3rd COG-MHEAR Workshop on Audio-Visual Speech Enhancement (AVSEC)
1 September 2024, Kos, Greece

RecognAVSE: An Audio-Visual Speech Enhancement Approach using Separable 3D convolutions and Deep Complex U-Net

João R. Manesco[1], *Leandro A. Passos*[1], *Rahma Fourati*[2,3], *João P. Papa*[1], *Amir Hussain*[4]

[1]School of Sciences, São Paulo State University, Bauru, São Paulo, Brazil
[2]REGIM-Lab.: REsearch Groups in Intelligent Machines, University of Sfax,
National Engineering School of Sfax (ENIS), BP 1173, Sfax, 3038, Tunisia.
[3]Université de Jendouba, Faculté des Sciences Juridiques, Economiques et de Gestion de Jendouba,
8189 Jendouba, Tunisie
[4] School of Computing, Engineering and The Built Environment, Edinburgh, EH9 3FF, UK.
{joao.r.manesco, leandro.passos, joao.papa}@unesp.br, rahma.fourati@ieee.org,
a.hussain@napier.ac.uk

Abstract

Audio-visual speech enhancement concerns a multimodal task that aims to provide a clean reconstruction of a speech given visual information and noisy audio signals. The assignment is particularly attractive for medical purposes since it might provide resources to aid deaf individuals, besides being useful for distinct contexts like social interactions in uproarious environments. This paper proposes RecognAVSE, an audio-visual speech enhancement solution developed for the AVSEC-3 challenge that combines Separable 3D CNNs and Deep Complex U-Nets to create an efficient alternative to tackle the problem. The model learns correlated features between the noisy audio signal and the visual stimuli, thus inferring meaning to the speech by amplifying relevant information and suppressing noise. Experiments over the AVSEC-3 dataset show that RecognAVSE can obtain outstanding results, outperforming the baselines in quantitative and qualitative results.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Introduction

Audio-visual (AV) speech enhancement (SE) concerns improving speech quality in a multimodal fashion by fusing (noisy) audio and visual inputs. The field shows itself essential in public health since it might benefit 2.5 billion people who will present some degree of hearing impairment by 2050 [1], also helping to ease social relationships [2] and psychological distresses [3, 4]. Moreover, the approach can be useful in daily life assignments, improving communication in boisterous environments by combining meaningful visual information to add context to the speech and improve its intelligibility.

Recently, several challenges have encouraged the development of new approaches in the field. The Clarity [5], Deep Noise Suppression [6], and the Hurricane [7] Challenges, for instance, promoted the development of audio-only-based methods for SE. At the same time, the COG-MHEAR Audio-Visual Speech Enhancement Challenge (AVSE Challenge) [8] fostered the development of novel architectures concerning the audio-visual paradigm, stimulating the growth of the field and the proposal of dozens of new models.

Regarding the latter, the challenge emboldens the participants to enhance a target speech signal given audio-visual samples whose audio is mixed with an interferer comprising competing speakers or environmental noises as input and a reverberation level since the samples are derived from TED talks.

In this context, many efforts have been made towards the development of new methods aiming to boost audio quality [9], reduce latency [10], and improve the energy efficiency [11]. Besides, the problem has been addressed by modeling biologically plausible methods [12], Neural Vocoders [13], knowledge distillation [14], and generative models [15].

Apart from these works, Deep Complex U-Net (DCU-Net) [16], a U-Net-based architecture specially developed for speech enhancement that deals with complex-valued spectrograms through well-defined complex-valued building blocks, obtained notorious popularity due to its outstanding achievements, outperforming state-of-the-art results over several metrics. The method implements complex convolutions as two different real-valued convolution operations with shared real-valued convolution filters. Further, the results are improved by applying a weighting mask to the spectrogram mixture.

Regarding the visual context, Separable 3D Convolutional Neural Network (S3DCNN) [17] emerges as an alternative to reduce the computational complexity of video classification systems by finding a trade-off between 3-dimensional convolutions, which usually generate better feature representation and temporal information at a more expensive computational cost, and a computationally lighter approach using 2D convolutional architectures. The method generates faster and more accurate results by replacing the 3D convolutions at the bottom of the network, concluding that semantic features extracted from temporal representation are more relevant on high-level layers.

This paper proposes the RecognAVSE (Fig. 1) for the AVSE Challenge 2024. The model comprises an audio-visual speech enhancement architecture built upon S3DCNN, employed for video extraction, and DCU-Net, used for the audio feature extraction. Further, the model fuses audio and visual features in the innermost layer of the UNet through a cross-attention mechanism. The model obtained competitive results with limited resources and reduced training epochs. Thus, the main contributions of this paper are:

- to propose the RecognAVSE, a multimodal approach for AVSE; and
- to foster the AVSE literature by providing a competitive method for the task.

The remainder of this paper is described as follows. Section 2 presents the proposed RecognAVSE, while Section 3 de-

10.21437/AVSEC.2024-3

scribes the methods and materials employed in the experiments. Further, Section 4 provides the results and discussions. Finally, Section 5 states the conclusions and future works.

2. RecognAVSE

RecognaAVSE is a method proposed to solve the multimodal speech enhancement problem by using visual information as context to improve audio quality. The architecture employs a DCU-Net to enhance audio signals, while an S3DNN encodes a speaker's visual information to capture relevant cues intrinsic to lip movements, facial expressions, and other contextual data that correlates with speech. The model proposed for the 3rd COG-MHEAR AVSE Challenge 2024 is depicted in Figure 1.

To effectively combine information from both modalities, following a commonly used strategy on multimodal works [18], the model employs a cross-attention mechanism that plays a crucial role in connecting the innermost layer of the DCU-Net with the latent embedding of the S3DCNN. Such connection is responsible for entangling the video and the audio signal, allowing it to integrate contextual features from the visual information with audio features. This integration ensures that the visual context enhances the audio features, improving speech clarity and intelligibility. Moreover, focusing only on the innermost layer of the DCU-Net reduces the number of required learning parameters, thus leading to a more computationally efficient method.

This cross-attention mechanism integrates audio and visual features using queries (Q), keys (K), and values (K) to ensure that data from different modalities are contextually aligned. Here, the queries Q are derived from the visual features, while K and V come from the audio features. The choice of using visual features as the query source is deliberate and aligns with our goal of enhancing audio signals based on visual context. Each audio feature attends to all visual features, allowing the model to integrate visual information effectively into the audio processing.

The attention scores are computed by taking the dot product of Q and K, scaling, applying a softmax function, and then using these scores to weigh the values V. As such, given the visual features $X_V \in \mathbb{R}^{n \times d_V}$ and the audio features $X_A \in \mathbb{R}^{n \times d_A}$, we first project both features sets into a common dimensional space d_A through the transformation matrices:

$$Q = X_V W_{query}, \quad K = X_A W_{key}, \quad V = X_A W_{value},$$

where $W_{query} \in \mathbb{R}^{d_V \times d_A}$, $W_{key} \in \mathbb{R}^{d_A \times d_A}$, and $W_{value} \in \mathbb{R}^{d_A \times d_A}$ are learned weight matrices. The resulting equation for cross-attention can then be expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \qquad (1)$$

where $\sqrt{d_k}$ stands for the keys' dimensionality. The output attention is then used to weigh the audio sequence as a way to carry contextual cues extracted from video features.

3. Methodology

This section describes the AVSE Challenge dataset and the setup employed to execute the experiments.

3.1. AVSE Challenge Dataset

The AVSE challenge dataset [8] is built upon the LRS3 dataset [19], composed of thousands of small video fragments containing sentences/phrases from lectures extracted from TED and TEDx. It randomly selects a set of speakers and post-processes the audio by resampling it at 16 kHz and 16 bits. Further, the authors introduce interference to audio, such that the enhancement task comprises removing or reducing this interference, composed of (i) a competing speaker, also extracted from LRS3, or (ii) noise.

Regarding the latter, the noise introduction process is performed by mixing audio extracted from three datasets: (i) Clarity Challenge dataset [5], which consists of domestic noises, (ii) DEMAND dataset [20], containing noise from soundscapes, and (iii) the DNS Challenge dataset [6], from which the authors extract environmental sounds[1].

Finally, the dataset is distributed into disjoint subsets for training and development. Table 1 provides an overview of such distribution. Notice that each mix concerns a single sentence mixed with an interference signal, i.e., noise, competing speaker, at a specific signal-to-noise ratio.

Sets	# Mixes	# Target Speakers	Interferers
Train	$34,524$	605	405 competing speakers and $7,346$ noise files.
Dev	$3,306$	85	30 competing speakers and $1,825$ noise files.

Table 1: *Training and development sets distribution.*

3.2. Experimental Setup

This work implements RecognAVSE, an audio-visual speech enhancement architecture built upon a Separable 3D Convolutional Neural Network and Deep Complex U-Net. It fuses audio and visual information using a cross-attention mechanism to produce high-quality enhanced audio outputs.

The S3DCNN [17] was trained using videos of 64 frames, resized to an image size of 224x224. The model architecture, designed to process video data efficiently, consists of an initial layer of separable temporal convolutions, followed by 9 Separable Inception Blocks. These blocks utilize kernels of size $3 \times 3 \times 3$, with padding set to 1 and strides of 2 for efficient downsampling and feature extraction. At the end, a projection layer is employed to map the output embeddings of the network to a 512-dimension feature space.

The audio was divided into random clips of $40,800$ samples, randomly clipped for each video during training, and sampled at 16 kHz. The audio waveform was then converted to the time domain using the Fast Fourier Transform (FFT), utilizing a Hann window of size 400, a window shift of 160, and a hop length of 512. The DCU-Net was built with five layers of downsampling and five layers of upsampling. The last downsampling layer of the network was flattened and then used to combine and integrate video features through the cross-attention layer.

The experiments used the Scale-Invariant Signal-to-Noise Ratio (SI-SNR) loss [21] as the loss function since it denotes a popular metric used to evaluate and optimize the performance of audio enhancement models. This metric measures the similarity between the clean target signal and the enhanced signal produced by the model by decomposing the enhanced signal into two components: one aligned with the target signal (the desired

[1]The DNS dataset's environmental sounds were obtained from `https://freesound.org/`.

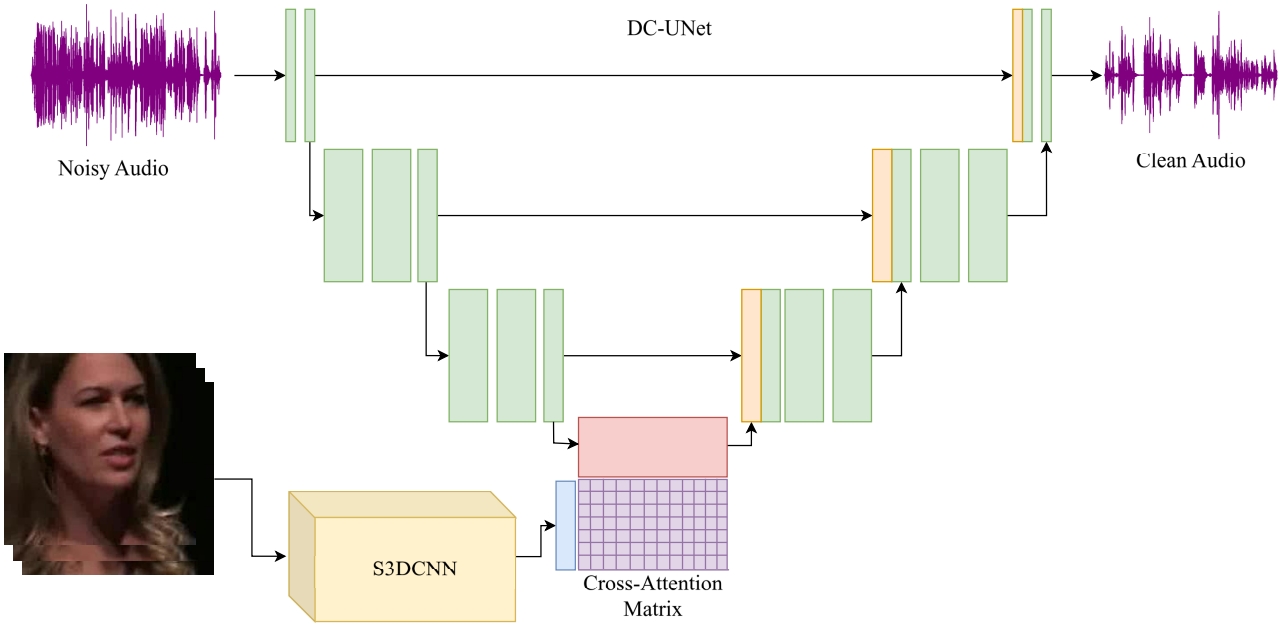


Figure 1: *Depiction of the overall pipeline of our method: RecognAVSE. An S3D network extracts video embeddings while a DCU-Net filters noisy audio. To combine both features, a cross-attention mechanism is used between both embeddings to correlate contextual information.*

component) and one orthogonal (the noise component). The SI-SNR loss is then defined as the ratio between the magnitude of each element, providing a robust measure of the model's performance. Concerning the evaluating metrics, this work presents the results considering the Short-Time Objective Intelligibility (STOI) [22], which measures the intelligibility of degraded speech signals, and the Perceptual Evaluation of Speech Quality (PESQ), which measures audio quality considering its sharpness, background noise, clipping, and audio interference.

Our results were compared to the baseline provided by the 3rd COG-MHEAR Audio-Visual Speech Enhancement Challenge, which will be referred to as *baseline* in our comparisons. In addition, we also provide a comparison to the original noisy audio without any processing, referred to as *noisy*, during our analysis.

Further, the parameters are optimized using Adam with a learning rate of 10^{-3}, reduced by a factor of 0.5 on a plateau with the patience of 2 during 20 epochs considering a batch size of 4. RecognAVSE is implemented in Python using Pytorch framework [23] and the code is available in GitHub[2]. The model comprises 164M parameters, 12.81 GFlops, and presents an average training time of 3h16m per epoch running on an Intel® Xeon® Bronze 3204 CPU with 1.90GHz, 48GB of RAM, and a Tesla T4 Nvidea GPU with 16GB of memory.

4. Results and Discussion

Table 2 details the experiments conducted over the development set, where RecognAVSE consistently outperforms the baseline regarding the STOI and SISDR metrics, displaying scores of 0.70 and 2.35, compared to baselines of 0.68 and 2.13, respectively, highlighting the method's effectiveness in noise reduction. Despite a slightly lower PESQ score than the baseline, the model still significantly enhances the noisy dataset, reaching a value of 1.16.

Method	PESQ↑	STOI↑	SISDR↑
Noisy	1.16	0.62	−4.33
Baseline	**1.33**	0.68	2.13
RecognAVSE	1.32	**0.70**	**2.35**

Table 2: *Quantitative results obtained on the development set for the AVSE Challenge dataset, comparing the performance of RecognAVSE against the noisy audio and the baseline using STOI, SISDR, and PESQ metrics. Bold indicates the best results.*

Regarding the test set, Table 3 showcases that RecognAVSE performed slightly lower than the baseline, obtaining almost similar results. It is worth noting that the test set includes a higher presence of competing speaker interference, which denotes a more challenging task for RecognAVSE.

Method	PESQ↑	STOI↑	SISDR↑
Noisy	1.17	0.61	−4.88
Baseline	**0.65**	**1.29**	**0.80**
RecognAVSE	**0.65**	1.28	0.40

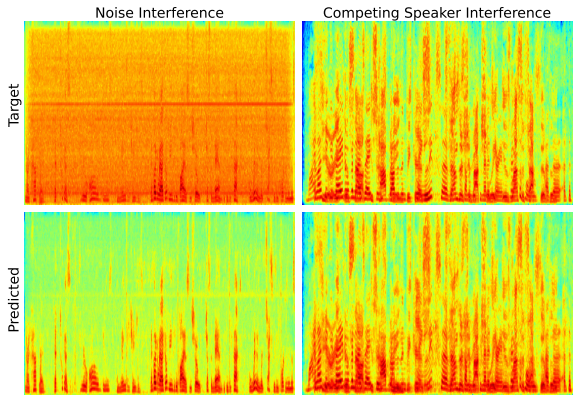
Table 3: *Quantitative results obtained on the test set for the AVSE Challenge dataset, comparing the performance of RecognAVSE against the noisy audio and the baseline using STOI, SISDR, and PESQ metrics. Bold indicates the best results.*

The observed limitations in competing speaker scenarios can be attributed to the design upon which RecognAVSE is built, i.e., the DCU-Net model. This architecture, which relies on spectrogram analysis in the complex space, excels at noise

[2]Available at: `https://github.com/jrjoaorenato/avse_challenge3`.

reduction but faces challenges when separating multiple overlapping speakers. Figure 2 emphasizes this difficulty in dealing with this particular interference, showing that our method effectively reduces noise but struggles with additional dialogue scenarios.

Figure 2: *Spectrograms illustrating the qualitative results of our method on different types of interference present in the AVSE Challenge dataset.*



Despite these challenges, it is crucial to note that our method achieves results similar to or better than the baseline in many scenarios. It does so with a significantly faster convergence rate, as exemplified in Figure 3, which illustrates the convergence plot of our method on the training and development sets. This rapid convergence to optimal results within just 20 epochs is significantly faster than the baseline's 100 epochs and demonstrates the potential for further improvement and efficiency in the audiovisual speech enhancement task.

Figure 3: *Loss convergence plot of our proposed method the training and development sets.*

5. Conclusions

This paper proposed the RecognAVSE, a novel audio-visual speech enhancement method submitted to the 3rd COG-MHEAR Audio-Visual Speech Enhancement Challenge. The model fuses audio features obtained at the bottom-most layer of a Deep Complex U-Net with the visual features extracted via a Separable 3D Convolutional Neural Network through a cross-attention mechanism.

Our experiments over the AVSE challenge dataset demonstrate that RecognAVSE not only surpassed the baseline on the development set, based on the STOI and the SISDR metrics but also delivered similar results on the testing set. Notably, these results were achieved in a significantly smaller number of epochs. Additionally, one can notice that RecognAVSE obtained better results in handling noise interference, a result that was not consistent when observing interference in the form of competing speakers. Such a behavior was expected due to the properties of Deep Complex U-Net.

Regarding future works, we aim to employ diffusion transformers for the task of audio-visual speech enhancement. Additionally, we expect to study the use of metaheuristic optimization techniques to improve the model's performance.

6. Acknowledgements

João R. Manesco, Leandro A. Passos, João P. Papa are grateful to the São Paulo Research Foundation (FAPESP) grants $2023/10823-6$, $2024/00789-8$, $2013/07375-0$, $2023/14427-8$, and $2023/01374-3$, as well as the National Council for Scientific and Technological Development (CNPq) grant $308529/2021-9$ for their financial support.
Rahma Fourati has received funding from the Ministry of Higher Education and Scientific Research of Tunisia under grant agreement number LR11ES48.
Professor Hussain acknowledges the support of the UK Engineering and Physical Sciences Research Council (EPSRC) (Grants No. EP/M026981/1, EP/T021063/1, EP/T024917/1).

7. References

[1] W. H. Organization *et al.*, "Hearing screening: considerations for implementation," 2021.

[2] W. Noble, *Self-assessment of hearing and related function*. Wiley-Blackwell, 1998.

[3] A. R. Huang, J. A. Deal, G. W. Rebok, J. M. Pinto, L. Waite, and F. R. Lin, "Hearing impairment and loneliness in older adults in the united states," *Journal of Applied Gerontology*, vol. 40, no. 10, pp. 1366–1371, 2021.

[4] A.-S. Helvik, G. Jacobsen, and L. R. Hallberg, "Psychological well-being of adults with acquired hearing impairment," *Disability and rehabilitation*, vol. 28, no. 9, pp. 535–545, 2006.

[5] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, and R. Viveros Munoz, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2, 2021.

[6] C. K. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Icassp 2021 deep noise suppression challenge," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6623–6627.

[7] J. Rennies, H. F. Schepker, C. Valentini-Botinhao, and M. Cooke, "Intelligibility-enhancing speech modifications-the hurricane challenge 2.0." in *INTERSPEECH*, 2020, pp. 1341–1345.

[8] A. L. A. Blanco, C. Valentini-Botinhao, O. Klejch, M. Gogate, K. Dashtipour, A. Hussain, and P. Bell, "Avse challenge: Audio-visual speech enhancement challenge," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 465–471.

[9] S. Ahmed, C.-W. Chen, W. Ren, C.-J. Li, E. Chu, J.-C. Chen, A. Hussain, H.-M. Wang, Y. Tsao, and J.-C. Hou, "Deep complex u-net with conformer for audio-visual speech enhancement," *arXiv preprint arXiv:2309.11059*, 2023.

[10] M. Gogate, K. Dashtipour, and A. Hussain, "Towards real-time privacy-preserving audio-visual speech enhancement," *algorithms*, vol. 2, p. 3, 2022.

[11] L. A. Passos, J. P. Papa, J. Del Ser, A. Hussain, and A. Adeel, "Multimodal audio-visual information fusion using canonical-correlated graph neural network for energy-efficient speech enhancement," *Information Fusion*, vol. 90, pp. 1–11, 2023.

[12] L. A. Passos, J. P. Papa, A. Hussain, and A. Adeel, "Canonical cortical graph neural networks and its application for speech enhancement in audio-visual hearing aids," *Neurocomputing*, vol. 527, pp. 196–203, 2023.

[13] R. Mira, B. Xu, J. Donley, A. Kumar, S. Petridis, V. K. Ithapu, and M. Pantic, "La-voce: Low-snr audio-visual speech enhancement using neural vocoders," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[14] R.-C. Zheng, Y. Ai, and Z.-H. Ling, "Incorporating ultrasound tongue images for audio-visual speech enhancement through knowledge distillation," *arXiv preprint arXiv:2305.14933*, 2023.

[15] A. Golmakani, M. Sadeghi, and R. Serizel, "Audio-visual speech enhancement with a deep kalman filter generative model," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[16] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations*, 2018.

[17] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 305–321.

[18] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer *et al.*, "Perceiver io: A general architecture for structured inputs & outputs," in *International Conference on Learning Representations*, 2021.

[19] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[20] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics*, vol. 19, no. 1. AIP Publishing, 2013.

[21] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[22] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[23] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.