Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
AI for Content Creation Workshop (AI4CC),
Vancouver, Canada, June 2023.

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

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, mouth patterns, head, and body movements) signs to a target video in a photo-realistic manner. Our method can be used for Sign Language Anonymization, Sign Language Production (synthesis module), as well as for reenacting other types of full body activities (dancing, acting performance, exercising, etc.). We conduct detailed qualitative and quantitative evaluations and comparisons, which demonstrate the particularly promising and realistic results that we obtain and the advantages of our method over existing approaches.*

1. Introduction

One of the most challenging open problems of Sign Language (SL) technologies is the generation of synthetic SL videos that allow SL users to experience natural and fluid communication, similar to human-to-human communication. Prior to the deep learning era, the SL Production (SLP) problem was historically tackled using animated avatars (*e.g.* VisiCast [10], Tessa [17], eSign [55] and DictaSign [19]). However, in terms of the avatars' appearance and motion, this typically resulted in a low level of realism, reducing the plausibility and engagement of users with such technologies.

With the advent of deep learning, novel methods have been introduced that build upon the latest advances in photo-realistic neural rendering and synthesize SL videos with avatars that have the appearance of real persons. Initial approaches (*e.g.* [39, 40, 52]) dealt with this problem by concatenating isolated signs disregarding the natural co-articulation between them. In addition, other works (*e.g.* [34, 36, 52]) used skeleton pose representations rather than photo-realistic videos, which was shown to reduce Deaf understanding [44]. To improve sign comprehension, more recent approaches go one step further and apply human motion retargeting techniques to transform the predicted skeleton pose sequences into a photo-realistic human actor video. Human motion retargeting is an emerging topic at the intersection of computer vision and graphics due to its extensive potential for content creation. Over the last years, a plethora of deep learning-based methods has been introduced in this field. Some of them require high-fidelity 3D pose estimation or reconstruction [27–29, 45]. Retargeting motion from 2D inputs has also been studied in several works [7, 8, 16, 50, 54]. SignGAN [35] was the first SLP model to produce photo-realistic continuous SL videos by conditioning synthesis on the predicted skeletal pose sequence and the style image of a reference signer. The closest work to this paper is that of Saunders *et al.* [37], 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, their generated frames include artifacts and the synthesized human appearances are not always convincing as being real. This work overcomes the aforementioned

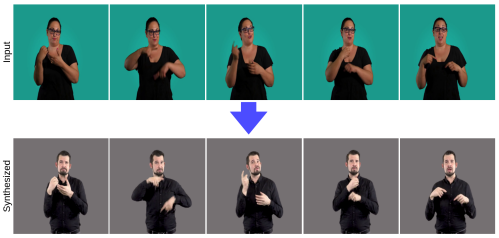

Figure 1. Given an input sign language video, our *Neural Sign Reenactor* synthesizes a photo-realistic and temporally coherent video of a target signer imitating the source signer's body and facial movements. Please also refer to Suppl. Video [1].

limitations and synthesizes videos of unprecedented realism that include the upper body movements and facial expressions of a virtual signer who is almost indistinguishable from a real person. The motivation of our work is discussed in further detail in the Suppl. Material. Our contributions can be summarized as follows: **1)** We build upon an effective combination of two different body trackers for implementing high-fidelity body and face tracking. **2)** We propose a novel scheme for conditioning the neural renderer. **3)** We introduce a novel pose retargeting step that enables our model to work reliably across signers of different genders and body structures. **4)** We conduct detailed qualitative and quantitative evaluations and user studies to evaluate our method and compare it with previous human motion transfer methods. The experiments demonstrate the particularly promising and realistic results that we obtain under challenging continuous signing scenarios.

2. Methodology

Given an input video $\mathbf{Y}$, our method generates a photo-realistic and temporally coherent video $\tilde{\mathbf{Y}}$ of a target actor imitating the source actor's upper body movements and facial expressions. An overview of the proposed pipeline is presented in Fig. 2. It consists of four main components:

1) Upper Body Detection: We first extract the skeleton pose sequences from the SL videos (source and target) using the MediaPipe (MP) [30] Pose and Holistic modules. More specifically, we use MP Holistic to track the head and hands, inferring 520 landmarks in total, while for the torso we use 9 from the 33 3D landmarks detected by the MP Pose model (since the Holistic module does not capture the depth of the pose landmarks). After preprocessing, every frame is represented by a pose vector that stores the 3D coordinates of $K = 529$ tracked joints. Hereafter, the joint in the middle of the shoulders will be referred to as the *root* joint. Finally, for every video, we crop every frame with a fixed-size and fixed-position bounding box that surrounds all locations of the skeleton's joints over all frames, leaving at each side (left/right/top/bottom) a margin whose size is 5% of the corresponding average dimension (width/height). The cropped frames are then resized to the constant resolution of **256×256 pixels**.

2) Pose Retargeting: We propose a novel pose retargeting algorithm based on **Procrustes Analysis** for transferring the motion from a source character to a target. It is applied separately for two parts of the upper body, namely the head and the torso along with the hands, taking into consideration possible differences between their body shapes. For the **head**, similarly to [11], we use a subset of $n = 94$ facial landmarks from the most rigid area of the face that are less affected by facial deformations during facial expressions and mouth motions. For the sake of simplicity, we call these landmarks **rigid** and all the rest **non-rigid**. For every frame and every video (either source or target), we consider the set of rigid 3D landmarks and perform Procrustes Analysis [42] to rigidly align them to a common reference face template, which is defined in an anatomical coordinate system with axes aligned to the axial, coronal and sagittal planes. For each video (either source or target), we consider the aligned rigid landmarks and apply **geometric median** [43] over all frames to extract a median face that robustly approximates the subject's facial geometry. To account for cross-subject anatomical differences in the facial shape, we find the non-uniform per-axis scaling $\mathcal{S}$ that optimally registers the median face of the source to the median face of the target. Note that both median faces live in the anatomical coordinate system, therefore considering per-axis scaling only provides a satisfactory approximation. Finally, for each frame of the source video, we consider all facial landmarks (rigid and non-rigid) and apply the following transformations: $\mathcal{T}1$) the already estimated Procrustes transformation from the source domain to the anatomical coordinate system, $\mathcal{T}2$) the non-uniform scaling $\mathcal{S}$, $\mathcal{T}3$) the inverse of transformation $\mathcal{T}1$. For the **remaining part** of the upper body, we follow a similar procedure to that outlined for the head, ending up with two independent skeletons: one for the target subject's head pose and the other for his/her torso and hands movements. However, since the final sequence of retargeted skeletons must match the target actor's upper body movements, additional translations are required to combine the two separate skeletons into one and then adjust its overall position. To achieve this, every head skeleton in the output sequence is first attached to the nose joint of the corresponding torso skeleton. As a final step, a global translation and scaling are applied to the unified skeleton (head, torso, hands) to align it with the target subject's median scale and position at the target video's domain. This helps the neural renderer during reenactment since it ensures that the retargeted skeleton is as similar to the skeleton of the training video as possible.

3) Color-coded Conditioning: Having adjusted the motion of the source person subject to the body shape and location of the target person, we follow [18, 21] and generate convenient for neural rendering semantic representations of the body pose in the 2D image space, which we term **color-coded body representations**, **CCBR** $\in \mathbb{R}^{256\times256\times3}$. 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 color based on a novel coloring scheme of a template skeleton. Please refer to Suppl. Material for more details. 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, where both the color and number of interpolated points along a certain bone are fixed. For their coloring, we interpolate between the RGB colors of the tracked joints that define each bone. Similarly to [18],

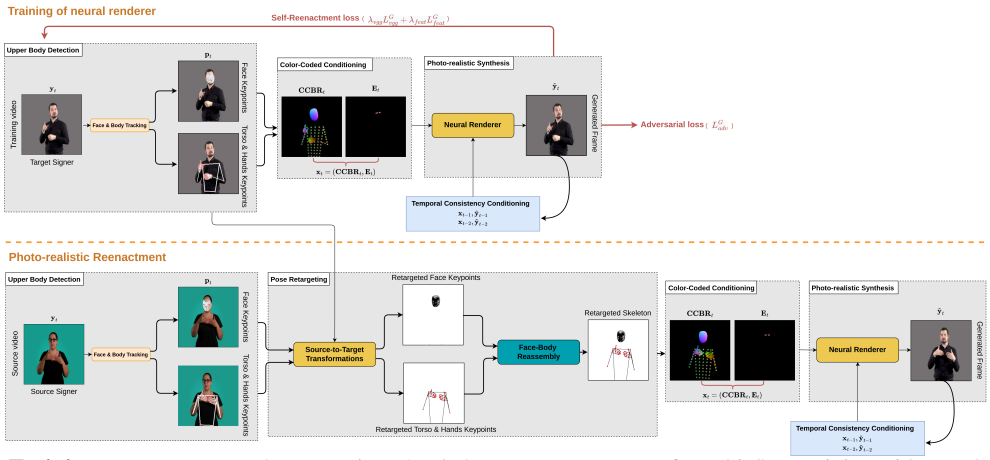

Figure 2. (Top) **Training**: We extract each target signer's skeleton pose sequence from his/her training video and use it to generate the corresponding color-coded body representations and eye gaze images, which are concatenated and fed into the neural renderer as conditional input. (Bottom) **Reenactment**: We extract the source signer's skeleton pose sequence from his/her source video and then transform the estimated landmarks to match the target actor's body shape and location within each frame. The output frames are generated by the neural renderer from the corresponding conditional inputs using the previously trained model for the specific target signer.

we also condition our video rendering network to **eye gaze images**, $\mathbf{E} \in \mathbb{R}^{256\times256\times3}$, which are generated by drawing the left and right pupils as disks of fixed radius and connecting the eyes' contour landmarks. At each time step t, the CCBR is concatenated with the corresponding eye gaze image and fed to the neural renderer as conditional input, $\mathbf{x}_t = (\mathbf{CCBR}_t, \mathbf{E}_t) \in \mathbb{R}^{256\times256\times6}$.

4) Photo-realistic Synthesis: We build upon the publicly available video rendering network of Head2Head++ [18] for producing photo-realistic, temporally coherent videos. Our neural renderer is person-specific, which indicates that it is trained separately for every target actor using his/her reference video as the only training data. 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. The network consists of: **a)** a *Generator* G, **b)** an *Image Discriminator* D_I, and **c)** a *Dynamics Discriminator* D_D. In contrast to [18], we also use a body segmentation model to prevent some artifacts in the background of the generated images. In terms of the network's architecture and training process, we follow Head2Head++. Please refer to Suppl. Material for more details.

3. Comparison with other methods

We compare our method with two previous human motion transfer methods, namely Everybody Dance Now (EDN) [16] and Video-to-Video Synthesis (Vid2Vid) [47]. It is important to note that these approaches have been tested for reenacting full body activities, 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 Suppl. Video [1].

Qualitative Results: Fig. 3 displays the qualitative results of the three methods for a few representative frames of a male and female source actor signing in Greek SL (GSL). As can be seen, our method is capable of efficiently transferring the source person's head, torso, and hands movements, facial expressions, and eye gaze to the target subject. It also works reliably for different body types, generating frames with respect to the target subject's body structure. Moreover, it is 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. Compared to [16] and [47] that even use a specialized GAN to add realism to a certain region (*e.g.* Face GAN in EDN), we also result in a more accurate transfer of the source actor's facial expressions and handshapes to the target subjects. In general, our method generates photo-realistic videos of the target actor signing in the source actor's SL even though he/she has never used the particular SL and has never seen or performed the retargeted motions, which are determined by the input video.

Quantitative Results: To assess the performance of the various methods, we conduct a **cycle reenactment** experiment, where the signing of a source actor is transferred to a target subject and then back to the same source. Ideally, the final video at the end of the experiment should be a reconstruction of the input one, so we can measure the per-pixel differences and calculate performance metrics. In particular, we use the **Average Pixel Distance (APD)** metric, which is computed as the average l_2 distance of RGB values across all pixels and frames, between the ground truth and final synthesized video. Table 1 shows the APD values for the three methods over the entire test sequence of our male and female target actors ($1,000$ frames each). As can be seen, our method outperforms EDN [16] and Vid2Vid [47] overall. Examples from our cycle reenactment experiments' results are displayed in Fig. 4. As already mentioned, our

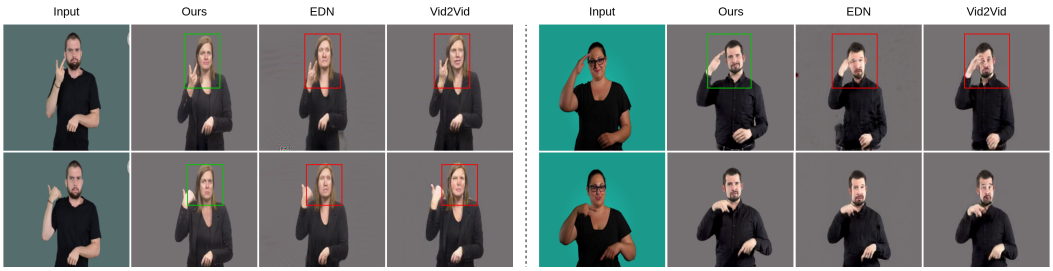

Figure 3. Visual comparison with EDN [16] and Vid2Vid [47] on different reenactment examples. We illustrate some erroneous results with red boxes and some successful examples of preserving the original mouth patterns and handshapes with green boxes. Please zoom in for details and refer to Suppl. Video [1].

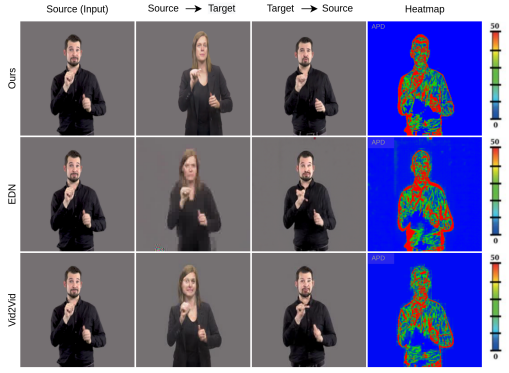

Figure 4. Cycle reenactment comparisons with EDN [16] and Vid2Vid [47]. From left to right: source actor, intermediate-target actor, original source actor driven by the manipulated target actor in the column before, and per-pixel differences between the first and third column.

method synthesizes highly realistic frames, as opposed to the blurry and substantially distorted images that the other methods produce.

	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 comparison of the three methods.

User Studies: We designed and implemented [25] two user studies to evaluate the realism and faithful reenactment of different glosses from human users of GSL. The **first study** was a **Realism Study** which consisted of four questions, each including a pair of synthesized videos, one from our method and one from EDN or Vid2Vid, and asking the user to pick the one that seemed more realistic to him/her. The study was completed by 21 users and the preference results are presented in Table 2. As can be seen, the overwhelming majority of users have rated our method as more realistic than the other two.

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.

In our **second study**, which was a **Sign Classification Study**, we evaluated how faithfully each method reenacted a number of different GSL glosses. For that, we carefully selected based on the guidance of an SL expert 14 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 source videos) and asked them which gloss was being signed, from a list of 7 choices (including "None of the above"). A total of 23 users completed this study, and the results are shown in Table 3, where we can see that all methods achieve high accuracy regardless of their realism, in the cost however of the user experience. The small discrepancies between the different methods are statistically insignificant and can be attributed to: **a)** the random sampling from the question bank leading to a slightly different distribution of glosses between the various methods and **b)** the fact that some participants might not have identified the specific signing style of the source actor for some glosses, leading them to mistakenly select "None of the above" if the source video had a different signing style from the one they are familiar with.

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

Table 3. 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).

4. 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 SL 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 for human motion retargeting that dramatically alter the target subject's appearance, it produces highly realistic and natural-looking results. We believe that our work paves the way for the development of novel SLP systems that go beyond computer-generated avatars and produce photo-realistic SL videos increasing the appeal and engagement of the users.

References

[1] https://youtu.be/xKAfguacOkE. 1, 3, 4, 9, 10

[2] Association of Greek Sign Language Interpreters. https://sdeng.org.gr/. 10

[3] Hellenic Federation of the Deaf. https://www.omke.gr/. 10

[4] MediaPipe Face Mesh. https://google.github.io/mediapipe/solutions/face_mesh.html. 8

[5] OTC-CTA Canada. https://www.youtube.com/user/otccta. 9

[6] Dictionary of Greek Sign Language. www.keng.gr, 2022. 10

[7] 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. 1

[8] 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. 1

[9] 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. 8

[10] 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. 1

[11] Thabo Beeler and Derek Bradley. Rigid stabilization of facial expressions. *ACM Transactions on Graphics (TOG)*, 33(4):1–9, 2014. 2

[12] Danielle Bragg, Oscar Koller, Naomi Caselli, and William Thies. 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, 2020. 8

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

[14] 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. 8

[15] 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. 8

[16] 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. 1, 3, 4, 10, 11

[17] 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. 1

[18] 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. 2, 3, 9

[19] 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. 1

[20] 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. 8

[21] 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. 2

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 10

[23] Oscar Koller. Quantitative survey of the state of the art in sign language recognition. *arXiv preprint arXiv:2008.09918*, 2020. 8

[24] 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. 8

[25] Kosmas Kritsis, Aggelos Gkiokas, Aggelos Pikrakis, and Vassilis Katsouros. Danceconv: Dance motion generation with convolutional networks. *IEEE Access*, 10:44982–45000, 2022. 4

[26] Sooyeon Lee, Abraham Glasser, Becca Dingman, Zhaoyang Xia, Dimitris Metaxas, Carol Neidle, and Matt Huenerfauth. American sign language video anonymization to support online participation of deaf and hard of hearing users. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–13, 2021. 8

[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. 1

[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. 1

[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. 1

[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. 2

[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. 9

[32] Ilias Papastratis, Christos Chatzikonstantinou, Dimitrios Konstantinidis, Kosmas Dimitropoulos, and Petros Daras. Artificial intelligence technologies for sign language. *Sensors*, 21(17):5843, 2021. 8

[33] 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. 8

[34] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Adversarial training for multi-channel sign language production. *arXiv preprint arXiv:2008.12405*, 2020. 1, 8

[35] 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. 1

[36] 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. 1, 8

[37] 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. 1

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 9

[39] 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. 1

[40] 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. 1, 8

[41] 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. 8

[42] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(04):376–380, 1991. 2

[43] 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. 2

[44] 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. 1

[45] 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. 1

[46] 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. 8

[47] 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. 3, 4, 9, 10, 11

[48] 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. 9

[49] World Health Organization. Deafness and hearing loss. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss, 2021. 8

[50] 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. 1

[51] 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. 8

[52] 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. 1

[53] 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. 8

[54] 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. 1

[55] 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. 1

Supplementary Material

A. Motivation

In this section, we provide a more detailed discussion of the motivation of our work.

Tens of millions of Deaf worldwide use Sign Language (SL) as their native language [9, 13, 20, 49]. 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 [32]. This has been an active research area during the last three decades, but it was only in the last years that it started maturing, thanks to the introduction of novel deep learning methods that yielded highly robust and promising results on the challenging tasks of Sign Language Recognition (SLR) [15, 23, 24, 33, 41, 53], Translation (SLT) [14, 46, 51] and Production (SLP) [34, 36, 40].

Deep learning approaches require the availability of large-scale SL corpora which is very limited due to participants' concerns over privacy and video misuse [12]. Therefore, there is an urge to increase the amount of publicly available data and thereby further improve the performance of SL systems. In addition, special attention must be paid to cases of videos of SL datasets that refer to third-party personal information (*e.g.* names or personal data of other people). At the same time, one of the important barriers that the Deaf are currently facing is related to their ability for online participation, especially in cases where the option of anonymity is a valuable tool for constructing a safe space to discuss sensitive, controversial or personal topics in social media or other online platforms [26]: In contrast to the users of spoken languages who can easily communicate anonymously by just typing a text, the SL users can only communicate in their native language by using a camera capturing their hands, body, and face during signing, which reveals their identity. Since all these body parts convey cues that are important for SL communication [9], it becomes evident that there is no easy way to conceal the signers' identity through simple video editing approaches.

The aforementioned problems have recently attracted the interest of the research community, resulting in some specialized systems that seek to anonymize SL videos. This is a particularly difficult task due to the challenges in capturing, representing, and retargeting the human motions during signing, for example: extremely fast motion and articulation of the hands, complex interactions between the different body parts (*e.g.* between the two hands or between each of the hands and the face), large variability and complexity of hand configurations, and inter-signer variations due to anatomical differences. Our method can conceal the identity of the original signers by reproducing their videos using other actors who have given their informed consent for their recordings to be shared publicly and therefore can support the Deaf in increasing their online participation.

Regarding the applications of our framework, it can also be beneficial for the following purposes: 1) It can be readily used as the backend module in SLP systems, offering the option to have virtual interpreters with the appearance of real persons, going well beyond the traditional graphics-generated avatars. 2) Although it is developed and tested on the especially challenging problem of SL reenactment, it can be readily applied to other types of full body activities (dancing, exercising, etc.).

B. Color-coded Conditioning

In this section, we provide more details about our novel color-coding scheme, which is used in the *Color-coded Conditioning* module of our pipeline for generating the color-coded body representations and eye gaze images.

Regarding the **CCBRs**, the joints are colored using the following scheme, which assigns each joint a **unique color**: The Red and $Green$ channels are given values directly from the x and y coordinates, respectively, of a template body's joints in the 2D image space, after being normalized between 0 and 1. Similarly, for the coloring of the face, we are based on its UV visualization and 2D texture coordinates by MediaPipe Face Mesh [4]. The $Blue$ channel has predefined and independent of the landmarks values for the torso, left hand, right hand, and face (see Fig. 5). Because we give each joint a unique, fixed 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 also referred to as semantic and they have generally been shown to help neural renderers learn the mapping to the output images since they are both in the RGB space.

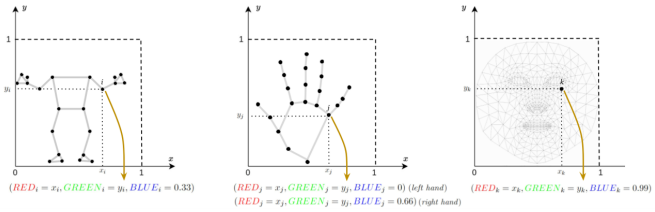

Figure 5. Visualization of our color-coding scheme for the torso, hands, and face.

As also mentioned in the main paper, in addition to the CCBRs, we condition our video rendering network to **eye gaze** images, which are generated by drawing the left and right pupils as disks of fixed radius and connecting the eyes'

contour landmarks. For their coloring, we follow [18] and tint the contour landmarks white and the pupils red. An illustrative example of all types of conditional inputs that we feed our neural renderer with is provided in Fig. 6.

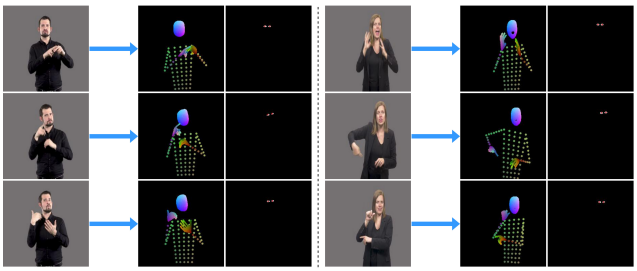

Figure 6. Examples of conditional inputs generation for some representative frames of the target actors' training videos. For each section, we illustrate from left to right: input frame, color-coded body representation, and eye gaze image. Please zoom in for details and refer to Suppl. Video [1].

C. Photo-realistic Synthesis

As stated in the main paper, our video rendering network's components and training objectives are identical to Head2Head++ [18], thus they are briefly described below.

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

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

  The output video $\tilde{\mathbf{Y}}_{1:T}$ shows the target subject performing the source signer's manual and non-manual signs, as determined by the conditional inputs sequence $\mathbf{X}_{1:T}$. The Generator consists of two identical encoders, operating in parallel, as well as a decoder. The first encoder receives the conditional inputs $\mathbf{x}_{t-2:t}$, while the second is given the two previously generated frames $\tilde{\mathbf{y}}_{t-2:t-1}$. The two extracted feature maps are first added and then passed through the decoder, which brings the output $\tilde{\mathbf{y}}_t$ in a normalised $[-1, +1]$ range, using a tanh activation function.

- **Image Discriminator** D_I: The image discriminator is used during training and aims at telling real and synthesized frames apart. At time step t, it 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 used during training to enforce the temporal coherence of the output video. It receives a set of three consecutive real frames $\mathbf{y}_{t:t+2}$ or fake frames $\tilde{\mathbf{y}}_{t:t+2}$ along with the optical flow $\mathbf{w}_{t:t+1}$, computed from the target's subject training video $\mathbf{Y}_{1:T}$, and should learn to distinguish the fake data $(\mathbf{w}_{t:t+1}, \tilde{\mathbf{y}}_{t:t+2})$ from real data $(\mathbf{w}_{t:t+1}, \mathbf{y}_{t:t+2})$. In this way, the generator tries to synthesize fake frames with the same flow/dynamics as the corresponding real ones in order to fool the discriminator.

- **Objective function**: The total objective for G consists of three terms:

$$L^G = L^G_{adv} + \lambda_{vgg} L^G_{vgg} + \lambda_{feat} L^G_{feat} \tag{2}$$

  with $\lambda_{vgg} = \lambda_{feat} = 10$ as in [18].
  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 $a = 0$):

$$L^G_{adv} = \frac{1}{2}\mathbb{E}_t[(D_I(\mathbf{x}_t, \tilde{\mathbf{y}}_t) - 1)^2] + \frac{1}{2}\mathbb{E}_t[(D_D(\mathbf{w}_{t:t+1}, \tilde{\mathbf{y}}_{t:t+2}) - 1)^2] \tag{3}$$

  The second term is the **VGG loss**, which is computed as in [48] and [47], by using the VGG network [38] to extract feature representations in different layers for both the ground truth $\mathbf{y}_t$ and the synthesized frame $\tilde{\mathbf{y}}_t$ and then calculating their euclidean distance.
  The final loss in the generator's objective function is the overall **feature matching loss** which is equal to:

$$L^G_{feat} = L^{G-D_I}_{feat} + L^{G-D_D}_{feat} \tag{4}$$

  The first sub-loss, $L^{G-D_I}_{feat}$, is computed by extracting the activations on an intermediate layer of the image discriminator D_I for a fake frame $\tilde{\mathbf{y}}_t$ and the corresponding ground truth $\mathbf{y}_t$ and then computing their l_2 squared distance. Similarly, $L^{G-D_D}_{feat}$ is computed using the Dynamics Discriminator D_D instead of D_I.

D. Experimental Setup

In this section, we describe the experimental setup of our method including the collected datasets and some implementation details.

D.1. Datasets

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

1) Target Actors dataset: We selected 2 publicly available YouTube videos [5] for **training** our person-specific neural renderer. More specifically, we chose two individuals as our target subjects, a male and a female with different body types, signing in American Sign Language

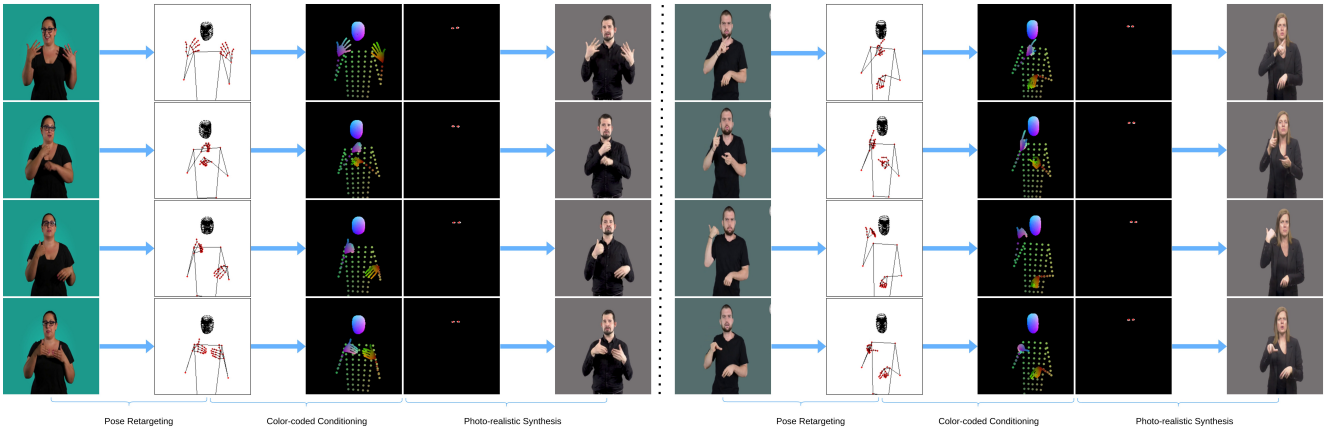

Figure 7. Visualization of intermediate and retargeted results for some representative frames of a female and male source actor from the Continuous Signing dataset. From left to right: input frame, retargeted skeleton, conditional inputs, and output frame.

(ASL) and Quebec Sign Language (QSL), respectively. 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 [6], which we used to assess the performance of the various methods in our **Sign Classification Study**. 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 training videos, we only require decent pose detection on the source footage.

3) Continuous Signing dataset: We chose 4 publicly available videos [2, 3] of two male and two 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 opposite gender's target subject, resulting in a total of four synthesized videos. These videos were included in our **Realism Study**.

We'd like to clarify that our neural renderer is trained separately for every target actor and the only training data is his/her training video. Therefore, each trained model at the end is dedicated to a specific target subject from the training dataset, which is the Target Actors dataset. There is no need to train the neural renderer using the videos from the remaining two datasets, i.e., Source Actors and Continuous Signing, because they only serve as source videos in our experiments.

D.2. Implementation Details

Our person-specific video rendering network requires a few minutes of footage for each target actor. In particular, for every subject in our Target Actors dataset, we used a ≈10-minute video and the training task (100 epochs) was completed in approximately 4 days on two NVIDIA GeForce GTX 1080 Ti GPUs. The networks were optimized using Adam [22] with an initial learning rate $\eta = 2 \cdot 10^{-4}$, $\beta_1 = 0.5$ and $\beta_2 = 0.999$.

E. Additional Visualizations

We show in Fig. 7 a more detailed view of our method's intermediate steps in a reenactment setting, including pose retargeting and color-coded conditioning. Also, Fig 8 provides additional qualitative results of our method in comparison with EDN [16] and Vid2Vid [47] in the form of static frames for two actors of our Continuous Signing dataset.

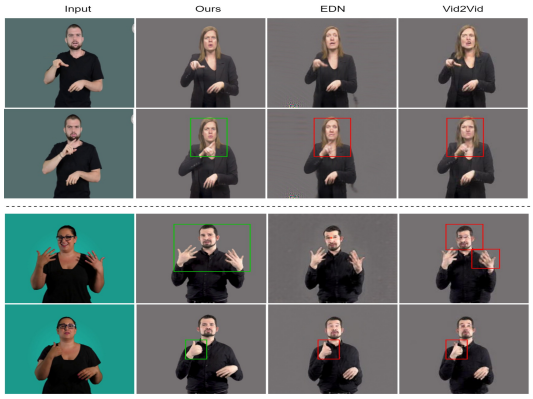

Figure 8. Visual comparison with EDN [16] and Vid2Vid [47] on different reenactment examples. We illustrate some erroneous results with red boxes and some successful examples of preserving the original mouth patterns and handshapes with green boxes. Please zoom in for details and refer to Suppl. Video [1].

As reported in the main paper, our method performs better in preserving the source signer's facial expressions and handshapes without distorting the characteristics of the specific identity. Lastly, in Fig. 9, we extend the qualitative results of cycle reenactment presented in the main paper by providing more comparisons of the various approaches in the $Female \rightarrow Male \rightarrow Female$ cycle reenactment experiment.

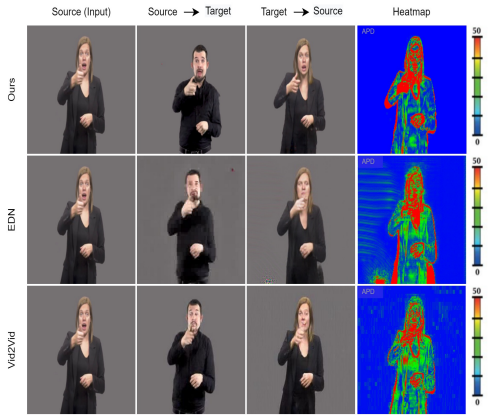

Figure 9. Cycle reenactment comparisons with EDN [16] and Vid2Vid [47]. From left to right: source actor, intermediate-target actor, original source actor driven by the manipulated target actor in the column before, and per-pixel differences between the first and third column.

F. 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.