Research of Different Neural Network Architectures for Audio and Video Denoising

Anton Kanev
Bauman Moscow State Technical University
Moscow, Russian Federation
aikanev@bmstu.ru

Maksim Nazarov
Bauman Moscow State Technical University
Moscow, Russian Federation
maxzbox1@gmail.com

Daniel Uskov
Bauman Moscow State Technical University
Moscow, Russian Federation
UskovDanek@gmail.com

Vladislav Terentyev
Bauman Moscow State Technical University
Moscow, Russian Federation
webman.com@gmail.com

***Abstract*—Recently, neural networks have attracted considerable attention due to their better denoising quality compared to the previously used frequency-time filters for audio denoising and algorithmic methods for visual denoising. In this paper different neural network architectures used for denoising from audio and video files are discussed. The architectures of two neural networks RNNoise and PoCoNet are compared to audio denoising. The main source of visual noise in video is Gaussian noise, which occurs during digital imaging of real objects, such as sensor noise caused by poor light or high temperature, and electronic noise. Two solutions to neural network architectures are compared to remove Gaussian noise from video. In the first approach, each frame is processed as a separate image. For this approach, the DnCNN and Restormer neural networks have been considered. In the second approach, frames are considered as a sequence of images over time. In this approach, FastDVDnet and PaCNet neural networks have been considered. The aim is to compare different neural networks architectures for removing audio and video noise from public room camera recordings.**

Keywords— Deep learning, neural network, denoising, audio noises, gaussian noise.

I. INTRODUCTION

The task of removing audio and video noise remains one of the urgent tasks of video processing. Until recently, the main way to remove noise from audio were frequency filters, which were pre-prepared filters that remove certain frequencies from the audio. Recently, neural networks have been used for this task due to their advantage in noise removal over frequency filters. Neural networks for the removal of visual noise have already surpassed such a method as Block-matching and 3D filtering (BM3D), which is one of the best ways to remove noise without machine learning [1].

A. About Audio Noise Classifications

Noise is found in many areas of human activity: construction, aerospace, music and other industries. Due to the large number of industries in which noise affects human activity, there are many different classifications of noise, for example, according to the nature of noise is divided into mechanical, aerodynamic, hydraulic, or electromagnetic; according to frequency characteristics noise is divided into low frequency (<300 Hz), medium frequency (300-800 Hz) and high frequency noise (>800 Hz); according to the nature of the spectrum noise is divided into broadband noise, which is a continuous spectrum of width over one octave, and tones The classification according to temporal characteristics is taken as the basic one due to the fact that it most accurately reflects the noises that occur in audio files and need to be removed. The most standardized classification of noise according to temporal characteristics is a division into three main groups: stationary noise (white noise), impulse noise (various claps, sneezes, creaks of chairs and doors and other sounds) and non-stationary noise, which in turn are divided into intermittent noise (alarms, phone beeps and other sounds) and fluctuating noise (wind, engine sounds and others).

This classification of noise is most often used because it allows you to distinguish audio noise by the difficulty of removing it. One of the main difficulties associated with noise removal lies in the fact that noises are unpredictable, that is, it is impossible to say exactly where and when a certain type of noise appears in the received audio signal. If we know what type of noise appears in the audio signal, then the task of noise removal becomes a lot easier. It is also quite easy to get rid of the stationary noises since they will be evenly distributed throughout the audio signal and at times when there are no extraneous sounds, voices in the audio signal it will be possible to clearly determine the amplitude of the noise [2]. If to rank noises according to their difficulty of removal from the easiest types to the most difficult to define and to remove, the following ranking is obtained: stationary noises, fluctuating noises, intermittent noises, impulse noises.

B. B. Digital Noise in Video

Digital noise is random variations in color information that are not present in the real object in digital imaging and are usually part of electronic noise. This noise can be produced by the photographic matrix and the electrical circuitry of the digital camera. It can also occur as the unavoidable shot noise of a perfect photon detector. Thus, digital noise is an unwanted byproduct of digital imaging that adds false and extraneous information.

The main source of digital noise is Gaussian noise that occurs during digital imaging of real objects, such as sensor noise caused by poor lighting and/or electronic circuit noise caused by high temperature, capacitor reset noise or transmission noise. [3] Gaussian noise is a type of signal noise that has a probability density function equal to the normal distribution function (also known as the Gaussian distribution). In other words, the values that such noise can take have a Gaussian distribution. The

979-8-3503-9952-3/23/$31.00 ©2023 IEEE

standard model of this noise is additive, independent at each pixel and independent of signal intensity. Such a noise model is known as additive white Gaussian noise.

Digital noise is particularly evident in low light conditions. In low light conditions, proper exposure requires the use of slow shutter speeds to increase the amount of light (photons) captured. But if the shutter speed limits are reached and the resulting image is still not bright enough, then higher gain (higher ISO sensitivity) is used. And amplifier noise is the main source of image sensor readout noise, that is, a constant level of noise in the dark areas of the image.

II. Methods

A. A. Gaussian Noise Removal Approaches

There are two approaches to removing Gaussian noise from video. In the first approach, the neural network removes noise from each frame of video as from separate images, independent of each other. In the second approach, the neural network inputs a sequence of frames, which is treated as a sequence of interconnected images over time, from which noise is removed.

B. RNNoise

RNNoise is a neural network for noise removal from audio that uses a Gated Recurrent Unit (GRU) because it performs slightly better than Long Short-Term Memory (LSTM) in the noise removal task and requires fewer resources (both CPU and memory for weights) [4]. Compared to simple recurrence blocks, GRUs have two additional gates. The reset element controls whether the state (memory) is used in computing the new state, while the update gate controls how much the state will change based on the new input. This update element (when off) allows the GRU to remember information over a long period of time and is the reason why GRUs (and LSTMs) perform much better than simple recurrence units. The architecture of this neural network is shown in Fig. 1.

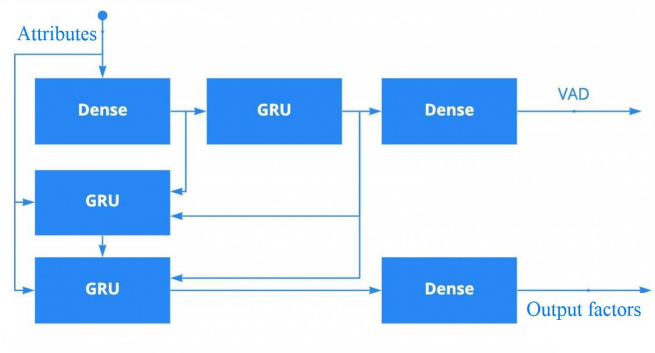

Fig. 1. RNNoise architecture.

C. PoCoNet

PoCoNet is a neural network for removing noise from audio, which uses a full-link convolutional 2D U-Net architecture [5] with self-monitoring layers and 4-layer DenseNet blocks [6] on each layer for the neural model. The top two layers of the PoCoNet architecture are shown in Fig. 2, shown with frequency-positioning blocks and inputs with real and imaginary parts in the short-term Fourier transform. The authors of the architecture believe that the convolutions are causal in the time direction but not in the frequency direction, which means that padding is applied symmetrically in the frequency direction, as is usual in two-dimensional convolutional networks, but asymmetrically in the time direction in the sense that it is only used at the edge of each layer. This helps preserve quality at the end of the output, for use in low-latency applications, since fill tends to degrade quality near edges and borders. Note that the advance is provided by averaging filtering layers, which are used instead of maximum filtering. The self-attention blocks used use an information aggregation mechanism only in the temporal direction to improve learning and inference efficiency.

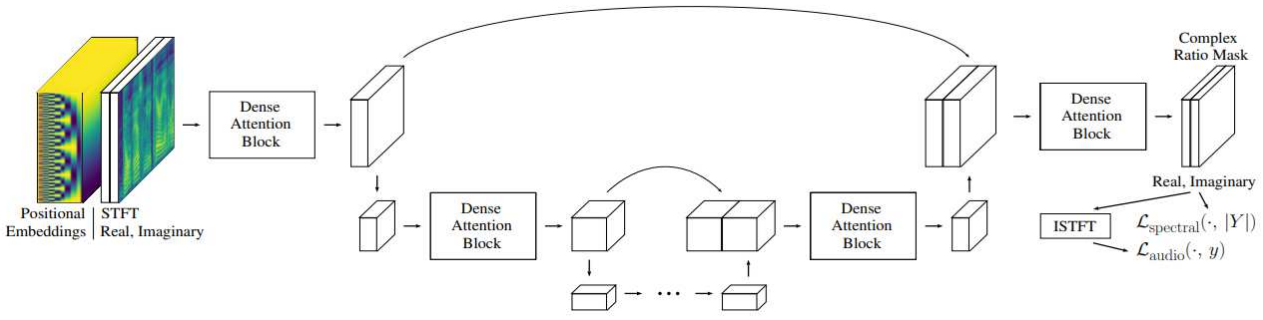

Fig. 2. Two upper layers of PoCoNet architecture.

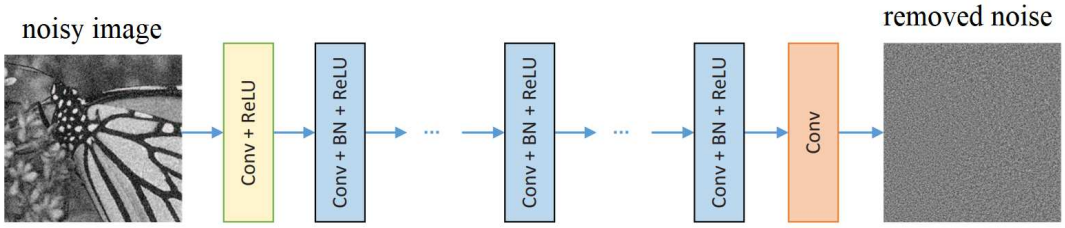

Fig. 3. DnCNN architecture.

D. *Restormer*

The Restormer neural network [2] is a Transformer architecture introduced in 2017 by researchers at Google Brain [7], with two key changes (Fig. 4). A new Multi-Dconv Head Transposed Attention (MDTA) block has been introduced (Fig. 4a), which has linear complexity, instead of the standard self-attention with quadratic complexity. The new Gated-Dconv Feed-Forward Network (GDFN) (Fig. 4b) instead of the usual feed-forward network (FN) performs controlled feature transformation, i.e., suppresses less informative features and allows only useful information to pass further along the network hierarchy.

E. *FastDVDnet*

FastDVDnet is a neural network for removing noise from video, whose architecture consists of a number of modified U-Net blocks, which take three frames as input [8]. The first module of FastDVDnet consists of 3 modified U-Net blocks whose outputs are fed to the input of the next U-Net block in the second module. Thus 5 frames are fed to the input and the output of the last U-Net block is the result of central frame noise removal. In the modified U-Net blocks, the encoder was adapted to accept three frames and a noise map as input. Upsampling in the decoder is done with a PixelShuffle layer, which helps reduce grid artifacts. Combining the encoder and decoder features is done using a pixel-by-pixel addition operation instead of combining by channel. This results in lower memory requirements.The architecture of the neural network is presented in Fig. 5.

F. *PaCNet*

In the PaCNet neural network architecture, a new concept of patch-craft frames has been introduced to remove noise from video a new notion of patch-craft frames was introduced [9]. It is artificial frames, similar to the real ones, constructed by overlaying the same look and feel as real frames are built by overlaying the same fragments of frames. They are added to the original sequence of frames and then fed to the input of the convolutional neural Thereby greatly increasing the quality of noise reduction. More specifically, this synthetic frame can be used as an additional feature map representing the real frame.

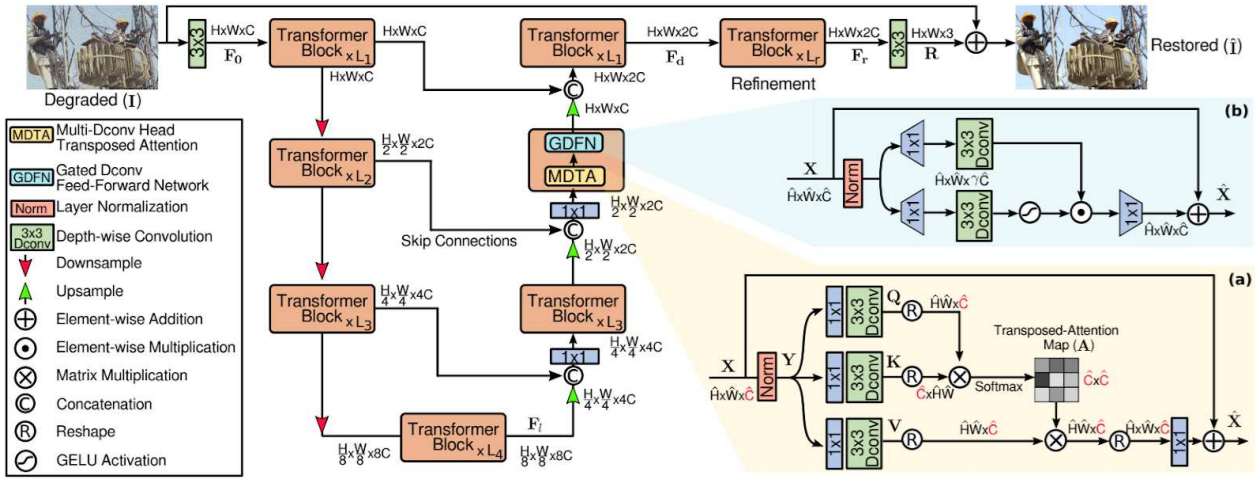

Fig. 4. Restormer Architecture. The main modules of the Transformer block are (a) Multi-Dconv Head Transposed Attention (MDTA) and (b) Gated-Dconv Feed-Forward Network (GDFN).

Fig. 5. FastDVDnet architecture

III. EXPERIMENTS

A. Dataset to Compare RNNoise and PoCoNet Neural Networks

To compare different neural networks, a dataset from the Deep Noise Suppression Challenge [10] conducted by Microsoft was chosen. The dataset consists of two folders, enrollment_speech (178 audio files) and testclips (859 audio files). It contains 638 audio files for mobile devices and 399 audio files for desktops/laptops. The duration of each testclips is 10 seconds. This dataset is excellent for comparison, because it contains completely cleaned audio tracks with voices and the same audio tracks with different noises added. The noises used in this dataset cover the whole classification of noises according to their temporal characteristics, therefore the comparison results can further show the advantages of neural network architectures with different categories of noises.

B. Dataset to Compare Dncnn And Restormer, Fastdvdnet And Pacnet Neural Networks

To compare neural network architectures, researchers in their research works used datasets containing images or videos to which additive white Gaussian noise was applied. In this way, pairs of noisy and clean data were obtained. The following are descriptions of these datasets.

DIV2K is a 2K resolution image dataset that contains 1000 images with different scenes and is divided into 800 for training, 100 for validation, and 100 for testing [11].

Flickr2K is a dataset that contains 2000 images collected from the Flickr website, along with 5 reference sentences provided by human annotators [12].

BSD is a dataset often used to remove noise and enhance image resolution. The set consists of all sorts of images ranging from natural images to images of specific objects such as plants, people, food, etc. [13].

WED is a dataset that contains 4744 real images and 94880 distorted images created from them [14].

RENOIR - dataset of color images corrupted by natural noise in low-light conditions, together with spatially and intensely aligned images of the same scenes with low noise [15].

Densely Annotated Video Segmentation (DAVIS) is a high-quality dataset of densely annotated high resolution video segmentation in two resolutions, 480p and 1080p [16]. There are 50 video sequences with 3455 densely annotated frames at the pixel level. Thirty videos with 2079 frames are for training and twenty videos with 1376 frames are for validation.

These datasets contain a variety of images or videos of all kinds of orientation.

C. Comparison of Neural Network Architectures

To compare the architectures of neural networks we used a dataset containing recordings of video cameras from public spaces, which were overlaid with audio noise and visual Gaussian noise. Based on the performance of neural networks, the results of quality metrics for RNNoise and PoCoNet when dealing with audio noise were obtained, recorded in Table 1, and the results of neural networks to remove visual noise for each approach: for the first approach DnCNN and Restormer, for the second approach FastDVDnet and PaCNet, and the results for both approaches are shown in Table 2.

D. Metrics to Compare RNNoise and PoCoNet Neural Networks

To compare RNNoise and PoCoNet neural networks we chose the following metrics: PESQ (Perceptive evaluation of speech quality) - speech quality evaluation algorithm [17]; CBAK - MOS predictor of background noise intrusiveness [17]; COVL - MOS predictor of overall signal quality [17]; CSIG - MOS predictor of signal distortion [17]. For the metrics listed above, the higher the score obtained by the neural networks, the better they performed.

E. Metrics to Compare DnCNN and Restormer, FastDVDnet and PaCNet Neural Networks

The processing time per frame is calculated for processing on the central processing unit (CPU). This is done to ensure that all models are evaluated equally, since for some models there is not enough available memory of the graphics processing unit (GPU) to work.

PSNR shows the ratio between the maximum possible signal power and the distorting noise power. PSNR is expressed in decibels and can be defined through the root mean square error (MSE). The higher the value of this metric, the better the quality of noise removal.

TABLE I. RNNOISE AND POCONET COMPARISON RESULTS

Metrics	PESQ	CBAK	COVL	CSIG
Noisy	1,582	2,533	2,351	3,186
RNNoise	1,973	3,463	2,789	2,692
PoCoNet	2,745	3,04	3,422	4,08

TABLE II. DNCNN AND RESTORMER, FASTDVDNET AND PACNET COMPARISON RESULTS

	Image sequence		Video	
	DnCNN	***Restormer***	***FastDVDnet***	***PaCNet***
Processing time per frame, sec	0,1	0,45	0,27	180
PSNR (DB)	27,51	29,80	31,90	32,2

F. Results

Based on the results from the experiment, we can conclude that the PoCoNet neural network performed better in the task of removing noise from audio. In CBAK metric RNNoise neural network showed better result than PoCoNet. Neural networks that remove noise from video as a sequence of frames over time showed better results than networks that remove noise from video as a sequence of individual images. PacNet showed better noise removal results than FastDVDnet, but processing time per frame was significantly longer than FastDVDnet.

IV. CONCLUSION

The authors compared neural network architectures for noise removal from audio and video. The methods used achieved acceptable results for noise removal from video camera recordings from public premises. At the same time, inaccuracies in the results of the experiment can be caused by specific initial data, where audio was dominated by one type of background noise, while video, in addition to Gaussian noise, may also

contain other noises (e.g., Poisson noise), whose removal was not considered in this experiment. In the future, the authors plan to use several types of noise to increase the realism of the experimental model.

REFERENCES

[1] Kai Zhang, Yunjin Chen, Deyu Meng, and Lei Zhang, "Beyond a Gaussian Denoiser: Residual learning of deep CNN for image denoising," IEEE Transactions on Image Processing, 2016, 13 p.

[2] Syed Zamir, Aditya Arora, and Salman Khan, "Restormer: Efficient transformer for high-resolution image restoration, CCVPR 2022, 2022, 12 p.

[3] Tudor Barbu, "Variational image denoising approach with diffusion porousmedia flow," Abstract and Applied Analysis, 2013, 8 p.

[4] Jean-Marc Valin, "A hybrid DSP/Deep learning approach to real-time full-band speech enhancement," IEEE MMSP Workshop, 2018, 5 p.

[5] Umut Isik, Ritwik Giri, Jean-Marc Valin, and Karim Helwani, "PoCoNet: Better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss," Interspeech, 2020, 2020, 5 p.

[6] Ritwik Giri and Andrew H. Song, "Umut Isik: Channel-Attention Dense U-Net for Multichannel Speech Enhancement," in 2020 IEEE ICASSP, 2020, 5 p.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, "Attention is all you need," Conference on Neural Information Processing Systems, 2017, 15 p.

[8] Matias Tassano, Julie Delon, and Thomas Veit, "FastDVDnet: Towards real-time deep video denoising without flow," CVPR 2020, 2020, 13p.

[9] Gregory Vaksman, Michael Elad, and Peyman Milanfar, "Patch craft: Video denoising by deep modeling and patch matching," ICCV 2021, 2021, 16 p.

[10] Chandan K A Reddy, Harishchandra Dubey, Kazuhito Koishida, Arun Nair, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan, "Interspeech 2021 deep noise suppression challenge," [Dataset]. Interspeech 2021, 2021, 3 p.

[11] Radu Timofte, Eirikur Agustsson, Shuhang Gu, Jiqing Wu, Andrey Ignatov, and Luc Van Gool, "DIVerse 2K resolution high quality images as used for the challenges," [Dataset]. CVPR 2017, 2017, 8 p.

[12] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars, "Guiding long-short term memory for image caption generation," ICCV 2015, 2015, p. 5-6.

[13] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," IEEE TPAMI, vol. 33, no. 5, 2011, pp. 898-916.

[14] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang, "Waterloo exploration database, New challenges for image quality assessment models," [Dataset]. IEEE Transactions on Image Processing, vol. 26, no. 2, February 2017, 13 p.

[15] J. Anaya and A. Barbu, "RENOIR – A dataset for real low-light image noise reduction," Journal of Visual Communication and Image Representation, 2014, 27 p.

[16] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool, "The 2019 DAVIS challenge on VOS: Unsupervised multi-object segmentation," [Dataset]. Computer Vision and Pattern Recognition 2019 Workshop, Challenge 2 May 2019, 4p.

[17] Philipos Loizou, "Speech enhancement: Theory and practice," CRC press, 2007, 716 p.