Proceedings of the 10th Workshop on the Representation and Processing of Sign Languages (sign-lang@LREC 2022), pages 202–211
Language Resources and Evaluation Conference (LREC 2022), Marseille, 20-25 June 2022

Sign Language Video Anonymization

Zhaoyang Xia[1], Yuxiao Chen[1], Qilong Zhangli[1]
Matt Huenerfauth[2], Carol Neidle[3], Dimitris N. Metaxas[1]
[1] Rutgers University, [2] Rochester Institute of Technology, [3] Boston University
[1] 110 Frelinghuysen Road, Piscataway, NJ 08854,
[2] 152 Lomb Memorial Drive, Rochester, NY 14623
[3] Boston University Linguistics, 621 Commonwealth Ave., Boston, MA 02215
zx149@rutgers.edu, yc984@cs.rutgers.edu, qz181@scarletmail.rutgers.edu
matt.huenerfauth@rit.edu, carol@bu.edu, dnm@cs.rutgers.edu

Abstract

Deaf signers who wish to communicate in their native language frequently share videos on the Web. However, videos cannot preserve privacy—as is often desirable for discussion of sensitive topics—since both hands and face convey critical linguistic information and therefore cannot be obscured without degrading communication. Deaf signers have expressed interest in video anonymization that would preserve linguistic content. However, attempts to develop such technology have thus far shown limited success. We are developing a new method for such anonymization, with input from ASL signers. We modify a motion-based image animation model to generate high-resolution videos with the signer identity changed, but with preservation of linguistically significant motions and facial expressions. An asymmetric encoder-decoder structured image generator is used to generate the high-resolution target frame from the low-resolution source frame based on the optical flow and confidence map. We explicitly guide the model to attain clear generation of hands and face by using bounding boxes to improve the loss computation. FID and KID scores are used for evaluation of the realism of the generated frames. This technology shows great potential for practical applications to benefit deaf signers.

Keywords: Motion Estimation, ASL, Sign Language, Anonymization, Image Animation

1. Introduction

We present here a new method for anonymizing American Sign Language (ASL) videos. Our approach is based on a state-of-art image animation model (Siarohin et al., 2021) to generate a video expressing the linguistic message of the original signer, as articulated by the hands, arms, and face, but retargeted to appear as though the signing is produced by a different person whose image is used as the source for generating the new video. In order to generate high-resolution videos with articulate hand gestures and accurate facial expressions, we enhance the model by using an asymmetric encoder-decoder structured image generator for high resolution image generation and designing a new Hand & Face Focused Loss function for better generation of hand gestures and facial expressions. Our method generates promising results for sign language video anonymization.

2. The Need for Video Anonymization

American Sign Language (ASL) is the natural language that serves as the primary means of communication within the Deaf Community in the United States and parts of Canada. In parallel with manual signing, signed languages use the nonmanual channel—facial expressions and movements of the head and upper body—to express many types of linguistic information, including syntactic marking of, e.g., negation, topics, question status, and clausal type (Baker-Shenk, 1985; Kacorri and Huenerfauth, 2016; Neidle et al., 2000; Coulter, 1979; Valli and Lucas, 2000). Thus, the face, in particular, cannot be obscured to achieve anonymity (e.g., for communicating about sensitive topics, such as medical, legal, or controversial issues) without loss of critical linguistic information.

Although there have been a number of writing systems developed for sign language (Arnold, 2009), there is no standard written form for ASL. Communicating in written English is, in principle, an option for preservation of privacy; however, this is often dispreferred by native signers, who may be less proficient and less comfortable in English than in ASL.

Many Deaf signers have expressed interest in a tool that would preserve linguistic information while allowing the signer's identity to be disguised (Lee et al., 2021). This could be used to enable, for example: anonymous peer review for academic submissions in ASL; neutrality in a range of multimodal ASL-based tools, making possible anonymized definitions in an ASL dictionary; or neutrality in interpreting situations, including messaging. It could also increase participation in video-based AI databases (Bragg et al., 2020), which are quite valuable for research.

3. Previous Approaches to Privacy Preservation

Various approaches to enabling preservation of privacy in ASL videos have been explored. See Isard (2020) for a detailed overview.

3.1. Concealment of Part or All of the Face

There have been attempts to disguise the face in various ways. They all suffer from the same unavoidable problem: that facial expressions convey essential linguistic information, which is degraded or lost by concealment. For example, in Bragg et al. (2020), a tiger-shape filter was used to disguise the signer's face, as shown in Figure 1. However, the absence of facial expression resulted, unsurprisingly, in severely diminished comprehension. Likewise, blocking out certain regions of the face, as in Figure 2, results in loss of critical linguistic information (Bleicken et al., 2016).

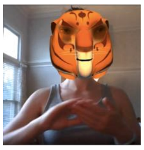

Figure 1: Tiger-shape filter used to protect signer privacy, taken from Figure 2 of Bragg et al. (2020)

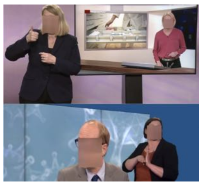

Figure 2: Anonymized video frames from Swiss broadcast footage, from Figure 1 of Camgöz et al. (2021)

3.2. Sign Animation Controlled by Users

Heloir and Nunnari (2016) and Efthimiou et al. (2015) explored providing instructions to enable signers to manipulate virtual humans to generate anonymous messages in sign language. However, these technologies are difficult to master, and it usually takes a long time for non-experts to produce reasonable messages.

3.3. Reproduction of the Original Signing

Several approaches have also been taken to reproduce the original signer's production, to preserve anonymity.

3.3.1. Actors

Use of actors as a way to share information from signed videos when privacy must be preserved has been considered . However, as Isard (2020) points out (§3.2.1):

> For total anonymity, short examples from a corpus can be reproduced by a human actor. In this case complete anonymity is assured, but there are several disadvantages as a result. The process is very labour-intensive, requiring not only the time of the signer but also of a studio and technicians to carry out the recording. In addition, no matter how well the second signer copies the original, some information will be lost. Performativity is a vital part of sign language and it is impossible to fully separate the affective and grammatical functions of facial expressions.

3.3.2. Avatars

In principle, a signer can be replaced by a cartoon-like character replicating what the original person had signed. However, the state-of-the-art in avatar generation (see, e.g., Bragg et al. (2019)) does not make it possible to automate this process; human intervention is required. Furthermore, there are serious technical difficulties such that the use of avatars usually results in dispreferred, unrealistic results (Kipp et al., 2011).

3.3.3. Skeleton-based AI Approaches to Image Generation

As deep learning technology has developed, some researchers have used image-to-image transformation models for sign language anonymization. Recent work such as AnonSign (Saunders et al., 2021) uses a combination of VAEs and GANs to generate sign language frames with different identities. Accurate skeleton keypoints are used as constraints for image generation. A style loss is proposed to generate human appearance of different identities. The results are encouraging.

However, the generation quality and accuracy of the hand gestures and facial expressions largely depend on the skeleton keypoints. In sign language, hand movements are often rapid, causing blurring in the video frame. Occlusions of the face happen frequently. Pretrained human pose estimation models are trained on datasets unrelated to sign language, which have different statistical properties. As a result, they may not transfer well when applied to sign language videos (a problem known as *domain gap*). All these problems make it difficult to obtain accurate keypoint information. State-of-the-art models, such as AlphaPose (Li et al., 2018; Fang et al., 2017; Xiu et al., 2018), can get a rough estimation over bounding boxes. But the accuracy and robustness of handshape estimation remain questionable. See figure 3 for failed cases of AlphaPose on the ASLLRP DSP dataset (see Section 6.1).

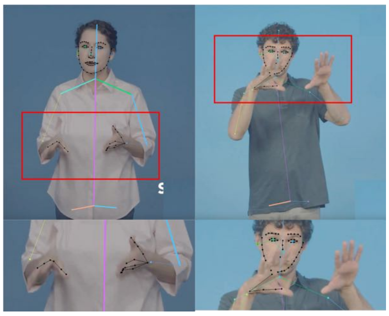

Figure 3: Failed cases of the AlphaPose human pose estimator on ASLLRP DSP dataset.

3.3.4. Facial Expression Transfer using Motion-based Animation Models

Recent work (Lee et al., 2021) applies the facial expression transfer method of Siarohin et al. (2019) for sign language anonymization. The signer's face in videos is replaced by another person's face with the facial expressions transferred. Thus, linguistic meanings are preserved while the identity of the signer in the video changes. Signers provided positive feedback with respect to this application. However, since in Lee et al. (2021) only the face has been transferred, the extent of the anonymization is limited.

3.3.5. Unsupervised Image Animation

Another way to reproduce signing is by using an unsupervised image animation model (Siarohin et al., 2021) to transfer the whole body in sign language videos. We have been the first to explore this approach to sign language anonymization. We present here enhancements to our method that we have introduced to overcome some of the problems we had encountered previously. For example, we are now able to generate high-resolution anonymized videos, with good visual representations of the hands and face, in a computationally efficient and robust way. This method has advantages over other approaches in that (1) it enables anonymization of the whole body (including clothing), without being limited to the face; and (2) it does not require (error-prone) extraction of skeletons from videos.

4. Challenges

Sign language video anonymization is extremely challenging. Accurate hand configurations and movements and detailed facial expressions are essential to preservation of linguistic meaning. Although video animation with retargeting offers many advantages over the other approaches described above, there are several difficult challenges that would need to be overcome.

4.1. Resolution & Computing Cost

The linguistically essential information is concentrated in the face, hands, and arms, although these regions make up only a small portion of the entire video frame. Rapid hand movements can result in blurring if the resolution is not optimal. Thus, high-quality videos are required as input to the model in order to preserve important information. The generated frames also need to be of high resolution for high image quality. However, for the unsupervised animation model, if we directly use the entirety of the high-resolution frames as input, the computation cost is high, and this may not be feasible if the training set contains a large number of videos. Furthermore, generating high-resolution images in one stage is not stable and degrades the image quality.

4.2. Information Density from Hands & Face

In sign language videos, hand movements and facial expressions carry important linguistic meanings. However, they occupy a relatively small part of the total frame as compared to the torso. Therefore, the information density is unbalanced in sign language videos. During training, generative neural network models calculate the difference between the generative frames and the ground truth to compute the loss function, which is important to enable the model to produce better results. Although loss function designs vary, the loss computation usually focuses on the whole image and neglects the small parts. This makes anonymization of sign language videos very challenging because the hands and face, with significantly higher information density than the torso, are neglected by the model.

4.3. 3D Hand Gesture Estimation

Hand gestures in sign language videos are complex movements with a high degree of freedom. For generation of sign language handshapes, 2D hand information is not adequate because of self-occlusion, blurriness from rapid hand movements, and the complex structure of the hands. Therefore, obtaining accurate 3D handshapes would greatly benefit anonymization of sign language videos. However, estimating 3D hand gestures from 2D images or videos is an extremely difficult problem. Pretrained models do not work well on videos in the wild because of the domain gap (explained in Section 3.3.3). The absence of 3D hand information makes hand generation in anonymized sign language videos very challenging.

5. Model Overview & Innovations to Address some of these Challenges

An overview of our methodology for sign language anonymization is presented in Section 5.1. In 5.2.1, we address the problem of computing efficiency for high-resolution image generation by using an asymmetric encoder-decoder structured image generator. In 5.2.2, we introduce a new focused L_1 loss function, which focuses on the the hands and face to improve their appearance in the generated images (given their importance in sign language and the challenges just mentioned with respect to information density). In Section 7, we will demonstrate that these innovations improve the quality of the generated videos.

5.1. Model Overview

Our approach uses 2 inputs. The first is a video sequence of a person signing an utterance (the driving video), while the second is the source image to be used for anonymization; we retarget the movements of the ASL signer in the driving video to a new video sequence based on the source image. The result is an anonymized video of the input utterance. To achieve this goal, we use a novel deep learning methodology that consists of training and inference phases.

5.2. Training Phase

As shown in Figure 4(A), during training, a pair of frames, S_H and D_H, is randomly chosen from the input utterance video sequence, which we term the high-resolution video sequence. To improve the efficiency and the quality of the generated anonymized video, we use a multiresolution approach. In the first stage, S_H and D_H are downsampled to half-size resolution images, S_L and D_L. To obtain an improved motion representation in latent space, we define an intermediate frame R. This conceptual frame is used to improve

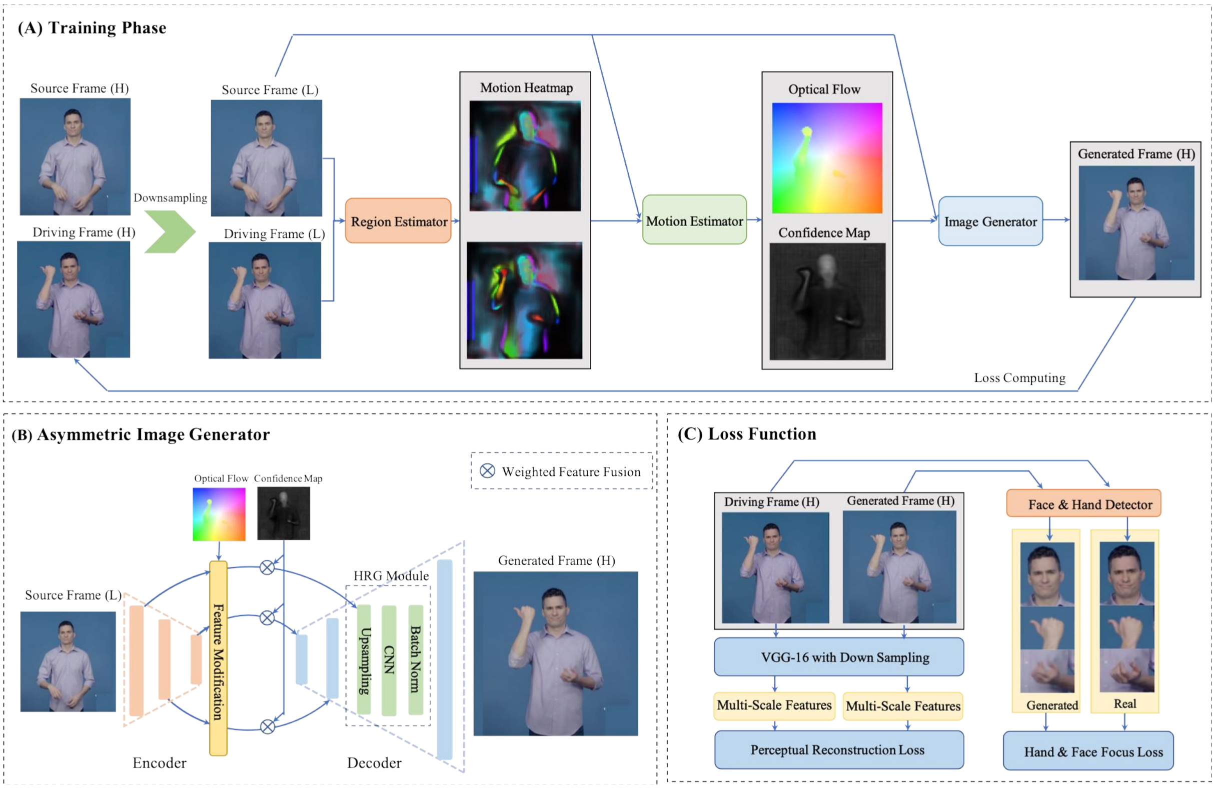

Figure 4: **Method Overview: Phase I - Training.** (A) Model is trained to generate the driving frame from the source frame. First, 2 high-resolution (H) frames are downsampled to low-resolution (L). Then the region motion estimator predicts the motion heatmap and coarse motion representation between these 2 frames. The dense motion estimator estimates the dense optical flow and confidence map from the heatmaps, coarse motion representation, and source image. The image generator outputs a high-resolution generated frame. (B) The image generator is an asymmetric encoder-decoder structured network with a High-Resolution Generation (**HRG**) module. The encoder takes the low-resolution source frame to obtain multiscale latent feature maps. The estimated optical flow is used to modify the feature maps. The confidence maps serve as the weights for latent feature fusions in the skip connections. The decoder along with the **HRG** module generates a high-resolution frame. (C) Multiscale perceptual loss based on VGG-16 and the Hand & Face Focused Loss ($\boldsymbol{L}_{\mathbf{HF}}$), designed for better generation of face and hands, are computed between the high-resolution generated frames and the driving frames.

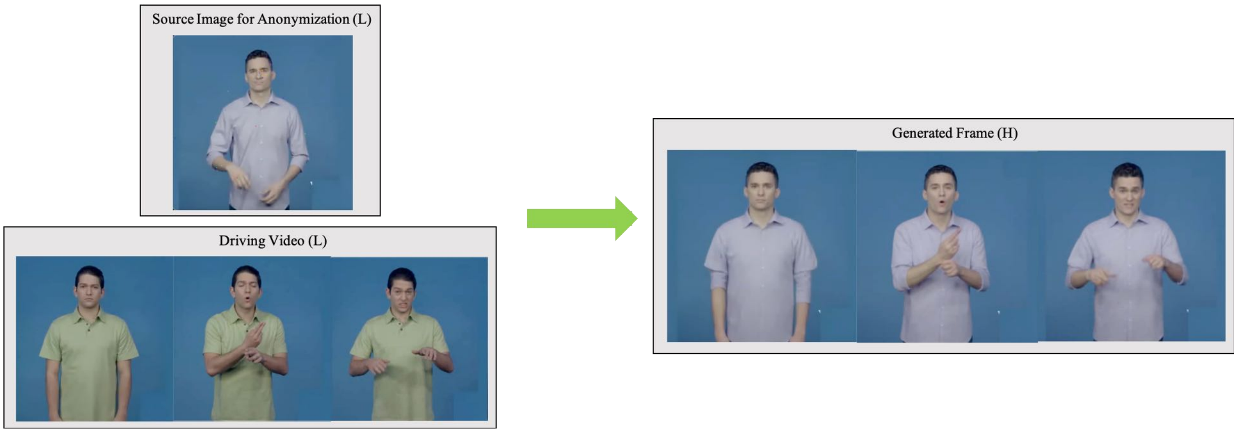

Figure 5: **Method Overview: Phase II - Inference.** In the inference phase, the source image for anonymization and the extracted frames from the driving video are input to the model. All the images are of low resolution. The model predicts the optical flow and confidence map between the source image and each frame in the driving video. The encoder-decoder takes the source image as input and outputs the high-resolution generated frames with the help of the estimated optical flow and confidence map.

the estimation of the foreground motion difference between frames S_L and D_L. The region motion estimator is used to estimate the motion heatmap M_K of k regions between the reference frame R and these 2 frames, S_L and D_L. The affine transformation matrix $A^k_{D_L \leftarrow R}$, $A^k_{S_L \leftarrow R} \in R^{2\times3}$ of the region k between the reference frame R and D_L, S_L is computed using principal component analysis (PCA) (Wall et al., 2003) on the heatmaps M_k.

Thus, the foreground motion between the anonymization source frame S_L and the frame from the driving video D_L is modeled as an affine transformation and can be computed by equation 1.

$$A^k_{S_L \leftarrow D_L} = A^k_{S_L \leftarrow R} \begin{bmatrix} A^k_{D_L \leftarrow R} \\ 0 \quad 0 \quad 1 \end{bmatrix}^{-1} \tag{1}$$

In addition, we predict another affine transformation matrix $A^k_{S_L \leftarrow D_L}$, $k = 0$ for the background motion by taking S_L and D_L into an encoder and predict the six parameters of the affine matrix using a network-based regression operation.

The motion estimator takes the source image, the motion heatmap representation, and the foreground and background affine transformation matrices to predict the dense optical flow $O(z)$ between S_L and D_L. $O(z)$ is considered as a weighted summation of all the affine transformations, given by following formula 2, where z represents the (x, y) coordinates of a pixel, $W^k(z)$ is the weight matrix predicted by motion estimator:

$$O(z) = \sum_{k=0}^{K} W^k(z) A^k_{S_L \leftarrow D_L} \begin{bmatrix} z \\ 1 \end{bmatrix} \tag{2}$$

The network also outputs a confidence map C for each pixel, indicating the pixels that need to be inpainted during the image generation stage.

The last step in the training phase is to use the image generator to reconstruct the high-resolution driving frame D_H based on the source image S_L, the estimated dense optical flow, and the confidence map between S_L and D_H. The loss function is computed between D_H and the generated frame $\hat{D}_H$.

5.2.1. Asymmetric Image Generator with HRG

As already mentioned, in the first stage of our approach, the input frames are downsampled to half-resolution for initial estimation of the motion representation between the 2 selected video sequence frames. However, our goal is to generate a high-resolution image (of the same resolution as the driving video) for the final generated ASL video sequence. Therefore, the information from the high-resolution input video frames is crucial during the image generator phase.

To generate the desired high-resolution video images, we design our image generator as an asymmetric encoder-decoder structured network. Figure 4(B) gives details of our proposed asymmetric image generator. The source frame S_L is input to the encoder, so that the multiscale latent feature maps can be obtained. The estimated optical flow is used to modify each feature map. The multiscale deformed feature maps contribute to the input of each layer in the decoder through skip connections. Therefore, for each layer of the decoder, the input is a weighted summation of the output features from the previous layer and the multiscale deformed feature maps through the skip connections. The weight matrix for this feature fusion approach is decided by the predicted confidence maps.

The High-Resolution Generation (**HRG**) module—which contains an upsampling layer, a convolutional layer, and a batch norm layer—is added before the final output layer in the decoder. This **HRG** module increases the width and height in the latent feature space. Therefore, our decoder does not learn a trivial solution to increase the resolution, such as interpolation. The generated frame $\hat{D}_H$ is of the same resolution as D_H.

The loss function is computed between these 2 high-resolution frames. This asymmetric encoder-decoder structure with the **HRG** module addresses the problem of computation cost and improves the quality of the generated images.

5.2.2. Loss Function

The model is trained to minimize 3 loss functions, including: the multiresolution perceptual loss $\boldsymbol{L_{\mathrm{MP}}}$, the equivalence loss $\boldsymbol{L_{\mathrm{Eq}}}$, and the Hand & Face Focused $\boldsymbol{L_1}$ Loss ($\boldsymbol{L_{\mathrm{HF}}}$); see equation 3. Figure 4(C) demonstrates the computation of $\boldsymbol{L_{\mathrm{MP}}}$ and $\boldsymbol{L_{\mathrm{HF}}}$. In particular, $\boldsymbol{L_{\mathrm{HF}}}$ is designed to explicitly guide the model to generate fine-grained and accurate hand movements and facial expressions:

$$L = \lambda_1 L_{\mathrm{MP}} + \lambda_2 L_{\mathrm{Eq}} + \lambda_3 L_{\mathrm{HF}} \tag{3}$$

where λ_1, λ_2 and λ_3 are the loss function weights.

Multiresolution Perceptual Loss (L_{MP}): This loss function forces the model to reconstruct images with similar high-level features extracted from a pretrained VGG-19 network (Johnson et al., 2016; Wang et al., 2018). The generated frame $\hat{D}_H$ and driving frame D_H are input to a downsampling operator F_l. The differences between the feature maps extracted by the i-th layer of the VGG-19 network are calculated and serve as the reconstruction perceptual loss.

$$L_{\mathrm{MP}}(\hat{D_H}, D_H) = \sum_l \sum_i |V_i(F_l \cdot \hat{D_H}) - V_i(F_l \cdot D_H)| \tag{4}$$

Equivariance Loss (L_{Eq}): This loss function is used to improve the model's robustness and stability for estimating the affine transformation matrix. $\tilde{X}$ is image X transformed by $\tilde{A}$, and $\tilde{A}$ is some random geometric transformation used for data augmentation.

$$L_{\mathrm{Eq}} = |A^k_{X \leftarrow R} - \tilde{A} A^k_{\tilde{X} \leftarrow R}| \tag{5}$$

Hand & Face Focused L_1 **Loss** (L_{HF}): In sign language videos, accurate and clear hand gestures and facial expressions are essential for expression of linguistic meaning. Therefore, the model needs to focus especially on the area around the hands and face, which suggests that the generation quality of these areas needs to contribute more to the loss function. To achieve that, we explicitly guide the model to focus more on the hand and face areas by computing the loss within the hand and face bounding boxes. The bounding boxes are produced using AlphaPose (Li et al., 2018; Fang et al., 2017; Xiu et al., 2018). This loss computation is implemented by constructing weighted masks of both the hands H_r, H_l and the face F based on bounding boxes, calculated in equation 6:

$$L_{\text{HF}} = |(\hat{D_H} - D_H) * (H_l + H_r + F)| \tag{6}$$

The Hand & Face Focused L_1 Loss allows for capture of more details of the hands and face, thereby improving generation of hand gestures and facial expressions.

5.3. Inference Phase

Figure 5 illustrates the Inference Phase. Our model is capable of generating high-resolution videos using the low-resolution source image and driving videos. First, we extract the frames from the driving video. We estimate the bounding box of the human body and make sure the body pose in the source image is roughly aligned with those in the driving video for best results. Then, we input each pair of source image and driving frames to the model and estimate the optical flow and confidence map between them. The encoder takes the source image as input for anonymization and obtains the latent feature map. The latent feature map is then modified using the estimated optical flow and confidence map in the same manner as in the training phase. The decoder outputs the high-resolution generated frames, which preserve the identity of the source image but have the motion in the driving frames.

6. Experiments

6.1. Datasets

We trained our model on the American Sign Language Linguistic Research Project (ASLLRP) Continuous Signing Corpora (Neidle et al., 2018; Neidle and Opoku, 2021) `https://dai.cs.rutgers.edu/dai/s/dai`: (1) the BU SignStream® 3 Corpus, and (2) the DSP dataset, generously contributed by DawnSignPress (DawnSign Press, 2022). We selected 527 videos from each of the 2 ASLLRP datasets. We use 90% of the data for training and 10% for testing. Each video contains a continuous signed sentence. We trained our model on each dataset separately.

6.2. Implementation Details

We trained our model on 8 RTX6000 GPUs with a batch size of 24. The input images are cropped and resized to $(768, 768)$. For both models, we set the training epoch numbers at 100. The region number parameter k is set to 30. The learning rate is set to be $2e^{-4}$ at the beginning, and decreases at the epoch of 60 and 90.

7. Results

Figures 6 and 7 illustrate some of our results from the 2 ASLLRP corpora.

EFFECTS ON REALISM FROM ENHANCED HIGH-RESOLUTION IMAGE GENERATION (**HRG**) AND THE HAND & FACE FOCUSED L_1 LOSS FUNCTION ($\boldsymbol{L}_{\textbf{HF}}$)

HRG	$\boldsymbol{L}_{\textbf{HF}}$	FID ↓	KID ↓
✓	✓	**50.10**	**0.026**
✓		51.30	0.027
	✓	58.95	0.042
		57.47	0.038

HRG	$\boldsymbol{L}_{\textbf{HF}}$	FID ↓	KID ↓
✓	✓	**91.54**	**0.060**
✓		92.33	0.062
	✓	98.56	0.063
		97.98	0.068

Table 1: KID and FID scores on the ASLLRP DSP dataset (top) and the ASLLRP SignStream® 3 Corpus (bottom). The ✓ marks the modifications used with the model. The best results are highlighted in bold.

7.1. Quantitative Evaluation

We used Fréchet Inception Distance (FID) (Parmar et al., 2022; Heusel et al., 2017) and Kernel Inception Distance (KID) (Bińkowski et al., 2018) as metrics to evaluate the quality of the generated images. These metrics measure the discrepancy between the generated and real frames.

For each of the 2 datasets, we randomly sample frames from test videos to construct the test image dataset. For each model, we select driving video and source image pairs to generate anonymized videos. Then, we randomly sample frames from these anonymized videos and construct the generated image set. Finally, we calculated the FID and KID scores between the test image set and generated image sets. The table 1 shows the effects of 2 aspects of our model on the realism of the results. Lower FID and KID scores indicate that the generated image is more similar to the real image set, and thus is considered to be more realistic. Our modifications improve the realism of the generated images. In particular, with the **HRG** module, our method is able to generate high-resolution images and achieve more realistic results. The improvements from using the Hand & Face Focused L_1 Loss are not reflected in the KID and FID scores. This is reasonable because this modification is intended to improve the quality of small parts in the image, which the KID and FID scores neglect.

FID and KID scores measure the overall generation quality of the images. They do not reflect the image details and cannot assess the preservation of linguistic meaning. In order to show that our method improves the generation of facial expressions and hand gestures, we compare the generated images in the next section.

7.2. Qualitative Evaluation

We focus on the quality of facial expressions and hand gestures in the generated frame. First, we compare hand gesture generation with and without the

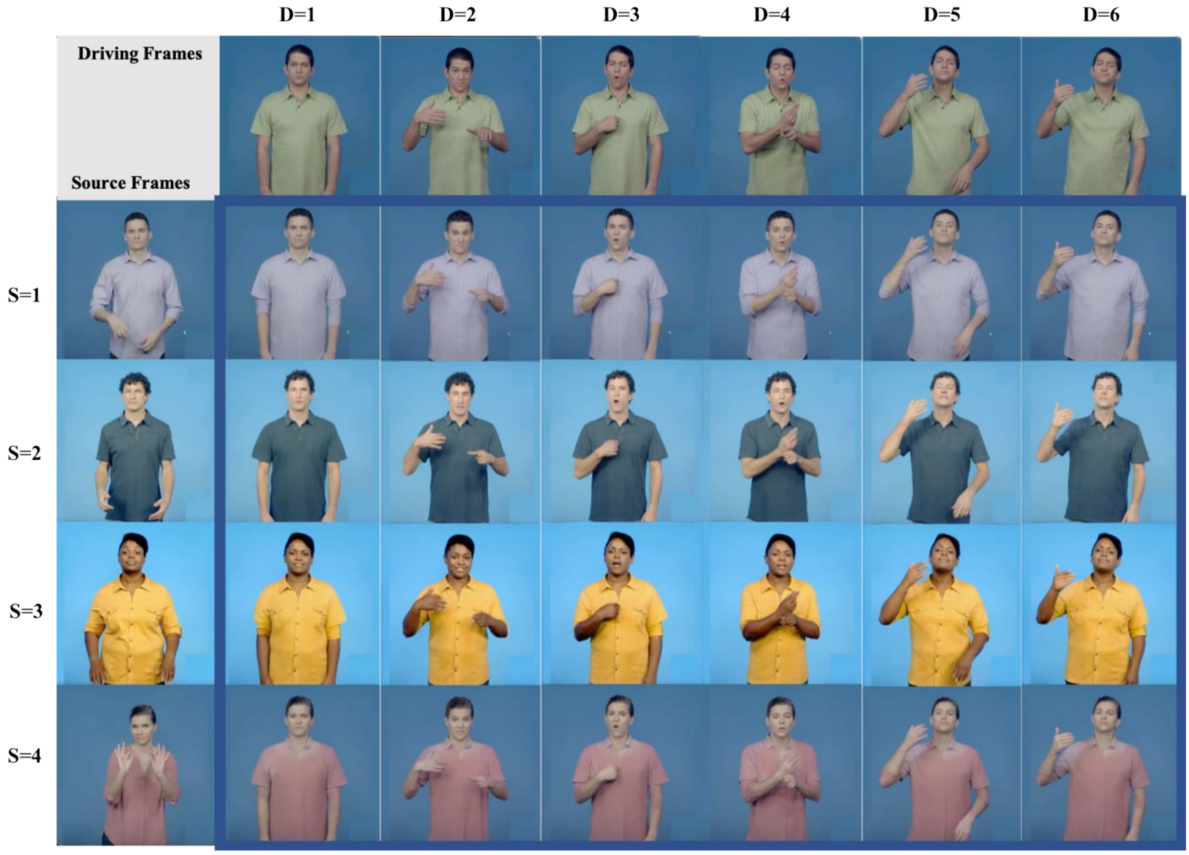

Figure 6: **Anonymization examples from the ASLLRP DSP dataset**. Our method takes the driving frames from a sign language video sequence as the motion reference and generates a new sign language video with the human appearance and body pose taken from the designated source frame. In this example, 6 driving frames $D = 1...6$ are selected from a video sequence. We use four source frames $S = 1, 2, 3, 4$ to provide the human appearance and body pose for the generated anonymized video sequence. So, Row 1 shows frames from the original signer; Column 1 shows four different source images providing the appearance to be used for the anonymized versions driven by the signing from Row 1; those generated images are shown in the rest of the frames in each row.

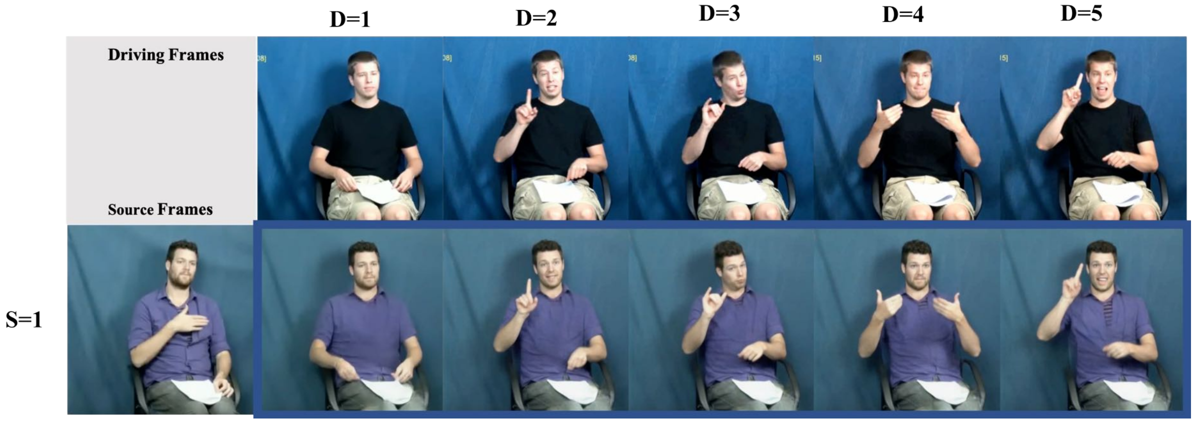

Figure 7: **Anonymization examples from the ASLLRP SignStream® 3 Corpus**. Display is similar to Figure 6

HRG modification and Hand & Face Focused L_1 Loss ($\boldsymbol{L}_{\mathbf{HF}}$) function. As seen in the sample results shown in Figure 8, the **HRG** modification improves the quality and clarity of the images. The $\boldsymbol{L}_{\mathbf{HF}}$ Loss adds more details to the hand configuration and further improves the hand appearance in the generated videos.

In Figure 9, our full model gives the clearest face generation. The **HRG** modification increases accuracy of details around the eyes and mouth in the generated videos. In particular, the improved model is able to generate wrinkles and teeth. Moreover, the generated faces have higher resolution than the low-resolution input videos. The $\boldsymbol{L}_{\mathbf{HF}}$ Loss helps with preservation of facial expressions and alleviates possible facial distortions.

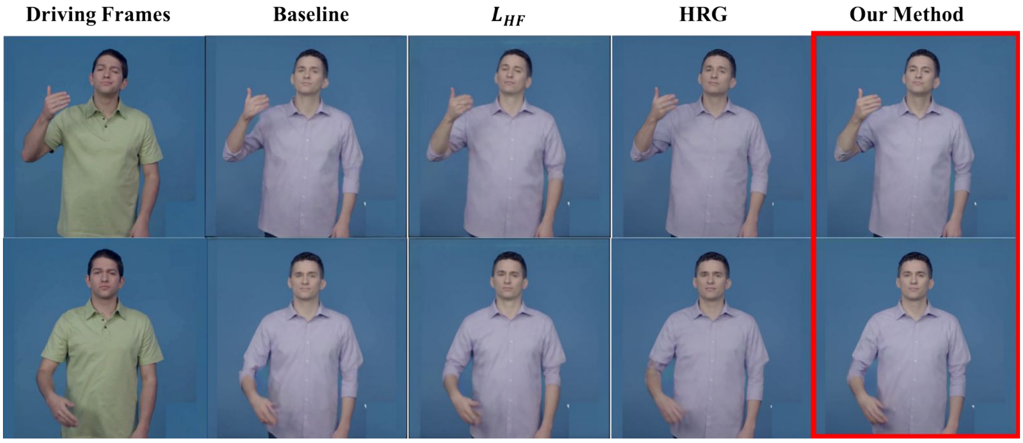

Figure 8: Comparison of Hand Gesture Generation for different models. In particular, the signer's handshape is generated best in the images shown in the red box, where the model incorporates both **HRG** and Hand & Face Focused L_1 Loss ($\boldsymbol{L_{\mathrm{HF}}}$).

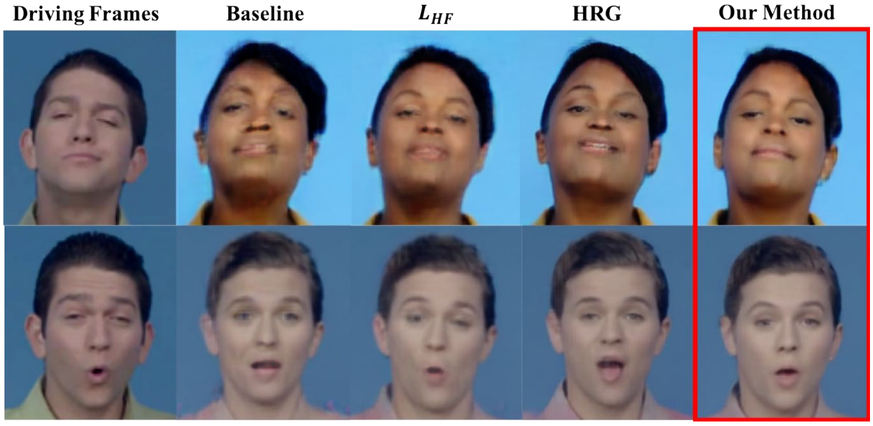

Figure 9: Comparison of Facial Expression Generation for different models. The best results are those in the red box, where the model incorporates both **HRG** and Hand & Face Focused L_1 Loss ($\boldsymbol{L_{\mathrm{HF}}}$).

8. Conclusions & Directions for Future Research

We have been developing methods for anonymizing ASL videos with input from Deaf signers. For example, Lee et al. (2021) reports on user studies with Deaf signers who evaluated our earlier experiments using facial expression transfer for purposes of video anonymization. They evaluated the extent to which the anonymized videos looked natural, succeeded in transmitting the linguistic information, and disguised the identity of the original signer. They also commented on the extent to which they would find it useful to be able to anonymize videos in this way for various purposes. Although the overall feedback was quite positive, the fact that only the face was anonymized presented a serious drawback to that technology. We have thus been developing new methods, as presented here, to enable full-body anonymization. Here we described several innovations that we developed in order to overcome challenges involved in image animation with retargeting to anonymize sign language videos. Preliminary user interviews indicate that our new method is extremely promising. With respect to the success in disguising the identity of the signer, one Deaf signer commented: "Unrecognizable – amazing work! I could not recognize the original signer, yet it kept the signing style. Impressive!" However, there are still some cases where the handshape or facial expression is not generated perfectly in certain frames, because of issues just discussed, including the fact that we are not doing any explicit 3D modeling. We will continue to refine our methods to improve these remaining glitches, after which we will conduct another set of comprehensive user studies with Deaf signers for quantitative evaluation of the degree to which the identity has been successfully disguised and of the comprehensibility and naturalness of the resulting anonymized signing.

9. Acknowledgments

We are grateful to Augustine Opoku, Carey Ballard, Indya-loreal Oliver, Lutece Dekker, Gregory Dimitriadis & Douglas Motto for important contributions. This research has been supported in part by NSF grants 2040638, 1763486, 1763523, 2125362 & 1763569. Any opinions, findings, or conclusions expressed here are those of the authors and do not necessarily reflect the views of the National Science Foundation.

10. References

Arnold, R. W. (2009). *A proposal for a written system of American Sign Language*. Gallaudet University.

Baker-Shenk, C. (1985). The facial behavior of deaf signers: Evidence of a complex language. *American Annals of the Deaf*, 130(4):297–304.

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. (2018). Demystifying MMD GANs. *arXiv preprint arXiv:1801.01401.*

Bleicken, J., Hanke, T., Salden, U., and Wagner, S. (2016). Using a language technology infrastructure for German in order to anonymize German Sign Language corpus data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3303–3306, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., et al. (2019). Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st international ACM SIGACCESS conference on computers and accessibility*, pages 16–31.

Bragg, D., Koller, O., Caselli, N., and Thies, W. (2020). Exploring collection of sign language datasets: Privacy, participation, and model performance. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–14.

Camgöz, N. C., Saunders, B., Rochette, G., Giovanelli, M., Inches, G., Nachtrab-Ribback, R., and Bowden, R. (2021). Content4all open research sign language translation datasets. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–5. IEEE.

Coulter, G. R. (1979). *American Sign Language typology*. Ph.D. thesis, University of California, San Diego.

DawnSign Press. (2022). DawnSign-Press (2022) About Us. *DawnSignPress Website*. `http://www.dawnsign.com/about-us`.

Efthimiou, E., Fotinea, S.-E., Goulas, T., and Kakoulidis, P. (2015). User friendly interfaces for sign retrieval and sign synthesis. In *International Conference on Universal Access in Human-Computer Interaction*, pages 351–361. Springer.

Fang, H.-S., Xie, S., Tai, Y.-W., and Lu, C. (2017). RMPE: Regional multi-person pose estimation. In *ICCV*.

Heloir, A. and Nunnari, F. (2016). Toward an intuitive sign language animation authoring system for the deaf. *Universal Access in the Information Society*, 15(4):513–523.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Isard, A. (2020). Approaches to the anonymisation of sign language corpora. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 95–100.

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer.

Kacorri, H. and Huenerfauth, M. (2016). Continuous profile models in ASL syntactic facial expression synthesis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2084–2093.

Kipp, M., Nguyen, Q., Heloir, A., and Matthes, S. (2011). Assessing the deaf user perspective on sign language avatars. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, pages 107–114.

Lee, S., Glasser, A., Dingman, B., Xia, Z., Metaxas, D., Neidle, C., and Huenerfauth, M., (2021). *American Sign Language Video Anonymization to Support Online Participation of Deaf and Hard of Hearing Users*. Association for Computing Machinery, New York, NY, USA.

Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.-S., and Lu, C. (2018). Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *arXiv preprint arXiv:1812.00324.*

Neidle, C. and Opoku, A. (2021). Update on Linguistically Annotated ASL Video Data Available through the American Sign Language Linguistic Research Project (ASLLRP). In *ASLLRP Project Report No. 19*, Boston University, Boston, MA.

Neidle, C. J., Kegl, J., Bahan, B., MacLaughlin, D., and Lee, R. G. (2000). *The syntax of American Sign Language: Functional categories and hierarchical structure*. MIT press.

Neidle, C., Opoku, A., Dimitriadis, G., and Metaxas, D. (2018). New shared & interconnected ASL resources: SignStream® 3 software; DAI 2 for web access to linguistically annotated video corpora; and a Sign Bank. In *8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community, Miyazaki, Language Resources and Evaluation Conference 2018.*

Parmar, G., Zhang, R., and Zhu, J.-Y. (2022). On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*.

Saunders, B., Camgoz, N. C., and Bowden, R. (2021). Anonysign: Novel human appearance synthesis for sign language video anonymisation. *arXiv preprint arXiv:2107.10685.*

Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., and Sebe, N. (2019). First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32:7137–7147.

Siarohin, A., Woodford, O. J., Ren, J., Chai, M., and Tulyakov, S. (2021). Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662.

Valli, C. and Lucas, C. (2000). *Linguistics of American Sign Language: An introduction*. Gallaudet University Press.

Wall, M. E., Rechtsteiner, A., and Rocha, L. M. (2003). Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pages 91–109. Springer.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. (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.

Xiu, Y., Li, J., Wang, H., Fang, Y., and Lu, C. (2018). Pose Flow: Efficient online pose tracking. In *BMVC*.