Pose-Guided Sign Language Video GAN with Dynamic Lambda

Christopher Kissel, Christopher Kümmel, Dennis Ritter, Kristian Hildebrand
Beuth University of Applied Sciences Berlin

Abstract

We propose a novel approach for the synthesis of sign language videos using GANs. We extend the previous work of Stoll et al. (2020a) by using the human semantic parser of the Soft-Gated Warping-GAN from (Dong et al., 2018) to produce photorealistic videos guided by region-level spatial layouts. Synthesizing target poses improves performance on independent and contrasting signers. Therefore, we have evaluated our system with the highly heterogeneous MS-ASL dataset (Vaezi Joze and Koller, 2019) with over 200 signers resulting in a SSIM of 0.893. Furthermore, we introduce a periodic weighting approach to the generator that reactivates the training and leads to quantitatively better results.

1 Introduction

The WHO states that 430 million people in the world are deaf or suffer from significant hearing loss. Affected people often use one of the many different sign languages to communicate. In both directions of translation, the first approaches to barrier-free communication between signers and non-signers have emerged in recent years (Bangham et al., 2000; Koller et al., 2015; Camgoz et al., 2018). Our approach refers to the generation of personalized sign videos that can be synthesized in a pose-guided pipeline. Our network architecture is based on the findings of the Text2Sign (Stoll et al., 2020a) paper, which presented the first non-graphical avatar approach to continuous gesture video generation. However, we only consider here the generation of glosses in videos. Our sign language GAN generates a video over the space of all signs from the MS-ASL dataset (Vaezi Joze and Koller, 2019) from a set of input-target pairs. Our goal is to improve the approach by Stoll et al. and enhance the quality of the generated videos to support a variety of signers and different scene conditions. Hence, we apply our approach to the challenging MS-ASL dataset. In this way, we hope to drive the development of real-world sign language applications with personalized generated signers.

When training on highly variable data, such as the MS-ASL dataset, we have observed that our approach first finds a stable equilibrium, but after some time the discriminator loss starts to increase, which means that it is too easily fooled (see Figure 2). It seems, however, that with a dataset as diverse as MS-ASL, not enough uniform criteria can be learned to judge the authenticity of an image when the generator has exceeded a certain threshold. If the discriminator is too weak, the L1 loss of the generator will prevail and the resulting images will be blurred. We have evaluated different strategies to stabilize the training and found that a dynamic lambda, a value that periodically weights the generator error against the discriminator, reactivates the training and leads to an improvement in image quality as shown in Figure 3.

2 Related Work

Sign language related tasks can be divided into three subcategories *Sign Language Recognition (SLR)* (Kasukurthi et al., 2019; Fang et al., 2017; Bheda and Radpour, 2017), *Sign Language Translation (SLT)* (Camgoz et al., 2018, 2020) and *Sign Language Production (SLP)* (Stoll et al., 2020a). Our approach relates to SLP tasks to produce glosses and signs of performed sign language. Sign language datasets mostly consist of annotated videos. Recent approaches like "Word Level American Sign Language" (WLASL) (Li et al., 2020) or "Microsoft - American Sign Language" (MS-ASL) (Vaezi Joze and Koller, 2019) contain annotated high quality videos showing many signers performing various glosses. For our approach we use the MS-ASL dataset as it contains over 25000 videos of more than 200 signers with very different scene conditions, which comes closest to video data in the wild. Other datasets like SMILE (Ebling

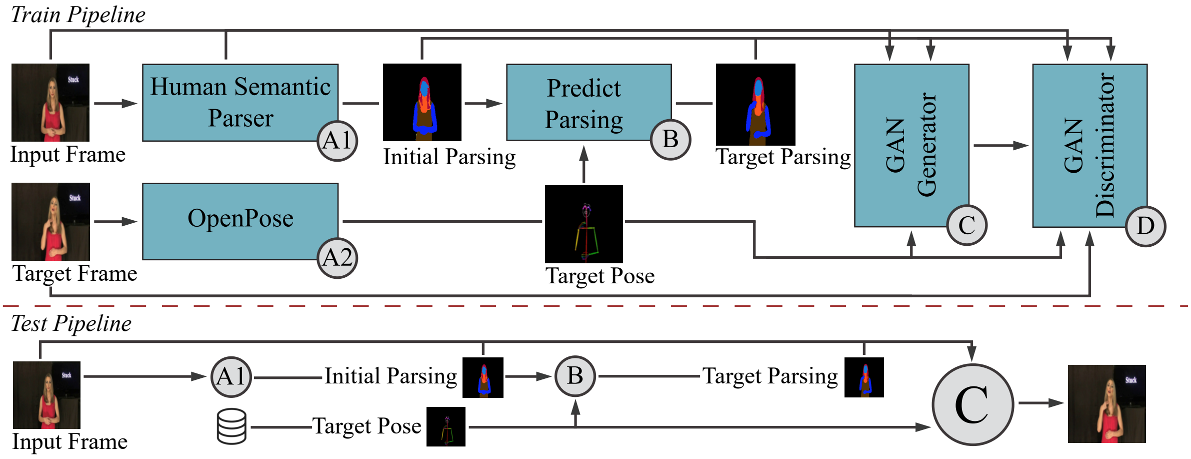


Figure 1: (Top:) During training, an input frame and a target frame are selected from a sign language video to serve as ground truth to train our models. The human semantic parser (A1) creates an initial parsing output of the input frame, while OpenPose (A2) extracts key joints from the target frame. These outputs are concatenated and the target parsing is predicted (B). Finally, input frame, initial and predicted parsing and key joints are used as input for generator (C). It generates an image that is as similar as possible to the target frame input. Whereas the discriminator (D) uses the initial frame, initial parsing, predicted parsing, target pose and the ground truth target frame to check the generated image of C for authenticity. (Bottom:) For testing, the input of C is the same as described above but the target pose is taken from prerecorded key joints. Finally, C creates an image of the person shown in the input frame in the desired pose.

et al., 2018) used by (Stoll et al., 2020a) or (Duarte et al., In Press) even contain multi-sensor data but fewer signers and unfortunately have not been published yet. The use of GANs to tackle SLP problems is still rather new. An important contribution from 2016 is the pix2pix (Isola et al., 2017) framework which uses a U-Net (Ronneberger et al., 2015) generator with skip-connections. It is a key component in our SLP approach as well. Synthesizing sign language videos share some challenges with tasks of Warping-GANs. They aim to transfer a person into previously unseen poses. There have been some remarkable developments in this area in the last years (Dong et al., 2018; Balakrishnan et al., 2018; Ma et al., 2017; Liu et al., 2019). In early 2020, Stoll et al. released a comprehensive approach to translate spoken language into sign language videos. The pipeline called *Text2Sign* (Stoll et al., 2020a) solves the task in two major stages. First the spoken glosses are translated into a sequence of motion graphs by an encoder-decoder architecture (Text2Pose). These poses are then passed to a sign language production GAN (Pose2Video), which is also based on a pix2pix architecture and an extension (Wang et al., 2018) of it. Our work picks up on the results of the *Pose2Video* part and further develops the approach. Later in 2020, more GAN applications for SLP have been released (Saunders et al., 2020; Stoll et al., 2020b; Ventura et al., 2020; Jiang et al., 2020). Which is an indication of the great interest in the issue and show improved results. Contrary to all this work, our approach concentrates on the synthesis of sign language in more real life conditions using the MS-ASL dataset.

3 Method

3.1 Approach & Architecture

Our *SignLanguageGAN* pipeline combines the approaches from Soft-Gated Warping-GAN (Dong et al., 2018) and Text2Sign (Stoll et al., 2020a) and is displayed in Figure 1. At the first stage the human semantic parser (A1), a U-Net (Ronneberger et al., 2015) autoencoder trained on the CIHP dataset, a subset of the LIP dataset (Liang et al., 2019) creates an abstraction of the input frame. It has been shown that a semantic translation can be beneficial for spatial transformation of a person in an image (Dong et al., 2018; Balakrishnan et al., 2018). The human semantic parser provides such a mapping from limb categories to colour values. This initial parsing is now transformed by a pix2pix network (Isola et al., 2017) (B) using the initial parsing and the OpenPose (Cao et al., 2019) generated target pose (A2) representing the desired gloss. Finally, the result is the target parsing, a semantic color map of the person in the input frame who takes the pose represented by the OpenPose skeleton. The second stage forms the actual adversarial model. Starting with a generator (C) that was expanded by a second convolutional encoder network. The latent spaces of both encoders are concatenated and a decoder learns to reproduce the target image from them. The encoder-decoder with skip connections is derived from U-Net architecture (Ronneberger et al., 2015), which was utilized in pix2pix. However, unlike in the original U-Net architecture, down-sampling is performed from two separate networks, while up-sampling is executed by a single one which merges the information of its counterparts. In its setup, the first encoder processes the generated target parsing from the first stage together with the OpenPose generated target pose. Encoder two, on the other hand, receives the input frame together with the respective initial parsing. Both encoders learn their individual mapping into a low dimensional representation. From the concatenation of these two latent spaces the decoder learns how to create a new image of the person seen in the initial frame performing the desired sign language gloss. This output is passed to the SignLanguageGAN discriminator (D) for the adversarial training. The discriminator determines whether the image originates from the generator or if a real sample from the dataset was used. In particular, it is important that the discriminator receives all information for assessment that was processed in the generation.

3.2 Training with a Dynamic Lambda

While our architecture is a composite of several technologies, the training process differs in certain aspects from the frameworks used. We apply a least squares objective function in the discriminator as Mao et al. show that the L2 loss reduces vanishing gradients due to its sensitivity (Mao et al., 2016). This effect is particularly noticeable with highly heterogeneous data such as the MS-ASL (Vaezi Joze and Koller, 2019) collection. Another challenge in working with such dynamic data became apparent as the training has advanced. The equilibrium between discriminator and generator becomes increasingly unstable, presumably because the discriminator has more and more difficulty learning uniform criteria to determine. This results in images with more artifacts and lower variance σ, as the generator contribution becomes stronger, for which we choose a L1 loss in accordance with pix2pix. σ influences contrast and structure in SSIM. To break this trend, we apply a dynamic weighting between generator and discriminator to cyclically amplify the impact of the discriminator. For this we use the lambda parameter, which in pix2pix is a constant value. Several approaches have been evaluated to adjust lambda periodically or adaptively. However, strategies that interfered too much with the balance of the GAN were counterproductive. It have become evident that a steady fluctuation yielded the best results (see Figure 2). Compared to a static weighting the dynamic lambda actively manipulates the loss leading to more updates for the generator. Those updates encourage the generator to increase the output quality in order to fool the discriminator. With the aim of having a high generator contribution at the beginning of the training, we use a cosine period with an amplitude span of 1.5 epochs $\approx \pi$ and a value range of 50 to 150. Our objective function consists of the conditional least squares discriminator objective $V(D, G)$ and the $L1$ loss with the dynamic term where i and j determine the amplitude range.

$$\lambda = i + \frac{cos([0, \pi]) + 1}{2} \cdot j$$

$$\min_{D} \min_{G} L_{SgnlGAN} = V(D, G) + L_1(G) \cdot \lambda$$

By periodically manipulating the loss we observe an overall better performing generator with fewer artifacts and higher variance. This leads to a slight improvement in the SSIM and improved image results as shown in Figure 3(a,b).

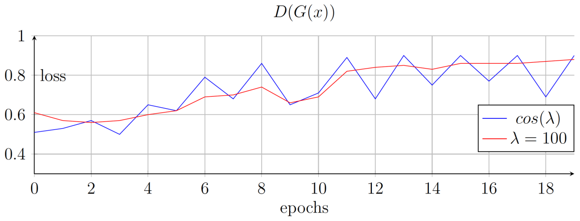
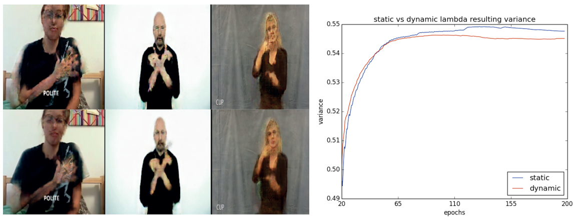


Figure 2: SignLanguageGAN discriminator losses for 20 epochs of training with dynamic lambda (blue) and static lambda (red).

4 Evaluation

4.1 Setup

We evaluate our approach on the MS-ASL (Vaezi Joze and Koller, 2019) dataset, the most diverse sign language dataset in terms of number of signers and scene conditions. We



(a) (b)

Figure 3: (a) (Top:) Static lambda during training. (Bottom:) Dynamic lambda during training. (b) Curve shows increased variance (in blue) when dynamic weighting is applied which leads to higher SSIM.

compare our generated frames with ground truth images and determine the MSE, PSNR and SSIM to assess their quality. Note, unfortunately, we cannot compare our results directly to Text2Sign (Stoll et al., 2020a), since the SMILE dataset (Ebling et al., 2018) is not publicly available.

Architecture	PSNR	MSE	SSIM
pix2pix	20.546	783.376	0.786
Soft-Gated Warping	**22.033**	**622.683**	**0.813**

Table 1: pix2pix and Soft-Gated Warping architectures tested on 50 videos of the MS-ASL dataset after three epochs of training.

4.2 Multi-Signer Generation Challenges

In order to evaluate the Soft-Gated Warping extension in our architecture we have compared it with a network arrangement using pix2pix (Isola et al., 2017) instead. Table 1 shows that the human semantic parser extension leads to improved results for all evaluated metrics using 50 randomly selected videos from the MS-ASL dataset. We also have compared these multi-signer MS-ASL results to results on homogenious *Gebärdenlernen*[1] videos, which show a single-signer in the same scene performing german sign language glosses. In this way, we wanted to assess how the number of signers and the conditions of the scene affect SLP quality. The results show that single signer setups (with PSNR=25.634, MSE=194.174, SSIM=0.882) can easily outperform multi-signer datasets (with PSNR=22.033, MSE=622.683, SSIM=0.813). That underlines our assumption that our pose-guided approach would perform well in real-world scenarios.

Since SignLanguageGAN is the first SLP approach to use the MS-ASL dataset it is difficult to make a meaningful comparison. Nevertheless, we use a subset of the MS-ASL dataset with 184 videos consisting of a total of 16601 frames as a benchmark resulting in a PSNR of 26.044, MSE of 260.363 and SSIM of 0.893 after 300 epochs of training. We compare our results to the results for three individual signers from Text2Sign (Stoll et al., 2020a, p. 901 Table 3), even though our test setup is significantly different from the Text2Sign setup. First, we did not have access to the SMILE dataset, which makes the comparison much more difficult and also prevents us from using the pix2pix HD (Wang et al., 2018) architecture. Second, Text2Sign performs a fine-tuning for the three signers. Since pix2pix HD performs better and SLP for only three signers is not as difficult as for more than 200 we are not surprised that Text2Sign results, with an average PSNR of 25.604, MSE of 166,614 and SSIM of 0.942 are better than ours in terms of pure numbers. Nevertheless, we show that our architecture performs well on extremely diverse data that also contains many low quality samples and without special fine-tuning for individual signers. Figure 4 shows five examples of our results.

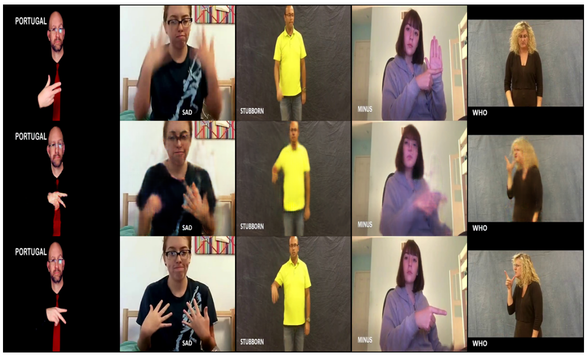


Figure 4: Five examples of our produced sign language glosses. First line shows the input image, second the image produced by our SignlanguageGAN, third the ground truth.

5 Conclusion

In this paper, we present the SignLanguageGAN, a novel approach for photo-realistic synthesis of sign language videos. In addition to previous work on SLP we extended our GAN with a human semantic parser. This allows our system to learn region-level spatial layouts in order to guide the generator to produce more realistic appearing videos on a large

[1] https://gebaerdenlernen.de/index.php?article_id=72

variety of signers. The introduction of a periodic lambda for weighting the discriminator is worth to investigate further since it allows to create images with less artifacts in long training sessions.

References

Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. 2018. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8340–8348.

J. A. Bangham, S. J. Cox, R. Elliott, J. R. W. Glauert, I. Marshall, S. Rankov, and M. Wells. 2000. Virtual signing: Capture, animation, storage and transmission-an overview of the ViSiCAST project. In *IEE Seminar on Speech and Language Processing for Disabled and Elderly People (Ref. No. 2000/025)*, pages 6/1–6/7.

Vivek Bheda and Dianna Radpour. 2017. Using deep convolutional networks for gesture recognition in american sign language. *CoRR*, abs/1710.06836.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10023–10033.

Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. 2018. Soft-gated warping-gan for pose-guided person image synthesis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 472–482, Red Hook, NY, USA. Curran Associates Inc.

Amanda Duarte, S. Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, F. Metze, Jordi Torres, and Xavier Giró i Nieto. In Press. How2sign: A large-scale multimodal dataset for continuous american sign language. In *CVPR 2021*.

Sarah Ebling, Necati Camgoz, Penny Braem, Katja Tissi, Sandra Sidler-Miserez, Stephanie Stoll, Simon Hadfield, Tobias Haug, Richard Bowden, Sandrine Tornay, Marzieh Razavi, and Mathew Magimai-Doss. 2018. Smile swiss german sign language dataset. In *LREC*.

Biyi Fang, Jillian Co, and Mi Zhang. 2017. Deepasl: Enabling ubiquitous and non-intrusive word and sentence-level sign language translation. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*, SenSys '17, New York, NY, USA. Association for Computing Machinery.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Dehao Jiang, Mingqi Li, and Chunling Xu. 2020. Wigan: A wifi based gesture recognition system with gans. *Sensors*, 20(17):4757.

Nikhil Kasukurthi, Brij Rokad, Shiv Bidan, and Aju Dennisan. 2019. American sign language alphabet recognition using deep learning. *CoRR*, abs/1905.05487.

Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1459–1469.

Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. 2019. Look into person: Joint body parsing pose estimation network and A new benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:871–885.

Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. 2019. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5904–5913.

Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. 2017. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, and Zhen Wang. 2016. Multi-class generative adversarial networks with the L2 loss function. *CoRR*, abs/1611.04076.

O. Ronneberger, P.Fischer, and T. Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer. (available on arXiv:1505.04597 [cs.CV]).

Ben Saunders, Richard Bowden, and Necati Cihan Camgöz. 2020. Adversarial training for multi-channel sign language production. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press.

Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. 2020a. Text2sign: Towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, 128(4):891–908.

Stephanie Stoll, Simon Hadfield, and Richard Bowden. 2020b. Signsynth: Data-driven sign language video generation. In *European Conference on Computer Vision*, pages 353–370. Springer.

Hamid Vaezi Joze and Oscar Koller. 2019. Ms-asl: A large-scale data set and benchmark for understanding american sign language. In *The British Machine Vision Conference (BMVC)*.

Lucas Ventura, Amanda Cardoso Duarte, and Xavier Giró-i-Nieto. 2020. Can everybody sign now? exploring sign language video generation from 2d poses. *CoRR*, abs/2012.10941.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807.