ECCV'20 ONLINE
23-28 AUGUST 2020

16TH EUROPEAN CONFERENCE ON
COMPUTER VISION

WWW.ECCV2020.EU

SIGNSYNTH

Data-Driven Sign Language Video Generation

Stephanie Stoll, Simon Hadfield, Richard Bowden

UNIVERSITY OF SURREY

ECCV'20 ONLINE
23-28 AUGUST 2020

SIGNSYNTH

Why is sign language generation a challenge?

- Sign languages are an integral part of the lives of the Deaf and Hard of Hearing.
- Translating spoken languages to sign languages is difficult as sign languages use multiple channels to convey meaning.
- Sign languages are not limited to hands and their shape, but also: facial expression, body pose, hand position, mouthings.
- However sign languages can be transcribed to gloss form (each gloss is the closest word representation of a given sign).

e.g. What is the weather today?

TODAY WEATHER WHAT

UNIVERSITY OF SURREY
CVSSP
Centre for Vision, Speech and Signal Processing

ECCV'20 ONLINE
23-28 AUGUST 2020

SIGNSYNTH

Our approach

We propose using this very simplified version to generate **information-rich, life-like sign language** content using advancements in machine learning.

ECCV'20 ONLINE
23-28 AUGUST 2020

SIGNSYNTH

Literature review

Traditionally avatars are used to generate synthetic sign language content which rely on manually annotated, and/or motion capture data, making their scalability questionable.

Recently, Stoll et al.[1a&b] introduced an approach that uses a sequence-to-sequence model to translate spoken language text to gloss. Whilst scalable, there are two main issues with this work:

1. The lookup-table restricts this approach in terms of co-articulation between signs, and expressiveness.
2. The GAN in this work is only capable of low-resolution output.

[1a] Stoll, S., Camgoz, N.C., Hadfield, S., Bowden, R., Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks. Int J Comput Vis 128, 891–908 (2020).

[1b] Stoll, S., Camgoz, N.C., Hadfield, S., Bowden, R., Sign Language Production using Neural Machine Translation and Generative Adversarial Networks, In Proceedings of the British Conference on Machine Vision (BMVC), BMVA Press, 2018.

UNIVERSITY OF SURREY
CVSSP
Centre for Vision, Speech and Signal Processing

ECCV'20 ONLINE
23-28 AUGUST 2020

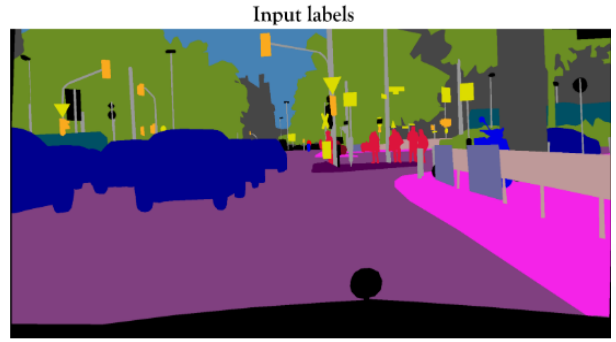
SIGNSYNTH

Literature review

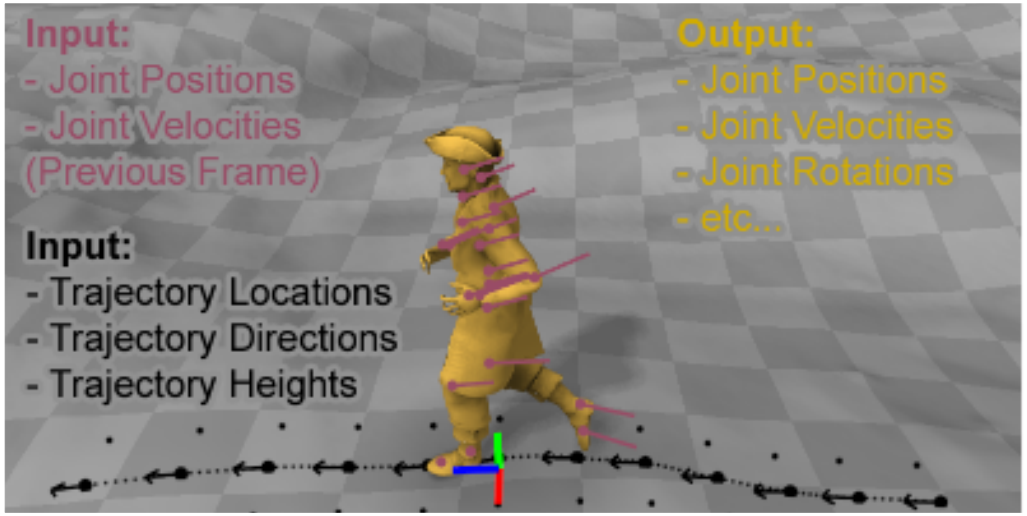
We base our solution on advancements in **Motion Graphs**, a popular concept in **Computer Graphics** and **image-to-image translation**, improving those points significantly.

1

2



3



- Zhang, H., Starke, S., Komura, T., Saito, J.: Mode-adaptive neu- ral networks for quadruped motion control. ACM Trans. Graph. 37(4), 145:1–145:11 (2018).[1]
- Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018) [2]
- Holden,D.,Komura,T.,Saito,J.: Phase-functioned neural networks for character control. ACM Trans. Graph. 36(4), 42:1–42:13 (2017). [3]

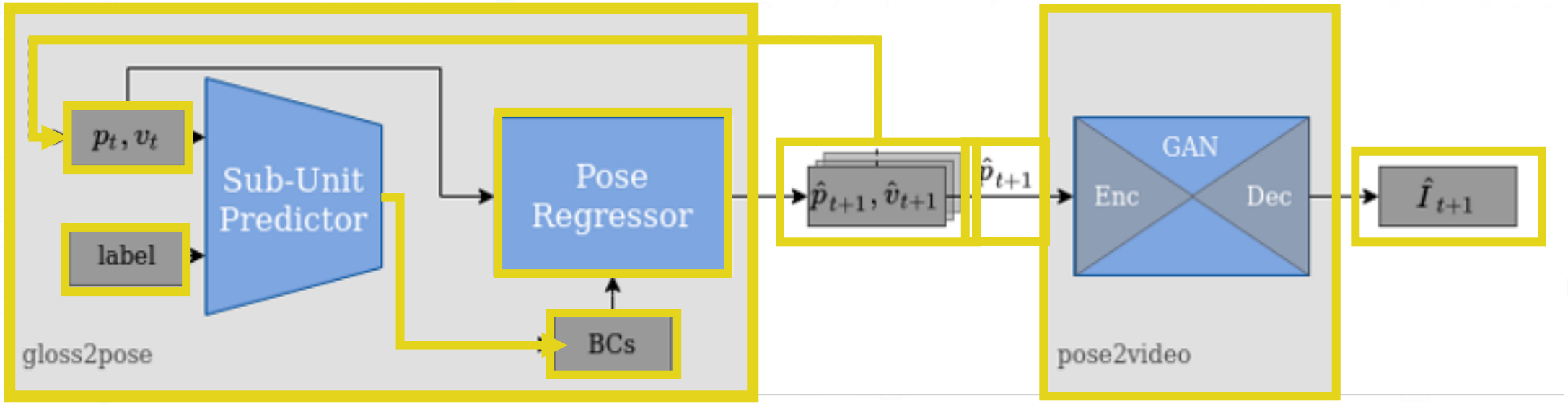
UNIVERSITY OF SURREY
CVSSP
Centre for Vision, Speech and Signal Processing

ECCV'20 ONLINE
23-28 AUGUST 2020

SIGNSYNTH

Methodology

Our approach to generating sign language video from glosses works in two stages. First glosses are translated into a human pose sequence by our gloss2pose network. The acquired poses are then used to condition a generative network called pose2video.



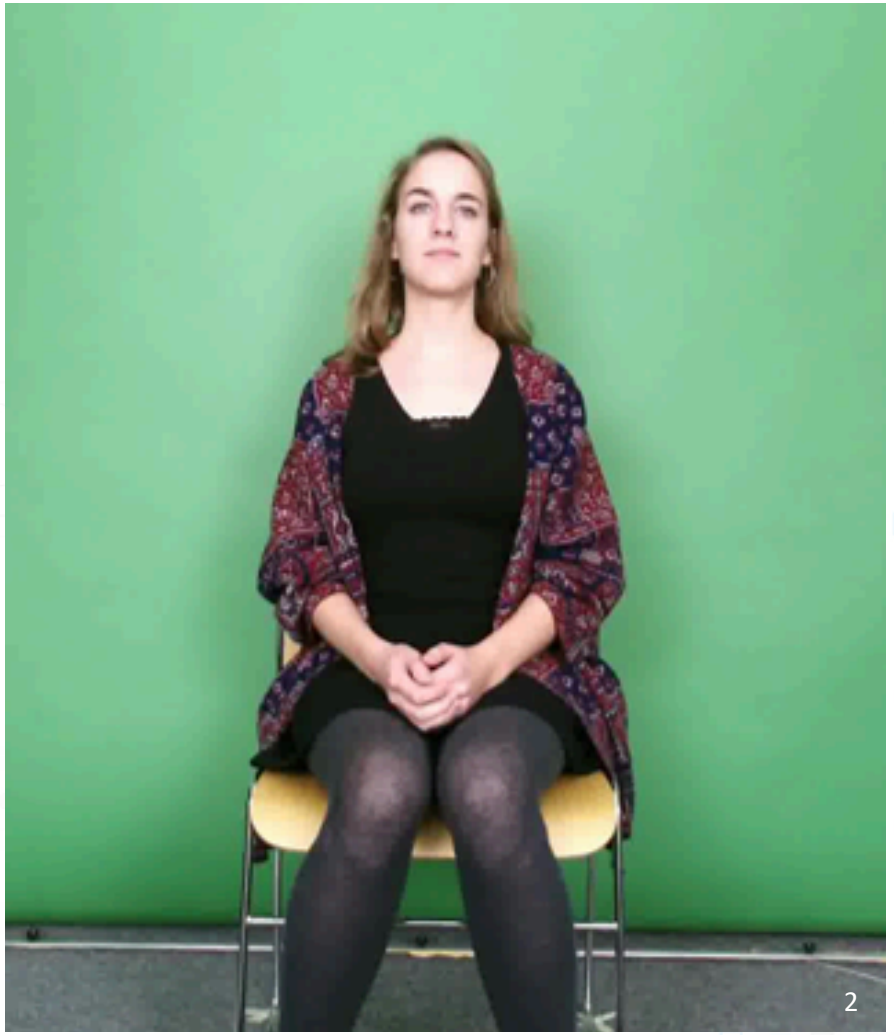
UNIVERSITY OF SURREY
CVSSP
Centre for Vision, Speech and Signal Processing

SIGNSYNTH

Experiments & Results – Dataset

- **We use the SMILE Sign Language Assessment Dataset for training and testing** which consists of 42 signers performing 105 signs in isolated form in Swiss-German Sign Language (DSGS).

- **We extract 2D human pose estimations from the video data**, using OpenPose for the upper body, face and hands.

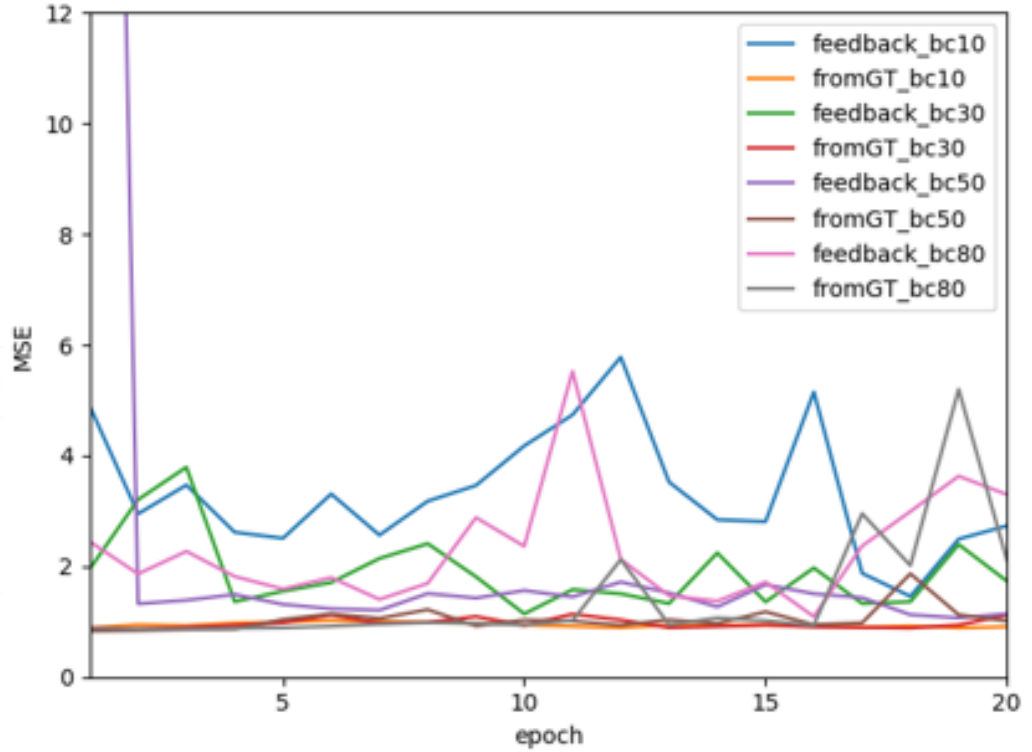


- Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei and Y. A. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," in IEEE Transactions on Pattern Analysis and Machine Intelligence.[1]
- SMILE Sign Language Data Set, in 11th edition of the Language Resources and EvaluationConference (LREC),2018, S. Ebling, N. C. Camgz, P. Boyes Braem, K. Tissi, S. Sidler-Miserez, S.Stoll, S. Hadfield, T. Haug, R. Bowden, S. Tornay, M. Razavi, M. Magimai-Doss.[2]

SIGNSYNTH

Experiments & Results – Gloss2Pose

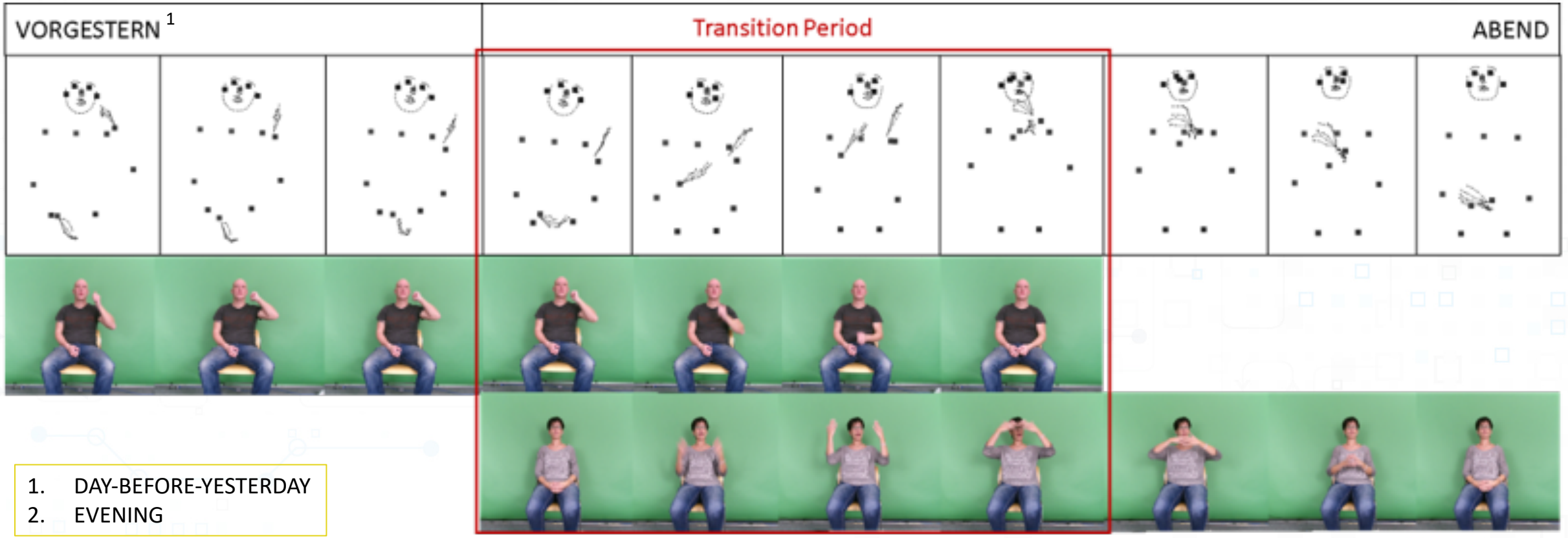
- Most stable performance was achieved with 50 blending coefficients.
- Our subunit-predictor can dissect sign motions into sub-motions.
- We next compare the MSE of our best performing network against the lookup table approach of Stoll et al.
- Our network outperforms the SotA by a factor of 18.



	LUT (Stoll et al)	G2p (ours)
MSE	19.2886	**1.0673**

SIGNSYNTH

Experiments & Results – Gloss2Pose



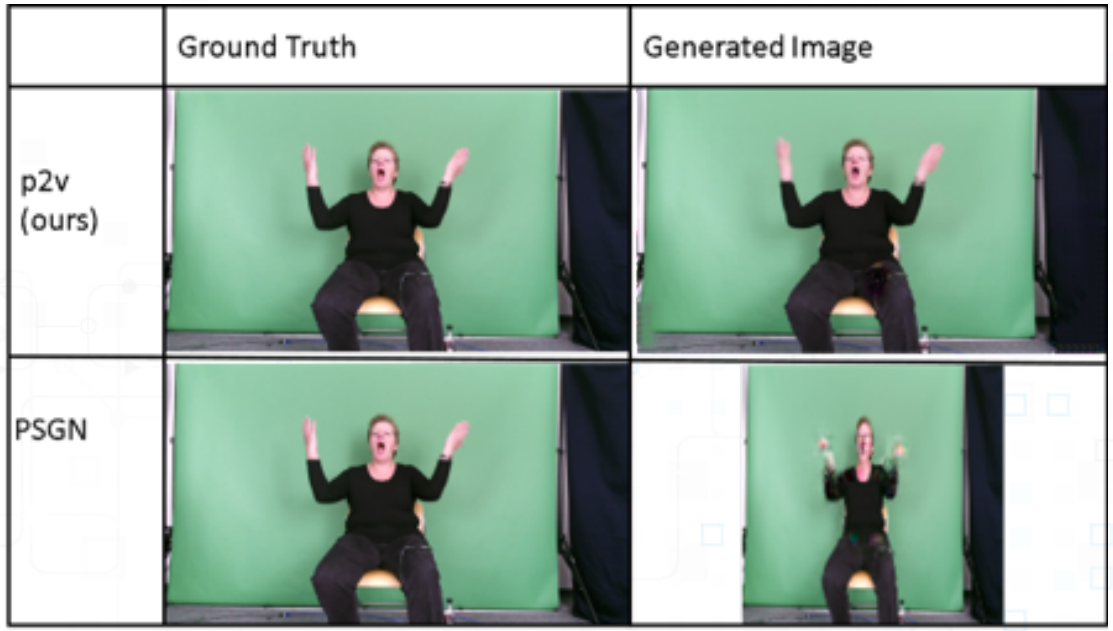
1. DAY-BEFORE-YESTERDAY
2. EVENING

UNIVERSITY OF SURREY
CVSSP
Centre for Vision, Speech and Signal Processing

SIGNSYNTH

Experiments & Results – Pose2Video

- We evaluate the performance of pose2video against the Pose-Conditioned Sign Generation Network (PSGN) by Stoll et al.
- We evaluate the image generation using SSIM, PSNR and MSE.
- PSGN scores marginally higher for SSIM.
- In contrast pose2video beats PSGN by a large margin for both PSNR and MSE.



	PSGN (Stoll et al)	P2v (ours)
SSIM	**0.9434**	0.9428
PSNR	24.7248	**29.0181**
MSE	226.2474	**86.0553**

SIGNSYNTH

Experiments & Results – Synthetic Sign Video Generation from Gloss

We devise two metrics to evaluate the quality of generated sign language videos:

1. A confidence score that assesses the level of detail and human characteristics of the generated videos.
2. A distance measure that assesses how closely generated sign videos resemble the ground truth videos for that gloss.

We refer to the paper for details on both metrics.

UNIVERSITY OF SURREY
CVSSP
Centre for Vision, Speech and Signal Processing

SIGNSYNTH

Experiments & Results – Synthetic Sign Video Generation from Gloss

- For our first metric we assess confidences on specific body parts, such as the hands and the face, for both our approach and Stoll et al.

- For our second experiment we analyse the behaviour of keypoints over time.

- We report the mean distance between clusters of signs and synthetic samples.

	C_{pose}	C_{face}	C_{handl}	C_{handr}	C_{total}
Stoll et al. [34]	0.499	0.000	0.026	0.025	0.138
SignSynth	**0.791**	**0.766**	**0.120**	**0.266**	**0.485**

	SignSynth			Stoll et al. [34]			Reference		
Clusters	ABEND	ABER	ERZ	ABEND	ABER	ERZ	ABEND	ABER	ERZ
ABEND	**12.51**	14.49	16.17	19.82	**16.36**	19.29	**12.28**	14.61	14.69
ABER	15.58	**14.72**	17.02	18.13	**16.29**	19.35	14.86	**13.34**	15.43
ERZ	15.51	**15.32**	15.93	20.68	**17.04**	19.78	14.70	15.03	**13.54**

UNIVERSITY OF SURREY
CVSSP
Centre for Vision, Speech and Signal Processing

ECCV'20 ONLINE
23-28 AUGUST 2020

SIGNSYNTH

Experiments & Results – Synthetic Sign Video Generation from Gloss

SIGNSYNTH
Conclusions

1. We presented a **novel approach** to Sign Language Production that only requires **minimal user input**.
2. We produce sign language video with a **natural variance** in speed, expressiveness, distinctive hand shape, and **non-manuals**.
3. We smoothly and **automatically transition** between glosses, without manually enforcing co-articulation.
4. We **surpass the SotA** for pose and video generation, as well as generating videos from gloss.
5. We developed **two new metrics** to assess the quality of sign language videos.

UNIVERSITY OF SURREY
CVSSP Centre for Vision, Speech and Signal Processing

ECCV'20 ONLINE
23-28 AUGUST 2020

SIGNSYNTH

Future work

In the future, we want to:

1. Extend our approach to 3D data
2. Address complete spoken language to sign language translation
3. Undertake a user study with members of the Deaf community.

UNIVERSITY OF SURREY
CVSSP
Centre for Vision, Speech and Signal Processing

Stephanie Stoll

Data-Driven Sign Language Video Generation

s.m.stoll@surrey.ac.uk

For any questions

Live session on the 28th of August
8.00 – 10.00 UTC+1

UNIVERSITY OF SURREY