2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

SynthSL: Expressive Humans for Sign Language Image Synthesis

Jilliam M. Díaz Barros[1,2*] Chen-Yu Wang[2*] Jameel Malik[1,3] Abdalla Arafa[1] Didier Stricker[1,2]
[1] German Research Center for Artificial Intelligence (DFKI) [2] RPTU Kaiserslautern
[3] NUST School of Electrical Engineering and Computer Science

***Abstract*— Around 5% of the world's population live with disabling hearing loss. Despite recent advancements to improve accessibility to the Deaf community, research on sign language is still limited. In this work, we introduce a large-scale synthetic dataset on sign language, SynthSL, targeted to sign language production, recognition and translation. Using state-of-the-start methods for human body modelling, SynthSL aims to augment current datasets by providing additional ground truth data such as depth and normal maps, rendered models, segmentation masks and 2D/3D body joints. We additionally explore a generative architecture for the synthesis of sign images and propose a new generator based on Swin Transformers, conditioned on given body poses and appearance. We believe that an increase on the publicly available data on sign language would boost research and close the performance gap with related topics on human body synthesis. Our code, models and dataset are available at https://github.com/jilliam/SynthSL.**

I. Introduction

Sign language (SL) is one of the main communication channels for people with severe or profound deafness. It is considered a natural language, where signs result from the combination of body postures, hand gestures and facial expressions, giving a representation in the spatio-temporal domain. It differs from spoken language both in grammar and word order, without a one-to-one mapping from sign to word. Instead, spoken sentences can be translated to glosses, which represent the written form of the signs [9].

Since SL perception is visual, it is actively investigated in computer graphics and vision, particularly for three main tasks: production, recognition and translation. SL production aims at generating sign videos or human pose sequences from written or spoken sentences, while recognition and translation refer to the interpretation of sign videos. Sign recognition targets the identification of glosses, whereas in translation, signs are mapped to written or spoken language words, in some cases using glosses in an intermediate step.

Communication with the Deaf community is a bidirectional process, and both recognition and production are essential to increase accessibility. Production, in particular, is a crucial step for them to understand live speeches, news, cultural activities, among others, specially when no interpreters are available. Signs are preferred over closed captions or subtitles, as SL users are accustomed to their own natural language, while written text is not always easily comprehensible for them [17], [80].

Although SL is used worldwide, it is not universal, since signs can vary widely in different regions, and several types

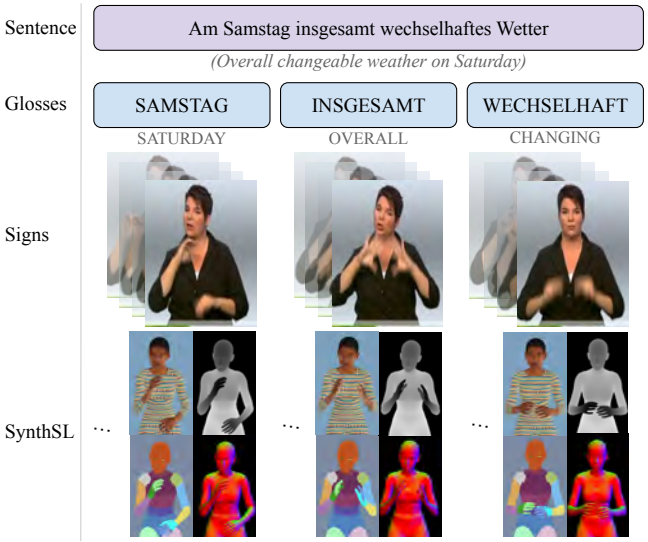

Fig. 1. In sign language, the representation of sentences is encapsulated by glosses, which map sentences to morphemes. Existing datasets usually provide RGB or RGB+D sign sequences and the respective translation to sentences and glosses, which can be exploited for sign language recognition, translation and production. We introduce a rendering pipeline and our dataset SynthSL, which extends the *PHOENIX 2014T* dataset [9] to include additional ground truth such as rendered RGB images, depth and normal maps, segmentation masks and SMPL-X [59] model parameters.

and dialects might co-exist in the same country [17], [45], [46]. Each variation has a defined structure and its own grammar. This poses a challenge in sign recognition and translation, since datasets cannot be used cross-linguistically.

In the era of deep learning, large-scale datasets are of vital importance for training and testing. However, collecting such datasets is usually expensive and time consuming. Additionally, SL datasets require expert signers and annotators to label the data, in order to avoid interpretation errors due to similarity in the signs [46]. To address this limitation, some datasets have been collected from multiple sources, such as learning channels for SL [46] or from news or weather broadcasts [26], usually providing only RGB videos. Acquisition of additional ground truth data such as 2D/3D skeletons, keypoints and body pose usually requires specialized systems such as 3D motion capture (MoCap) data or custom setups with hundreds of cameras and sensors [21]. Unfortunately, these systems are not easily accessible.

In this paper, we introduce a large-scale synthetic dataset on SL with expressive humans, SynthSL, based on the SMPL-X [59] model. Inspired by realistic synthetic datasets in human analysis such as SURREAL [77], we propose a pipeline for augmenting real RGB-based sign language

* Authors contributed equally.
This work was partially funded by the German Ministry of Education and Research (BMBF) under Grant Agreement 01IW20002 (SocialWear).

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

datasets to include other modalities and add rich ground truth data such as depth information, rendered models, pose, body-part segmentation, normal maps and 2D/3D body joints. With the proposed data generation framework, we aim to extend existing datasets and advance research on SL production and interpretation. Furthermore, our rendering pipeline can contribute to close the gap between animated avatars used for SL production and synthetic sign videos. SynthSL and the synthetic sign generation code are publicly available for research purposes.

The use of synthetic data offers multiple advantages on SL research: *(i)* overcome the limitations of monocular RGB-based videos with extended ground truth data, especially for learning models; *(ii)* enable inter-signer variation. Body shapes and textures (with identity and clothing) are randomized, to increase the diversity in our dataset; *(iii)* ensure enough number of instances per word/sentence. Each sentence can be generated with different signers, adding random noise to the signs to increase variability; *(iv)* eliminate motion blurry, intrinsic to fast motions in video sequences; *(v)* control over the camera position and the acquisition of multiple views; and *(vi)* control of external factors such as illumination and background.

Finally, we propose a new pipeline for synthesizing sign language images, targeted to sign language production (SLP). This pipeline is based on Generative Adversarial Networks (GANs), where a base signer (or *style*) image, the current and a target pose are used as input. The model then generates an output image of the base signer in the target pose. We leverage the improved performance of transformers on vision tasks and introduce SwinGenerator, a new generator based on Swin Transformer V2 [50].

The main contributions of this work are:

- A new rendering pipeline based on the SMPL-X model, which integrates the shape and pose of the body, hands and face, in addition to facial expressions.
- SynthSL, a large-scale synthetic dataset for research on sign language. Our dataset expands the *PHOENIX 2014T* dataset [9] with ground truth on rendered models, segmentation masks, depth and normal maps, 2D/3D body joints, body pose and shape, which are provided on a frame basis.
- A new deep-learning based model for sign image synthesis, with a novel SwinGenerator. Our model takes as input a source and target pose, plus a style image, and generates an image of the signer in the given style and target pose.

II. Related work

In this section, we review the state of the art in SLP and analyze current SL large-scale datasets. We additionally describe related work on human analysis with synthetic datasets. For a comprehensive study on methods for SL recognition, we refer the reader to [1].

Sign Language Production. Traditionally, SLP has been mainly focused on the animation of 3D avatars, using MoCap data [17], [29], [30] or pre-designed synthetic sequences [31], [40], [56]. Recent works have moved towards the generation of skeleton poses or realistic video sequences, using deep neural networks (DNN). Some approaches rely in an intermediate step to translate spoken sentences to skeletal poses, using glosses. [71] produces sign videos from spoken language, by combining Neural Machine Translation (NMT) and a generative model. The sentences are translated to glosses using an encoder-decoder, which in turn are mapped to skeletal pose sequences using a dictionary. [72] overrides the look-up table and generates skeletal pose sequences via Motion Graph, which is conditioned by a sequence of gloss probabilities extracted from the written or spoken sentences. [73] maps glosses to signs using a tailored DNN.

Other works explore the direct transformation from text to skeletal poses, without an intermediate gloss representation [25]. [83] used an architecture based on neural networks to generate fixed-length sign pose sequences. [68] introduced progressive transformers for SLP, from discrete spoken sentences to continuous sign pose sequences. [67] extended the framework to integrate a Mixture Density Network (MDN) to the transformer. [66] proposed an adversarial approach to exploit the multi-channel features of SL, to include facial expressions and mouthing in the sign pose sequences.

More relevant to our work is the pose-conditioned human synthesis for SLP. PSGN [71] proposed a hybrid architecture composed of a Variational Autoencoder (VAE) [42] generator and DCGAN [60], to synthesize sign sequences conditioned on the skeletal information and appearance. [78] produces realistic sign videos using the GAN architecture for human motion transfer in *Everybody Dance Now* [13], while the synthesis pipelines in [72], [73] are based on pix2pixHD [79], with an encoder-decoder-based generator. Similarly, SignGAN [67], [69] generates realistic sign videos using a conditional GAN based on [13] and [79], given a style image and a sign pose sequence. It additionally introduces a keypoint-based loss to improve the synthesis of hands. SignDIFF in [25] is based on dual-condition diffusion models for SLP on American Sign Language (ASL).

Sign Language Datasets. Datasets can be classified according to the language, if they were captured in a controlled or unconstrained (in the wild) scenario and if they are targeted to finger-spelling [20], word (isolated) or sentence level (continuous) sign language. Table I lists a set of large-scale datasets for isolated and continuous sign language, from different regions around the world.

Some datasets were collected from weather forecasts [10], [26], [27], TV broadcast [3], [10], videos on YouTube [75], or different online channels of the Deaf community [46]. Others datasets comprise videos recorded in controlled environments, with multiview RGB [43], or multimodal systems [22], [24], [57], and provide a large variety of ground truth data, e.g. depth maps, 2D/3D keypoints, facial landmarks, body and hand pose, usually extracted from Kinect v1/v2 and OpenPose [11]. *How2Sign* [21] includes data captured in a custom setup with hundreds of cameras and sensors placed in a geodesic dome. Other datasets in-

TABLE I
LARGE-SCALE WORD AND SENTENCE-LEVEL SIGN LANGUAGE DATASETS.
**Not publicly available. **This information was inferred from [72], where all signers perform each sign 3 times.*

Name	Language	Videos	Sentences	Signers	Vocab. size	Data type	Level	In the wild
DEVISIGN [12]	Chinese	24000	-	8	2000	RGB-D	Word	-
SMILE* [24]	Swiss German	12600**	-	42	100	RGB-D	Word	-
MS-ASL [75]	American English	25513	-	222	1000	RGB	Word	✓
WLASL [46]	American English	21083	-	119	2000	RGB	Word	✓
BosphorusSign22k [57]	Turkish	22542	-	6	744	RGB-D	Word	-
PHOENIX 2012 [26]	German	190	1980	7	911	RGB	Sentence	-
PHOENIX 2014 [27]	German	645	6861	9	1558	RGB	Sentence	-
PHOENIX 2014T [9]	German	8257	8257	9	1066	RGB	Sentence	-
CSL [37]	Chinese	25000	100	50	178	RGB-D	Sentence	-
KETI* [43]	Korean	14672	105	14	524	RGB	Sentence	-
How2Sign [22]	American English	2500	38611	10	4000	RGB-D	Sentence	-
GSL [1]	Greek	10290	10290	7	310	RGB-D	Sentence	-
C4A - SWISSTXT-WEATHER [10]	Swiss German	183	811	-	1248	RGB	Sentence	-
C4A - SWISSTXT-NEWS [10]	Swiss German	181	6031	-	10561	RGB	Sentence	-
C4A - VRT-NEWS [10]	Flemish	120	7174	-	6875	RGB	Sentence	-
BOBSL [3]	British English	1962	1.2M	39	2281	RGB	Sentence	-

troduce complementary annotations, e.g. [44] extends *RWTH PHOENIX Weather 2014* [27] with hand shapes annotation, while [2] provides annotations to analyze facial expressions and emotions on the same dataset. [9] reformatted the dataset with new segmentation boundaries, to ease SL recognition and translation.

Synthetic Datasets for Human Analysis. Synthetic datasets have been widely used in recent years to compensate the scarcity of large-scale data for training deep learning models [74], [14]. Many human-based datasets synthesize images either from virtual characters generated with computer graphic software [6], [28], [49], or by using parametric body models such as SCAPE [4] or SMPL [52]. The latter has gained special attention for synthetic human datasets [77], [76], [34], [62], [15], due to its realistic appearance and its compatibility to existing rendering engines. SMPL is a statistical body model learned from thousand of 3D body scans and is based on blend skinning.

[77] introduced SURREAL (Synthetic hUmans foR REAL tasks), the fist synthetic dataset with realistic rendered humans performing different actions. Pose and motion were capture using MoCap data and additionally to the 2D/3D body pose, the dataset provides depth map, segmentation masks, surface normals and optical flow. They demonstrated that CNNs trained on synthetic data can be used for human body segmentation and depth estimation on real RGB images. [76] presented SURREACT (Synthetic hUmans foR REal ACTions), a dataset intended to investigate the scope of temporal CNNs trained on synthetic data for human action recognition. They exploited the synthetic data to include augmentation, such as multiple viewpoints and variations in appearance and motion. [34] explored synthetic data for hand-object manipulation. They introduced a synthetic dataset, ObMan, which rendered the hands and body using the SMPL+H model [64], to produce synthetic RGB images, depth maps and segmentation masks. SMPL+H integrates a parametric hand model, MANO [64] to the SMPL model. The images were rendered using Blender [16] and its python API. [61] and [62] explore synthetic datasets with optical flow fields for single and multi-human motion analysis (MHOF). The output was rendered using a similar approach to [34]. For MHOF, additional ground truth such as depth maps, segmentation masks, normal maps, and SMPL+H shape and pose parameters are provided. This ground truth was created using the SMPL+H model and MoCap data.

[36] investigated multi-person pose estimation networks trained on purely synthetic data and on augmented data based on real images. Based on [23], they additionally presented a style-transfer method to map synthetic to real humans. BEDLAM [7] introduced a dataset of synthetic humans based on the SMPL-X model, with high level of realism. The bodies are animated using datasets collected with MoCap. Their experiments show that simple models trained on this synthetic data can improve more elaborated SOTA approaches in human pose and shape (HPS) estimation.

Previous work demonstrates that human-based synthetic datasets help improve DNN performance in vision-related tasks. With our dataset and rendering pipeline, we aim to help advance the research on SL, by providing additional types of ground-truth information to existing datasets.

III. SYNTHSL AND RENDERING PIPELINE

We introduce a pipeline to render synthetic sign images, used to generate our large-scale dataset, SynthSL. Our rendering framework, depicted in Fig. 2, leverages the parametric model SMPL-X [59] to extend RGB-based sign videos with additional ground-truth annotations. To that end, we exploit a state-of-the-art model fitting pipeline that regresses the SMPL-X parameters from RGB images. These parameters are then used to animate the human model and generate synthetic data. Our pipeline can be applied to videos from different sources, such as existing SL datasets, broadcasts and YouTube videos. Nonetheless, its use is not limited to SL-related tasks and can be employed in the

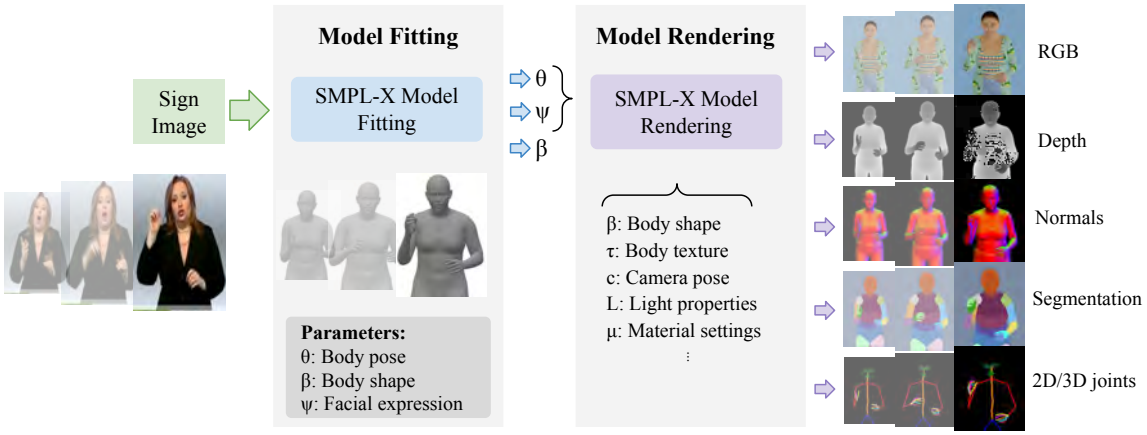

Fig. 2. Rendering pipeline for generating synthetic data. Given an RGB sequence, an SMPL-X model fitting method estimates the body pose, shape and facial expression parameters per frame. The body pose and facial expression are passed to the animation framework, where multiple types of ground-truth data are rendered. Additional parameters such as body shape and texture are randomized per sequence. Input frames taken from *PHOENIX 2014T* [9].

generation of other human-related datasets. We note that our pipeline is compatible with other rendering pipelines and animation software.

SMPL-X Model. The SMPL-X model, taking its name from *SMPL eXpressive*, improves the SMPL model [52] by integrating detailed face and hand models to the given body template. The model was trained on multiple datasets of 3D human scans, incorporating the additional parametric models to increase the resolution of hands and face. Given a pose θ, shape β and facial expression ψ, the SMPL-X model is formulated as $M(\theta, \beta, \psi) : \mathbb{R}^{|\theta| \times |\beta| \times |\psi|} \rightarrow \mathbb{R}^{3N}$, with $N = 10475$ being the number of vertices.

The pose parameters θ embody the pose for K joints in the whole body, including jaw and finger joints. In total, SMPL-X has $K = 55$ joints, with one joint corresponding to the global orientation and the 54 left to body joints. 30 of these joints belong to the fingers, which are modelled based on MANO [64]. Such model uses a PCA pose space to reduce the dimension of the axis-angle rotations from 90 (3 angles per joint) to 24 parameters. Similarly, the facial expression is represented using a lower dimension PCA space with 10 parameters based on FLAME [47].

SMPL-X has in total 119 parameters: 99 for the pose (75 for { body, eyes and jaw } and 24 for the hands), 10 parameters for the shape and 10 for the facial expression. In addition, it has the option to use a female, male or neutral body model.

Augmented dataset. SynthSL was generated from the training set of the publicly available dataset *PHOENIX 2014T* [9]. This dataset is targeted to continuous recognition and translation on German SL, and comprises sign videos describing the weather forecast. Even though this dataset has a large vocabulary size (1000+), the number of signers is relatively low (9) and is only RGB-based.

Shape, pose and expressions. Our dataset is based on the SMPL-X model, which comprises body, face and hands. It differs from related human body rendering pipelines such as SURREAL [77] and Obman [34], which use the SMPL (body) and SMPL+H (body and hands), respectively. The model parameters are estimated using SMPLer-X [8], a state-of-the-art fitting approach based on vision transformers (ViT) [19]. This method has currently the best performance for body modelling in the AGORA benchmark [58], which evaluates SMPL-X model fitting on a synthetic dataset with high realism. In SynthSL, we estimate the pose θ and facial expression ψ from SMPLer-X with ViT-Small. These parameters are then passed to the rendering step. The body shape, on the other hand, is randomly generated for every sequence to increase the diversity of the signers.

Texture and background. The body textures are randomly sampled from the BEDLAM dataset [7]. It includes 50 skin textures for male and 50 for female models, with large diversity in terms of identity and skin tones. In addition, the dataset provides 1738 clothing textures which are overlaid on the given skins.

Our rendering pipeline offers the option to randomly select backgrounds from ImageNet [18] or LSUN [82]. In our dataset, we fixed the background to only one specific color, to simulate the controlled recording scenarios in sign language videos.

Rendering. The RGB images are rendered using the Python API from Blender, similarly to [34]. However, our pipeline generates multiple types of ground truth data, such as depth, segmentation and normal maps, following SURREAL and MHOF. In contrast to these pipelines, the facial expression and jaw pose are also integrated, to render expressive models. The RGB images, segmentation masks, depth and normal maps are rendered with a default resolution of 512×512×3. SMPL-X model parameters such as body pose, shape, facial expression and 2D/3D body joints are provided in SynthSL as well. SMPL-X is available in Autodesk's Filmbox (FBX) file format, which enables the animation in software such as Blender and Maya [5].

Our rendering pipeline provides multiple benefits to SL research: (1) to augment existing RGB-based datasets; (2) to assist in the design of DNN architectures, where various

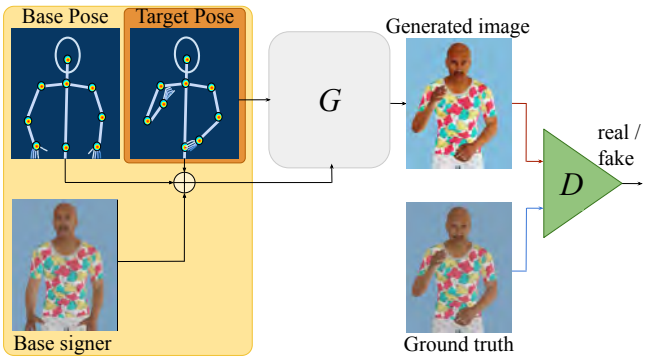

Fig. 3. Proposed framework for the synthesis of sign language images. Note that the poses (keypoints) are simplified for demonstration purposes.

types of data could be integrated into SL-based models; and (3) to guide in the collection of new datasets, where different technologies and algorithms could be evaluated beforehand.

IV. Sign Language Image Synthesis

In addition to the rendering pipeline and our synthetic dataset SynthSL, we investigate the synthesis of sign images for SLP. This task can be divided into three main stages: *(i)* transcription of spoken to written words, *(ii)* translation from written words to glosses and *(iii)* mapping from glosses to signs for production. In this work, we focus on the third stage, the generation of sign images and videos from glosses. More specifically, we investigate the generation of sign images from skeleton poses representing the glosses.

Inspired by [71], we designed a GAN architecture to generate sign images from an input base signer and a target pose. This problem is often termed as image to image translation, where the generator G maps the input data to a given visual domain, conditioned on a given input data. The discriminator D decides whether the generated image is fake or not, and trains in tandem along with the generator. Our pipeline introduces the pose of the base signer as an additional input to G, as shown in Fig. 3. We additionally propose a new generator architecture for image synthesis, based on Swin Transformers [50], [51].

A. Generator and Discriminator Training

The generator G and discriminator D are trained in intervals with an adversarial training, where G tries to maximize the number of images that deceives D, and D penalizes G for generating unrealistic images [32].

D is trained using the adversarial loss [32], given by:

$$\mathcal{L}_{adv} = \mathbb{E}_y[\log(D(y))] + \mathbb{E}_{x,z}[\log(1 - D(G(x,z)))], \quad (1)$$

where $\mathbb{E}[\cdot]$ is the expected value and $D(\cdot)$ is the probability estimated by the discriminator that the given input corresponds to real or generated data. In the first term in (1), D takes as input the ground-truth sign image, y. This term enhances the ability of D to correctly identify real images. In the second term, D assesses the image reconstructed by the generator, $G(x,z)$, and is used to improve the capability of D to distinguish $G(x,z)$ as fake or real.

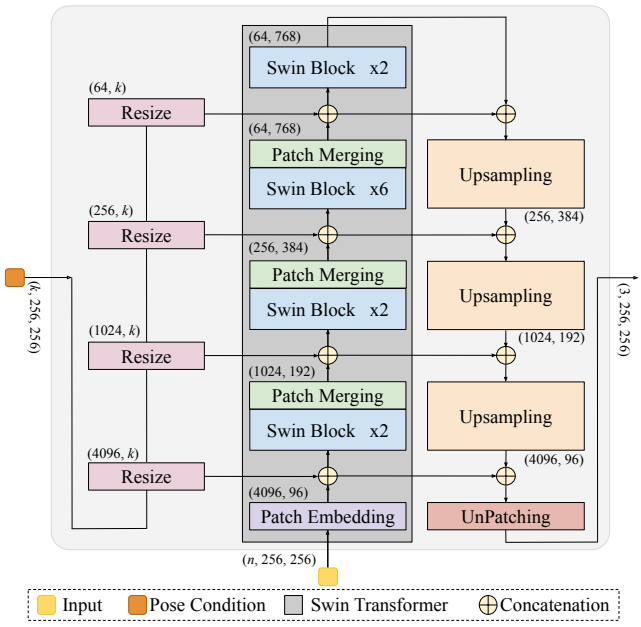

Fig. 4. Architecture of the proposed SwinGenerator.

In our architecture, the input of G corresponds to the concatenation of the base signer, base pose and target pose, x, in addition to the target pose z, while the output $G(x,z)$ is an RGB image depicting the base signer in the given target pose. We propose to train G with the loss function:

$$\mathcal{L}_G = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{perc}\mathcal{L}_{perc}, \quad (2)$$

with $\mathcal{L}_{rec}$ being the reconstruction loss given by the $L1$ pixel-wise distance between the ground truth and reconstructed image at $G(x,z)$; $\mathcal{L}_{adv}$ is the adversarial image loss in (1); $\mathcal{L}_{perc}$ is the perceptual loss in [39]; and $\lambda_{(\cdot)}$ is the weight of each term.

The final objective is then formulated as:

$$\mathcal{L} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{adv} \arg\min_G \max_D \mathcal{L}_{adv} + \lambda_{perc}\mathcal{L}_{perc}. \quad (3)$$

Generator Network. We introduce a new generator model, SwinGenerator, as depicted in Fig. 4. The backbone of our generator G is based on Swin Transformer V.2 [50], which improves previous models such as ViT, by integrating shifting windows [51]. Since Swin Transformer applies the attention mechanism to patches of the window instead of the full image, the model complexity is reduced, with improved computational performance. This enhancement allows the use of this model on higher-resolution sign images.

In our SwinGenerator, the Swin Transformer V.2 serves as the encoder, with each block containing paired W-MSA (Windows Multi-Head Attention) and SW-MSA (Shifted Windows Multi-Head Attention) for feature extraction. Patch merging is used for down-sampling, reducing the size of the output feature map by half and doubling the number of channels.

To enable the generation of images conditioned on a given pose and appearance, we propose to add upsampling and un-patching layers, which act as a decoder and map

the embedded code to the synthesized sign image. The upsampling layers are based on transposed convolutions, followed by two convolution layers, and the un-patching layer recovers the original image structure from the patches.

We additionally integrate skip connections between the corresponding down-sampling and up-sampling layers, to improve the convergence of the model. Furthermore, the target pose is resized and concatenated to the output of every block in the encoder and to the input of the corresponding blocks in the decoder.

V. Experiments and Results

A. Implementation

The proposed synthesis pipeline was developed in PyTorch and trained on a Nvidia RTX A6000 GPU. Inference time was computed on an Nvidia RTX2070.

B. Datasets

We trained and evaluated our SL synthesis pipeline on SynthSL, *PHOENIX 2014T* and *BosphorusSign22k*, using separate models for each dataset. Images in *PHOENIX 2014T* have resolution of 210×260×3, while *BosphorusSign22k* has full HD images of 1920×1080×3 pixels. Since *BosphorusSign22k* is very large, we used 4% of the training data, which we found was sufficient for generating realistic sign language images. Nonetheless, we used the complete test set for evaluation. In SynthSL, we selected 1000 sequences, with 600 for training, 100 for validation and 300 for testing. The training set alone has 79362 frames in total, with an average of ∼132 frames per sequence. The test set has 35848 frames, with ∼119 frames per sequence.

C. Data pre-processing and model architecture

During pre-processing, input images were resized to 256×256×3. For *BosphorusSign22k*, we trimmed 1/4 from each side, to have square images. The face and body keypoints were resized accordingly.

SwinGenerator uses a Swin-T V2, with number of channels C = 96 and block = $\{2, 2, 6, 2\}$. Its input corresponds to the concatenation of the base signer RGB image, base and target poses, in addition to the target pose as pose condition. The output is an RGB image of the signer in the new pose, with size $256 \times 256 \times 3$. The base and target poses have each $k = 96$ 2D keypoints extracted with MediaPipe [53], where 42 belong to the hands, 46 to the face, and 8 to the body. Although there are many recent works on pose estimation from images [81], [54], we employ MediaPipe due to its highly robust estimations from RGB data and real-time performance. We map the extracted keypoints to Gaussian heatmaps, with one channel per keypoint.

The discriminator is composed of two convolutional layers and three residual blocks. It takes as input the ground truth image and the generated image, as depicted in Fig. 3. The discriminator is updated one time for every generator update. We used the Adam optimizer [41] with learning rate α_G = 2e$-$3 for G and α_D = 2e$-$4 for D. In (2), we set λ_{rec} = 5, λ_{adv} = 2 and λ_{perc} = 0.5. All the models are trained for 30 epochs, with batch size of 32. For the perceptual loss, we used layers $\{1, 3, 5, 9, 13\}$ from VGG19 [70].

D. Baseline architectures

Since existing sign language synthesis models are not publicly available, we implemented a baseline architecture inspired by [71]. We introduced five main changes to the architecture in [71]: (i) the generator is based on U-Net [65] instead of VAE, similarly to pix2pix [38]; (ii) the keypoints are mapped to Gaussian heatmaps, instead of binary heatmaps; (iii) we used the same keypoints $k = 96$ as in our architecture, and not 10; (iv) we added convolutional blocks in the input of the encoder and output of the decoder, to generate higher quality images of size 256×256×3, instead of 128×128×3; and (v) we introduced the perceptual loss to train the generator. We refer to this model as "U-Net".

Additionally, we implemented a second baseline generator model, named as "ResNet". This models is composed of deep residual blocks [35] in the encoder, following related works on image synthesis [39], [79], [85] and SLP [69], [72], while the decoder is given by transposed convolutions.

In both models, the input is given by the concatenation of the base signer and target pose, following [69], [71], [72]. We use the same loss function in (3) and discriminator as in our model for training and evaluation, for a fair comparison.

E. Metrics

The performance of the proposed SL synthesis pipeline is evaluated using the structural similarity index metric (SSIM), peak-signal-to-noise ratio (PSNR), position errors (PE) and Fréchet Inception Distance (FID), following related work on SLP [69] and human body synthesis [84]. In the SSIM, PSNR and FID the generated RGB image is compared to the respective ground truth RGB image in the given pose. For the PE, we extracted the keypoints from the ground truth and generated images using MediaPipe, and summed the Euclidean distance between the respective keypoints.

F. Evaluation

The quantitative results of the proposed synthesis approach on *PHOENIX 2014T*, SynthSL and *BosphorusSign22k* are reported in Tables II, III and IV, respectively, under SwinGenerator. Qualitative results for the three datasets are shown in Fig. 5 and in the Supplementary material. In contrast to *PHOENIX 2014T* and *BosphorusSign22k*, SynthSL has large diversity in terms of signer and clothes. Furthermore, the images do not suffer from blurriness, thereby enabling the synthesis of sharp hands without additional datasets.

Generator architecture. We evaluated the influence of the proposed SwinGenerator and compared it to the generator architectures based on U-Net and ResNet, described in Section V-D. The results for *PHOENIX 2014T*, SynthSL and *BosphorusSign22k* are shown in Tables II, III and IV, respectively, and in Fig. 5.

In all datasets, the ResNet-based architecture showed an improved performance with respect to U-Net in terms of SSIM and PE. For the PSNR and FID, the results differ

Fig. 5. Sign images synthesized with the proposed pipeline and baseline architectures on SynthSL, PHOENIX 2014T and BosphorusSign22k.

TABLE II
QUANTITATIVE EVALUATION ON PHOENIX 2014T

Method	SSIM↑	PSNR↑	PE↓	FID↓
U-Net	0.684	30.999	0.884	42.624
ResNet	0.745	28.692	0.832	47.624
SwinGenerator	**0.835**	**33.971**	**0.675**	**31.678**

TABLE III
QUANTITATIVE EVALUATION ON SYNTHSL

Generator Architecture	SSIM↑	PSNR↑	PE↓	FID↓
U-Net	0.883	35.071	0.489	16.301
ResNet	0.897	33.050	0.469	14.160
SwinGenerator	**0.911**	**36.382**	**0.391**	**7.544**

TABLE V
EVALUATION OF DIFFERENT SWINGENERATOR ARCHITECTURES ON SYNTHSL

Generator Architecture	SSIM↑	PSNR↑	PE↓	FID↓
Swin-T$_{noPC}$	0.892	36.170	0.511	16.569
Swin-T$_{Res}$	0.895	36.362	0.468	12.633
SwinGenerator	0.911	36.382	0.391	7.544

TABLE VI
EVALUATION OF DIFFERENT SIZES OF SWIN V2 ON SYNTHSL

Swin Architecture	SSIM↑	PSNR↑	PE↓	FID↓
Swin-T	0.911	36.382	0.391	7.544
Swin-S	0.912	36.810	0.387	7.346
Swin-B	0.918	37.114	0.374	6.304

in each dataset for U-Net and ResNet. Our SwinGenerator outperforms both architectures in every dataset and metric, as shown in Tables II, III and IV. Furthermore, the image quality is degraded for both U-Net and ResNet using the same discriminator and loss function as in our model. In many cases, SwinGenerator provides steadier hands reconstructions (see 1st column of *PHOENIX 2014T* in Fig. 5), specially if the hands suffer from blurriness. However, in some cases ResNet generates slightly better hands as in the last column of *BosphorusSign22k* in Fig. 5.

G. Ablation Study

The following experiments were conducted on SynthSL.

Generator design. In this ablation study, we investigated the influence of different modules within SwinGenerator for synthesizing images. We first evaluated the SwinGenerator without the pose condition, i.e., without concatenating the resized target pose in the residual connections. This architecture is shown in Table V as Swin-T$_{noPC}$. The results show that adding the target pose in the residual connections is advantageous for SwinGenerator, with improvements in every metric, specially the PE and FID.

We additionally implemented a pipeline with only the source image as input, and the residual connection incorporating the difference between target pose and base pose. This model is shown in Table V as Swin-T$_{Res}$. This architecture reduces the model size, but compromises some level of quality compared to the proposed SwinGenerator.

TABLE IV
QUANTITATIVE EVALUATION ON BOSPHORUSSIGN22K

Generator Architecture	SSIM↑	PSNR↑	PE↓	FID↓
U-Net	0.769	29.673	0.420	34.161
ResNet	0.865	32.150	0.302	21.356
SwinGenerator	**0.930**	**37.589**	**0.211**	**15.824**

TABLE VII
EVALUATION ON SYNTHSL AFTER TRAINING WITH DIFFERENT ADVERSARIAL LOSSES

Loss function	SSIM↑	PSNR↑	PE↓	FID↓
Vanilla	0.911	36.382	0.391	7.544
LS-GAN [55]	0.910	36.716	0.391	7.845
Hinge [48]	0.909	36.628	0.386	8.856
WGAN-GP [33]	0.913	36.803	0.394	7.927

TABLE VIII

EVALUATION ON PHOENIX 2014T AND BOSPHORUSSIGN22K OF MODELS WITH AND WITHOUT PRE-TRAINING ON OTHER DATASETS

(Pre-)Trained			Fine-Tuned & Evaluated		SSIM↑	PSNR↑	PE↓	FID↓
SynthSL	PHOENIX 2014T	BosphorusSign22k	PHOENIX 2014T	BosphorusSign22k				
-	✓	-	✓	-	0.835	33.971	0.675	31.678
✓	-	-	✓	-	0.842	34.221	0.659	26.921
-	-	✓	✓	-	0.839	34.143	0.659	27.928
-	-	✓	-	✓	0.930	37.589	0.211	15.824
✓	-	-	-	✓	0.931	37.676	0.205	11.792

Transformer backbone. Our architecture was also evaluated with different backbone sizes of Swin V2 on SynthSL. More specifically, we compared the tiny Swin-T with the small (Swin-S) and base (Swin-B) versions of Swin Transformers. The quantitative results are shown in Table VI.

As expected, enhanced results were observed when using larger Swin Transformer models. However, this comes at a computational cost as the model size and complexity increase 2 and 4 times respectively, when using Swin-S and Swin-B. In Swin-S and Swin-B, the number of layers are $\{2, 2, 18, 2\}$ while the channel number of the hidden layers in the first stage C is 96 for Swin-S and 128 for Swin-B.

Loss function. We additionally compared different adversarial loss functions to train the generator and discriminator on SynthSL. The results in Table VII show that the choice of the adversarial loss does not have a significant impact on the performance of our pipeline.

Pre-trained model. Finally, we evaluated the effect on image synthesis when pre-training the networks in other datasets. Results are shown in Table VIII. Overall, pre-training on other datasets result on an improved performance in every metric in *PHOENIX 2014T* and *BosphorusSign22k*.

H. Runtime Analysis

We estimated the average inference time for the different generator architectures after 300 iterations. For U-Net, ResNet and SwinGenerator a single frame was generated on 43.59ms, 112.60ms and 46.34ms, respectively. Our SwinGenerator has a frame rate of ∼22 FPS.

I. Limitations and Failure Cases

In SynthSL, we observed some artifacts if the clothes of the signer had similar texture to the skin, as shown in Fig. 6. This problem did not occur in *PHOENIX 2014T* and *BosphorusSign22k*, as the signers always wear dark clothes. These artifacts were also present in images generated with U-Net and ResNet. For *PHOENIX 2014T* and *BosphorusSign22k*, some synthesized images had noisy regions in the face or clothes of the signers, as shown in Fig. 6. We believe this noise is due to low variability in these datasets regarding skin color and clothes, as it was not observed in SynthSL.

J. Discussion

Our evaluation shows that the proposed SwinGenerator introduces significant improvements with respect to the U-Net and ResNet-based generators, widely used for image synthesis, including SLP. These improvements were observed in all the evaluated datasets in Tables II, III and IV.

Fig. 6. Failure cases.

We additionally found that the ResNet-based generator tends to suffer from mode collapse with the current pipeline. In the literature, ResNet-based architectures for SLP [69], [72] are usually trained with the multi-scale discriminator from pix2pixHD [79]. In our experiments, we used a simple discriminator model to focus mainly on the performance of the evaluated generators. The multi-scale discriminator could also be used as an alternative for an improved performance on the synthesis of high-resolution images.

Finally, we observed that on real datasets for SLP hand synthesis suffers from blurriness, due to the inherent fast motion on sign gestures. In that regard, [69] provides a framework and keypoint-based loss with outstanding hand reconstruction. However, a private dataset with high quality hands and manually selected frames without blurriness were used. Since this dataset is not publicly available, SynthSL could be used as an alternative.

VI. CONCLUSIONS

In this paper, we introduced SynthSL, a large-scale synthetic dataset for research on sign language. We also proposed a rendering pipeline, to extend RGB-based sign videos to include rich ground truth data such as body, hands and head poses, segmentation masks, depth and normal maps. In addition, we presented a SL synthesis architecture conditioned on a given body pose and appearance. To that end, we leverage a GAN to generate realistic sign sequences with a newly introduced SwinGenerator for image synthesis. For future work, we would like to investigate the effect of additional input data such as depth map and segmentation masks for image synthesis. Furthermore, we aim to integrate additional ground truth data in our rendering pipeline such as optical flow, which can be exploited in SL-related tasks as in [63].

REFERENCES

[1] N. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G. T. Papadopoulos, V. Zacharopoulou, G. J. Xydopoulos, K. Atzakas, D. Papazachariou, and P. Daras. A comprehensive study on sign language recognition methods. *arXiv preprint arXiv:2007.12530*, 1:1–12, 2020.

[2] M. Alaghband, N. Yousefi, and I. Garibay. Facial expression phoenix (feph): An annotated sequenced dataset for facial and emotion-specified expressions in sign language. *Image*, 20:27, 2020.

[3] S. Albanie, G. Varol, L. Momeni, H. Bull, T. Afouras, H. Chowdhury, N. Fox, B. Woll, R. Cooper, A. McParland, and A. Zisserman. BOBSL: BBC-Oxford British Sign Language Dataset. In *arXiv preprint arXiv:2111.03635*, 2021.

[4] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. *ACM TOG*, 24(3):408–416, 2005.

[5] Autodesk, INC. Maya.

[6] S. Bak, P. Carr, and J.-F. Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. In *ECCV*, pages 189–205, Cham, Switzerland, 2018. Springer.

[7] M. J. Black, P. Patel, J. Tesch, and J. Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, pages 8726–8737, June 2023.

[8] Z. Cai, W. Yin, A. Zeng, C. Wei, Q. Sun, Y. Wang, H. E. Pang, H. Mei, M. Zhang, L. Zhang, C. C. Loy, L. Yang, and Z. Liu. Smpler-x: Scaling up expressive human pose and shape estimation. *NeurIPS*, 1, 2023.

[9] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden. Neural sign language translation. In *CVPR*, pages 7784–7793, Piscataway, NJ, 2018. IEEE.

[10] N. C. Camgöz, B. Saunders, G. Rochette, M. Giovanelli, G. Inches, R. Nachtrab-Ribback, and R. Bowden. Content4all open research sign language translation datasets. In *FG*, pages 1–5. IEEE, 2021.

[11] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, pages 7291–7299, Piscataway, NJ, 2017. IEEE.

[12] X. Chai, H. Wang, and X. Chen. The devisign large vocabulary of chinese sign language database and baseline evaluations. Technical Report VIPL-TR-14-SLR-001, Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, 2014.

[13] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros. Everybody dance now. In *ICCV*, pages 5933–5942, Piscataway, NJ, 2019. IEEE.

[14] Y. Chen, W. Li, X. Chen, and L. V. Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *CVPR*, pages 1841–1850, Piscataway, NJ, 2019. IEEE.

[15] H. M. Clever, Z. Erickson, A. Kapusta, G. Turk, K. Liu, and C. C. Kemp. Bodies at rest: 3d human pose and shape estimation from a pressure image using synthetic data. In *CVPR*, pages 6215–6224, Piscataway, NJ, 2020. IEEE.

[16] B. O. Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018.

[17] S. Cox, M. Lincoln, J. Tryggvason, M. Nakisa, M. Wells, M. Tutt, and S. Abbott. Tessa, a system to aid communication with deaf people. In *Proceedings of the 5th International ACM Conference on Assistive Technologies*, pages 205–212, New York, USA, 2002. ACM Press.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[20] P. Dreuw, T. Deselaers, D. Keysers, and H. Ney. Modeling image variability in appearance-based gesture recognition. In *ECCV Workshop on Statistical Methods in Multi-image and Video Processing*, pages 7–18, Cham, Switzerland, 2006. Springer.

[21] A. Duarte, S. Palaskar, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, and X. Giro-i Nieto. How2sign: A large-scale multimodal dataset for continuous american sign language. In *Sign Language Recognition, Translation and Production (SLRTP) Workshop-Extended Abstracts*, pages 1–4, Cham, Switzerland, 2020. Springer.

[22] A. C. Duarte. Cross-modal neural sign language translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1650–1654, New York, USA, 2019. ACM.

[23] A. Dundar, M.-Y. Liu, T.-C. Wang, J. Zedlewski, and J. Kautz. Domain stylization: A strong, simple baseline for synthetic to real image domain adaptation. *arXiv preprint arXiv:1807.09384*, 1:1–10, 2018.

[24] S. Ebling, N. C. Camgöz, P. Boyes Braem, N. Calzolari, et al. Smile swiss german sign language dataset. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18)*, pages 4221–4229, Miyazaki, Japan, 2018. European Language Resources Association (ELRA).

[25] S. Fang, C. Sui, X. Zhang, and Y. Tian. Signdiff: Learning diffusion models for american sign language production. *arXiv preprint arXiv:2308.16082*, 2023.

[26] J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. H. Piater, and H. Ney. RWTH-PHOENIX-Weather: A large vocabulary sign language recognition and translation corpus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, volume 9, pages 3785–3789, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).

[27] J. Forster, C. Schmidt, O. Koller, M. Bellgardt, and H. Ney. Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-Weather. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1911–1916, Reykjavik, Iceland, 2014. European Language Resources Association.

[28] M. F. Ghezelghieh, R. Kasturi, and S. Sarkar. Learning camera viewpoint using cnn to improve 3d body pose estimation. In *2016 fourth international conference on 3D vision (3DV)*, pages 685–693, Piscataway, NJ, 2016. IEEE.

[29] S. Gibet, N. Courty, K. Duarte, and T. L. Naour. The signcom system for data-driven animation of interactive virtual signers: Methodology and evaluation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 1(1):1–23, 2011.

[30] S. Gibet, F. Lefebvre-Albaret, L. Hamon, R. Brun, and A. Turki. Interactive editing in french sign language dedicated to virtual signers: requirements and challenges. *Universal Access in the Information Society*, 15(4):525–539, 2016.

[31] J. Glauert, R. Elliott, S. Cox, J. Tryggvason, and M. Sheard. Vanessa–a system for communication between deaf and hearing people. *Technology and Disability*, 18(4):207–216, 2006.

[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *NeurIPS*, 27, 2014.

[33] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. *NeurIPS*, 30, 2017.

[34] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, pages 11807–11816, Piscataway, NJ, 2019. IEEE.

[35] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE, 2016.

[36] D. T. Hoffmann, D. Tzionas, M. J. Black, and S. Tang. Learning to train with synthetic humans. In *German Conference on Pattern Recognition (GCPR)*, pages 609–623, Cham, Switzerland, Sept. 2019. Springer.

[37] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li. Video-based sign language recognition without temporal segmentation. In *AAAI*, pages 2257–2264, Palo Alto, California, 2018. IAAAI, AAAI Press.

[38] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017.

[39] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016.

[40] K. Karpouzis, G. Caridakis, S.-E. Fotinea, and E. Efthimiou. Educational resources and implementation of a greek sign language synthesis architecture. *Computers & Education*, 49(1):54–74, 2007.

[41] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR*, 2014.

[42] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *ICLR*, 2014.

[43] S.-K. Ko, C. J. Kim, H. Jung, and C. Cho. Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 9(13):2683, 2019.

[44] O. Koller, H. Ney, and R. Bowden. Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In *CVPR*, pages 3793–3802, Piscataway, NJ, 2016. IEEE.

[45] J. G. Kyle, J. Kyle, B. Woll, G. Pullen, and F. Maddix. *Sign language: The study of deaf people and their language*. Cambridge University Press, Cambridge, UK, 1988.

[46] D. Li, C. Rodriguez, X. Yu, and H. Li. Word-level deep sign language recognition from video: A new large-scale dataset and

methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1459–1469, Piscataway, NJ, 2020. IEEE.

[47] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4d scans. *ACM TOG*, 36(6):194–1, 2017.

[48] J. H. Lim and J. C. Ye. Geometric GAN. *arXiv preprint arXiv:1705.02894*, 2017.

[49] J. Liu, H. Rahmani, N. Akhtar, and A. Mian. Learning human pose models from synthesized data for robust rgb-d action recognition. *IJCV*, 127(10):1545–1564, 2019.

[50] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, pages 12009–12019, 2022.

[51] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *CVPR*, pages 10012–10022, 2021.

[52] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 34(6):248:1–248:16, Oct. 2015.

[53] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann. Mediapipe: A framework for perceiving and processing reality. In *CVPR-W*, 2019.

[54] J. Malik, S. Shimada, A. Elhayek, S. A. Ali, C. Theobalt, V. Golyanik, and D. Stricker. Handvoxnet++: 3d hand shape and pose estimation using voxel-based neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8962–8974, 2021.

[55] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley. Least squares generative adversarial networks. In *ICCV*, pages 2794–2802, 2017.

[56] J. McDonald, R. Wolfe, J. Schnepp, J. Hochgesang, D. G. Jamrozik, M. Stumbo, L. Berke, M. Bialek, and F. Thomas. An automated technique for real-time production of lifelike animations of american sign language. *Universal Access in the Information Society*, 15(4):551–566, 2016.

[57] O. Özdemir, A. A. Kındıroğlu, N. C. Camgöz, and L. Akarun. Bosphorussign22k sign language recognition dataset. *arXiv preprint arXiv:2004.01283*, 1:1–8, 2020.

[58] P. Patel, C.-H. P. Huang, J. Tesch, D. T. Hoffmann, S. Tripathi, and M. J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[59] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, pages 10975–10985, Piscataway, NJ, 2019. IEEE, IEEE.

[60] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[61] A. Ranjan, D. T. Hoffmann, D. Tzionas, S. Tang, J. Romero, and M. J. Black. Learning multi-human optical flow. *IJCV*, 128:873–890, Jan. 2020.

[62] A. Ranjan, J. Romero, and M. J. Black. Learning human optical flow. In *BMVC*, pages 1–13, Durham, UK, Sept. 2018. British Machine Vision Association, BMVA Press.

[63] J. Rodriguez, J. Chacon, E. Rangel, L. Guayacan, C. Hernandez, L. Hernandez, and F. Martinez. Understanding motion in sign language: A new structured translation dataset. In *Proceedings of the Asian Conference on Computer Vision*, pages 1–16, Cham, Switzerland, 2020. Springer.

[64] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM TOG*, 36(6):245, 2017.

[65] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Med. Image Comput. Comput. Assist. Interv.*, pages 234–241, 2015.

[66] B. Saunders, N. C. Camgöz, and R. Bowden. Adversarial training for multi-channel sign language production. In *The 31st British Machine Vision Virtual Conference*, pages 1–15, Durham, UK, 2020. British Machine Vision Association, BMVA Press.

[67] B. Saunders, N. C. Camgoz, and R. Bowden. Everybody sign now: Translating spoken language to photo realistic sign language video. *arXiv preprint arXiv:2011.09846*, 1:1–11, 2020.

[68] B. Saunders, N. C. Camgoz, and R. Bowden. Progressive transformers for end-to-end sign language production. In *ECCV*, pages 687–705, Cham, Switzerland, 2020. Springer.

[69] B. Saunders, N. C. Camgoz, and R. Bowden. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *CVPR*, pages 5141–5151, 2022.

[70] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

[71] S. Stoll, N. C. Camgöz, S. Hadfield, and R. Bowden. Sign language production using neural machine translation and generative adversarial networks. In *BMVC*, pages 1–12, Durham, UK, 2018. British Machine Vision Association, BMVA Press.

[72] S. Stoll, N. C. Camgoz, S. Hadfield, and R. Bowden. Text2sign: towards sign language production using neural machine translation and generative adversarial networks. *IJCV*, 128:891–908, 2020.

[73] S. Stoll, S. Hadfield, and R. Bowden. Signsynth: Data-driven sign language video generation. In *8th International Workshop on Assistive Computer Vision and Robotics*, pages 1–17, Cham, Switzerland, 2020. Springer.

[74] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, and S. Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *CVPRW*, pages 969–977, Piscataway, NJ, 2018. IEEE.

[75] H. R. Vaezi Joze and O. Koller. MS-ASL: A large-scale data set and benchmark for understanding american sign language. In *BMVC*, pages 1–16, Durham, UK, 2019. British Machine Vision Association, BMVA Press.

[76] G. Varol, I. Laptev, C. Schmid, and A. Zisserman. Synthetic humans for action recognition from unseen viewpoints. *Int. Journal of Computer Vision (IJCV)*, 129:2264—2287, 2021.

[77] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *CVPR*, pages 109–117, Piscataway, NJ, 2017. IEEE.

[78] L. Ventura, A. Duarte, and X. Giro-i Nieto. Can everybody sign now? exploring sign language video generation from 2d poses. In *Sign Language Recognition, Translation and Production (SLRTP) Workshop - Extended Abstracts*, pages 1–4, Cham, Switzerland, 2020. Springer.

[79] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *CVPR*, pages 8798–8807, 2018.

[80] D. Wood, A. Griffiths, H. Wood, and I. Howarth. *Teaching and talking with deaf children*, volume 10. Wiley, New York, USA, 1986.

[81] Y. Xu, J. Zhang, Q. Zhang, and D. Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022.

[82] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[83] J. Zelinka and J. Kanis. Neural sign language synthesis: Words are our glosses. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 3395–3403, Piscataway, NJ, 2020. IEEE.

[84] P. Zhang, L. Yang, J.-H. Lai, and X. Xie. Exploring dual-task correlation for pose guided person image generation. In *CVPR*, pages 7713–7722, 2022.

[85] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017.