INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Another Point of View on Visual Speech Recognition

Baptiste Pouthier[1,2], *Laurent Pilati*[1], *Giacomo Valenti*[1], *Charles Bouveyron*[2], *Frederic Precioso*[2]

[1]NXP Semiconductors, France
[2]Université Côte d'Azur, Inria, CNRS, LJAD, I3S, Maasai, France

{baptiste.pouthier, laurent.pilati, giacomo.valenti}@nxp.com,
{charles.bouveyron, frederic.precioso}@univ-cotedazur.fr

Abstract

Standard Visual Speech Recognition (VSR) systems directly process images as input features without any apriori link between raw pixel data and facial traits. Pixel information is smartly sieved when facial landmarks are extracted from pictures and repurposed as graph nodes. Their evolution through time is thus modeled by a Graph Convolutional Network. However, with graph-based VSR being in its infancy, the selection of points and their correlation are still ill-defined and often bound to aprioristic knowledge and handcrafted techniques. In this paper, we investigate the graph approach for VSR and its ability to learn the correlation between points beyond the mouth region. We also study the different contributions that each facial region brings to the system accuracy, proving that more scattered but better connected graphs can be both computationally light and accurate.
Index Terms: visual speech recognition, graph convolutional network, point cloud definition

1. Introduction

Machine learning advancements in the last decade turned challenging everyday life applications into reasonable and achievable objectives, and the Automatic Speech Recognition (ASR) community attention has moved increasingly towards low Signal-to-Noise Ratio scenarios. In this context, with video becoming more accessible, recent studies [1, 2, 3, 4, 5] were able to demonstrate the benefits of including visual content into the classic ASR pipeline. Facial traits can therefore be considered as speech features to the same extent Visual Speech Recognition (VSR), the task of transcribing spoken words by processing only the visual content, is considered a branch of ASR. Our paper focuses on challenging the data structure and input features of the relatively recent VSR pipeline, following previous work involving approaches like optical flow [6] or graph data structure [7, 8].

Deep learning-based VSR methods rely on decoding the visual information of the video source, focusing on the mouth region [1, 2, 3, 4, 5, 9]. The pertaining picture area is often located, cropped, and aligned using a face landmark detector, then the sequence is fed to convolutional and/or attention-based neural networks. This processing pipeline is allegedly redundant as the pixel information is processed twice, once to find the landmarks and again by the network in order to find the visual cues, making the whole process also ill-suited for embedded applications. Moreover, the information present in detected landmarks revealing potentially important face contours is ignored, as those are discarded after the mouth area is located [1, 2, 3, 4, 9]. Finally, using raw images as input could be prone to biases such as gender, age, skin tone, and illumination, reflecting the distribution of training data [10, 11].

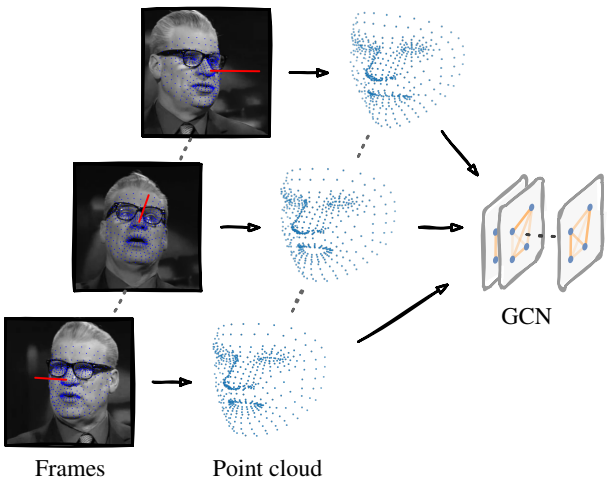


Figure 1: *Graph-based VSR system pipeline: landmark detection on video frames is applied to extract the point cloud. Following head pose estimation (red line) and normalization, the point cloud is processed by a GCN model outputting speech transcription.*

To overcome these issues, an alternative approach is to build graphs employing the extracted landmarks as a set of nodes. These graphs contain information well suited to represent the evolution of facial traits through time and can be processed by apt Graph Convolutional Networks (GCN) [12]. GCN have proven their efficiency in several applications [12, 13, 14], and have recently been applied to VSR in tandem with the standard visual information pipeline [7, 8]. The system described in [7] relies upon a handcrafted graph derived from the extracted facial landmarks fed to a Spatial Temporal GCN [15], while in [8] a set of adjacency matrices is learned to capture the complex relationship between such points using an Adaptive Graph Convolutional Network (AGCN) [14]. These systems rely on the face-alignment detector in [16] or dlib [17] which extract 68 facial landmarks. The common graph-based VSR pipeline is exemplified in Figure 1.

However, when using only landmark data for the learning process, systems fail to match the performance of the standard approaches. While previous work [18] made use of extra-oral

10.21437/Interspeech.2023-969

pixels of the source image, GCN studies in [7, 8] only focused on improving the design of the network without challenging the assumption that only mouth-related data is useful for the task.

In this paper we analyze the benefits of the extra-oral facial landmarks and prove those points can, perhaps surprisingly, greatly improve the accuracy of a GCN system, progressively closing the performance gap with the standard approaches. The use of point clouds as data also allows each pose to be normalized with the aid of depth estimation. This produces a time series of 3-D face contours with a constant orientation, dimming the side effects of the original picture perspective. In contrast to pixel data, point clouds are inherently already sparser but we can yet apply landmark sub-sampling and observe its effects on performance. Our best configuration is nevertheless achieved by using the entire point cloud. When compared to other graph-based VSR systems on the LRW dataset [1], our approach improves upon the state-of-the-art accuracy by a 2% absolute margin.

2. Graph-based Visual Speech Recognition

2.1. Point Cloud Definition

Landmark detection is applied to each video frames to extract a point cloud. In our experiments, the full set of points per frame consists of 478 extracted landmarks from the detector in [19] via the MediaPipe framework [20]. This model estimates 3-D coordinates from the frames. The extracted landmarks coordinates are normalized and aligned using the tip of the nose as a fixed reference point. We denote $V \in \mathbb{R}^{C \times T \times N}$ as the point cloud representing the input sequence with T being the temporal length, N the number of landmarks, and C the number of channels. As depicted in Figure 1, we leverage the depth of the extracted 3-D landmarks to estimate the head pose, solving the Perspective-n-Point problem [21]. We then define two ways of selecting input points. The first is by location: we decompose the point cloud in Regions Of Interest (ROIs) which represent different face areas to assess their contribution to VSR accuracy; their boundaries are detailed in Figure 3a. The second is by sub-sampling: we observe the relevance of the point cloud resolution by using fewer landmarks. We apply sub-sampling either on the entire point cloud (Figure 3b) or after selecting only the lips ROI. Finding an automated method of sub-sampling point clouds for VSR is not straightforward, therefore in our current work the subset selection task is still performed by hand.

2.2. Graph Neural Network

Standard GCN models require handcrafted graph topologies defined a priori. It is laborious and sub-optimal to manually define such a graph from the face point cloud as there is no obvious connection between pairs of points. Furthermore, the fixed graph topology lacks the flexibility and ability to model the multilevel relations contained in different layers. The Adaptive Graph Convolutional Network (AGCN) proposed in [14] overcomes these limitations and demonstrates its efficiency for the graph-based VSR task in [8].

Similarly to the Transformer's multi-head structure [22], each AGCN layer is composed of K parallel heads which encode the point cloud information spatially by learning a set of Global and Adaptive adjacency matrices. Within each layer

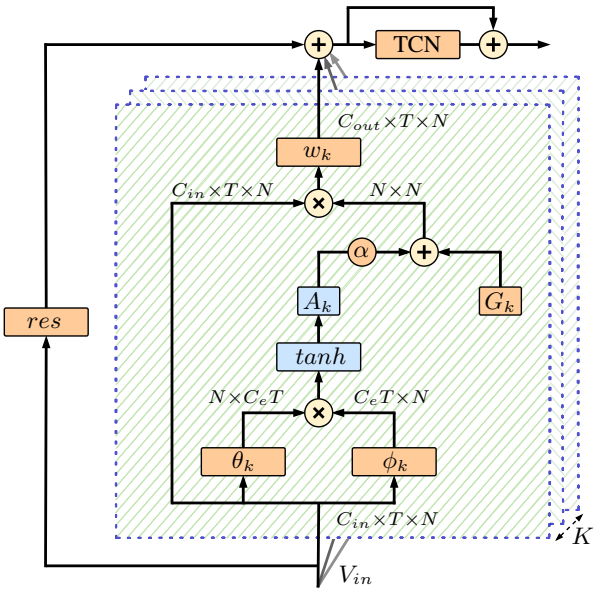


Figure 2: *AGCN layer with a single TCN layer on top. Orange boxes indicate the learned components, and the $\oplus$ and $\otimes$ symbols denote element-wise addition and matrix multiplication, respectively. Further details in Section 2.2. The figure is adapted from [14].*

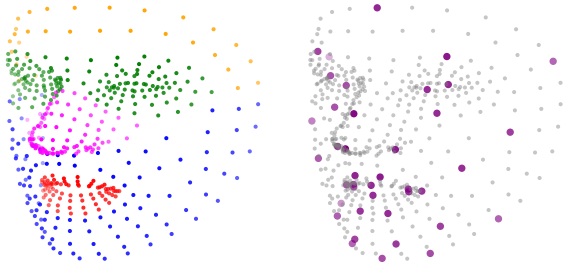
and each head k, the Global graph is determined through the adjacency matrix $G_k \in \mathbb{R}^{N \times N}$, whose elements are actively learned during the training process along with other network parameters. Besides, the Adaptive adjacency matrix $A_k \in \mathbb{R}^{N \times N}$ is computed for each input using the soft self-attention mechanism to define graph nodes associations using their similarity. The input feature map $V_{in} \in \mathbb{R}^{C_{in} \times T \times N}$ is first embedded by two separated 1×1 convolutional layers θ_k and ϕ_k in the embedding space $\mathbb{R}^{C_e \times T \times N}$. The embeddings are then reshaped to $\mathbb{R}^{N \times C_e T}$ and $\mathbb{R}^{C_e T \times N}$ spaces, respectively, and multiplied to build the adjacency matrix of the Adaptive graph.

Thus, the graph convolution operation is defined in Equation 1:

$$V_{out} = \sum_{k}^{K} W_k V_{in} (G_k + \alpha A_k) \tag{1}$$

where $V_{out} \in \mathbb{R}^{C_{out} \times T \times N}$ is the output feature map of the spatial graph convolution, W_k is the parameter matrix from the 1×1 projection function w_k, and α is a parameter weighting the linear combination of the adjacency matrices G_k and A_k. We use Temporal Convolutional Network (TCN) [23] layers with a kernel of size 9 after each of the stacked AGCN layers to encode the temporal evolution of each node. Multiple residual connections are also added to the pipeline to ease the training phase and address the over-smoothing problem [24]. A supplementary attention mechanism equivalent to the one in [14] is used to further calibrate the spatial and temporal layers. The whole block is depicted in Figure 2.

In our setup, the complete architecture of the network is composed of six stacked AGCN+TCN layers of $K = 6$ parallel heads, where all Global adjacency matrices are initialized with



(a) *Regions Of Interest* (b) *Sub-sampling*

Figure 3: *(a) details the different face regions. We distinguish five regions: the lips (in red with 80 points), the jaw (in blue with 113 points), the nose (in pink with 100 points), the eyes (in green with 155 points), and the forehead (in orange with 30 points). (b) shows the application of sub-sampling with a 1/12 ratio. The 40 resulting points are in purple. Note that this figure should be viewed in color.*

the constant value 10^{-6}. The output dimensions of each layer are 64, 64, 128, 128, 256, and 256. A batch-normalization layer is added to process the input data. Global average pooling is applied to the final graph prior to the softmax layer.

The complexity of the graph network follows a quadratic growth with respect to the number of landmarks, because the Global adjacency matrices values are part of the learned parameters. This issue can be mitigated by sub-sampling the point cloud, which will be discussed in Section 3.3.

3. Experiments and Discussions

3.1. Dataset

Our experiments are conducted on the same dataset as the referenced graph-based VSR systems [7, 8]. Lip Reading in the Wild (LRW) [1] is a large and challenging dataset consisting of more than 1000 utterances of 500 different English words, spoken by hundreds of different speakers in a wide diversity of situations. In total, 540k videos are divided into training (490k), validation (25k) and test (25k) sets. Each video lasts 1.16 seconds at 25 frames per second. Each word is surrounded by its sentence context and thus influenced by co-articulation effects.

To assess the gender neutrality of the proposed algorithm, we hand-labeled the gender of test samples. Test data is found to be comprised of approximately 60% men and 40% women.

3.2. Experimental Setup

The setup described in 2.2 is common to all reported experiments. During the training phase, Adam optimizer [25] is employed with a batch-size of 32 sequences and parameters $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$. Following [22], the learning rate increases linearly with the first 25000 steps, reaching a peak value of 0.0003 and then decreases proportionally to the inverse square root of the step amount. Each run, the whole network is initialized without any pre-training and trained for 15 epochs using Cross-Entropy loss function. On a machine equipped with an Nvidia RTX 3090 GPU and Intel I9 processor, each training takes between 10 and 48 hours depending on the point cloud definition.

Table 1: *An ablation study over ROIs. Accuracy is reported for the LRW test set, with last column detailing the gap in accuracy between genders. Top section concerns performances and complexity when only one ROI is considered; middle section reports several region combinations,"face" indicating the combination of all ROIs; bottom section assesses the impact of point cloud sub-sampling. Further details are described in Section 2.*

ROI	#Points	#Params	ACC(%)	Δ_G(%)
Lips	80	3.3M	54.4	5.7
Jaw	113	3.6M	44.2	5.5
Nose	100	3.5M	43.3	5.5
Eyes	155	4.0M	40.3	5.7
Forehead	30	3.1M	37.0	6.1
Combined ROIs				
Face	478	11.7M	62.7	4.5
Lips + Jaw	193	4.6M	60.5	5.2
Lips + Eyes	235	5.2M	60.2	4.6
Lips + Nose	180	4.4M	59.6	5.0
Lips + Forehead	110	3.6M	57.6	5.7
Face w/o Lips	398	9.1M	48.6	6.0
Sub-sampling				
Face /6	80	3.3M	61.6	5.3
Face /12	40	3.1M	60.5	4.6
Face /24	20	3.1M	57.8	5.3
Lips /2	40	3.1M	53.6	5.0
Lips /4	20	3.1M	53.3	5.1

Since LRW is usually employed in words classification tasks [1, 2, 7, 8, 9], performance is reported with the accuracy metric. The measure of the gender bias is obtained through predictive parity [26] calculating the difference between gender accuracy values, defined in Equation 2:

$$\Delta_G = ACC_{women} - ACC_{men} \tag{2}$$

3.3. Experimental Results

Experimental results presented in Table 1 show the accuracy of different point cloud ROIs, the combination of the lips ROI with every other region and the impact of sub-sampling. While it is expected of the lips and jaw regions to be the most contributing [8], it is interesting to see that every other ROI is beneficial to the accuracy. Notably, the eyes and nose regions are almost as useful as the jaw points despite being not commonly perceived as correlated to mouth movements. The complementary aspect of the eyes ROI is noteworthy: it ranks as penultimate on its own (40.3%) but when combined, it places as a close second contributor (60.5%). Overall, every isolated region yields significant accuracy, with the forehead ROI performing the worst (37%) but nevertheless still puzzling, given its anatomic location.

It seems no face area is completely devoid of correlation to mouth movements, and with the training data consisting of hundreds of speakers, it is safe to assume this is a common trait. To get a first non-comprehensive insight into this trend, we crafted a visual representation of the correlation of each ROI for a 29-frame example. This was achieved by taking the tip of the nose as origin and calculating the distance with a

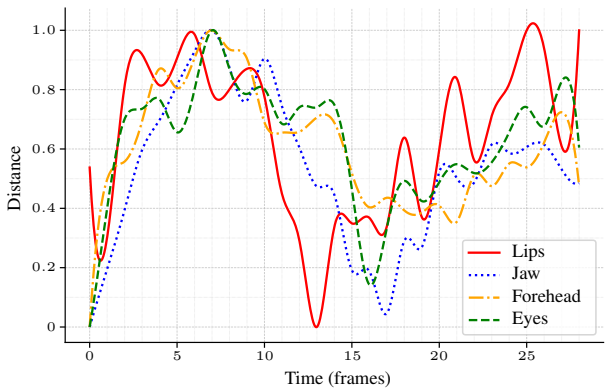


Figure 4: *Distance evolution from a selected point of each ROI with respect to the tip of the nose, over time. The dynamics of each distance curve are normalized between 0 and 1.*

hand-picked point from each ROI, in time. This produced the line graph shown in Figure 4. Depending on their position relative to the nose, we observe that regions such as lips and jaw move in parallel, while others like lips and forehead move oppositely. The two types of movements represent simultaneous events and hence denote some correlation. Probably, both the landmark extractor and anatomic traits concur to the correlation, although it remains unclear to which extent. As these remarks stem from initial observations, a throughout analysis with sounder metrics is left for future work.

The highest accuracy is nevertheless achieved when using the entire point cloud: the model reaches its best configuration with all 478 points yielding an accuracy of 62.7% on the LRW test set, which is itself an improvement of 2% over the best graph-based system in [8] (see Table 2 for comparison). This configuration experiences a drop of 0.8% accuracy when pose normalization is not applied. The advantage of filtering out the pose information is especially evident in extremely skewed poses, in which cases the burden of learning to discard orientation information is relieved from the network.

Sub-sampling results follow an analog trend, with the whole face area yielding better results (61.6%) even when reduced to the same amount of points of the sub-sampled lips ROI (60.5% versus 53.6%). Indeed, the favorable comparison still holds when considering the full-resolution lips subset, showing that 40 or even 20 points scattered over a wider area can be more valuable than 80 limited to the lips. This quality-over-quantity principle is also very advantageous when the network complexity is of crucial importance: notably, the 80-point sub-sampled model in Table 2 surpasses both state-of-the-art graph-based systems with significantly fewer parameters. Overall, sub-sampled setups are always in the order of 3 million parameters, while the only network structure that manages to surpass the top sub-sampling result comes at a cost of 8.4 additional million parameters.

Hand-labeling the speaker gender for the test set allowed for an approximate gender fairness assessment: according to the bias measure defined in Equation 2 our models yield Δ_G scores favoring women, reported in Table 1. We performed this benchmark with the image-based system in [9] and the same trend is present, while less pronounced, with a Δ_G score of 2%.

Table 2: *Comparison of our system in its most efficient configurations with the two existing graph-based VSR approaches and state-of-the-art models of similar complexity. Accuracy is reported on the test set of the LRW dataset. *Parameter number unavailable from original source, values are estimated by re-implementation.*

Method	Input Type	#Params	ACC (%)
Liu et al. [7]	Graph	30M*	49.3
Sheng et al. [8]	Point cloud	45M*	60.7
Ours (80 pts.)	Point cloud	3.3M	**61.6**
Ours (478 pts.)	Point cloud	11.7M	**62.7**
Ma et al. [9]	Image	2.9M	79.9
Ma et al. [9]	Image	9.3M	85.3

3.4. Limitations and perspective of current approach

Performance-wise all graph-based systems reported in Table 2 are still far from the most competitive image-based convolutional approaches. Gender neutrality is also not up to par with such systems, whether the deficiencies are in landmark detection or in the actual graph structure. While a given amount of points could possibly solve the VSR task as well as image-based convolutional systems, the current state of the art for GCN is still striving for competitive fruition. In perspective, our current approach manages to slightly surpass the image-based baseline introduced along with the LRW dataset in 2016 [1].

There are nevertheless some undeniable advantages in a graph-based framework: texture information is discarded early in the pipeline for a compact and better-correlated representation of facial traits which is, as shown above, prone to more systematic sub-sampling. Along with the consistent ROI analysis and related improvement of graph-based state of the art, standalone graph systems are proving themselves as a promising alternative approach for VSR.

4. Conclusions and future work

This paper investigated graph-based VSR using estimated face landmarks as node data. The current work confirms previous findings and brings additional evidence that every extra-oral region improves upon the system performance. In particular, selecting fewer points on facial ROIs can drastically reduce network complexity and concurrently yield superior performance to lips-focused setups. Furthermore, we managed to achieve better accuracy than existing graph-based VSR methods by exploiting landmarks across the whole face. We believe that our conclusions will serve as a basis for future studies. A more in-depth analysis of time correlation between ROIs is a topic we deem worth investigating in future work. To further improve the efficiency and explainability of the system, research into introducing a point cloud tailored sub-sampling strategy is already underway.

5. Acknowledgements

This work has been supported by the French government, through the 3IA Côte d'Azur, Investment in the Future, project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002.

6. References

[1] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision*, 2016.

[2] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6548–6552.

[3] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7613–7617.

[4] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," in *10th International Conference on Learning Representations, (ICLR)*, 2022.

[5] D. Serdyuk, O. Braga, and O. Siohan, "Transformer-Based Video Front-Ends for Audio-Visual Speech Recognition for Single and Muti-Person Video," in *Proc. Interspeech*, 2022, pp. 2833–2837.

[6] X. Weng and K. Kitani, "Learning spatio-temporal features with two-stream deep 3d cnns for lipreading," in *British Machine Vision Conference*, 2019.

[7] H. Liu, Z. Chen, and B. Yang, "Lip graph assisted audio-visual speech recognition using bidirectional synchronous fusion," in *Proc. Interspeech*, 2020, pp. 3520–3524.

[8] C. Sheng, X. Zhu, H. Xu, M. Pietikäinen, and L. Liu, "Adaptive semantic-spatio-temporal graph convolutional network for lip reading," *IEEE Transactions on Multimedia*, vol. 24, pp. 3545–3557, 2022.

[9] P. Ma, B. Martinez, S. Petridis, and M. Pantic, "Towards practical lipreading with distilled and efficient models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7608–7612.

[10] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *FAT*, 2018.

[11] A. Das, A. Dantcheva, and F. Brémond, "Mitigating bias in gender, age and ethnicity classification: A multi-task convolution neural network approach," in *ECCV Workshops*, 2018.

[12] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *5th International Conference on Learning Representations, (ICLR)*, 2017.

[13] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[14] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 9532–9545, 2020.

[15] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, 2018.

[16] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *International Conference on Computer Vision*, 2017.

[17] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[18] Y. Zhang, S. Yang, J. Xiao, S. Shan, and X. Chen, "Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition," in *15th IEEE International Conference on Automatic Face and Gesture Recognition*, 2020, pp. 356–363.

[19] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, "Real-time facial surface geometry from monocular video on mobile gpus," in *CVPR Workshop on Computer Vision for Augmented and Virtual Reality*, 2019. [Online]. Available: https://arxiv.org/abs/1907.06724

[20] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "Mediapipe: A framework for perceiving and processing reality," in *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR)*, 2019.

[21] E. Marchand, H. Uchiyama, and F. Spindler, "Pose Estimation for Augmented Reality: A Hands-On Survey," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, 2016, pp. 2633 – 2651.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[23] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *Computer Vision – ECCV Workshops*, 2016, pp. 47–54.

[24] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, 2018.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, (ICLR)*, 2015.

[26] S. Verma and J. Rubin, "Fairness definitions explained," in *Proceedings of the International Workshop on Software Fairness*, 2018.