Neural Sign Reenactor: Deep Photorealistic Sign Language Retargeting

Christina O. Tze[1], Panagiotis P. Filntisis[1], Athanasia-Lida Dimou[2], Anastasios Roussos[3], Petros Maragos[1]
[1]NTUA, Greece [2]Athena RC, Greece [3]ICS-FORTH, Greece

Demo Video

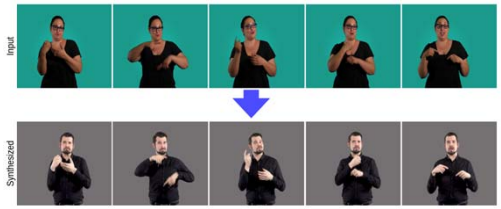
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

1. Introduction

- **Background:** The generation of synthetic Sign Language (SL) videos is historically tackled using computer-generated avatars [1].
- **Problem:** Low level of realism in terms of the avatars' appearance and motion reduces the plausibility and engagement of users with such technologies.
- **Our solution:** Novel neural rendering pipeline that generates highly realistic human actor videos.

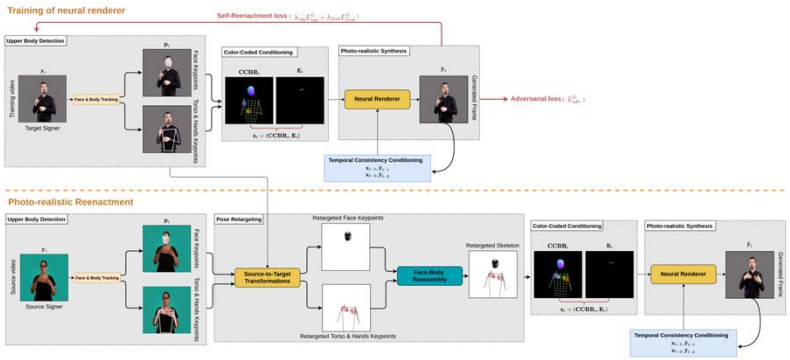
2. Proposed Method

- A **neural rendering pipeline** for transferring the body and facial movements of a source actor to a target one.
- It is applied to the challenging case of SL videos.
- It can be particularly beneficial for SL Anonymization [2], SL Production [1], and reenactment of full-body activities [3].



- Our contributions can be summarized as follows:
  1) Effective **combination** of **two different body trackers**
  2) Novel scheme for **conditioning** the **neural renderer**
  3) Novel **pose retargeting** step
  4) Detailed **qualitative** and **quantitative evaluations** and **user studies**

3. Pipeline



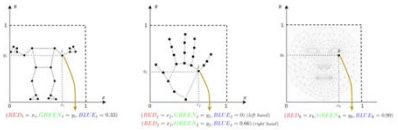
4. Reenactment Methods

- **Self-reenactment:** Used during training. The source actor coincides with the target one.
- **Reenactment:** Main functionality of our method. The source and target actors are different from one another.
- **Cycle reenactment:** Used for quantitative comparisons. The movements of a source actor are transferred to a target and then back to the same source.

5. Color-coded Conditioning

We condition our video rendering network to:

1) Color-coded body representations: These are generated using our novel color-coding scheme.

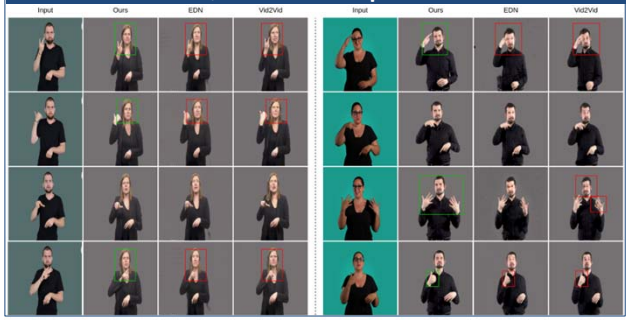


2) **Eye gaze images:** We tint the contour landmarks white and the pupils red.

6. Comparison with other methods

We compare our method with Everybody Dance Now (EDN) [4] and Video-to-Video Synthesis (Vid2Vid) [5].

7. Qualitative Comparisons



8. Quantitative Comparisons

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

9. User Studies

- **Realism Study:**

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%

- **Sign Classification Study:**

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

References

[1] Rastgoo et al. **Sign Language Production: A Review**. CVPR 2021
[2] Saunders et al. **AnonySIGN.** FG 2021
[3] Liu et al. **Neural Rendering and Reenactment of Human Actor Videos**. TOG 2019
[4] Chan et al. **EDN**. ICCV 2019
[5] Wang et al. **Vid2Vid**. NeurIPS 2018