SynLibras: A Disentangled Deep Generative Model for Brazilian Sign Language Synthesis

Wellington Silveira, Andrew Alaniz, Marina Hurtado, Bernardo Castello da Silva and Rodrigo de Bem
Center of Computational Sciences (C3)
Federal University of Rio Grande (FURG), Brazil
Emails: {wellingtonfs, girafandrewalaniz, marina_hurtado, becastellosilva, rodrigobem}@furg.br

***Abstract*—Recent advances regarding deep generative models have strengthened a realm of approaches in which discriminative and generative tasks are tackled jointly in an *analysis-by-synthesis* manner. In this category, variational autoencoders (VAEs) and generative adversarial networks (GANs) aim for learning latent data representations from which sampling of synthetic images may be performed. However, sampling in such models normally does not allow for independent control of diverse factors of variation. Despite general efforts to overcome this issue, deep generative models tailored for sign language with disentangled factors of variation are yet not vastly explored in the literature. In this work, we introduce the SynLibras, a novel model that allows for disentangling appearance and gestural communication (i.e. body, hands and face poses) on image synthesis. Our model is capable of performing cross-language pose-transfer while maintaining the appearance of the source signer. We perform experiments on the RWTH-PHOENIX-Weather dataset and evaluation using the PSNR and the SSIM metrics. To our knowledge, the SynLibras is the first method for Brazilian sign language (Libras) synthesis in images. We compare our model with the EDN, a well-known general pose-transfer method, achieving better results on Libras synthesis. Finally, we also introduce the SynLibras-Pose, a dataset with annotated poses of Libras signers performing single words.**

I. Introduction

According to recent information from the World Health Organization (WHO) [1], approximately 430 million people in the world present some degree of disabling hearing loss. In addition to the personal issues caused by such a disability, the WHO estimates that it has an annual global cost of US$ 980 billion, which includes health sector costs, educational support, loss of productivity, and societal costs.

People who are not able to hear well face a significant communication barrier that may prevent them to take part in several activities, from social interaction to education. Such restrictions are frequently overcome with the use of sign language. Despite their enormous importance, gestural languages are still spoken by a small portion of countries' overall population, as shown by studies about American sign language (ASL) speakers in the United States [2]. Therefore, even in their own countries, people with hearing loss frequently depend on translation to communicate.

Assistive technologies may play a relevant role in the mitigation of the personal and economic consequences of this problem. One may find a variety of related approaches, from gloves that turn signs into speech [3] to computer graphics avatars that reproduce sign language from text or sound [4]. In this realm, computer vision approaches stand as non-intrusive methods for handling images and videos of people speaking in sign language. Efforts aiming for the recognition and translation of sign languages are already well-known in the literature. Nevertheless, the recent advent of deep generative models, such as GANs and VAEs, has opened a new branch of analysis-by-synthesis methods [5], [6], which aim for the recognition, translation, and synthesis of realistic images and videos of people speaking in sign languages.

Here, we propose the SynLibras, a novel conditional VAE-GAN model capable of generating disentangled synthetic images of Libras signers. To our knowledge, it is the first capable of producing synthetic images of people communicating in Libras. We also introduce the SynLibras-Pose dataset, a visual dataset for Libras with annotated poses of actors signing words in videos. We perform experiments on the RWTH-PHOENIX-Weather dataset [7] and evaluation using the well-known peak signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM) metrics [8]. We compare our model with the well-known and publicly available EDN [9] pose-transfer method, achieving better scores on the image quality metrics and better Libras pose-transfer performance. The SynLibras is capable of training and synthesizing images of multiple people with just one trained model, in contrast with the EDN method, in which a model only handles images of one single person at a time in training and testing. Our method and dataset represent relevant contributions to this challenging task, aiming for the automatic and controllable generation of realistic synthetic sign language speakers.

II. Related Work

Sign language recognition and translation may be considered as classical applications of computer vision [10]–[13]. Synthesizing sign language with computer graphics avatars is a well-known strategy [14]–[17]. Therefore, instead of aiming for a broad literature review of these topics, we highlight methods for sign language synthesis on images and videos based on deep learning, covering visual datasets for this purpose. In both cases, we also emphasize literature related to Libras. A recent and more general survey on deep learning sign language production is provided by Rastgoo et al. [18].

Early adopters of deep generative models for sign language generation in videos, Stoll et al. [19] propose sign language translation using neural machine translation (NMT) [20] and generative adversarial networks (GANs) [21]. They perform a sequence of translations, from spoken language to lexical entities (glosses), and finally, to sequences of body postures. In doing so, they produce faithful German sign language (DGS) videos sequences conditioned to appearance and pose. The authors extend their work using a motion graph in [22], which is also extended in [23]. More recently, Stoll et al. [24] and Saunders et al. [25] perform translation to German sign language videos, still relying on GANs. Also using GANs, Vasani et al. [26] and Krishna et al. [27] perform video frames generation of Indian sign language (ISL). Ventura et al. [28] apply the EDN method [9] for generating signing videos. Saunders et al. [29] propose an image-to-image method for anonymization of synthetic sign language videos, and very recently a pose-conditioned GAN model for video production [30]. Concerning Libras, despite the existence of machine learning methods for recognition and translation [31]–[35], to our knowledge, the SynLibras is the first method for synthesis on images. In contrast with previous art, our disentangled deep generative model, based on a conditional VAE-GAN architecture, was capable of performing cross-language pose-transfer with independent control over the pose and appearance.

Regarding visual datasets for sign languages, several benchmarks are found in the literature. The RWTH-PHOENIX-Weather dataset [7] has been widely used since it provides a large collection of German sign language sequences (8,257) performed by 9 signers. It does not provide body posture annotations, which are frequently generated with body pose estimators [36] when required during sign language image synthesis. A more recent dataset, the How2Sign [37] consists of a larger set of sequences (38,611) from 10 different ASL signers. It already includes postural annotations. Besides these, other collections of visual data are found in the literature, for diverse languages, such as British sign language [38], [39], ASL [40], [41], German sign language [42], Turkish sign language [43], Chinese sign language [44], Czech sign language [45], and, Libras [32], [46]–[48]. In contrast, our SynLibras-Pose contains annotations of the body, hands and faces of the signers, as well as the words they are performing.

III. Methodology: The SynLibras Architecture

Deep generative models based on VAEs [49] aim for maximizing the evidence lower bound (ELBO) of the log-likelihood over training data $\mathbf{x}$, marginalized over a latent variable $\mathbf{z}$. Such models may be turned into *conditional* VAEs (CVAEs) [50] by incorporating a conditioning variable $\mathbf{y}$ to the marginal log-likelihood, as per

$$\begin{aligned}\log p_\theta(\mathbf{x}|\mathbf{y}) \geq \mathcal{L}_{\text{CVAE}}(\phi, \theta; \mathbf{x}, \mathbf{y}) \\ = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})}\left[\log p_\theta(\mathbf{x}|\mathbf{z},\mathbf{y})\right] \\ - \text{KL}\left[q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y}) || p_\theta(\mathbf{z}|\mathbf{y})\right], \quad (1)\end{aligned}$$

where, $\log p_\theta(\mathbf{x}|\mathbf{y})$ is the conditional marginal log-likelihood and $\mathcal{L}_{\text{CVAE}}(\phi, \theta; \mathbf{x}, \mathbf{y})$ is the evidence lower bound. The ELBO is composed of two terms, which are learned simultaneously through an encoder-decoder neural network architecture. In such an architecture, following Eq. 1, an "inference network" (encoder), with parameters ϕ, minimizes the KL-divergence between the surrogate distribution $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})$ and the prior distribution $p_\theta(\mathbf{z}|\mathbf{y})$, while a "generative network" (decoder), with parameters θ, minimizes the expected reconstruction error of the generative model $p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{y})$. Additionally, such models may be extended with discriminators from GANs [21], which have their objective defined as a two-player minimax game

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(D, G) = \mathbb{E}_{p(\mathbf{x})}\left[\log D(\mathbf{x})\right] \\ + \mathbb{E}_{p_{\mathbf{z}}(\mathbf{z})}\left[\log(1 - D(G(\mathbf{z})))\right]. \quad (2)\end{aligned}$$

Here, the discriminator (D) is trained to maximize the probability of assigning the correct labels to real and fake samples, while simultaneously the generator (G) is trained to minimize $\log(1-D(G(\mathbf{z})))$. Finally, CVAE-GAN models [51]–[53] have their final objective given by

$$\mathcal{L} = \mathcal{L}_{\text{CVAE}} + \mathcal{L}_{\text{GAN}}. \quad (3)$$

Our SynLibras architecture, illustraded in Figure 1, is an CVAE-GAN model. During training, an input RGB image $\mathbf{x}$ is received by the Encoder (E) and reconstructed by the Decoder/Generator (G) aiming for the minimization of the L1-norm reconstruction loss. Meanwhile, the heatmap pose representation $\mathbf{y}$ is conditioning these two modules. It is also the input of a Prior (P) module correspondent to $p_\theta(\mathbf{z}|\mathbf{y})$, which is used in the minimization of the KL-divergence $KL[q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})||p_\theta(\mathbf{z}|\mathbf{y})]$. The Gaussian latent variable $\mathbf{z}$ is sampled using the reparametrization trick [49], as per $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y}) = \mathcal{N}(\mu, \sigma^2)$, and encodes the visual appearance of the signer, once the gestural communication is encoded by the conditioning pose representation $\mathbf{y}$. The intentional disentanglement of visual appearance $\mathbf{z}$ and posture $\mathbf{y}$ in the generative model $p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{y})$ aims for the independent manipulation of these factors of variation in the synthesis of images. Finally, a Discriminator module (D) classifies the output images as real or fake. Following Eq. 2, it minimizes the cross-entropy loss and contributes for image quality.

Since visual appearance and body posture are disentangled, at test time the SynLibras may be employed for pose-transfer. In this task, the Encoder is fed with an input RGB image which is reconstructed by the Decoder with a different body posture. Therefore, the model can synthesize images with appearance given by the input image and body pose given by the conditioning variable $\mathbf{y}$. The Prior and the Discriminator modules are not used during this task.

Finally, concerning the details of our architecture, all SynLibras modules are composed of residual layers [54], except the Decoder/Generator (G), which has a purely convolutional part. We also employ the Leaky ReLU [55] as the activation function of all the non-linear layers. In addition, we use equalized learning rate [56] and pixel-wise feature vector normalization [56] to maintain greater stability during training. Finally, our pose representation emphasizes hand poses and facial expressions, which are crucial for sign language.

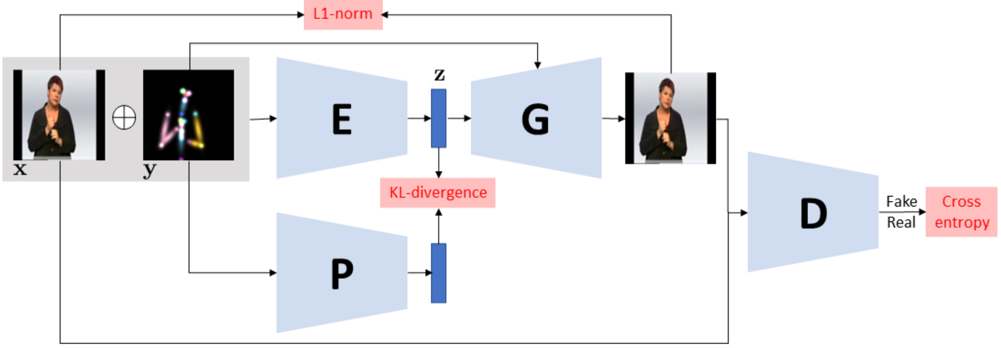

Fig. 1. **SynLibras architecture.** During training, an input RGB image $\mathbf{x}$ is received by the Encoder (E) and reconstructed by the Decoder/Generator (G) through the minimization of the L1-norm. Meanwhile, the pose representation $\mathbf{y}$ is conditioning these two modules. It is also the input of a Prior (P) module, used in the minimization of the KL-divergence $KL[q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})||p_\theta(\mathbf{z}|\mathbf{y})]$, between the surrogate distribution $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$ and the prior distribution $p_\theta(\mathbf{z}|\mathbf{y})$. The Gaussian latent variable $\mathbf{z}$ is sampled using the reparametrization trick [49], as per $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y}) = \mathcal{N}(\mu,\sigma^2)$, encoding the visual appearance of the signer. Finally, a Discriminator module (D) classifies the output images as real or fake and contributes to further improvement of image quality. At test time, for the pose-transfer task, a given input RGB image may be combined with different poses $\mathbf{y}$, allowing for the synthesis of a person with a certain appearance in diverse postural configurations. The Prior and the Discriminator modules are not used during testing.

We derive Gaussian heatmaps from body, hands and facial keypoints. The 2D heatmaps are employed as the conditioning pose representation $\mathbf{y}$ in the model (see Fig. 1).

IV. Experiments and Results

A. Datasets

We train our model on the RWTH-PHOENIX-Weather dataset [7]. This is a well-known sign language benchmark in which 9 subjects translate the news and the weather forecast into German sign language. The dataset consists of videos captured at 25 frames per second and an original resolution of 210×260 pixels. We employ the OpenPose [36] method to automatically estimate the position of the body, hands and face keypoints. In total, we use 30 keypoints: 12 for the upper body, 5 for each hand and 8 for the face. Each one of these keypoints is mapped to a 2D Gaussian heatmap [57], which together are the conditioning pose representation in our VAE-GAN model. After sampling images in time, and discarding corrupted and inconsistent estimates, we end up with a training set of 19,365 frames and a testing set of 2,957 frames.

Regarding Brazilian sign language, we introduce the SynLibras-Pose dataset, which is based on an open Libras-Portuguese video dictionary [58]. It consists of 1,133 videos with approximately 200 frames each with an original size of 1920×1080 pixels. The videos contain subjects performing signs corresponding to individual words related to several different themes and areas. We crop 1024×1024 windows, keeping the subjects centralized. We use the OpenPose [36] to automatically annotate the pose keypoints, manually correcting them to enforce consistency and ending up with 427 videos (approximately 85,400 frames). Samples of video frames with the corresponding annotations are shown in Figure 2.

B. Implementation and Training

Both, our implementation and dataset will be made publicly available for the sake of reproducibility[11]. We implement our model in PyTorch and all experiments were run on Google Colab. Regarding training, it has lasted for 15 epochs with our mini-batches consisting of 6 images on the RWTH-PHOENIX-Weather dataset [7]. We use the Adam [59] with a learning rate of 10^{-3}, $\beta1 = 0.0$ and $\beta2 = 0.999$. Network weights were initialized following Gaussian initialization. All images are normalized to 256×256 pixels, centered on the subject. No data augmentation was used except for image normalization to zero mean and unit variance.

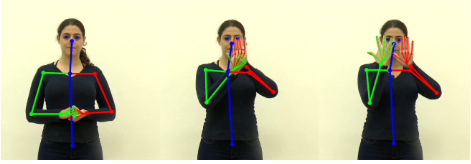

Fig. 2. **SynLibras-Pose.** A visual dataset of Libras with annotated poses, as illustrated above, based on an open Libras-Portuguese video dictionary [58]. We provide annotations for 30 keypoints: 12 for the upper body, 5 for each hand, and 8 for the face.

C. Baseline Method

To our knowledge, no other method in the literature tackles Libras synthesis in images. Therefore, to evaluate the SynLibras model we have adopted a well-known and publicly available method for pose-transfer, the EDN (FBF) [9]. Differently from the SynLibras, which is capable of learning a generative model of several people with diverse appearances and postures simultaneously, the EDN is trained and tested with sequences of images of a single person at a time. At test, the model is then capable of transferring postures to that particular person employed in training. In this context, we have performed specific experiments with the SynLibras for allowing a fair comparison with the EDN baseline. All experiments and results are presented in the following sections.

D. Qualitative and Quantitative Evaluation

Initially, we evaluate the image quality of our reconstructions on the RWTH-PHOENIX-Weather test set. For that, we

[11] https://github.com/ReplicAI/SynLibras

employ the well-known SSIM [8] and the PSNR metrics. The SynLibras achieves SSIM=0.89 and PSNR=25.2 on the full test set composed of multiple people. The reconstructions of the RWTH-PHOENIX-Weather are illustrated in Figure 3.

For a comparison between the SynLibras and the EDN baseline, we have trained ($\approx$ 12 epochs) and tested the EDN on a reduced version of the RWTH-PHOENIX-Weather dataset containing only a single person (PHOENIX_Single), since the EDN is not capable of learning the appearance of multiple people simultaneously. The quantitative results of both models on the reconstruction of the PHOENIX_Single test set, which contains 2,084 images, are shown in Table I. Our SynLibras model has presented better performance regarding both metrics. A qualitative comparison is shown in Figure 4.

TABLE I
COMPARISON ON THE TEST SET OF THE PHOENIX_SINGLE DATASET. LARGER SCORES MEAN BETTER RESULTS.

	SSIM	**PNSR**
SynLibras	0.87	24.12
EDN [9]	0.86	22.85

Employing our SynLibras-Pose dataset for testing, we perform cross-language pose-transfer. With the SynLibras model, we produce synthetic images of people presenting appearances from the RWTH-PHOENIX-Weather [7] subjects, yet performing signs in Libras. This shows the capability of the model to disentangle gestural communication and visual appearance. To our knowledge, the SynLibras is the first method to produce disentangled synthetic images of multiple Libras signers from a single model. Qualitative results of pose-transfer are shown in Figure 5. For the comparison between the SynLibras and the EDN [9] on the pose-transfer task, we also employ the PHOENIX_Single, as shown in Figure 6.

Lastly, we importantly emphasize that in contrast with related methods such as the EDN, the SynLibras is capable of training with images of multiple people simultaneously. Consequently, the same trained model is capable of synthesizing images of diverse people, as illustrated in Figure 5. The EDN, for instance, is only capable of training and testing with images of a single person at a time. Thus, it is needed to retrain the model for generating images of different people. Due to these differences, regarding the comparisons between the SynLibras and the EDN, the SynLibras was trained on the full RWTH-PHOENIX-Weather dataset and tested on the PHOENIX_Single, while the EDN was trained and tested on the PHOENIX_Single dataset.

V. CONCLUSIONS, LIMITATIONS AND FUTURE

We introduce the SynLibras, a disentangled deep generative model for tackling the challenging and relevant task of synthesizing Libras in images. Our deep learning architecture is a CVAE-GAN conditioned on gestural communication (body, hands and facial postures). We also introduce the SynLibras-Pose, a dataset of Libras signers performing single words in videos with body, hands and facial keypoints labeling. We have trained the SynLibras on the RWTH-PHOENIX-Weather, and tested on both the RWTH-PHOENIX-Weather and the SynLibras-Pose datasets. We compare our model with the EDN [9], a well-known and publicly available method for general pose-transfer. The SynLibras presents better scores on the SSIM and the PSNR metrics. Moreover, our model was capable of presenting better performance than the EDN on cross-language pose-transfer, i.e. to synthesize images of RWTH-PHOENIX-Weather dataset subjects performing Libras signs from the SynLibras-Pose dataset, emphasizing the disentanglement between appearance and gestural posture in the model.

Despite the good results, fine details of hair, faces and hands are still hard to reconstruct (Fig. 3), as well as the precise direction of the face (Fig. 5). Such problems may be overcome with better and more accurate pose representation. Finally, another direction of future improvement is the use of pose estimators over synthetic images as a surrogate measurement of image quality.

ACKNOWLEDGMENT

This work was supported by FAPERGS, grant ARD-21/2551-0000678-1. Wellington Silveira (PROBITI/FAPERGS), Andrew Alaniz (PIBITI/CNPq), Marina Hurtado (PIBIC/CNPq), and Bernardo Castello da Silva (EPEC/FURG) are scholarship holders.

REFERENCES

[1] World Health Organization (WHO), "Deafness and hearing loss," 2021, available at https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss.

[2] R. E. Mitchell, T. A. Young, B. Bachelda, and M. A. Karchmer, "How many people use asl in the united states? why estimates need updating," *Sign Language Studies*, vol. 6, no. 3, 2006.

[3] Z. Zhou, K. Chen, X. Li, S. Zhang, Y. Wu, Y. Zhou, K. Meng, C. Sun, Q. He, W. Fan *et al.*, "Sign-to-speech translation using machine-learning-assisted stretchable sensor arrays," *Nature Electronics*, 2020.

[4] M. Kipp, A. Heloir, and Q. Nguyen, "Sign language avatars: Animation and comprehensibility," in *International Workshop on Intelligent Virtual Agents*, 2011.

[5] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," *NIPS*, 2017.

[6] R. de Bem, A. Ghosh, T. Ajanthan, O. Miksik, A. Boukhayma, N. Siddharth, and P. Torr, "Dgpose: Deep generative models for human body analysis," *IJCV*, vol. 128, no. 5, 2020.

[7] J. Forster, C. Schmidt, O. Koller, M. Bellgardt, and H. Ney, "Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-Weather," in *LREC*, 2014.

[8] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE TIP*, vol. 13, no. 4, 2004.

[9] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," in *ICCV*, 2019.

[10] H. Cooper, B. Holt, and R. Bowden, "Sign language recognition," in *Visual analysis of humans*. Springer, 2011, pp. 539–562.

[11] S. C. Ong and S. Ranganath, "Automatic sign language analysis: A survey and the future beyond lexical meaning," *IEEE TPAMI*, 2005.

[12] M. J. Cheok, Z. Omar, and M. H. Jaward, "A review of hand gesture and sign language recognition techniques," *IJMLC*, 2019.

[13] O. Koller, "Quantitative survey of the state of the art in sign language recognition," *arXiv preprint arXiv:2008.09918*, 2020.

[14] J. Kennaway, J. R. Glauert, and I. Zwitserlood, "Providing signed content on the internet by synthesized animation," *ACM TOCHI*, 2007.

[15] R. Elliott, J. R. Glauert, J. Kennaway, I. Marshall, and E. Safar, "Linguistic modelling and language-processing technologies for avatar-based sign language presentation," *UAIS*, 2008.

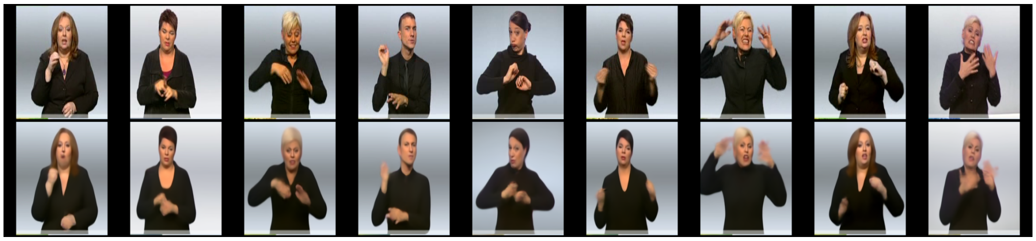

Fig. 3. **SynLibras.** Reconstructed images on the RWTH-PHOENIX-Weather [7] test set. The top row shows the original images and the bottom row shows the images reconstructed by the SynLibras model. We achieve SSIM=0.89 and PSNR=25.2. The images have 256 × 256 pixels.

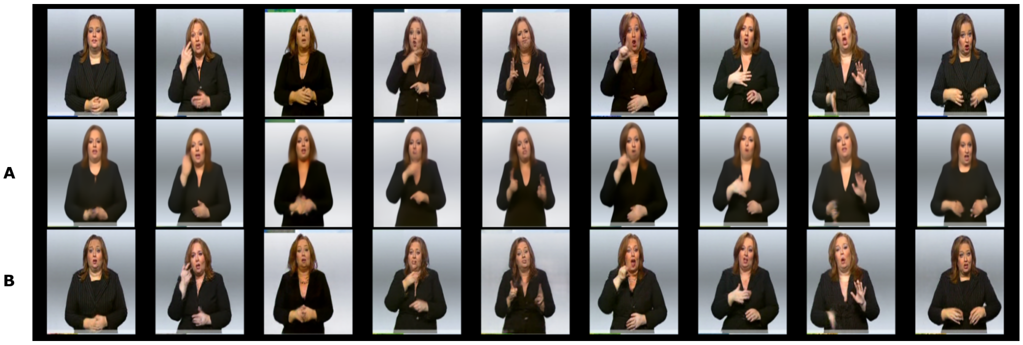

Fig. 4. **SynLibras vs. EDN.** Reconstructed images from PHOENIX_Single test set comparing the SynLibras and the EDN [9]. The top row shows the original images, followed by the reconstructions by the SynLibras (row **A**) and by the EDN [9] (row **B**). All the images have 256 × 256 pixels.

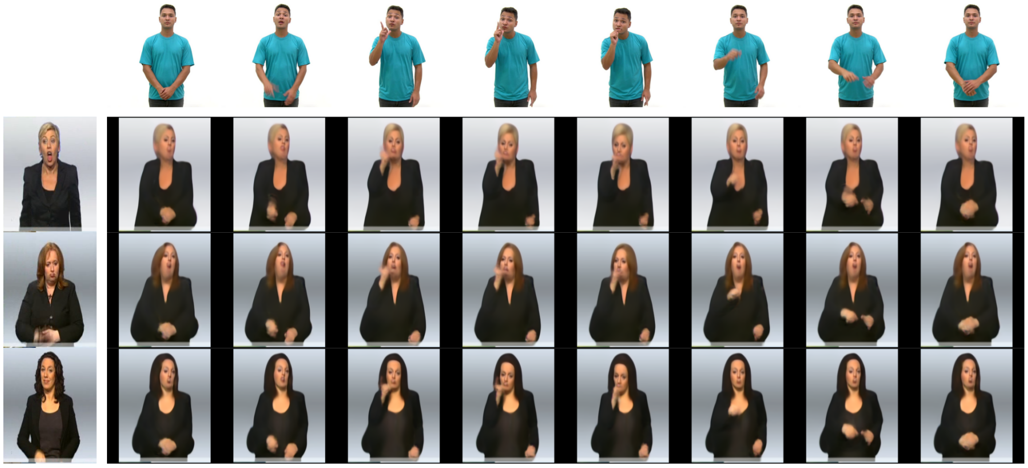

Fig. 5. **SynLibras cross-language pose-transfer.** In the top row, we show a man performing the sign correspondent to the word *water* in Libras. Following, we have the original appearance of subjects from the RWTH-PHOENIX-Weather [7] (first column), followed by pose-transfer images generated by the SynLibras. We highlight that the subjects in the RWTH-PHOENIX-Weather dataset have never performed signs in Libras, showing the generality of the method. We call attention to the changes in face orientation and hand shapes along the sequence. All images have 256 × 256 pixels.

[16] P. Cabral, M. Gonçalves, H. Nicolau, L. Coheur, and R. Santos, “Pe2lgp animator: A tool to animate a portuguese sign language avatar,” in *LREC Workshops*, 2020.

[17] Hand Talk, “Hand talk,” 2021, available at https://www.handtalk.me/en.

[18] R. Rastgoo, K. Kiani, S. Escalera, and M. Sabokrou, “Sign language production: A review,” in *CVPR Workshops*, 2021.

[19] S. Stoll, N. C. Camgöz, S. Hadfield, and R. Bowden, “Sign language production using neural machine translation and generative adversarial networks,” in *BMVC*, 2018.

[20] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *NIPS*, 2014.

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

[22] S. Stoll, N. C. Camgoz, S. Hadfield, and R. Bowden, “Text2sign: Towards sign language production using neural machine translation and generative adversarial networks,” *IJCV*, vol. 128, no. 4, 2020.

[23] C. Kissel, C. Kümmel, D. Ritter, and K. Hildebrand, “Pose-guided sign language video gan with dynamic lambda,” *arXiv preprint arXiv:2105.02742*, 2021.

[24] S. Stoll, S. Hadfield, and R. Bowden, “Signsynth: Data-driven sign language video generation,” in *ECCV*, 2020.

[25] 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*, 2020.

[26] N. Vasani, P. Autee, S. Kalyani, and R. Karani, “Generation of indian sign language by sentence processing and generative adversarial networks,” in *ICISS*, 2020.

[27] S. Krishna and J. Ukey, “Gan based indian sign language synthesis,” in *ICVGIP*, 2021.

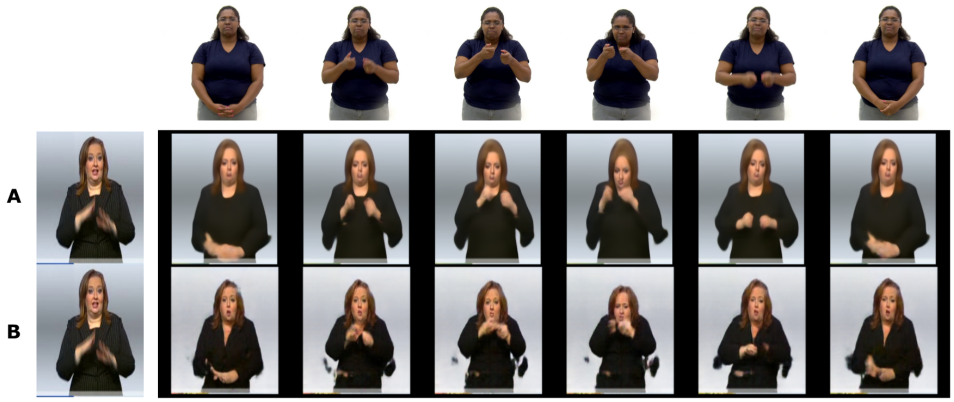

Fig. 6. **SynLibras vs. EDN cross-language pose-transfer.** In the top row, we show a woman performing the sign correspondent to the word *video game* in Libras. Following, we have the original appearance of the single subject from the PHOENIX_Single (first column), followed by pose-transfer images generated by the SynLibras (row **A**) and by the EDN [9] (row **B**). We highlight the absence of artifacts in the synthetic images obtained by the SynLibras in comparison with the EDN results. All images have 256×256 pixels.

[28] L. Ventura, A. Duarte, and X. Giró-i Nieto, "Can everybody sign now? exploring sign language video generation from 2d poses," *arXiv preprint arXiv:2012.10941*, 2020.

[29] B. Saunders, N. C. Camgoz, and R. Bowden, "Anonysign: Novel human appearance synthesis for sign language video anonymisation," in *FG*, 2021.

[30] ——, "Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production," in *CVPR*, 2022.

[31] S. G. M. Almeida, F. G. Guimarães, and J. A. Ramírez, "Feature extraction in brazilian sign language recognition based on phonological structure and using rgb-d sensors," *Expert Systems with Applications*, vol. 41, no. 16, 2014.

[32] C. F. F. Costa, R. S. d. Souza, J. R. d. Santos, B. L. d. Santos, and M. G. F. Costa, "A fully automatic method for recognizing hand configurations of brazilian sign language," *Research on Biomedical Engineering*, vol. 33, pp. 78–89, 2017.

[33] D. B. Dias, R. C. Madeo, T. Rocha, H. H. Biscaro, and S. M. Peres, "Hand movement recognition for brazilian sign language: a study using distance-based neural networks," in *IJCNN*, 2009.

[34] F. M. D. P. Neto, L. F. Cambuim, R. M. Macieira, T. B. Ludermir, C. Zanchettin, and E. N. Barros, "Extreme learning machine for real time recognition of brazilian sign language," in *IEEE ICSMC*, 2015.

[35] J. J. A. M. Junior, M. L. B. Freitas, S. L. Stevan, and S. F. Pichorim, "Recognition of libras static alphabet with myo tm and multi-layer perceptron," in *CBEB*, 2019.

[36] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE TPAMI*, 2019.

[37] A. Duarte, S. Palaskar, L. Ventura, 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 *CVPR*, 2021.

[38] J. Charles, T. Pfister, D. Magee, D. Hogg, and A. Zisserman, "Domain adaptation for upper body pose tracking in signed TV broadcasts," in *BMVC*, 2013.

[39] J. Charles, T. Pfister, M. Everingham, and A. Zisserman, "Automatic and efficient human pose estimation for sign language videos," *IJCV*, 2013.

[40] L. Jing, E. Vahdani, M. Huenerfauth, and Y. Tian, "Recognizing american sign language manual signs from rgb-d videos," *arXiv preprint arXiv:1906.02851*, 2019.

[41] B. Shi, A. M. Del Rio, J. Keane, J. Michaux, D. Brentari, G. Shakhnarovich, and K. Livescu, "American sign language fingerspelling recognition in the wild," in *IEEE SLT Workshop*, 2018.

[42] U. von Agris, "Signum database," 2013, available at https://www.phonetik.uni-muenchen.de/forschung/Bas/SIGNUM/.

[43] O. Özdemir, A. A. Kındıroğlu, N. Cihan Camgoz, and L. Akarun, "BosphorusSign22k Sign Language Recognition Dataset," in *LREC Workshop*, 2020.

[44] Q. Xiao, M. Qin, and Y. Yin, "Skeleton-based chinese sign language recognition and generation for bidirectional communication between deaf and hearing people," *Neural networks*, vol. 125, 2020.

[45] J. Zelinka and J. Kanis, "Neural sign language synthesis: Words are our glosses," in *WACV*, 2020.

[46] A. J. Porfirio, K. L. Wiggers, L. E. Oliveira, and D. Weingaertner, "Libras sign language hand configuration recognition based on 3d meshes," in *IEEE ICSMC*, 2013.

[47] L. R. Cerna, E. E. Cardenas, D. G. Miranda, D. Menotti, and G. Camara-Chavez, "A multimodal LIBRAS-UFOP brazilian sign language dataset of minimal pairs using a microsoft kinect sensor," *Expert Systems with Applications*, vol. 167, pp. 114–179, 2021.

[48] S. G. M. Almeida, T. M. Rezende, G. T. B. Almeida, A. C. R. Toffolo, and F. G. Guimarães, "Minds-libras dataset," Jul. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2667329

[49] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[50] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *NIPS*, 2015.

[51] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *ICML*, 2016.

[52] C. Wan, T. Probst, L. Van Gool, and A. Yao, "Crossing nets: Combining gans and vaes with a shared latent space for hand pose estimation," in *CVPR*, 2017.

[53] R. de Bem, A. Ghosh, A. Boukhayma, T. Ajanthan, N. Siddharth, and P. Torr, "A conditional deep generative model of people in natural images," in *WACV*, 2019.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[55] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.

[56] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[57] Z. Luo, Z. Wang, Y. Huang, L. Wang, T. Tan, and E. Zhou, "Rethinking the heatmap regression for bottom-up human pose estimation," in *CVPR*, 2021.

[58] Universidade Federal de Viçosa (UFV), "Dicionário online libras-português," 2021, available at https://sistemas.cead.ufv.br/capes/dicionario/.

[59] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.