulting data is time consuming. If our monocular method can approach the performance of mocap, it will be widely applicable.

5. Experiments

5.1. Quantitative Evaluation

We quantitatively evaluate SGNify, compare it with state-of-the-art methods, and quantify the improvement derived from each linguistic constraint. To emulate in-the-wild data, which might have very low resolution, low framerate, and an occluded lower body, we pre-processed our high-quality video data to a resolution of 514×300 at 30 fps, and we cropped the input images above the pelvis (see Row 2 of Fig. 3). We used the synchronized meshes captured from the observed Vicon markers [34] as ground truth for evaluation. Since all tested methods estimate SMPL-X meshes with the same topology, we compute the mean per-vertex error (TR-V2V) by considering the vertices above the pelvis. The prefix "TR" means that we translationally align the mesh reconstructed for each frame with the ground truth; *i.e.*, we center the meshes before computing these errors. Since the starting and ending transitions are not part of the sign itself, we manually annotate the expressive central portion of each sign from the raw videos and compute the quantitative results on only these central frames (in total, 2,872 RGB frames).

Table 3 shows the mean TR-V2V error across the 57 signs for four methods and three body regions. The columns labeled "Upper Body" report the error computed when considering the hands and the upper body (vertices above the pelvis). We include the head but not the face because our mocap system struggles to reconstruct face details using only 27 markers. We separately report the TR-V2V errors for the left and right hands because of the central role they play in SL. We provide a visualization of the vertices selected for each of these evaluations in Sup. Mat. This experiment compares SGNify with FrankMocap [44], PIXIE [17], PyMAF-X [56], and our baseline SMPLify-SL. SGNify achieves the lowest error for the upper body and both hands, beating the state-of-the-art methods.

Tables 4 and 5 show the improvements derived from each linguistic constraint. Table 4 reports the mean TR-V2V in symmetric signs (classes 1a, 1b, and 2a) when using no constraints, *i.e.*, SMPLify-SL (Row 1), only one constraint (Rows 2 and 3), or both linguistic constraints (Row 4). When one linguistic constraint is used, the TR-V2V of both hands decreases. We observe that the symmetry constraint has the overall greater effect of the two. When the non-dominant hand is passive, it is often rotated at an angle difficult to capture from a frontal camera; this behavior might explain the greater effect of the symmetry constraint on the left hand. When both constraints are applied, we observe a more substantial decrease in the TR-V2V error of the left hand and a slight increase in the error of the right hand compared to when only the hand-pose invariance is applied to the right hand. We believe this happens because the symmetry constraint enforces symmetry between the two hands without knowing which hand reconstruction is more correct, and, for the same reason as above, detecting an accurate RPS for the non-dominant hand is often more challenging; overall, however, using both constraints greatly benefits the reconstruction. In Tab. 5, we separately evaluate performance on signs that do not present symmetry between the two hands (classes 0a, 0b, 2b, 3a, and 3b). As expected, the invariance prior included in SGNify improves the performance on all metrics. These quantitative results indicate that our linguistic constraints improve the reconstructions.

5.2. Perceptual Study

We conduct an online perceptual study to 1) compare SGNify with the best-performing state-of-the-art method for SL and 2) evaluate the improvement derived from the linguistic constraints (SGNify vs. SMPLify-SL). Even though FrankMocap has a lower hand error than PyMAF-X and PIXIE, it has significant errors in the lower body when

the full body is not seen, and higher errors in the upper body; these errors greatly affect the perceptual experience, making it unsuitable for the SL task, as already noticed in [31]. We thus use PyMAF-X since its hand-pose estimates are more accurate than PIXIE (see Tab. 3).

Approval of all experimental procedures for this study was granted by the Ethics Council of the Max Planck Society under the Haptic Intelligence Department's framework agreement under protocol number F027A. No participants are employed by our institution, and all are compensated 20 USD for their time.

Our study involves 20 adult participants who all stated that they have an advanced level of proficiency (expert level) in ASL. We discarded responses from other potential participants who did not correctly recognize at least 70% of the signs presented via real-person video. 15 of the final 20 participants (75%) are Deaf, and one participant is left-handed. Participant ages are 46.75 ± 11.78.

We used SGNify, SMPLify-SL, and PyMAF-X to reconstruct avatars from 50 videos taken from The American Sign Language Handshape Dictionary [47]; see Row 3 of Fig. 3 for an example. After responding to demographic questions, each participant evaluates the same six training videos to calibrate their responses to the quality of the presented reconstructions. We divide the remaining 44 signs into four batches (real-person RGB video, SGNify, SMPLify-SL, and PyMAF-X) that are balanced by sign class and sign frequency. We include real-person video to obtain an upper-bound on recognition performance and, as mentioned, to filter participants. We assign each participant to one of four surveys. Each survey contains all 44 test signs, and the four methods are rotated through the four sign batches across surveys. We further shuffle the questions in each survey for each user.

The participant enters the sign they believe the avatar or the real person is articulating in each video. They rate their ease in recognizing the presented sign using a visual analog scale (VAS) ranging from 0 to 100 with five standard labels ranging from "very difficult" to "very easy." They also evaluate the naturalness of the articulation on a VAS-labeled scale from "very unnatural" to "very natural." Participants are able to replay each video. We provide additional space for comments for each sign and at the end of the study. The self-reported participant carefulness is 84.45 ± 11.44 on a scale from 0 to 100.

The sign annotations submitted by participants are graded as either "incorrect" (no credit), "partially correct" (half credit), or "completely correct" (full credit). We calculate the rates at which each participant recognized the signs for each method. We visualize the resulting 20×4 matrix of recognition rates with box plots in the left plot of Fig. 4. Participants recognize signs in real-person videos with an average accuracy of 90.9% and signs reconstructed by SGNify with 86.2% accuracy. Signs reconstructed with SMPLify-SL and PyMAF-X are recognized less accurately, at 74.8% and 62.0%, respectively. We evaluate the statistical significance of these recognition rates. Some distributions failed a Shapiro-Wilk normality test, so we used the non-parametric Friedman test which shows that the method used (real video, SGNify, SMPLify-SL, or PyMAF-X) has a statistically significant effect ($p < 0.001$) on recognition rate. Pairwise comparison with Wilcoxon signed-rank tests and a Bonferroni post-hoc correction reveal that the average sign recognition rate with real video and SGNify are both significantly higher than SMPLify-SL and PyMAF-X. Importantly, sign recognition rates with real video and SGNify are not significantly different from one another.

The central plot of Fig. 4 shows the participants' perceived easiness in recognizing the sign. These four distributions passed the normality test and were analyzed with a one-way repeated-measures ANOVA with a Bonferroni post-hoc correction for pairwise comparisons. As expected, real videos are perceived to be significantly easier to recognize than the three reconstruction methods. However, signs reconstructed with SGNify and SMPLify-SL are significantly easier to recognize than those by PyMAF-X.

The right plot of Fig. 4 shows the participants' perceived naturalness of the articulation of the signs. All four distributions passed the normality test, so they were analyzed in the same way as perceived easiness. All methods are statistically different one from another, with real video receiving the highest naturalness ratings, followed by SGNify, SMPLify-SL, and PyMAF-X. Nearly all participants reported in the comments that they want avatars with more expressive face motions and smoother hand and body motions.

These results show that SGNify outperforms the state-of-the-art, PyMAF-X, in all qualitative metrics. Incorporating SL linguistic priors into our baseline SMPLify-SL yields statistically significant perceptual improvements, as judged by 20 expert signers. Most importantly, SGNify achieves a high sign-recognition rate that is not different from that of real-person video on the selected set of ASL signs.

A few participants suggested having clothed (rather than "nude") avatars and introducing more human-like features. We thus conduct a second perceptual study to see whether these recommendations have a positive impact on the intelligibility of SGNify's reconstructions. Thirteen participants from the first study participated in this follow-up; the study design is the same. The tested methods are the real video and three different SGNify avatars: 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] (see Sup. Mat.). This study comprises 24 signs (including the four for participant training) from the same ASL dataset [47] but different from those of the first study. Our results reveal that adding the

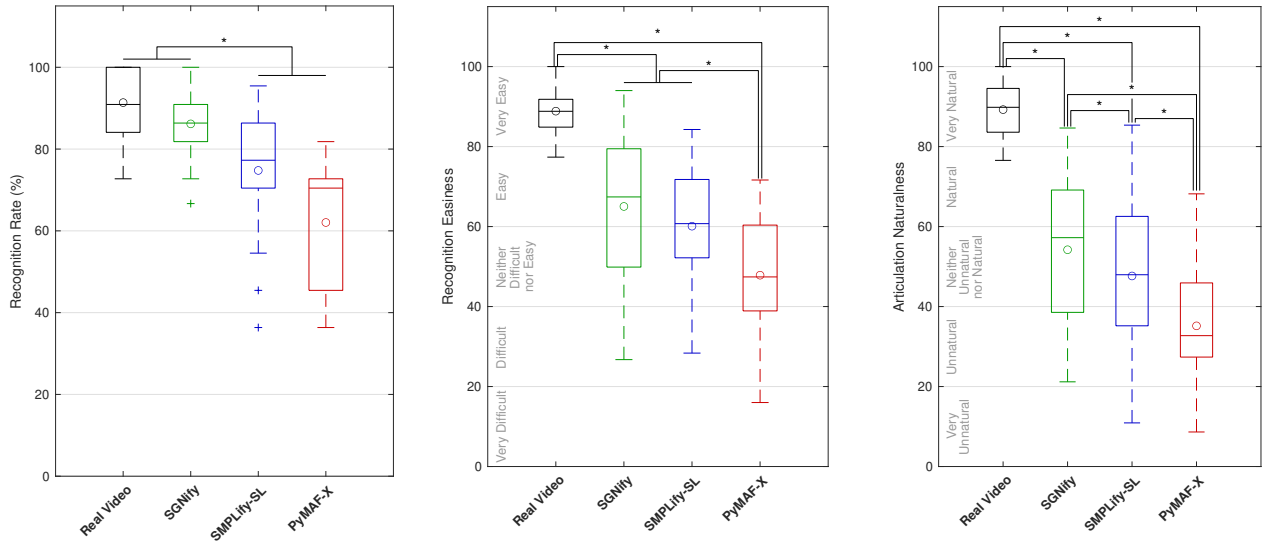

Figure 4. Box plots of the data from the perceptual study. The circle shows the mean, and the central line shows the median. The top and bottom box edges show the interquartile range (IQR), and the whiskers encompass the range up to 1.5 times the IQR. Outliers are marked with +. Statistically significant pairwise differences are indicated with a line and a ⋆. Left: Distribution of the rate at which participants recognized signs presented with each of the four methods. *Real Video* and *SGNify* achieve significantly higher recognition rates than both *SMPLify-SL* and *PyMAF-X*. Center: Distribution of the average easiness ratings participants assigned to recognizing signs presented by the four methods. All pairwise combinations except *SGNify–SMPLify-SL* are significantly different. Right: Distribution of the average articulation naturalness ratings participants assigned to the four methods. All pairwise combinations are significantly different.

t-shirt and using a fully textured avatar do not benefit actual or perceived sign recognition or the perceived naturalness of the reconstruction. As in the previous study, recognition of SGNify's reconstructions were not statistically different from real video.

6. Discussion

Our results show that SGNify performs quantitatively better than the state of the art, in particular due to the inclusion of our novel linguistic constraints. However, we believe that a per-frame metric is not ideal for SL. To recognize a sign, the temporal evolution is crucial, and this is not captured by V2V. For example, the few slightly inaccurate frames of the SMPLify-SL reconstruction of DOLL (see video on our project page) confused many signers during the perceptual study. Small changes over time can disrupt perception, while overall V2V error remains small. In the end, what matters is whether the meaning is clear to a human. We think a perceptual study provides key insights that complement metric evaluation. The perceptual study indicated that SGNify significantly outperforms the state of the art and, most importantly, produces the first 3D avatars to achieve a sign-recognition accuracy that is not statistically different from the source videos. Our perceptual study also highlights the next challenge for SLC: the need for improvements in the face including facial expressions, tongue and eye movements, mouth morphemes, and eyebrows.

7. Conclusions

We present SGNify, which estimates 3D avatars of isolated SL signs from monocular RGB video. Quantitative and qualitative experiments show that SGNify outperforms the state of the art in estimating challenging SL hand poses by leveraging constraints derived from linguistics. SGNify represents a step towards the capture of realistic 3D avatars from SL videos in the wild. Future work should explore the use of our constraints in training regression methods, real-time processing, and continuous signing.

Acknowledgments We thank Galina Henz and Tsvetelina Alexiadis for trial coordination; Matvey Safroshkin, Markus Höschle, Senya Polikovsky, Tobias Bauch, Taylor McConnell (TM), and Bernard Javot for the capture setup; TM for data-cleaning coordination; Leyre Sánchez Vinuela, Andres Camilo Mendoza Patino, and Yasemin Fincan for data cleaning; Nima Ghorbani and Giorgio Becherini for MoSh++; Joachim Tesch for help with Blender; Benjamin Pellkofer and Joey Burns for IT support; Yao Feng, Anastasios Yiannakidis, and Radek Daněček for discussions on facial methods; Haliza Mat Husin and Mayumi Mohan for help with statistics; and the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Maria-Paola Forte and Peter Kulits.

Disclosure: https://files.is.tue.mpg.de/black/CoI_CVPR_2023.txt

References

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

[2] Meshcapade GmbH, Tübingen, Germany. https://meshcapade.com, 2022. 7

[3] Ahmed H. Aliwy and A. Alethary Ahmed. Development of Arabic sign language dictionary using 3D avatar technologies. *Indonesian Journal of Electrical Engineering and Computer Science*, 21(1):609–616, 2021. 1

[4] Robbin Battison. *Lexical Borrowing in American Sign Language*. Education Resources Information Center (ERIC) , 1978. 2, 3, 4

[5] Sara Bilal, Rini Akmeliawati, Momoh Jimoh El Salami, and Amir A. Shafie. Vision-based hand posture detection and recognition for sign language – a study. In *International Conference on Mechatronics (ICOM)*, pages 1–6, 2011. 2

[6] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In *International Conference on Computer Vision (ICCV)*, pages 1021–1030, 2017. 2

[7] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(1):172–186, 2021. 3

[8] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Beyond static features for temporally consistent 3D human pose and shape from a video. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1964–1973, 2021. 2

[9] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision (ECCV)*, volume 12355, pages 20–40, 2020. 3

[10] Helen Cooper, Brian Holt, and Richard Bowden. Sign language recognition. In *Visual analysis of humans*, pages 539–562. Springer, 2011. 3

[11] Radek Daněček, Michael J. Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 20311–20322, 2022. 2, 3

[12] Qi Dang, Jianqin Yin, Bin Wang, and Wenqing Zheng. Deep learning based 2D human pose estimation: A survey. *Tsinghua Science and Technology*, 24(6):663–676, 2019. 2

[13] Adrian C. Davis and Howard J. Hoffman. Hearing loss: Rising prevalence and impact. *Bulletin of the World Health Organization*, 97(10):646, 2019. 1

[14] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2Sign: A large-scale multimodal dataset for continuous american sign language. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2735–2744, 2021. 3

[15] Petra Eccarius and Diane Brentari. Symmetry and dominance: A cross-linguistic study of signs and classifier constructions. *Lingua*, 117(7):1169–1201, 2007. 3

[16] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3D morphable face models - past, present and future. *ACM Transactions on Graphics*, 39(5), 2020. 2

[17] 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. 3, 6

[18] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 40(4):1–13, 2021. 2

[19] Panagiotis P. Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Visual speech-aware perceptual 3D facial expression reconstruction from videos. *arXiv preprint arXiv:2207.11094*, 2022. 2, 3

[20] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z. Li. Towards fast, accurate and stable 3D dense face alignment. In *European Conference on Computer Vision (ECCV)*, volume 12364, pages 152–168, 2020. 2

[21] 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. 1, 5

[22] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11807–11816, 2019. 2

[23] 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. 5

[24] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *International Conference on Computer Vision (ICCV)*, pages 3334–3342, 2015. 3

[25] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. In *International Conference on 3D Vision (3DV)*, pages 42–52, 2021. 2

[26] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8320–8329, 2018. 2

[27] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131, 2018. 2

[28] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *International Conference on Computer Vision (ICCV)*, pages 11127–11137, 2021. 2

[29] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision (ICCV)*, pages 2252–2261, 2019. 2

[30] Agelos Kratimenos, Georgios Pavlakos, and Petros Maragos. Independent sign language recognition with 3D body, hands, and face reconstruction. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4270–4274, 2021. 3

[31] Shyam Krishna, Vignesh P. Vijay, and Babu J. Dinesh. SignPose: Sign language animation through 3D pose lifting. In *International Conference on Computer Vision (ICCV)*, pages 2640–2649, 2021. 3, 7

[32] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M. Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[33] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3383–3393, 2021. 2

[34] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, pages 5441–5450, 2019. 2, 6

[35] Ian Marshall and Éva Sáfár. A prototype text to British Sign Language (BSL) translation system. In *Meeting of the Association for Computational Linguistics (ACL)*, pages 113–116, 2003. 3

[36] Gyeongsik Moon and Kyoung Mu Lee. I2L-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. In *European Conference on Computer Vision (ECCV)*, pages 752–768, 2020. 2

[37] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In *European Conference on Computer Vision (ECCV)*, volume 12365, pages 548–564, 2020. 2

[38] Amit Moryossef, Ioannis Tsochantaridis, Joe Dinn, Necati Cihan Camgoz, Richard Bowden, Tao Jiang, Annette Rios, Mathias Muller, and Sarah Ebling. Evaluating the immediate applicability of pose estimation for sign language recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3434–3440, 2021. 3

[39] 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. 2, 3

[40] Lucie Naert, Caroline Larboulette, and Sylvie Gibet. A survey on the animation of signing avatars: From sign representation to utterance synthesis. *Computers & Graphics (CG)*, 92:76–98, 2020. 1

[41] 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. 2, 3

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

[43] Katrin Renz, Nicolaj C. Stache, Neil Fox, Gül Varol, and Samuel Albanie. Sign segmentation with changepoint-modulated pseudo-labelling. In *Computer Vision and Pattern Recognition Workshops (CVPRw)*, 2021. 5

[44] 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. 3, 6

[45] Ken Shoemake. Animating rotation with quaternion curves. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 245–254, 1985. 4

[46] Breannan Smith, Chenglei Wu, He Wen, Patrick Peluse, Yaser Sheikh, Jessica K. Hodgins, and Takaaki Shiratori. Constraining dense hand surface tracking with elasticity. *ACM Transactions on Graphics*, 39(6):1–14, 2020. 2

[47] Richard A. Tennant, Marianne Gluszak, and Marianne Gluszak Brown. *The American Sign Language Handshape Dictionary*. Gallaudet University Press, 2010. 5, 7

[48] Stavros Theodorakis, Vassilis Pitsikalis, and Petros Maragos. Dynamic–static unsupervised sequentiality, statistical subunits and lexicon for sign language recognition. *Image and Vision Computing (IVS)*, 32(8):533–549, 2014. 3

[49] Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Taheri Omid, Michael J. Black, and Dimitrios Tzionas. 3D human pose estimation via intuitive physics. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[50] Manuel Vázquez-Enríquez, Jose L. Alba-Castro, Laura Docío-Fernández, and Eduardo Rodríguez-Banga. Isolated sign language recognition with multi-scale spatial-temporal graph convolutional networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, 2021. 2, 3

[51] Jinbao Wang, Shujie Tan, Xiantong Zhen, Shuo Xu, Feng Zheng, Zhenyu He, and Ling Shao. Deep 3D human pose estimation: A review. *Computer Vision and Image Understanding*, 210:103225, 2021. 2

[52] World Federation of the Deaf. Who we are. http://wfdeaf.org/who-we-are/. 1

[53] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10957–10966, 2019. 3

[54] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6183–6192, 2020. 2

[55] Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Neural descent for visual 3D human pose and shape. In *Computer Vision and Pattern Recognition (CVPR)*, pages 14484–14493, 2021. 3

[56] 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. 6

[57] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In *International Conference on Computer Vision (ICCV)*, pages 11426–11436, 2021. 2

[58] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In *International Conference on Computer Vision (ICCV)*, pages 813–822, 2019. 2