Neural Sign Reenactor: Deep Photorealistic Sign Language Retargeting

Christina O. Tze [1] Panagiotis P. Filntisis [1] Athanasia – Lida Dimou[4]
Anastasios Roussos [2,3] Petros Maragos[1]

[1]School of Electrical & Computer Engineering, National Technical University of Athens, Greece
[2]Institute of Computer Science (ICS), Foundation for Research & Technology - Hellas (FORTH), Greece
[3]College of Engineering, Mathematics and Physical Sciences, University of Exeter, UK
[4]Institute for Language and Speech Processing, Athena R.C., Greece

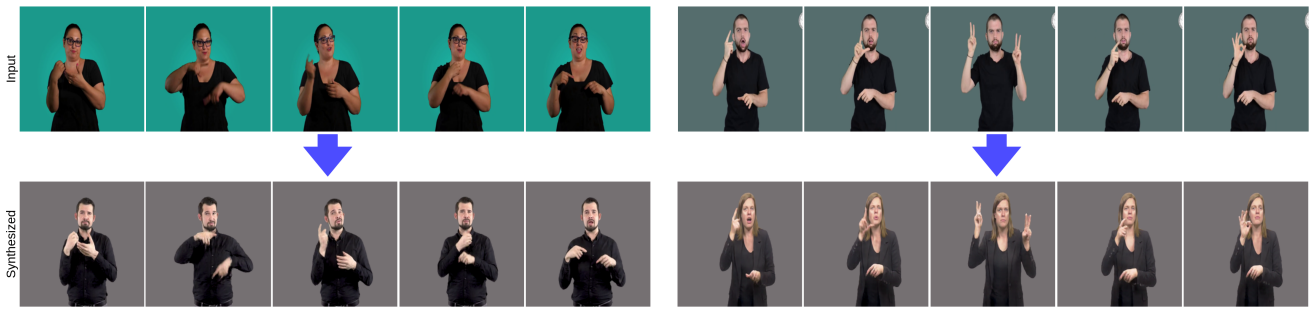

Figure 1: Given an input sign language video, our *Neural Sign Reenactor* synthesizes a photo-realistic and temporally coherent video of a target signer imitating the source signer's upper body movements and facial expressions. Please zoom in for details and refer to Supplementary Video [1].

Abstract

*In this paper, we introduce a neural rendering pipeline for transferring the facial expressions, head pose and body movements of one person in a source video to another in a target video. We apply our method to the challenging case of Sign Language videos: given a source video of a sign language user, we can faithfully transfer the performed manual (*e.g. *handshape, palm orientation, movement, location) and non-manual (*e.g. *eye gaze, facial expressions, head movements) signs to a target video in a photo-realistic manner. To effectively capture the aforementioned cues, which are crucial for sign language communication, we build upon an effective combination of the most robust and reliable deep learning methods for body, hand and face tracking that have been introduced lately. Using a 3D-aware representation, the estimated motions of the body parts are combined and retargeted to the target signer. They are then given as conditional input to our Video Rendering Network, which generates temporally consistent and photo-realistic videos. We conduct detailed qualitative and quantitative evaluations and comparisons, which demonstrate the effectiveness of our approach and its advantages over existing approaches. Our method yields promising results of unprecedented realism and can be used for Sign Language Anonymization. In addition, it can be readily applicable to reenactment of other types of full body activities (dancing, acting performance, exercising, etc.), as well as to the synthesis module of Sign Language Production systems.*

1. Introduction

Tens of millions of Deaf worldwide use Sign Language (SL) as their native language [55, 8, 17, 5]. At the same time, most of them have limited reading and writing skills in the spoken language, which for them is a foreign language with a fundamentally different grammatical structure. Because of that, the Deaf are still disadvantaged in many contexts of their daily life, such as social relations, education, work, usage of computers and the Internet. SL technologies can be a valuable ally of the Deaf community in their struggle to overcome these barriers, by building systems that facilitate their communication with the rest population [34]. This research area has witnessed many advances during the last three decades and, during the last years, the introduction of deep learning has resulted to especially ro-

bust and promising methods for the problems of Sign Language Recognition (SLR) [37, 46, 25, 10, 60, 24], Translation (SLT) [9, 57, 51] and Production (SLP) [45, 40, 38].

One of the most challenging open problems of SL technologies is the generation of synthetic SL videos that allow SL users to experience a natural and fluid communication, similar to human-to-human SL communication. Most existing SL synthesis techniques are based on animation of a computer-generated 3D avatar, followed by traditional 3D graphics rendering. However, when the computational efficiency constraints of real-world systems are taken into account, this typically results to a low level of realism, as far as the appearance and motion of the avatars are concerned. As in the case of immature speech synthesis technologies (with *e.g*. robot-like synthesized voices), this creates important problems from the side of the users in terms of the plausibility and engagement with such technologies.

A particularly promising alternative is offered by a few recent works that build upon the latest advances of photo-realistic neural rendering and synthesize SL videos with avatars that have the appearance of real persons. This type of frameworks have been incorporated in systems of SL Production that generate photo-realistic SL videos as the final output of spoken-to-sign language translation [39, 45, 38, 42]. Even though these methods open new pathways towards photo-realistic SL synthesis, their generated frames include artifacts and the synthesized human appearances are not always convincing as being real, having the risk of "falling" into the so-called "*uncanny valley*" [33].

This work overcomes the aforementioned limitations and synthesizes videos of high level of realism that include body, hands, head and face motions of a virtual signer who is almost indistinguishable from a real person. We are based on a novel 3D-based motion representation and conditioning of a neural renderer. Our contributions can be summarized as follows:

• Our method achieves results of unprecedented realism in the particularly challenging task of human motion retargeting in SL videos.

• We build upon an effective combination of two different body trackers for implementing high-fidelity body and face tracking.

• We propose a novel color-coding scheme for the conditioning and training of our neural renderer.

• We conduct detailed evaluations and comparisons of our method with other approaches to human motion retargeting that demonstrate the particularly promising and realistic results that we obtain under challenging continuous signing and across different genders and body structures.

2. Related Works

2.1. Motion Retargeting

Human motion retargeting is an emerging topic at the intersection of computer vision and graphics due to its extensive potential for content creation. Its goal is to transfer the motion of a source person in a driving video to a target person in a reference video.

Early approaches concentrated on the task of head reenactment, as opposed to the body reenactment that we are interested in. Kim *et al*. [22] presented Deep Video Portraits (DVP), a system that was capable of fully transferring the head motions and rotations, facial expressions and eye gaze of the source actor to a detailed portrait video of the target actor. However, their image-based model ignored the temporal dependencies between the synthesized frames. To overcome this limitation, Doukas *et al*. [15] proposed a video-based neural rendering network coupled with a multi-scale dynamics discriminator. Our work is based on [15], except we retarget full body motion. Over the last years, a plethora of deep learning-based methods have been introduced performing body reenactment. Some of them require high-fidelity 3D pose estimation or reconstruction [28, 50, 29, 27]. For each target subject, Liu *et al*. [28] used a 3D character model which was reconstructed from static posture images. Villegas *et al*. [50] employed the 3D pose estimator of [32] to estimate the 3D poses from human videos and transfer the motion to a virtual character in an unsupervised manner. Lim *et al*. [27] presented a 3D-based two-branch framework for unsupervised motion retargeting which learns frame-by-frame poses and overall movement separately. In [29], the authors proposed Liquid Warping GAN, a unified framework capable of handling human motion imitation, appearance transfer and novel view synthesis. Retargeting motion from 2D inputs has also been studied in several works [12, 3, 4, 56, 61]. Chan *et al*. [12] introduced a simple yet effective method that uses intermediate 2D skeleton representations and generates temporally coherent video results. The method proposed by Wang *et al*. [53] produced results of comparable quality to those of [12], but uses a more intricate shape representation ([11],[19]) and requires too much memory and computing power. Moreover, they do not address human body variations between the source and target subjects. Aberman *et al*. [3] used a two-branch framework to handle video-driven performance cloning, with one branch learning the target subject's appearance and the other one enhancing the temporal coherence. In a follow-up work, Aberman *et al*. [4] trained a deep neural network to decompose 2D pose input sequences into dynamic (motion) and static (skeleton and camera view-angle) components, which can then be recombined to generate new motions. Zhu *et al*. [61] presented Canonicalization Networks to address the chal-

lenging 2D-to-3D motion retargeting problem. Trained with two novel canonicalization operations, namely structure and view canonicalization, their method decomposes 2D skeleton sequences into three independent subspaces (i.e. motion, structure and view angle), similarly to [4]. Yang *et al*. [56] performed 2D motion retargeting by combining the extracted motion from the source sequence with the extracted structure from the target sequence.

2.2. Sign Language Production

Sign Language Production (SLP) is defined as the automatic translation from spoken language sentences to the corresponding sign language video. Research on the sign language field was initially focused on the challenging tasks of SLR and SLT. This was due to the misconception that deaf people are familiar with reading spoken languages. To bridge the communication gap between the hearing and the Deaf, novel deep learning methods have been introduced over the last years that yielded highly robust and promising results on the SLP task.

Prior to the deep learning era, the SLP problem was historically tackled using animated avatars. Parametrized glosses and motion capture (mocap) data were two methods for generating the sign avatars [36]. In the first case, spoken language is translated into sign glosses and afterwards sign language is produced by mapping each gloss to a parametric representation needed to animate the avatar. Such works (*e.g*. VisiCast [6], Tessa [14], eSign [62] and dictasign [16]) use the HamNoSys annotation system [35, 20] and the SiGML language [21] for encoding the sign language gestures. However, they fall into the "*uncanny valley*" [33], which has a negative effect on the audience's appeal and engagement. On the contrary, using motion capture data increases the realism in avatar animation. However, such approaches (*e.g*. Sign3D Project by MocapLab [18]) are limited to a small number of pre-recorded phrases due to the prohibitively high cost of producing mocap data.

Initial deep learning-based SLP methods concatenated isolated signs disregarding the natural co-articulation between them [44, 45, 58]. Stoll *et al*. [44] presented the first end-to-end spoken language to sign language video translation system based on a combination of Neural Machine Translation (NMT) and Generative Adversarial Networks (GANs). The proposed pipeline consists of three stages: 1) text-to-glosses translation (T2G), 2) glosses to skeletal sequences mapping and 3) pose-conditioned sign language video generation. In [45], the NMT network was combined with a motion graph to generate the human pose sequence which was then fed into the generative network frame by frame. Instead of using glosses, Zelinka *et al*. [58] suggested generating skeletal poses from words. Each input word was translated into a single 7-frame sign, producing sequences of fixed length and ordering. However, they focused solely on the manual features. More recently, Walsh *et al*. [52] introduced Text to HamNoSys (T2H) translation and demonstrated the advantages of phonetic representations over gloss representations for sign language translation. In addition, other works (*e.g*. [58, 40, 38]) used skeleton pose representations rather than photo-realistic videos, which has been shown to reduce deaf comprehension [49]. Saunders *et al*. [40] presented a novel transformer-based architecture that can translate from spoken language sentences to continuous 3D sign pose sequences, using a counter decoding to track the generation progress. Similarly to [58], their method ignored the non-manual features, which are fundamental components of all sign languages. In [38] the authors expanded production to include head motion and mouthing patterns. Over the last years, continuous SLP methods have been proposed [39, 42, 41]. SignGAN [39] was the first SLP model to produce photo-realistic continuous sign language videos directly from spoken language input. Recently, the same authors presented FS-NET [42], a novel frame selection network that models sign co-articulation by learning the optimal temporal alignment between interpolated dictionary signs and continuous signing sequences. The closest work to this paper is that of Saunders *et al*. [41], who presented a deep learning framework for the generation of photo-realistic retargeted videos, using novel synthesized human appearances instead of the original signer appearance. However, the generated frames include artifacts and the novel human appearances are not always convincing as being real.

3. Methodology

Our method enables the transfer of human body movements from a source actor to a target subject with realistic face synthesis. Formally, given an input SL video $\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_T)$, our method generates a new video $\tilde{\mathbf{Y}} = (\tilde{\mathbf{y}}_1, ..., \tilde{\mathbf{y}}_T)$, where the signing of the original signer is transferred to the target subject. An overview of the proposed pipeline is presented in Fig. 2. It consists of four main components: a) upper body detection, b) pose retargeting, c) color-coded conditioning and d) photo-realistic synthesis presented in the following sections.

3.1. Upper Body Detection

Detecting an actor's body in a video is the first step in our approach. Since only upper body videos are included in the collected dataset, we are focused on the detection of the head, hands and torso (hereafter referred to as the human body part apart from the head, neck, hands and legs). In terms of the more general problem of robust human pose estimation, recent advances in the field have made it possible even in the case of simple RGB input [13, 59]. OpenPose [11] is the first and one of the most popular bottom-up approach for multi-person human pose estimation. Medi-

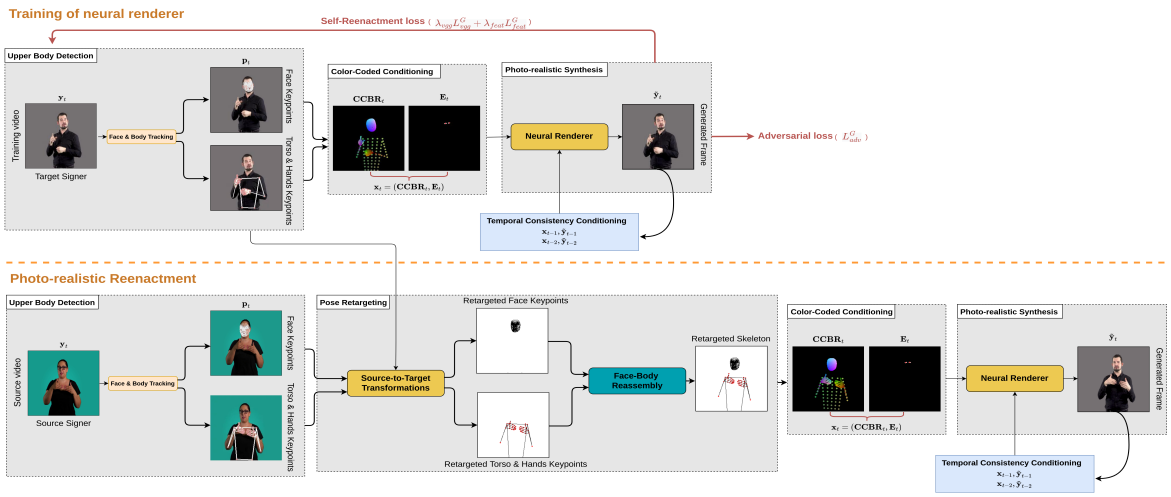

Figure 2: (Top) **Training**: First, for each target signer we perform upper body detection to obtain the 3D face, torso and hands keypoints. These are then used to create the color-coded body representations and the corresponding eye gaze images, which are concatenated and fed into a neural renderer (along with previously generated frames) as conditional input. The output frames are produced sequentially, one after the other, and ideally should be a reconstruction of the ground truth ones from the target's training video. (Bottom) **Reenactment**: To transfer the upper body motions and facial expressions of a source signer to a target signer, we begin by extracting the former's 3D upper body keypoints. Then, through our pose retargeting step, these are transformed to adapt to the target's body shape and location within each frame. Finally, the neural renderer generates the frames of the target signer, using the previously created color-coded semantic representations to drive synthesis.

aPipe [30] is an open-source, cross-platform framework for developing machine learning pipelines for multimodal data processing. Among others, it can be used to implement human face detection (478 landmarks in total), hand tracking (21 landmarks in total) and high-fidelity pose tracking (33 landmarks in total). However, there is currently no single MediaPipe module available that tracks the face, pose and hand landmarks while also being fully trained to predict their depth.

We first extract the skeleton pose sequences from the sign language videos using both MediaPipe (MP) [30] Pose and Holistic modules, because the Holistic model is not fully trained to predict the depth of the pose landmarks. Since upper body videos are used in this work, we only use 9 of the 33 3D landmarks that MP Pose infers, excluding those that correspond to the head (except the nose), hands, and lower body. We use MP Holistic to track the head and hands, inferring 520 landmarks in total. Thus, every frame $i \in [1, T]$ of an input video is represented by a pose vector $\mathbf{p}_i = (\mathbf{l}_1, ..., \mathbf{l}_K)$ of 3D landmarks coordinates $\mathbf{l}_j = (l_{j_x}, l_{j_y}, l_{j_z})$, where $K = 529$ is the number of tracked joints. After processing all T frames, the sequence of poses $\mathbf{P} = (\mathbf{p}_1, ..., \mathbf{p}_T)$ is extracted. For clarity, the joint in the middle of the shoulders shall henceforth be referred to as the root joint.

We crop the frames of a **target** actor's video using the minimum and maximum values of the x and y joint coordinates over the entire sequence $\mathbf{P}$. The cropped frames are then resized to **256×256 pixels**. We crop the **source** videos subject to the target person to whom we reenact the performed signs. We found that such cropping boosted our reenactment performance. More specifically, for each frame (cropped and resized) of the target actor's training video, we calculate the horizontal distance between the left and right shoulders and then take the median of these distances, h_m. Next, we determine the percentage $x\%$ of $dx = 255$ that h_m represents. Finally, we find the width, $dx' = x'_{max} - x'_{min}$, of a fixed bounding box, such that the source actor's median horizontal distance (calculated at the original frames) corresponds to the same percentage $x\%$ of dx'. Following the same procedure, we compute the vertical distances between the joints in the center of the shoulders and hips in order to determine the bounding box's height, $dy' = y'_{max} - y'_{min}$. The bounding box is cropped from the source actor's frames and the cropped images are then resized to **256×256 pixels**.

3.2. Pose Retargeting

When retargeting motion from one character (source) to another one (target), we must take into consideration possible differences between their body shapes (such as limb lengths and proportions) as well as their placement within the frames. Therefore, for every pair of source and target actors it is necessary to adapt the motion of the former in order to preserve the skeletal structure and location of the latter. To accomplish this, we suggest the following procedure, which is based on **Procrustes Analysis** and transfor-

mations of the source's landmarks. The proposed process is applied seperately for two parts of the upper body: i) head and ii) torso and hands.

For the head, similarly to [7], we use a subset of n face landmarks from the most rigid area of the face that are less affected by the facial deformations during facial expressions and mouth motions. First, at each frame, $i \in [1, K]$ and $j \in [1, N]$, of the source and training videos respectively, we consider the matrices $\mathbf{S}_r^i$, $\mathbf{T}_r^j \in \mathbb{R}^{n\times 3}$ with the 3D coordinates of the **rigid** face landmarks. Using Procrustes Analysis (and specifically the method of Umeyama [47]), we align the matrices $\mathbf{S}_r^i$ and $\mathbf{T}_r^j$ with the matrix $\mathbf{M}_r$ of a mean face template. Let *procrustes(***X**,**Y***)* denote the Procrustes transformation (rotation, translation, isotropic scaling) that transforms matrix **Y** to **X**. Moreover, let **Z** be the matrix resulting from performing the transformation on **Y**. Thus, at each time step, the following transformations are returned:

$$\mathbf{R}_s^i, \mathbf{T}_s^i, s_s^i = procrustes(\mathbf{M}_r, \mathbf{S}_r^i) \quad (1)$$

$$\mathbf{R}_t^j, \mathbf{T}_t^j, s_t^j = procrustes(\mathbf{M}_r, \mathbf{T}_r^j) \quad (2)$$

After processing all frames, we apply **geometric median** [48] to the aligned matrices $\{\mathbf{Z}_r^i\}_{i=1}^K$ and $\{\mathbf{Z}_r^j\}_{j=1}^N$ and extract the median source and target faces, $\mathbf{S}_m, \mathbf{T}_m \in \mathbb{R}^{n\times 3}$ respectively. Next, we identify the non-uniform scaling parameters (s_x, s_y, s_z) that adjust the dimensions of the median source face to match those of the median target face, by solving a least squares problem for each of the three spatial dimensions. We then use the following procedure for each frame of the source subject. First, we align the source person's 3D face landmarks $\mathbf{S}^i \in \mathbb{R}^{478\times 3}$ (**rigid plus non-rigid**) with the corresponding landmarks of the mean face template by performing the Procrustes transformation:

$$\mathbf{Z}^i = s_s^i \mathbf{S}^i \mathbf{R}_s^i + \mathbf{T}_s^i \quad (3)$$

Then, to match the dimensions of the median target face, the aligned matrix $\mathbf{Z}^i$ is multiplied by the previously determined non-uniform scaling parameters:

$$(\mathbf{Z}^i)' = s_x \mathbf{Z}_X^i + s_y \mathbf{Z}_Y^i + s_z \mathbf{Z}_Z^i \quad (4)$$

Finally, we un-align $(\mathbf{Z}^i)'$ according to the inverse of the target's median scaling parameter, $\bar{s}_t = median(\{s_t^j\}_{j=1}^N)$, as well as the inverse of the rotation matrix $\mathbf{R}_s^i$.

For the remaining part of the upper body (i.e. the torso along with the hands), we apply a similar procedure to the one we outlined for the head. In the end, for each frame of the source actor, we have two independent skeletons, one for the target subject's head pose and the other for his/her torso and hands pose. However, since the final sequence of retargeted skeletons must match the target subject's upper body movements, additional translations are required to combine the two separate skeletons into one and adjust its overall position. To achieve this, every head skeleton in the sequence is first attached to the nose joint of the corresponding torso skeleton. Then, a global translation is applied to the unified skeleton to align it with the target subject's median root joint, $\bar{\mathbf{l}}_{root} = median(\{\mathbf{l}_{root}^j\}_{j=1}^N)$ (see sec. 3.1).

3.3. Color-coded Conditioning

Having adapted the motion of the source person subject to the body shape and location of the target person, we follow [22, 15] and generate convenient for neural rendering semantic represenations of the body pose in the image space, which we term *Color-Coded Body Representations* (**CCBR** $\in \mathbb{R}^{256\times 256\times 3}$).

In more detail, these representations are 8-bit RGB images where each tracked joint is plotted as a disk of fixed radius and assigned a unique fixed color based on a predefined color coding scheme. The Red and Green channels are given values directly from the XY coordinates of a template body's joints in the 2D image space, after normalizing them to $[0, 1]$. The Blue channel has predefined and independent of the landmarks values for the torso, left hand, right hand, and head. Moreover, we found out that increasing the number of skeleton joints boosted our reenactment performance, and therefore we apply bone interpolation as a data augmentation technique. The color and number of interpolated points along a certain bone are both fixed. For their coloring, we interpolate between the RGB colors of the tracked joints that define each bone. Because we give each joint a fixed distinct color regardless of the signer, this indicates that all of them will have the exact same color in any such representation. This is why these representations are referred to as semantic and they have generally shown to help the renderer learn the mapping to the output images since they are both in the RGB space [15].

Similarly to [15], we also condition our video rendering network to they **eye gaze images**, $\mathbf{E} \in \mathbb{R}^{256\times 256\times 3}$. These are generated by drawing the left and right pupils as disks with fixed radius and connecting the eyes contour landmarks to form the outline of each eye. For each frame t, the body representation, $\mathbf{CCBR}_t$, and the corresponding eye gaze image, $\mathbf{E}_t$, are concatenated and fed to our video rendering network as conditional input, $\mathbf{x}_t = (\mathbf{CCBR}_t, \mathbf{E}_t) \in \mathbb{R}^{256\times 256\times 6}$ (see sec. 3.4).

During training, the joints' coordinates are directly mapped from the target actor's extracted pose sequence $\mathbf{P}_t$. On the other hand, in a reenactment scenario, where the source actor is different from the target, they are estimated by applying the pose retargeting step (see sec. 3.2) on the source subject's sequence $\mathbf{P}_s$.

3.4. Photo-realistic Synthesis

We build upon the publicly available video rendering network of Head2Head++ [15] for producing photo-realistic, temporally coherent videos. This network is person specific and is trained seperately on the training footage of each subject. During training, we follow a self-reenactment setting where the source signer coincides with the target, thus we have access to the ground truth frames. Ideally, the generated video $\tilde{\mathbf{Y}} = (\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_T)$ should be a reconstruction of the training target video $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_T)$. The network consists of: 1) a Generator G, 2) an Image Discriminator D_I and 3) a multi-scale Dynamics Discriminator D_D. In contrast to [15], we also use a body segmentation model in order to prevent cases where artifacts are introduced in the background of the generated images. The network's components and training objectives are identical to Head2Head++ [15], thus they are briefly described below.

Generator G: Given the conditional inputs $\mathbf{x}_t$ and $\mathbf{x}_{t-1}, \mathbf{x}_{t-2}$ of the current and the two preceding frames respectively as well as the two previously generated images $\tilde{\mathbf{y}}_{t-1}, \tilde{\mathbf{y}}_{t-2}$, the generator renders the frame of the output video at time step t:

$$\tilde{\mathbf{y}}_t = G(\mathbf{x}_{t-2:t}, \tilde{\mathbf{y}}_{t-2:t-1}) \tag{5}$$

The final output video $\tilde{\mathbf{Y}}$ shows the target subject performing the source signer's manual and non-manual signs, as determined by the conditional inputs sequence $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$.

Image Discriminator D_I: The discriminator is used during training and tries to determine if the input samples are real or fake. At time step t, the Image Discriminator D_I receives the real pair $(\mathbf{x}_t, \mathbf{y}_t)$ and the fake one $(\mathbf{x}_t, \tilde{\mathbf{y}}_t)$.

Dynamics Discriminator D_D: The Dynamics Discriminator is trained to detect videos with temporal incoherence between their frames. Given the optical flow $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_{T-1})$ of the ground truth video $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_T)$, the discriminator should learn to distinguish the fake data $(\mathbf{w}_{t:t+1}, \tilde{\mathbf{y}}_{t:t+2})$ from real data $(\mathbf{w}_{t:t+1}, \mathbf{y}_{t:t+2})$.

Objective function: The total objective for G is:

$$L^G = L^G_{adv} + \lambda_{vgg} L^G_{vgg} + \lambda_{feat} L^G_{feat} \tag{6}$$

with $\lambda_{vgg} = \lambda_{feat} = 10$ as in [15].

The first loss corresponds to the **adversarial objective** of the generator and is defined as in LSGAN [31] using the 0-1 binary coding scheme ($b = c = 1$ and $a = 0$). The second term is the **VGG loss** which is computed as in [54] and [53], by using the VGG network [43] to extract visual features in different layers for both the ground truth $\mathbf{y}_t$ and the synthesized frame $\tilde{\mathbf{y}}_t$. The final loss in the generator's objective function is the overall **feature matching loss** which is equal to:

$$L^G_{feat} = L^{G-D_I}_{feat} + L^{G-D_D}_{feat} \tag{7}$$

The first sub-loss, $L^{G-D_I}_{feat}$, is computed by extracting features with the Image Discriminator D_I and computing the l_1 distance of these features for a fake frame $\tilde{\mathbf{y}}_t$ and the corresponding ground truth $\mathbf{y}_t$. Similarly, $L^{G-D_D}_{feat}$ is computed using the Dynamics Discriminator D_D instead of D_I.

4. Experimental Setup

We describe the experimental setup including collected datasets and implementation details of our method.

4.1. Datasets

We used **three datasets** for our experiments, which are presented below:

1. Target Actors dataset: We selected two publicly available Youtube videos in order to train our person-specific video rendering network. More specifically, we chose two individuals as our target subjects, a male and a female with different body types. Each training video was at 30 fps and had approximately 10 minutes duration and 1280×720 spatial resolution. The frames of each subject were split into a training and a test set using a 90:10 split. It's crucial that the training videos show the target actors performing a wide range of upper body movements and facial expressions.

2. Source Actors dataset: We collected a small dataset of 14 source videos from an online Greek Sign Language (GSL) dictionary [2], which we used to assess the performance of the various approaches in our sign classification study (see sec 5.3). Six individuals, four men and two women, were included in our source footage and each of them performed a distinct GSL sign that lasted from one to three seconds. Each actor's frames from this dataset were kept as test data and used for our reenactment experiments. In contrast to the target training videos, we only require decent pose detection on the source footage.

3. Continuous Signing dataset: We chose 4 publicly available videos of 2 male and 2 female actors signing continuously for 30 seconds each. Every video in this dataset was used as source footage and the performed signs were retargeted at the target subject of the opposite gender, resulting in a total of four synthesized videos. These videos were included in our realism study (see sec. 5.3). We also show some representative frames from two generated video in the qualitative evaluation section of the experimental results.

4.2. Implementation Details

Our person-specific video rendering network requires a few minutes footage for each target actor. In particular, for every subject in our Target Actors dataset, we used a $\sim$10-minute video and the training task (100 epochs) was completed in approximately 4 days on two NVIDIA GeForce

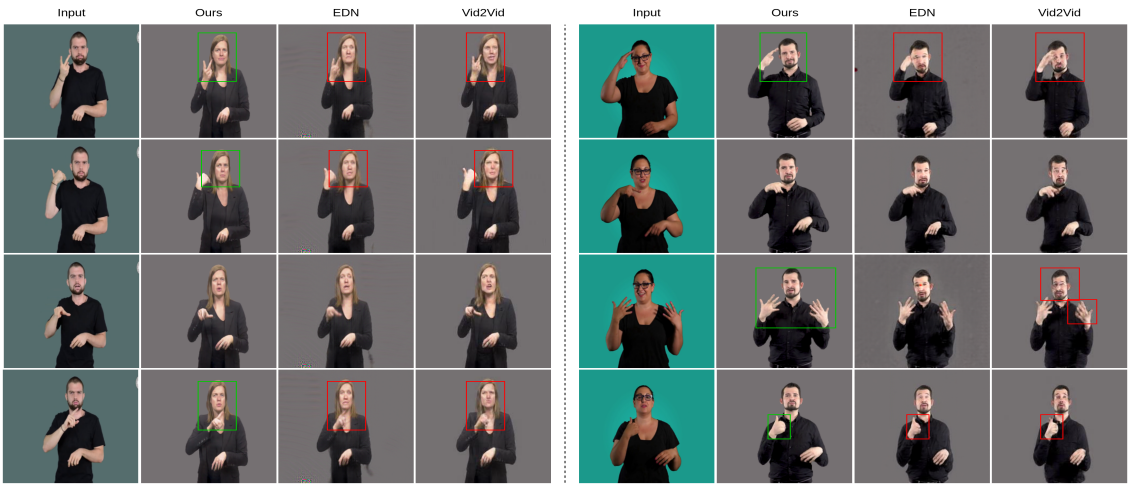

Figure 3: Visual comparison with other methods on reenactment examples for 2 actors from the Continuous Signing dataset. From left to right: input frame, ours, EDN [12], Vid2Vid [53]. We achieve better results in terms of realism and pose transfer. We also illustrate some erroneous results with red boxes and some successful examples of preserving the original mouth patterns and handshapes using green boxes. Please zoom in for details and refer to Supplementary Video [1] for additional results.

GTX 1080 Ti GPUs. The networks were optimized using Adam [23] with an initial learning rate $\eta = 2 \cdot 10^{-4}$, $\beta_1 = 0.5$ and $\beta_2 = 0.999$.

5. Comparison with other methods

In this section, we compare our method with recent approaches that solve the general problem of human motion transfer using qualitative and quantitative evaluations as well as user studies, to assess both their performance and realism. More specifically, we employ the publicly available implementations of Everybody Dance Now (EDN) [12] and Video-to-Video Synthesis (Vid2Vid) [53]. It is important to note that these approaches have been tested for reenacting full-body activities (dancing, exercising, etc.), but we were unable to find a method that addresses the same problem as us and also has source code available. For additional results and visualizations, please refer to the Supplementary Video [1].

5.1. Qualitative Results

Fig. 3 displays the qualitative results of the three methods (ours, EDN [12] and Vid2Vid [53]) for a few representative frames of a male and female source actor from our Continuous Signing dataset. It can be seen that our method is capable of transferring the source person's head, torso and hands movements, facial expressions and eye gaze to the target subject. Note also that it works reliably for different body types, generating frames with respect to the target subject's body structure. It is also evident that our approach outperforms the other two baselines in terms of both realism and pose transfer. In particular, we synthesize frames that look more realistic and natural, whereas EDN and Vid2Vid significantly distort the target's appearance. As shown in Fig. 3, our method results in a more accurate transfer of the source actor's handshapes and facial expressions to the target subjects, compared to [12] and [53].

5.2. Quantitative Results

To assess the performance of each method we conduct a **cycle reenactment** experiment which is a variant of the self-reenactment setting, where the source actor coincides with the target. In this experiment, the signing of a source actor is transferred to a target subject and then back to the same source. More specifically, we use every target subject from our Target Actors dataset as a source actor and transfer his/her performed signs (manual and non-manual) from the test data split to the other target subject. The upper body movements and facial expressions are then transferred back to the first actor in the cycle using the previously generated video as the source video.

For evaluating the performance of the various methods we use the **Average Pixel Distance (APD)** metric. APD is computed as the average l_2 distance of RGB values across all pixels and frames, between the ground truth and generated video (at the end of the experiment). Table 1 shows the values of the APD metric for the three methods over the entire test sequence, made up of $1,000$ frames.

As can be seen, our method outperforms EDN [12] and Vid2Vid [53] overall. It is worth mentioning that the extremely high APD value of Vid2Vid in the third row of Table 1 is attributed to some intense artifacts that were intro-

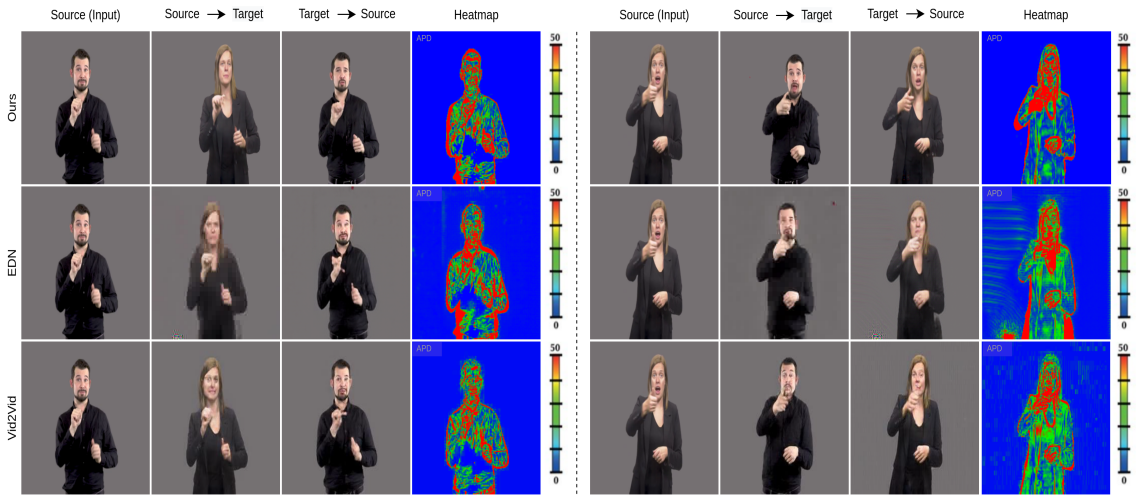

Figure 4: Cycle reenactment results. From left to right: source actor, intermediate-target actor, original source actor driven by the manipulated target actor in the column before, Average Pixel Distance (APD) between first and third column in the form of RGB heatmap. From top to bottom: ours, EDN [12], Vid2Vid [53]. Please zoom in for details and refer to Supplementary Video [1].

	Ours	EDN	Vid2Vid
Male	14.40	13.43	**10.99**
Female	**10.55**	13.60	108.42
Average	**12.48**	13.52	59.71

Table 1: Quantitative results for the cycle reenactment experiments.

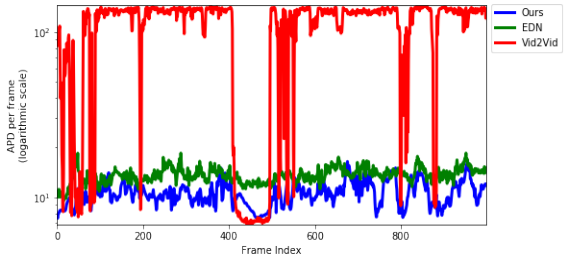

Figure 5: Cycle reenactment performance over time. Average Pixel Distance (APD) between the ground truth and generated video as a function of time (frame index) for each of the three methods in comparison.

duced in the background of the generated images. Examples from our cycle reenactment experiments on the test set of the two target actors are displayed in Fig. 4. As already mentioned, our method synthesizes highly realistic frames, as opposed to the blurry and substantially distorted images that the other methods produce. Fig. 5 further shows how the per-frame APD changes over time for the Female-Male-Female cycle experiment. Here, we observe that our method has the lowest temporal error variance.

5.3. User Studies

We have designed and implemented [26] two web-based studies in order to perceptually evaluate the realism and faithful reenactment of different glosses from human users of GSL.

Realism Study The first study consisted of four questions, each including a pair of videos ≈15 seconds long from our method and one of Vid2Vid or EDN and asking the user to pick the one that seems more realistic to him/her. The videos where chosen randomly from a pool of 4 videos we had rendered for each method (2 for each actor). The study was completed by 21 users and the preference results are presented in Table 2. As it can be seen, the overwhelming majority of users has rated our method as more realistic than the other two. This result is to be expected since as we also saw in visual comparisons, the other methods included multiple artifacts in both the face of the actor and the background.

Ours vs. EDN		**Ours vs. Vid2Vid**	
Ours	EDN	Ours	Vid2Vid
(39/42) 92.9%	(3/42) 7.1%	**(40/42) 95.2%**	(2/42) 4.8%

Table 2: Preference results on the realism of each method. Our method is **significantly** ($p \approx 10^{-9}$ and $p \approx 10^{-8}$, binomial test) more realistic compared to EDN and Vid2Vid and consistently preferred across participants.

Sign Classification Study In the second study, we evaluated how faithfully each method reenacted a number of different Greek sign-language (GSL) glosses. We carefully

selected based on the guidance of an SL expert 14 GSL glosses and reenacted them using our method, EDN and Vid2Vid. Then, we showed each user 12 glosses (3 for each method, plus 3 for the original source videos) and asked them which gloss was being signed, from a list of 7 choices (including "None of the above"). Note that we also included one of the source videos twice, as a control question. A total of 23 users completed this study.

The results of this second study can be seen in Table 3. We can see that all methods achieve high accuracy regardless of their realism, which is on par with the source videos as well. All methods faithfully reproduced the perception of different glosses, despite the evident difference in their realism, which shows that the recognition is possible even for non-realistic videos, in the cost however of the user experience. The small discrepancies between the different methods are not statistically significant (see Table 3) and can be attributed to: **a)** the random sampling from the question bank leading to slight different distribution of scores glosses in different methods and **b)** the fact that some participants might not have identified the specific signing style of the source for specific glosses, leading them to select "None of the above" if the source video had a different signing style with the one they are familiar with. It is characteristic that the real videos have a lower sign recognition rate.

Ours	EDN	Vid2Vid	Real video
(53/69) 76.8%	**(55/69) 79.7%**	(53/69) 76.8%	(51/69) 73.9%

Table 3: Results of sign recognition user study. Classification accuracy of each method on different GSL glosses. There is no significant difference between all methods (p=1 for all pairwise proportion tests with Bonferroni correction).

6. Conclusions

We proposed Neural Sign Reenactor, a novel neural rendering pipeline for transferring the body movements, head pose and facial expressions of a source actor in a driving video to a target subject in a reference video. We have applied our approach to the challenging case of Sign Language videos. Our extensive qualitative and quantitative evaluations have demonstrated that our method faithfully transfers the source signer's manual and non-manual signs to a target signer and works reliably across signers of different genders and body structures. Compared to earlier methods of human motion retargeting that dramatically alter the appearance of the target subject, it also produces highly realistic and natural looking results. We believe that our work paves the way for the development of novel Sign Language Production systems that go beyond avatars and produce photo-realistic continuous sign language videos increasing the appeal and engagement of the users.

Acknowledgments. A. Roussos was supported by the Greek Secretariat for Research and Innovation and the EU, Project SignGuide: Automated Museum Guidance using Sign Language T2EDK-00982 within the framework of "Competitiveness, Entrepreneurship and Innovation" (EPAnEK) Operational Programme 2014-2020. A. Roussos acknowledges also the support by an NVIDIA Academic Hardware Grant Program, which was beneficial in developing and testing the neural rendering models introduced in this paper.

References

[1] https://youtu.be/uWciU_Cqyd0.

[2] Dictionary of Greek Sign Language. www.keng.gr, 2022.

[3] Kfir Aberman, Mingyi Shi, Jing Liao, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Deep video-based performance cloning. In *Computer Graphics Forum*, volume 38, pages 219–233. Wiley Online Library, 2019.

[4] Kfir Aberman, Rundi Wu, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Learning character-agnostic motion for motion retargeting in 2d. *arXiv preprint arXiv:1905.01680*, 2019.

[5] Epameinondas Antonakos, Anastasios Roussos, and Stefanos Zafeiriou. A survey on mouth modeling and analysis for sign language recognition. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–7. IEEE, 2015.

[6] J Andrew Bangham, SJ Cox, Ralph Elliott, John RW Glauert, Ian Marshall, Sanja Rankov, and Mark Wells. 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. IET, 2000.

[7] Thabo Beeler and Derek Bradley. Rigid stabilization of facial expressions. *ACM Transactions on Graphics (TOG)*, 33(4):1–9, 2014.

[8] British Deaf Association. BSL statistics - British Deaf Association. https://bda.org.uk/help-resources/#statistics, 2019.

[9] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793, 2018.

[10] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 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, 2020.

[11] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[12] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the*

IEEE/CVF international conference on computer vision, pages 5933–5942, 2019.

[13] Yucheng Chen, Yingli Tian, and Mingyi He. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding*, 192:102897, 2020.

[14] Stephen Cox, Michael Lincoln, Judy Tryggvason, Melanie Nakisa, Mark Wells, Marcus Tutt, and Sanja Abbott. Tessa, a system to aid communication with deaf people. In *Proceedings of the fifth international ACM conference on Assistive technologies*, pages 205–212, 2002.

[15] Michail Christos Doukas, Mohammad Rami Koujan, Viktoriia Sharmanska, Anastasios Roussos, and Stefanos Zafeiriou. Head2head++: Deep facial attributes re-targeting. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(1):31–43, 2021.

[16] Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and François Lefebvre-Albaret. The dicta-sign wiki: Enabling web communication for the deaf. In *International Conference on Computers for Handicapped Persons*, pages 205–212. Springer, 2012.

[17] EU Think Tank. Sign languages in the EU: Think tank: European parliament. `https://www.europarl.europa.eu/thinktank/en/document/EPRS_ATA(2018)625196`, 2018.

[18] Sylvie Gibet, François Lefebvre-Albaret, Ludovic Hamon, Rémi Brun, and Ahmed Turki. Interactive editing in french sign language dedicated to virtual signers: Requirements and challenges. *Universal Access in the Information Society*, 15(4):525–539, 2016.

[19] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018.

[20] Thomas Hanke. Hamnosys-representing sign language data in language resources and language processing contexts. In *LREC*, volume 4, pages 1–6, 2004.

[21] Richard Kennaway. Avatar-independent scripting for real-time gesture animation. *arXiv preprint arXiv:1502.02961*, 2015.

[22] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018.

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[25] Oscar Koller, Necati Cihan Camgoz, Hermann Ney, and Richard Bowden. Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2306–2320, 2019.

[26] Kosmas Kritsis, Aggelos Gkiokas, Aggelos Pikrakis, and Vassilis Katsouros. Danceconv: Dance motion generation with convolutional networks. *IEEE Access*, 10:44982–45000, 2022.

[27] Jongin Lim, Hyung Jin Chang, and Jin Young Choi. Pmnet: Learning of disentangled pose and movement for unsupervised motion retargeting. In *BMVC*, volume 2, page 7, 2019.

[28] Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Hyeongwoo Kim, Florian Bernard, Marc Habermann, Wenping Wang, and Christian Theobalt. Neural rendering and reenactment of human actor videos. *ACM Transactions on Graphics (TOG)*, 38(5):1–14, 2019.

[29] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. 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, 2019.

[30] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

[31] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[32] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017.

[33] Masahiro Mori, Karl F MacDorman, and Norri Kageki. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, 19(2):98–100, 2012.

[34] Ilias Papastratis, Christos Chatzikonstantinou, Dimitrios Konstantinidis, Kosmas Dimitropoulos, and Petros Daras. Artificial intelligence technologies for sign language. *Sensors*, 21(17):5843, 2021.

[35] Siegmund Prillwitz and Hamburg Zentrum für Deutsche Gebärdensprache und Kommunikation Gehörloser. *HamNoSys: Version 2.0; Hamburg notation system for sign languages; an introductory guide*. Signum-Verlag, 1989.

[36] Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, and Mohammad Sabokrou. Sign language production: a review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3451–3461, 2021.

[37] Anastasios Roussos, Stavros Theodorakis, Vassilis Pitsikalis, and Petros Maragos. Dynamic affine-invariant shape-appearance handshape features and classification in sign language videos. *The Journal of Machine Learning Research*, 14(1):1627–1663, 2013.

[38] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Adversarial training for multi-channel sign language production. *arXiv preprint arXiv:2008.12405*, 2020.

[39] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Everybody sign now: Translating spoken language

to photo realistic sign language video. *arXiv preprint arXiv:2011.09846*, 2020.

[40] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Progressive transformers for end-to-end sign language production. In *European Conference on Computer Vision*, pages 687–705. Springer, 2020.

[41] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Anonysign: Novel human appearance synthesis for sign language video anonymisation. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021.

[42] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5141–5151, 2022.

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[44] Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. Sign language production using neural machine translation and generative adversarial networks. In *Proceedings of the 29th British Machine Vision Conference (BMVC 2018)*. British Machine Vision Association, 2018.

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

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

[47] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(04):376–380, 1991.

[48] Yehuda Vardi and Cun-Hui Zhang. The multivariate l 1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4):1423–1426, 2000.

[49] Lucas Ventura, Amanda Duarte, and Xavier Giró-i Nieto. Can everybody sign now? exploring sign language video generation from 2d poses. *arXiv preprint arXiv:2012.10941*, 2020.

[50] Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. Neural kinematic networks for unsupervised motion retargetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8639–8648, 2018.

[51] Andreas Voskou, Konstantinos P Panousis, Dimitrios Kosmopoulos, Dimitris N Metaxas, and Sotirios Chatzis. Stochastic transformer networks with linear competing units: Application to end-to-end sl translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11946–11955, 2021.

[52] HARRY THOMAS WALSH and Ben Saunders. Changing the representation: Examining language representation for neural sign language production. In *LREC 2022*.

[53] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018.

[54] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 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, 2018.

[55] World Health Organization. Deafness and hearing loss. `https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss`, 2021.

[56] Zhuoqian Yang, Wentao Zhu, Wayne Wu, Chen Qian, Qiang Zhou, Bolei Zhou, and Chen Change Loy. Transmomo: Invariance-driven unsupervised video motion retargeting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5306–5315, 2020.

[57] Kayo Yin and Jesse Read. Better sign language translation with stmc-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989, 2020.

[58] Jan Zelinka and Jakub Kanis. Neural sign language synthesis: Words are our glosses. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3395–3403, 2020.

[59] Ce Zheng, Wenhan Wu, Taojiannan Yang, Sijie Zhu, Chen Chen, Ruixu Liu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *arXiv preprint arXiv:2012.13392*, 2020.

[60] Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. Spatial-temporal multi-cue network for continuous sign language recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13009–13016, 2020.

[61] Wentao Zhu, Zhuoqian Yang, Ziang Di, Wayne Wu, Yizhou Wang, and Chen Change Loy. Mocanet: Motion retargeting in-the-wild via canonicalization networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3617–3625, 2022.

[62] Inge Zwitserlood, Margriet Verlinden, Johan Ros, Sanny Van Der Schoot, and T Netherlands. Synthetic signing for the deaf: Esign. In *Proceedings of the conference and workshop on assistive technologies for vision and hearing impairment (CVHI)*. Citeseer, 2004.