Proc. 14th IEEE Image, Video, and Multidimensional Signal Processing Workshop (IVMSP 2022), Nafplio, Greece, June 2022.

Cartoonized Anonymization of Sign Language Videos

Christina O. Tze[1], Panagiotis P. Filntisis[1], Anastasios Roussos[2] and Petros Maragos[1]
[1]School of ECE, National Technical University of Athens, 15773 Athens, Greece
[2]Institute of Computer Science, Foundation for Research & Technology - Hellas (FORTH), Greece

***Abstract*—In this paper, we propose a novel method for the anonymization of sign language footage: given a source RGB video of a sign language user, we conceal the identity of the original signer by reproducing the video using animated cartoon characters. Our method pays particular attention to the cues that are important for sign language communication, transferring the motion and articulation of the hands, upper body and head of the real signer to the cartoon character in a faithful manner. To effectively capture these cues, 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. Our system first extracts the skeleton pose sequence from the input video as well as the cartoon's skeleton from its reference figure. The extracted skeletons are then fed into our skeleton retargeting algorithm, which combines the bone lengths from the cartoon character with the pose information from the human signer. The recombined parameters are then used as input to a recursive kinematic tree-based algorithm, which retargets the input skeleton pose sequence to the cartoon's skeleton. Finally, the reproduced frames of the signing cartoon are generated from the retargeted skeleton pose sequence. To the best of our knowledge, our method is the first to implement video reproduction using cartoon characters as a solution to the challenging task of sign language video anonymization. We conduct qualitative evaluations to demonstrate the effectiveness of our approach and the promising results that we obtain.**

I. Introduction

It is estimated that tens of millions of Deaf worldwide use Sign Language (SL) as their native language [1]–[4]. At the same time, most of them have limited reading/writing skills in the spoken language, which for them is a foreign language with a fundamentally different grammatical structure. This puts them at a severe disadvantage in many contexts, including social life, education, work, usage of computers and the Internet. SL technologies can support the Deaf community by building systems that facilitate their communication with the rest population [5]. 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 [6]–[11], Translation [12]–[14] and Production [15]–[17].

Deep learning approaches require the availability of large-scale SL corpora which is very limited due to participants' concerns over privacy and video misuse [18]. 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 in 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 [19]: 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 [4], 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 to 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. Video anonymization methods can be divided according to whether they conceal all or part of a video, or they reproduce a video [20]. Regarding previous works on concealing, the techniques of blackening [21] and pixelation [22] have been used to protect third-parties anonymity by suppressing references to people or places [20]. In particular, Bleicken et al. [21] use one or more black rectangles to anonymize parts of the video data from the Public DGS Corpus where people's names are mentioned. In [22] the author inspects the video recordings of a small British SL Corpus for any signs that refer to third-party data. Then, throughout the duration of these signs, a pixellation filter is applied to the signer's hands and mouth to remove any personal data [20], [22]. Bragg et al. [18] address the privacy concerns by using greyscale and animoji filters to conceal the identity of the participants. While these methods can increase people's willingness to contribute to SL datasets, they have a negative impact on sign comprehension and imply an inevitable loss of information which in turn can deteriorate the performance of trained models. Other methods guarantee total anonymity of the signers by reproducing the videos using actors. However, such procedures are labour-intensive and require a significant amount of time and equipment [20]. Saunders et al. [23] propose a deep learning framework for the generation of photo-realistic retargeted videos of the source signer, using novel synthesized human appearances instead of the original signer appearance and building upon recent prior work on photo-realistic SL production using Generative Adversarial Networks (GANs) [15], [24]. Even though this method opens new pathways towards photo-realistic SL synthesis, the synthesized frames include artifacts and the synthesized human appearances are not always convincing as being real, having

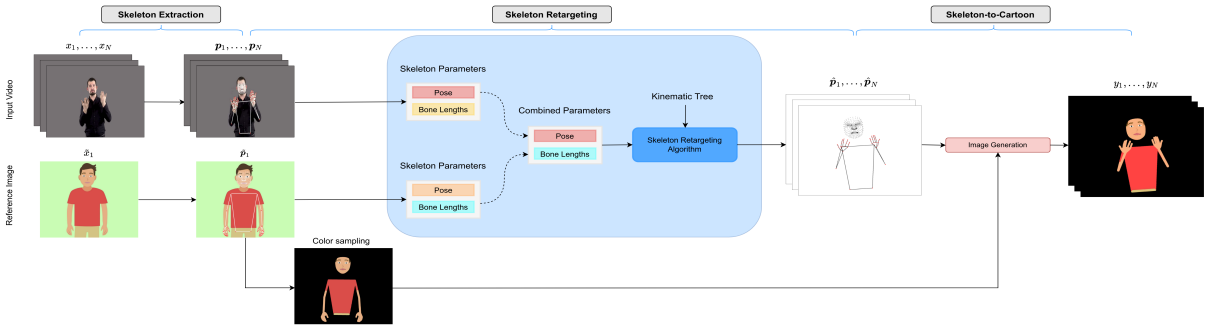

Fig. 1. Overview of the proposed pipeline: First, we extract the skeleton pose sequences from the input sign language video and the cartoon's reference image. These are then fed into our skeleton retargeting algorithm which combines the pose information from the original signer with the bone lengths from the cartoon character and generates the ordered sequence of retargeted skeletons. Finally, we generate every frame of the reproduced cartoon video.

the risk of "falling" into the so-called "*uncanny valley*" [25]. Especially in cases where the synthesized videos are produced for other humans to watch (e.g. the aforementioned case of anonymized online participation), this can have a negative effect in their engagement and solutions that do not target photo-realism might be preferable.

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 [26], [27]. OpenPose [28] is the first and one of the most popular bottom-up approach for multi-person human pose estimation. MediaPipe [29] 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, hand tracking and high-fidelity pose tracking. Unlike OpenPose, MediaPipe infers 3D landmarks rather than 2D. 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.

SL video anonymization tackled by this paper is effectively a special case of the more general problem of **motion retargeting**, which involves exploiting human pose estimation to transfer the pose sequence from a source to any target character. With the advent of deep learning, new methods of motion retargeting have been introduced to robustly solve this challenging problem [30]–[37]. Chan et al. [30] propose a simple yet effective method that generates temporal coherent video results and is capable of handling different body structures. Aberman et al. [31] train 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. Similarly to [31], Yang et al. [32] perform 2D motion retargeting by combining the extracted motion from the source sequence with the extracted structure from the target sequence. These methods all use 2D label maps, which are then fed into a pre-trained image-to-image generator [38], to render the retargeted video frames. Despite their particularly promising results, they do not focus on transferring the fine details of hand articulation, which makes them not suitable for SL anonymization since they cannot retain some of the most important cues of SL communication [39].

In this paper, we are the first to use cartoon avatars for SL anonymization. We argue that this offers advantages as compared to previous approaches, since it avoids falling into the "uncanny valley" and is more likely to increase the engagement of the human viewers. In addition, when cartoon-based anonymization is used, the majority of deaf and hard-of-hearing participants seem to be more willing to contribute in SL corpora, as evidenced by a recent user study exploring signers' privacy concerns [18]. Therefore, given the reliable retargeting results that we achieve, our method is anticipated to contribute in the solution to the SL data scarcity problem. Furthermore, in contrast to the recent prior art, our method is especially designed to be lightweight and capable of efficiently running on just a CPU, achieving near real-time performance and without requiring specialized hardware such as GPU. We present results of our method as applied to several different signers, demonstrating the particularly promising results that we obtain under challenging continuous signing and across different genders and body structures. Finally, we introduce an online demo that further showcases the potential and capabilities of our cartoon-based retargeting approach.

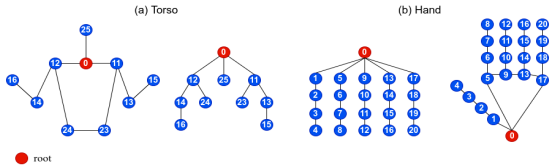

Fig. 2. The skeleton models (left) and the corresponding kinematic trees (right) for the torso and hand regions.

II. Methodology

Overview: Our method takes a source sign language video and a reference image of a cartoon character as inputs and retargets the human signer's performed manual (handshape, hand orientation, location, and movement) and non-manual signs (head, upper body motions) to the input cartoon character. The main steps are visualized in the pipeline of Fig. 1.

Formally, given a source SL video $X = (x_1, ..., x_N)$ and a reference cartoon image $\tilde{x}_1$, it generates a new video $Y = (y_1, ..., y_N)$ of an animated cartoon, where the signing of the original signer is retargeted to the cartoon character. The proposed pipeline consists of three main components: a) skeleton extraction, b) skeleton retargeting, and c) skeleton-to-cartoon image generation, presented in the following sections.

A. Skeleton Extraction

We first extract the skeleton pose sequence from the source SL video using both the MediaPipe (MP) Pose and Holistic

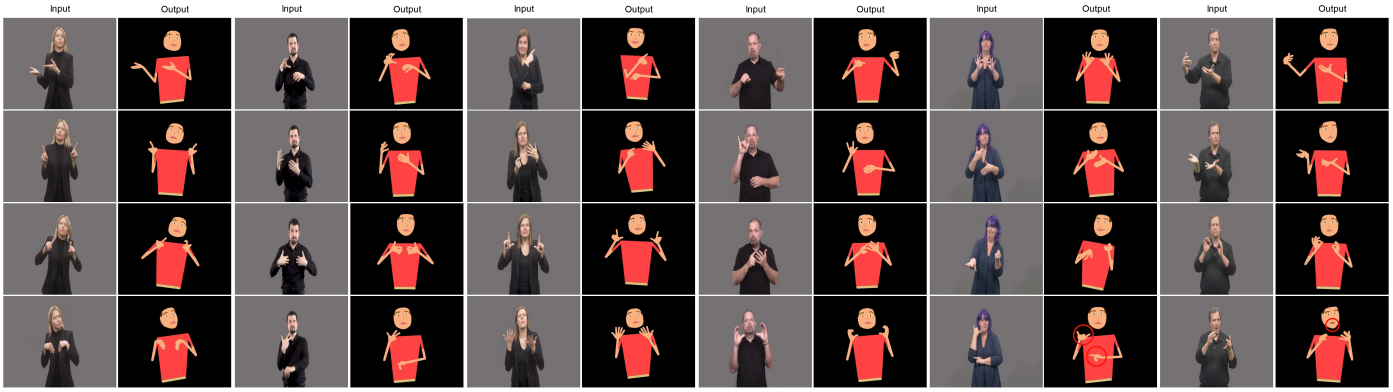

Fig. 3. Retargeting results. The first column of each section shows the original signer and the second column shows the retargeted frame of the cartoon character. Examples of failure cases are highlighted in red. Please zoom in for details. Video versions of the results can be found at: https://youtu.be/SXFjdMlKEhE

modules, since 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 in the form of image pixel coordinates. Specifically, we exclude those corresponding to the face (except the nose), hands, and lower body. We use MP Holistic to track the face and hands, inferring 510 landmarks in total. Thus, every frame $i \in [1, N]$ is represented by a pose vector $\boldsymbol{p}_i = (\boldsymbol{l}_1, ..., \boldsymbol{l}_K)$ of 3D landmarks coordinates $\boldsymbol{l}_j = (l_{j_x}, l_{j_y}, l_{j_z})$, where $K = 519$ is the number of tracked joints. After processing all N frames of the original input video, the sequence of skeleton poses $\boldsymbol{P} = (\boldsymbol{p}_1, ..., \boldsymbol{p}_N)$ is extracted. Similarly, the cartoon's skeleton pose, $\tilde{\boldsymbol{p}}_1 = (\tilde{\boldsymbol{l}}_1, ..., \tilde{\boldsymbol{l}}_K)$, is extracted from its reference image, $\tilde{x}_1$, in the exact same way as we did for the original signer. For clarity, we denote the joint in the middle of the shoulders as $\boldsymbol{l}_{root}$ and $\tilde{\boldsymbol{l}}_{root}$ for the original signer and cartoon character, respectively. Also, we will hereafter refer to torso as the human body part apart from the head, hands and legs.

B. Skeleton Retargeting

We propose a skeleton retargeting algorithm that generates a new skeleton sequence given the original signer's sequence and the cartoon character's skeleton representation (see Sec. A). Our method preserves the cartoon's skeletal structure and is based on the use of kinematic trees to separate bone length information from skeleton pose. The skeleton models and the corresponding kinematic trees for the torso and hand regions are shown in Fig. 2. For the construction of the head kinematic tree we assume that the root joint (i.e., nose) is connected to the remaining 467 facial landmarks and that there are no other edges. The retargeting process begins with the estimation of the cartoon's bone lengths, which is followed by the execution of the proposed 3D-based skeleton retargeting algorithm.

1) Bone Lengths Estimation: To apply the extracted pose sequence of the original signer to the cartoon character which may have very different bone lengths and proportions, we must first estimate its bone lengths. These are calculated from the joint positions in the cartoon's skeleton representation using the standard 3D euclidean distance formula. Let L_j denote the length of the bone that connects joint j with its parent at the kinematic tree (see Fig. 2). Then, the following sets of head, torso and left hand bone lengths are calculated:

$$L_{head} = (L_1, ..., L_{467}) \tag{1}$$

$$L_{left\ hand} = (L_1, ..., L_{20}) \tag{2}$$

$$L_{torso} = (L_{11}, L_{13}, L_{15}, L_{23}, L_{25}) \tag{3}$$

where we made the following symmetry assumptions:

$$L_{right\ hand} = L_{left\ hand} \tag{4}$$

$$L_{torso} : L_{12} = L_{11}, L_{14} = L_{13}, L_{16} = L_{15}, L_{24} = L_{23} \tag{5}$$

2) Retargeting Algorithm: Let $\boldsymbol{l}_j$ and $\boldsymbol{l}_{parent}$ be the 3D coordinates of joint j and its parent at a random frame of the input SL video. Also, let $\boldsymbol{D}_j$ denote the joint's direction with regard to its parent at the same frame. The following algorithm is applied to each of the four kinematic trees and recursively calculates the 3D coordinates, $\hat{\boldsymbol{l}}_j$, of their joints for every frame of the reproduced cartoon video:

INPUT: $\boldsymbol{P'} = (\boldsymbol{p'}_1, ..., \boldsymbol{p'}_N) \,|\, \boldsymbol{p'}_i = \{\boldsymbol{l}_j \in \boldsymbol{p}_i \wedge j \in \text{body part}\}$
for every $\boldsymbol{p'}_i \in \boldsymbol{P'}$ **do**
  for every $\boldsymbol{l}_j \in \boldsymbol{p'}_i$ **do**
    if joint = root **then**
      $\boldsymbol{T}_j \leftarrow 0$
    else
      $\hat{\boldsymbol{l}}_j \leftarrow \boldsymbol{l}_{parent} + \boldsymbol{D}_j \cdot L_j$
      $\boldsymbol{T}_j \leftarrow \hat{\boldsymbol{l}}_j - \boldsymbol{l}_j$
    end if
  end for
  for every joint **do**
    if joint = root **then**
      $\hat{\boldsymbol{l}}_j \leftarrow \boldsymbol{l}_j + \boldsymbol{T}_j$
    else
      $\boldsymbol{T}_j \leftarrow \boldsymbol{T}_j + \boldsymbol{T}_{parent}$
      $\hat{\boldsymbol{l}}_j \leftarrow \boldsymbol{l}_j + \boldsymbol{T}_j$
    end if
  end for
end for

Since the final ordered sequence of retargeted skeletons must match the cartoon's upper body motion, additional translations are needed to integrate the four separate skeletons (head, torso, left hand, right hand) into one and adjust its overall position. First, every head skeleton in the generated ordered sequence, is attached to the nose joint of the corresponding torso skeleton. Similarly, the left hand and right hand skeletons are attached to the left and right wrist joints, respectively. Then, a global translation is applied to the unified skeletons, in order to align them with the cartoon's root reference joint, $\tilde{\boldsymbol{l}}_{root}$ (see Sec. A).

C. Skeleton-to-Cartoon Image Generation

To generate the frames of the reproduced cartoon video, we combine the pose information from the ordered sequence of retargeted skeletons with the color sampling from the cartoon's reference image. The output frames are represented as 8-bit RGB images where each joint is plotted as a disk with fixed radius and assigned a fixed color. For a given joint j, its color is the same as the RGB value of the reference cartoon image at pixel coordinates $(\tilde{l}_{j_x}, \tilde{l}_{j_y})$. Moreover, bone interpolation is used as a data augmentation technique to include more skeleton joints. The color and number of interpolated points along a certain bone are both fixed. For their coloring, we first use linear interpolation to generate their pixel coordinates in the cartoon's reference image. Then, each interpolated point is given a fixed color in the same way that the tracked joints were. This process is applied to every bone in the MediaPipe's Pose and Hand connections. For the face, we are based on the triangular mesh representation used by MediaPipe and apply a coloring of the mesh's triangles, considering the transformed vertex positions after retargeting.

III. Experiments

In this section we conduct qualitative evaluations of our proposed pipeline using multiple signers and source videos. Please refer to the following link for a supplementary video with the results: https://youtu.be/SXFjdMlKEhE.

To validate the effectiveness of our method we have collected a small dataset of online American Sign Language (ASL) videos from the Canadian Transportation Agency's YouTube channel. These include 6 SL users (3 males and 3 females) that vary in terms of their body structure. Then, we proceeded to apply our method on the collected videos. Figure 3 shows our method's retargeting results for multiple indicative frames from the input SL videos. As it can be seen, our method succeeds in reliably transferring a wide range of motions and articulations of the real signer to the cartoon character. In particular, it is capable of successfully retargeting the head, torso and hand movements as well as the handshape and hand orientation. An important observation is also the fact that our method mitigates body structure biases and works reliably for a variety of different body types, generating a consistent skeleton sequence with the same body structure as the cartoon's reference skeleton. On the other hand, we also observe that our method does not always transfer the facial expressions and mouthings reliably and some times fails to preserve the correct spatial relation between the hands and the face in the signing space. Finally, we also show a more detailed view in Figure 4 where we also provide intermediate results of our method, including the original signer's extracted skeleton and the cartoon's retargeted skeleton.

Online demo: We have also implemented a lightweight near real-time version of our method, in which we render only the tracked facial landmarks and not the full face. Using this lightweight version we have created an online web-demo which can be found in https://robotics.ntua.gr/wp-content/demos/sign-avatar/

Runtime analysis: We include in Table I a run-time analysis of our method on a conventional CPU processor (Intel i3-4160 CPU). Note that neither skeleton tracking nor the retargeting algorithm require a GPU to run in real time. As we can see from the Table, the most computationally intensive step is the cartoon image generation. This computational bottleneck is significantly reduced in the lightweight version of our method, since it avoids the expensive rendering of the full face.

Step of the pipeline	Lightweight version	Full version
Face tracking & retargeting	12.3	12.3
Torso tracking & retargeting	0.8	0.8
Hands tracking & retargeting	0.7	0.7
Cartoon image generation	29.2	1464.4
Total	43.0	1478.2

TABLE I. Run time analysis (time in ms per frame) of the main steps of our cartoonized anonymization method on a conventional CPU processor.

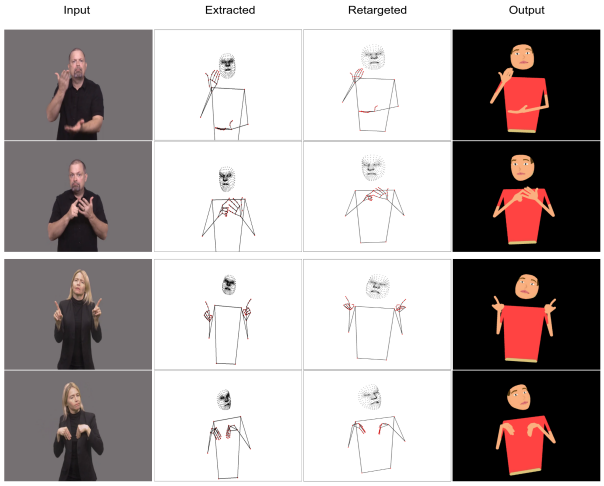

Fig. 4. Intermediate and retargeted results. From left to right: input frame, extracted skeleton, retargeted skeleton and reproduced frame.

IV. Conclusion

We proposed a novel approach for anonymizing SL footage based on the reproduction of their video data using a reference cartoon character. Our method can also be useful for other applications that require human to avatar motion retargeting such as video gaming, AR/VR, tele-presence and performance capture for animation movies. We have presented examples of applying our method to a multitude of signers, in lengthy videos of continuous signing. The presented results, supplementary video and online demo demonstrate that our method faithfully retargets the head, torso and hand movements from any signer to the animated cartoon, working reliably and consistently across signers of different genders and body structures. On the other hand, the results have also shown that our approach has some limitations, mostly in terms of transferring the facial actions and preserving the spatial relation between the hands and face, which can pave the way for future work. In the future, we also plan to conduct detailed quantitative evaluations and user studies involving native SL users to assess the performance of our method and compare it with other approaches to SL anonymization.

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.

REFERENCES

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

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

[3] 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.

[4] E. Antonakos, A. Roussos, and S. 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)*, vol. 1. IEEE, 2015, pp. 1–7.

[5] I. Papastratis, C. Chatzikonstantinou, D. Konstantinidis, K. Dimitropoulos, and P. Daras, "Artificial intelligence technologies for sign language," *Sensors*, vol. 21, no. 17, p. 5843, 2021.

[6] A. Roussos, S. Theodorakis, V. Pitsikalis, and P. Maragos, "Dynamic affine-invariant shape-appearance handshape features and classification in sign language videos," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1627–1663, 2013.

[7] S. Theodorakis, V. Pitsikalis, and P. Maragos, "Dynamic–static unsupervised sequentiality, statistical subunits and lexicon for sign language recognition," *Image and Vision Computing*, vol. 32, no. 8, pp. 533–549, 2014.

[8] O. Koller, N. C. Camgoz, H. Ney, and R. 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*, vol. 42, no. 9, pp. 2306–2320, 2019.

[9] N. C. Camgoz, O. Koller, S. Hadfield, and R. 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*, 2020, pp. 10 023–10 033.

[10] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial-temporal multi-cue network for continuous sign language recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 13 009–13 016.

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

[12] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7784–7793.

[13] K. Yin and J. Read, "Better sign language translation with stmc-transformer," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 5975–5989.

[14] A. Voskou, K. P. Panousis, D. Kosmopoulos, D. N. Metaxas, and S. 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*, 2021, pp. 11 946–11 955.

[15] S. Stoll, N. C. Camgoz, S. Hadfield, and R. Bowden, "Text2sign: Towards sign language production using neural machine translation and generative adversarial networks," *International Journal of Computer Vision*, vol. 128, no. 4, pp. 891–908, 2020.

[16] B. Saunders, N. C. Camgoz, and R. Bowden, "Progressive transformers for end-to-end sign language production," in *European Conference on Computer Vision*. Springer, 2020, pp. 687–705.

[17] ——, "Adversarial training for multi-channel sign language production," *arXiv preprint arXiv:2008.12405*, 2020.

[18] D. Bragg, O. Koller, N. Caselli, and W. Thies, "Exploring collection of sign language datasets: Privacy, participation, and model performance," in *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, 2020, pp. 1–14.

[19] S. Lee, A. Glasser, B. Dingman, Z. Xia, D. Metaxas, C. Neidle, and M. 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*, 2021, pp. 1–13.

[20] A. Isard, "Approaches to the anonymisation of sign language corpora," in *Proceedings of the LREC2020 9th workshop on the representation and processing of sign languages: Sign language resources in the service of the language community, technological challenges and application perspectives*, 2020, pp. 95–100.

[21] J. Bleicken, T. Hanke, U. Salden, and S. Wagner, "Using a language technology infrastructure for german in order to anonymize german sign language corpus data," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 3303–3306.

[22] L. A. Rudge, "Analysing british sign language through the lens of systemic functional linguistics," Ph.D. dissertation, University of the West of England, 2018.

[23] B. Saunders, N. C. Camgoz, and R. 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)*. IEEE, 2021, pp. 1–8.

[24] ——, "Everybody sign now: Translating spoken language to photo realistic sign language video," *arXiv preprint arXiv:2011.09846*, 2020.

[25] M. Mori, K. F. MacDorman, and N. Kageki, "The uncanny valley [from the field]," *IEEE Robotics & Automation Magazine*, vol. 19, no. 2, pp. 98–100, 2012.

[26] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," *Computer Vision and Image Understanding*, vol. 192, p. 102897, 2020.

[27] C. Zheng, W. Wu, T. Yang, S. Zhu, C. Chen, R. Liu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *arXiv preprint arXiv:2012.13392*, 2020.

[28] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[29] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[30] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5933–5942.

[31] K. Aberman, R. Wu, D. Lischinski, B. Chen, and D. Cohen-Or, "Learning character-agnostic motion for motion retargeting in 2d," *arXiv preprint arXiv:1905.01680*, 2019.

[32] Z. Yang, W. Zhu, W. Wu, C. Qian, Q. Zhou, B. Zhou, and C. C. Loy, "Transmomo: Invariance-driven unsupervised video motion retargeting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5306–5315.

[33] K. Aberman, P. Li, D. Lischinski, O. Sorkine-Hornung, D. Cohen-Or, and B. Chen, "Skeleton-aware networks for deep motion retargeting," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 62–1, 2020.

[34] J. Lim, H. J. Chang, and J. Y. Choi, "Pmnet: Learning of disentangled pose and movement for unsupervised motion retargeting." in *BMVC*, vol. 2, no. 6, 2019, p. 7.

[35] R. Villegas, J. Yang, D. Ceylan, and H. Lee, "Neural kinematic networks for unsupervised motion retargetting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8639–8648.

[36] H. Jang, B. Kwon, M. Yu, S. U. Kim, and J. Kim, "A variational u-net for motion retargeting," in *SIGGRAPH Asia 2018 Posters*, 2018, pp. 1–2.

[37] S. Guan, S. Wen, D. Yang, B. Ni, W. Zhang, J. Tang, and X. Yang, "Human action transfer based on 3d model reconstruction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8352–8359.

[38] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

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