Turkish Journal of Physiotherapy and Rehabilitation**; 32(3)**
ISSN 2651-4451 | e-ISSN 2651-446X

UNIVERSAL IMAGE AND AUDIO RESTORATION USING DEEP LEARNING

Omar H. Mohammed[1], Prof.Basil Sh. Mahmood[2]
[1]Department of Computer Engineering, Mosul University, Nineveh, Iraq
omerhatif@gmail.com
[2]Department of Computer Engineering, Mosul University, Nineveh, Iraq
basil.mahmood@uomosul.edu.iq

ABSTRACT

Restoring the original image or audio signal from a distorted version is a challenge in a real-world application; Traditional techniques such as the Wiener filter and statistical approach have been used, but recently deep learning has been widely found in many applications due to its high-performance quality. The main objective of this paper is to present a new algorithm to restore the images and sound signals suffering from different types of distortion, including blurring, noise, and other degradation processes. The degradation process is identified using the convolutional neural network VGG16, while the conditional-GAN is used for restoration due to the type of the identified distortion. The algorithm also successfully restored audio signals by converting the 1D audio signal to 2D image-like, using short-term fast Fourier transform (STFT). For training and evaluating both distortion identifier and restoration process, 31080 images and 1132 speech signals are tried.

Keywords: restoration, blurring, noise, STFT,conditional-GAN, VGG16.

I. INTRODUCTION

Digital signals such as images and audio signals can be degraded during the acquisition, transmission, or storage phase, and the degradation may result in the loss of important information. Images may be distorted by unwanted artifacts such as blurring and noise; blurred images maybe result from deterministic effects on the image process due to several reasons such as movement between the capturing device and the objects, atmospheric distortion, and optical aberration. Noise is another source of distortion that usually occurs during recording for example. Restoration is the process used to reverse the effect of the degradation process, and it attempts to restore the original clear image from the distorted one. A linear system can model the degradation and restoration processes, which are represented by equations 1 and 2, respectively [1]:

$$g\ (x,y) = A * f(x,y) + n(x,y) \qquad \ldots\ldots\ (1)$$

$$f\ (x,y) = A^{-1} * g(x,y) - n(x,y) \qquad \ldots\ldots\ (2)$$

Where $f(x,y)$ is the original non-distorted image, $g\ (x,y)$ represents the acquired image with blur and noise, A is the blurring matrix defined by a certain point spread function (PSF), and $n\ (x,y)$ is noise. There are two types of restorations, Non-blind and blind deconvolution Techniques[2]. In non-blind deconvolution, there is prior knowledge about the degradation process; in this type, several methods could be used, such as the Wiener filter, constraint least-square filter, and Lucy- Richardson Algorithm[3]. When there is no information about the distortion process, in this case, the blind deconvolution methods must be used, such as median filter, statistical approach, and deep neural network[4]. Deep learning-based image restoration techniques are widely used in recent years; there are two main approaches point spread function(PSF) estimation and the end-to-end approach[5]. The PSF approach using a fully convolutional deep neural network (FCN) to estimate the blurring kernel after training with a dataset of images that contain distorted images as input and blurring kernel as target[6]. The second approach which is end-to-end considers a deep deconvolutional neural network (DDNN) that processes distorted images and produces a non-distorted clear image at the output. The conditional-generative adversarial networks(GAN) are used for end-to-end restoration, the general GAN consists of generator and discriminator. The generator takes the distorted image as input and produces a clear image. The discriminator tries to figure out if the produced image is still distorted or not. By competition between generator and discriminator, the generator learns to produce a clear image that looks

very realistic[7]. In addition to the general GAN, the conditional-GAN content loss function must chosen so that the generator does not generate uncorrelated images with the sharp image during training[8]. Audio signals like images also suffer from distortions, and an unwanted modification in audio signals; the distortion may be during (or after) the acquisition process due to the noise coming from the microphone and amplifier[9]. Many methods could be tried for audio signal restoration, such as filtering approaches, transformation-based approaches like wavelet or fast Fourier transform, and neural networks. Another approach for audio signal restoration is by considering the problem as image restoration; that by converting the signal from 1D to 2D so that the signal becomes. Short-time Fourier transform (STFT) or Mel-frequency cepstral coefficients (MFCCs) can do that. After restoring the resultant image, then back to 1D which will be the restored audio signal[10, 11]. The STFT is a transformation tool usually used in the frequency-time analysis because it is unlike Fourier transform, it shows the signal in the frequency domain, and keeps time information[12].

II. THE PROPOSED ALGORITHM

The algorithm in this work, as shown in figure 1, could be used to restore images that are distorted by various types of distortions, such as several burring, noise, and other artifacts. The first stage is to classify the image as a normal image or a distorted image along with its types of distortions; if the input image is classified as a normal image, nothing happens, else according to distortion types, the image will be restored. Processes are repeated on the restored image until it gets rid of all types of distortion, this helps to restore an image that suffers from multiple artifacts at the same time.

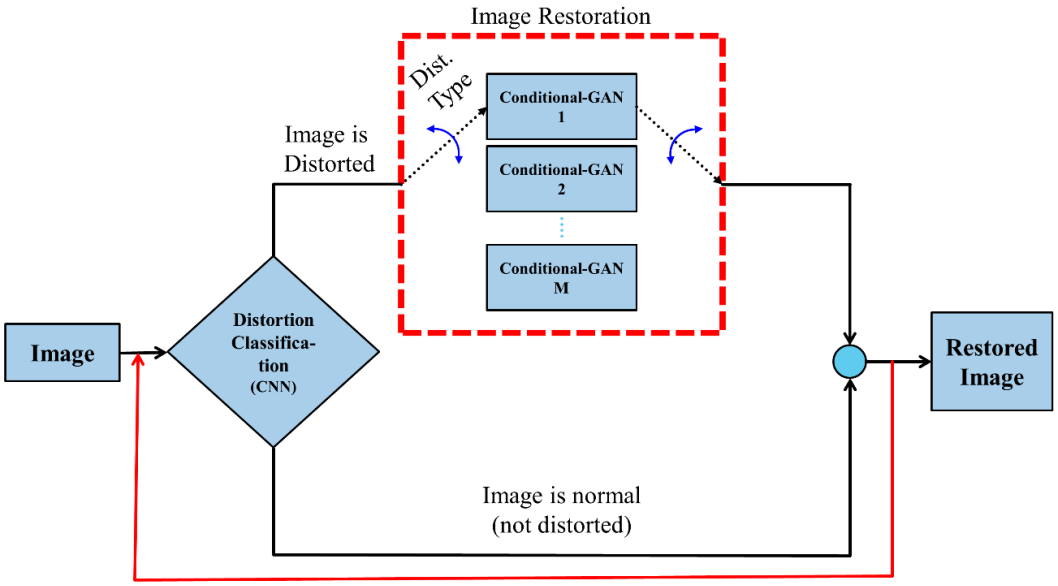


Figure 1. The proposed algorithm

Audio signals are restored by converting them into 2D using The short-time Fourier transform (STFT), as illustrated in figure 2, which means that audio signals considered as image-like.

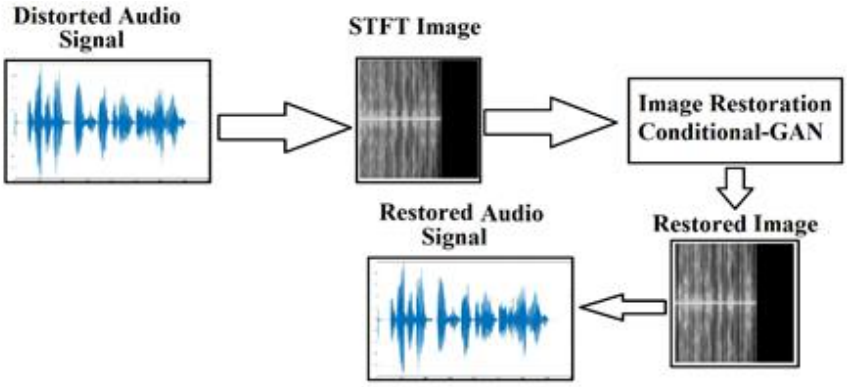


Figure 2. The proposed algorithm for audio restoration.

In the last decade convolutional neural networks (CNN) outperform show efficient way of image classification algorithms, so CNN is used as an image distortion classifier in this work. Instead of building a neural network from scratch, pre-trained deep neural networks are used; since a pre-trained network results in less time for training. Usually, the last layer of the network(classification layer) is modified to fit the new classification problem; this process is called transfer learning, after that the neural network is retrained in order to adjust the current weights of the pre-trained neural network[13]. The VGG16 is a CNN trained on ImageNet dataset to classify 1000 objects (cats, dogs, cars, etc.)[14]. In this work, as shown in figure 3, by using the transfer learning technique, the VGG16 is used as the image distortion classifier with 15 classes: 14 classes specify types of distortions and one class for the non-distorted image.

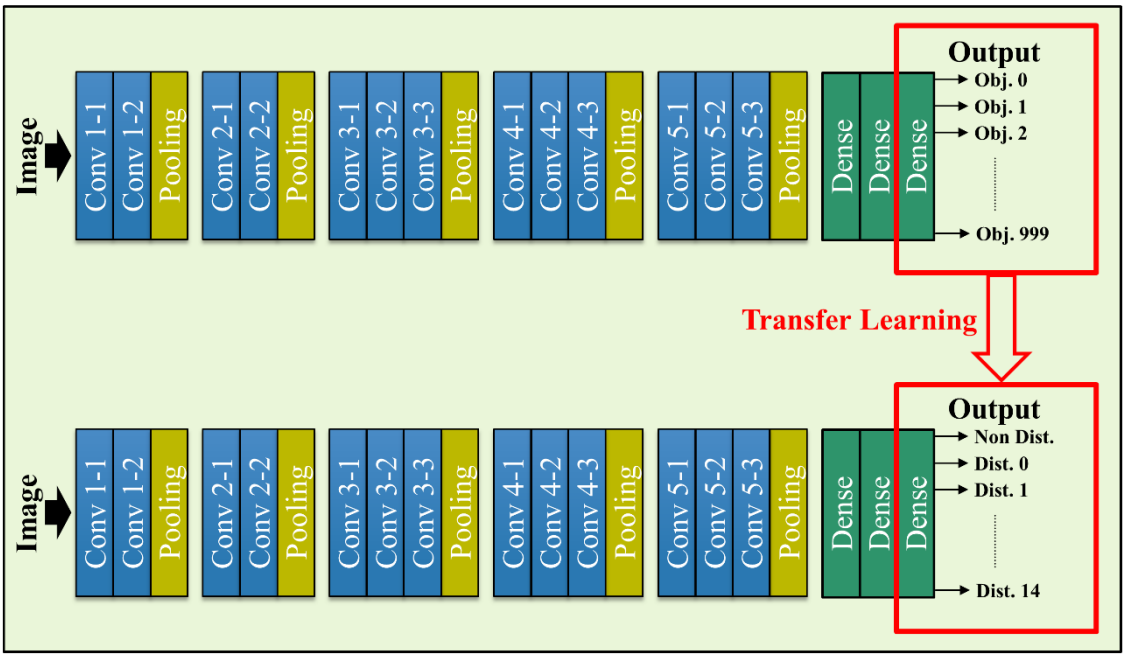


Figure 3. VGG16 transfer learning for using it as a distortion classifier.

The image is considered to be restored to the approximate original image after the classifier figures out the distortion type in the image. Here, a conditional-GAN is used for image restoration as in DeblurGAN[15]. As shown in figure 4, the DeblurGAN is a conditional-GAN and consists of two networks. The generator takes the distorted image and produces an approximation of the sharp image; the discriminator tries to determine if it is produced by a generator(fake image) or a real image. They use two-loss functions to train the network, one for the adversarial loss function that focuses on texture details by using Wasserstein GAN and the latter for the content loss function for producing a restored image with the similar general content of the distorted image[15]. After the conditional-GAN is trained, only the generative model will be used to restore images. In this work, the same conditional-GAN is trained several times; each time is trained to restore images for a specific distortion type and save the generative model's weights. When images input the system, as in figure 1, the classifier determines the distortion types, then the corresponded weights of the generative model are loaded into the conditional-GAN for image restoration. This procedure is also applied for the audio signal restoration by treating the audio signal as an image using the STFT. The distorted audio signals are restored using the same conditional-GAN using signal's STFT as restoring other image distortion, then converting the restored image to audio form.

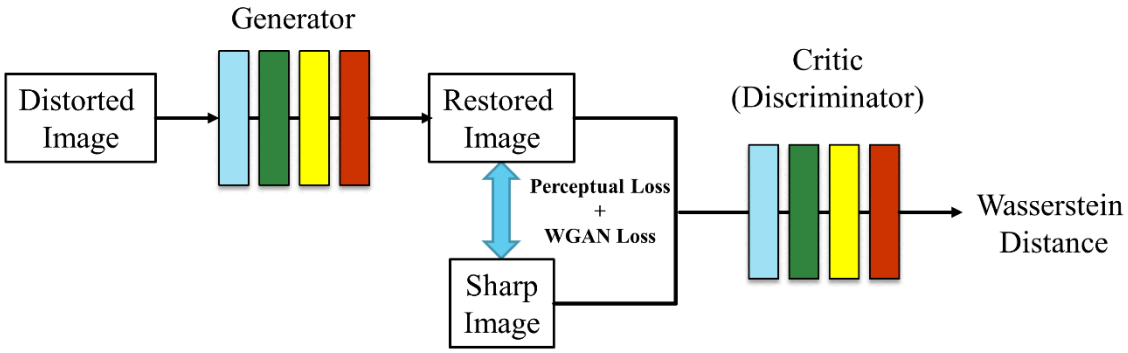
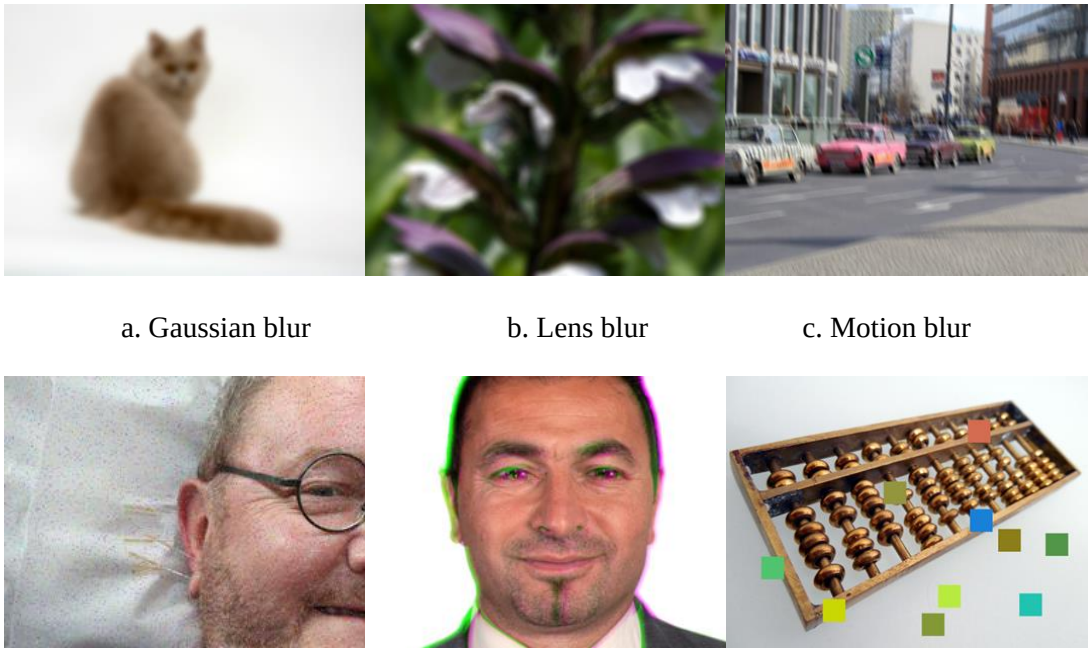


Figure 4. DeblurGAN Architecture[15]

III. DATASET AND RESULTS

Deep learning-based approaches require gigantic dataset size for both training and evaluating the model; for this reason, the KADIS-700K dataset is used in this work. The KADIS-700K contains 140,000 original non-distorted images, supplied with an algorithm to synthesis distortion images of 25 distortion types in 5 levels. Level 1 is the weakest distortion, and level 5 is the strongest one[16]. In this research, only 31,000 reference images were selected randomly from the KADIS-700K dataset and using 14 types of distortion with levels 4 and 5. The distortion types include blurring(Gaussian blur, lens blur, and motion blur), color distortions(color shift, color quantization), compression(JPEG2000, and JPEG), noise(white noise, white noise in color component, impulse noise, multiplicative noise, denoise which is produced by adding Gaussian white noise to RGB image and then applying a denoising DnCNN to each channel separately), and spatial distortions(Jitter_bicubic interpolation and Color block). Figure 5 shows some distortion types used in this work.



a. Gaussian blur b. Lens blur c. Motion blur

d. Impulse noise e. Color shift f.Color block

Figure 5. Some of the distortion types are used in this work.

For training the distortion classifier(VGG16), the trainset is 31080 images of 15 classes (14 classes of distortion types and 1 class for non-distorted image), each class consists of 2072 images, the validation set, and test set is 1080 images for each set (72 images for each class). After training the classifier for about 5 hours, the validation set accuracy is 95.5% and 93.4% for the test set, as shown below.

The same 2072 image for each distortion class (used in training the distortion classifier) is also processed here for image restoration considering DeblurGAN. These 2072 blurred images are treated as input, and the corresponding

2072 original images are treated as a target. Figure 6 shows some results for image restoration with their metrics measurements, Figure 7 shows an image restoration that suffers from two distortion types.

Distored image	Restored image	Original image
MSE=967.45	MSE=1079.93	
SNR=9.532	SNR=9.0545	
PSNR=18.27	PSNR=17.79	
SSIM=0.717	SSIM=0.727	
FSIM=0.7055	FSIM=0.8553	

(a)

Distored image	Restored image	Original image
MSE=596.54	MSE=331.19	
SNR=15.09	SNR=17.64	
PSNR=20.37	PSNR=22.93	
SSIM=0.825	SSIM=0.866	
FSIM=0.6652	FSIM=0.8126	

(b)

Distored image	Restored image	Original image
MSE=498.10	MSE=204.10	
SNR=16.35	SNR=20.22	
PSNR=21.15	PSNR=25.03	
SSIM=0.942	SSIM=0.947	
FSIM=0.9178	FSIM=0.9747	

(c)

Distored image	Restored image	Original image
MSE=440.18	MSE=227.18	
SNR=14.50	SNR=17.37	
PSNR=21.69	PSNR=24.56	
SSIM=0.947	SSIM=0.923	
FSIM=0.9413	FSIM=0.9654	

(d)

Figure 6. Some of the original, distorted, and restored images are used in this work.

Sharp image	Add Gaussian blur	Remove Gaussian
MSE=334.18	MSE=116.19	
SNR=17.46	SNR=21.64	
PSNR=22.76	PSNR=27.47	
SSIM=0.840	SSIM=0.854	
FSIM=0.8124	FSIM=0.9060	

Add two distortions

Add to the sharp image: Gaussian blur&Color block	Remove color block distortion	Remove Gaussian
MSE=958.78	MSE=462.41	MSE=347.76
SNR=13.06	SNR=16.09	SNR=16.92
PSNR=18.31	PSNR=21.48	PSNR=22.71
SSIM=0.788	SSIM=0.797	SSIM=0.806
FSIM=0.7994	FSIM=0.8086	FSIM=0.84609

Figure 7. An image restoration for an image suffering from two distortion types.

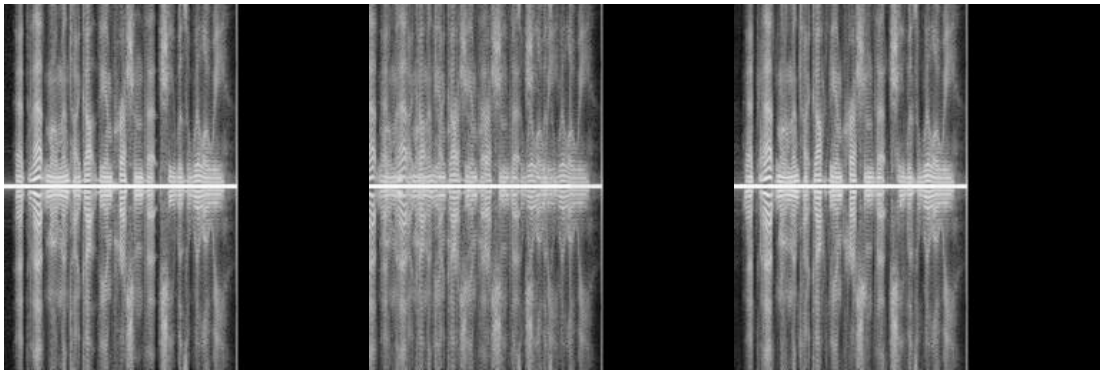
For audio signals restoration, an online speech dataset of 1132 speech signals is used[17]. From this dataset, another dataset(audio_dataset) is produced such that part of it contains the clear version of the audio signals and another part is distorted to become the distorted version. Each audio signal has a duration of 5s (46662 samples) and is sampled at 8192 Hz. When the audio signal is less than five-second length, it is extended by zeros(silent); but if it is larger than 5s, it is cut to only 5s. As described in equation 3 below, the distorted version is created by applying the echo effect by overlapping the audio signal with its self delayed by 0.5s. This distortion could be seen as the equivalent of motion blur in the image distortion.

$$y(t) = x(t) + x(t - 0.5) \ldots\ldots \quad (3)$$

where $x(t)$ is the original signal and $y(t)$ is the distorted version. An image dataset(STFT_dataset) is created by taking STFT for each audio signal in the audio dataset (audio_dataset) for both original and distorted signals. The STFT parameters: windows hamming(256,'symmetric'); overlapping length = 150 samples, and the audio signal length = 256 samples. The STFT matrix($STFT\_abs\_log$) is created, as shown in equation 4, by applying the natural logarithm on the magnitude of the STFT($STFT\_abs$).

$$STFT\_abs\_log = log(STFT\_abs + 0.0001) \quad \ldots\ldots \quad (4)$$

The STFT matrix($STFT\_abs\_log$) has a size of 256 × 438, then encoded as the gray image of 8-bit depth, then reshaping it to 256 × 438×3 image (by repeating the image two times). Finally, it is resized as an image of the size of 256 × 256 ×3. So the STFT_dataset consists of 1132 pairs of clear and distorted images of speech signals. One thousand pairs are used as the trainset and the remaining as the test set. After training the image restoration model(conditional-GAN), the model is tested by restoring the distortion audio(its STFT) then back to audio to be evaluated. Figure 8 shows an example of STFT images including original, distorted, and the restored signals. The restored audio signals are evaluated subjectively and objectively using two metrics: mean absolute error(MAE) and SNR relative to the clear audio signal, the restoration performance is shown in table 1.



(a) Orginal STFT (b) Distorted STFT (c) Restorted STFT

Figure 8. STFT images for orginal, distorted and restorted audio signal.

Table1. Restoration performance of audio signals.

Distorted signal		Restored signal		Restoration improvement	
MAE	SNR	MAE	SNR	MAE decreased	SNR increased
0.34	8.56	0.22	12.47	64.38%	145.74%

IV. CONCLUSIONS

Images and audio signals suffer from distortions due to several factors; the restoration process plays a fundamental role in this problem. There is no perfect way to recover the exact non-distorted version in real-world problems, but in recent years using deep learning, the approach has become prevalent, giving restored results very close to the original version. This paper presented a new hybrid algorithm for both image and signal restoration of many types of distortion. The distorted image first passes to the distortion classifier(CNN). Based on the classified result, the image restored using conditional-gan by selecting the corresponding network weights; the procedure may be repeated to solve another distortion type. The same approach is used to restore audio signals using STFT technology.

REFERENCES

[1] D. Perrone and P. Favaro, "A clearer picture of total variation blind deconvolution," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, pp. 1041-1055, 2015.

[2] S. Athar and Z. Wang, "A Comprehensive Performance Evaluation of Image Quality Assessment Algorithms," *Ieee Access*, vol. 7, pp. 140030-140070, 2019.

[3] S. Jain and M. S. Goswami, "A Comparative Study of Various Image Restoration Techniques With Different Types of Blur," *International Journal of Research in Computer Applications and Robotics*, vol. 3, pp. 54-60, 2015.

[4] Y. Chang, "Research on de-motion blur image processing based on deep learning," *Journal of Visual Communication and Image Representation*, vol. 60, pp. 371-379, 2019.

[5] X. Xu, J. Pan, Y.-J. Zhang, and M.-H. Yang, "Motion blur kernel estimation via deep learning," *IEEE Transactions on Image Processing*, vol. 27, pp. 194-205, 2017.

[6] J. Sun, W. Cao, Z. Xu, and J. Ponce, "Learning a convolutional neural network for non-uniform motion blur removal," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 769-777.

[7] D. W. Kim, J. R. Chung, J. Kim, D. Y. Lee, S. Y. Jeong, and S. W. Jung, "Constrained adversarial loss for generative adversarial network-based faithful image restoration," *ETRI Journal*, vol. 41, pp. 415-425, 2019.

[8] Z. Hong, X. Fan, T. Jiang, and J. Feng, "End-to-End Unpaired Image Denoising with Conditional Adversarial Networks," in *AAAI*, 2020, pp. 4140-4149.

[9] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535-557, 2017.

[10] J. SIMON, R. GODSILL, and J. PETER, *DIGITAL AUDIO RESTORATION*: SPRINGER LONDON Limited, 2013.

[11] D. M. Nogueira, C. A. Ferreira, E. F. Gomes, and A. M. Jorge, "Classifying heart sounds using images of motifs, MFCC and temporal features," *Journal of medical systems*, vol. 43, p. 168, 2019.

[12] D. Hepsiba and J. Justin, "Role of deep neural network in speech enhancement: A review," in *International Conference of the Sri Lanka Association for Artificial Intelligence*, 2018, pp. 103-112.

[13] L. Xiao, F. Heide, W. Heidrich, B. Schölkopf, and M. Hirsch, "Discriminative transfer learning for general image restoration," *IEEE Transactions on Image Processing*, vol. 27, pp. 4091-4104, 2018.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[15] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "Deblurgan: Blind motion deblurring using conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8183-8192.

[16] H. Lin, V. Hosu, and D. Saupe, "KADID-10k: A large-scale artificially distorted IQA database," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, 2019, pp. 1-3.

[17] Available: http://tts.speech.cs.cmu.edu/awb/cmu_arctic/

AUTHORS

Omar Hatif Mohammed received the B.Sc. and M.Sc. degree in Computer Engineering Technology from Northern Technical University, Iraq, in 2012 and 2015. He is currently a Ph.D. candidate in computer engineering at the University of Mosul, Iraq. His researches interests include pattern recognition, image restoration, and signal processing.

Basil Sh. Mahmood was born in 1953 in Mosul/ Iraq, graduated in 1976 from the University of Mosul/ Electrical department, and the M.Sc. degree in Electronics and Communications in 1979. Then he joined in Computer Center of the same university as an assistant lecturer then after he got the degree of Ph.D. on microprocessors architecture in 1996. Now, he is a microprocessors and computer architecture professor in the Computer Engineering Department/the University of Mosul. He published with others four books and more than 50 research papers in many journals and conferences. He supervised more than 22 M.Sc. and 14 Ph.D. Students. His interests are in microprocessors, computer architectures, image and signal processing, modern methods of Artificial Intelligence. He was awarded many prizes and Medals.