TENCON 2023 - 2023 IEEE Region 10 Conference (TENCON)
31 Oct - 3 Nov 2023. Chiang Mai, Thailand

Mixed Noise Suppression Using UNET and Its Variants

Milan Tripathi and Toshiaki Kondo
School of Information and Communication Technology
Sirindhorn International Institute of Technology, Thammasat University
Pathum Thani, Thailand
Email: m6522040036@g.siit.tu.ac.th, tkondo@siit.tu.ac.th

Abstract— **Image denoising holds significant importance in the realm of image processing due to the potential distortions caused by environmental factors and technical problems. Consequently, it is logical to consider image denoising as a critical research domain as it aids in addressing various other image processing challenges. Although numerous techniques for image denoising have emerged in recent years, a majority of them primarily focus on restoring images afflicted by a single source of noise. In this study, the effectiveness of UNET and its variant in denoising facial images with mixed noises is examined. Furthermore, traditional filtering techniques are investigated for the purpose of comparison. The experimental results indicate the insufficiency of conventional filtering techniques in effectively mitigating mixed noise in facial images. Conversely, employing UNET-based architectures yields promising outcomes, characterized by facial images exhibiting commendable values of peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). Furthermore, the denoised images produced by employing the proposed residual attention UNET exhibit notable enhancements in terms of clarity and intricate details.**

Keywords—Image Denoising, Deep Learning, Traditional Filtering, UNET, Residual, Attention.

I. Introduction

Image denoising is a crucial task in image processing, aiming to recover the original image by reducing noise in a noisy version. It helps solve other image processing problems and is divided into two main techniques: traditional filtering [1]-[3] and deep learning [4]. Traditional filtering uses mathematical operations and spatial filters like mean or median filters, but it has limitations in complex scenarios. Deep learning utilizes convolutional neural networks to learn noise patterns from pairs of noisy and clean images, resulting in more accurate denoising. Image denoising is an important area of research with the potential to enhance various image processing applications.

Deep learning, a subset of machine learning, employs models like transformers [5], LSTM [6], and CNN [7] to mimic human learning. The encoder-decoder [8] method excels in image denoising. Training deep learning systems demands substantial data to adjust parameters, yet the same architecture can handle diverse noise types without modifications. By utilizing input and target, the encoder-decoder generates a mathematical function. Iterative learning allows the model to approximate the desired output, resulting in target-resembling images upon reaching a threshold.

In this paper, our objective is to design experiments and models for denoising mixed-noise facial images. Our main contributions are: (1) We introduce novel combinations of Gaussian and Salt & Pepper noise for experimentation. (2) We propose two UNET-based architectures with attention and simplified residual blocks, enhancing the extraction of relevant information. (3) We compare our methods with other deep learning and traditional denoising approaches to evaluate their performance.

II. Literature

The popularity of deep learning in image processing has grown due to abundant data availability and powerful processors. Numerous neural network-based techniques have been developed specifically for image denoising. Ghose et al. [9] proposed a CNN model for image denoising, achieving better qualitative and quantitative results compared to traditional filtering methods. However, their comparison did not include current methods and used a limited number of noise levels. Ramos et al. [10] introduced RDUNet, a residual dense neural network, for image denoising. Their approach demonstrated competitive results without requiring knowledge of the noise level. However, the study employed constrained noise levels and necessitated separate training for each noise type. Limshuebchuey et al. [11] compared traditional and deep learning-based denoising algorithms using PSNR on Gaussian and Salt and pepper noise conditions. The comparisons showed that the deep learning algorithm yielded superior PSNR values. However, the evaluation lacked other metrics and the dataset size was limited. Olaf Ronneberger [12] introduced a UNET structure for biological image segmentation, incorporating a skip connection between the encoding and decoding layers. This innovation improved image formation by allowing selective data transfer. O. Oktay et al. [13] proposed the attention gate (AG) model for medical imaging. AGs automatically focus on target structures of varying sizes and shapes, eliminating the need for external localization modules. They enhance CNN architectures like U-Net, improving sensitivity and accuracy without significant computational overhead. Evaluation on CT abdominal datasets shows consistent performance improvement and computational efficiency. Zhang et al. [14] introduced RatUNet, an enhanced deep convolutional U-Net framework for image denoising. It improves network depth, down-sampling, up-sampling, skip-connection, and utilizes depthwise and polarized self-attention mechanisms. RatUNet achieves better performance than existing methods, although it focuses on removing AWGN noise only.

979-8-3503-0219-6/23/$31.00 ©2023 IEEE

III. Image Noises

A multitude of factors can be responsible for producing image noise. The primary focus of this investigation is on Gaussian and Salt & Pepper noise. Gaussian noise is generated when a random Gaussian function is introduced into an image function, as illustrated in Equation (1), which depicts the process of adding Gaussian noise to a denoised image.

$$y = x + w \times G \qquad (1)$$

The noised image is represented by the variable $'y'$, whereas the noise-free original image is represented by the variable $'x'$. The Gaussian noise is denoted by the variable $'G'$, and the noise factor is denoted by the variable $'w'$.

Salt and pepper noise is generated by randomly adding bright and dark areas to an image. To add salt and pepper noise to an image, two parameters are used. The first parameter, $'S'$, determines the amount of noise added to the image, while the second parameter controls the proportion of salt and pepper noise. This second parameter is always set to 0.5, indicating that the impact of salt and pepper noise will be equal.

Mixed noise occurs when an image is distorted by more than one type of noise. In this study, images corrupted with seven distinct types of mixed noise are utilized, and a visualization of these images is displayed in Fig. 8.

IV. Denoising Methods

UNET and its variants are investigated in this study to denoise images.

A. UNET

The Autoencoder model can maintain the dimensions of an image, but its linear evaluation of the input creates a bottleneck that restricts the complete transmission of information. However, the UNET [12] model overcomes this limitation by integrating a skip connection that enables feature representations to bypass the bottleneck. Fig. 1 provides a graphical representation of the UNET model's architecture.

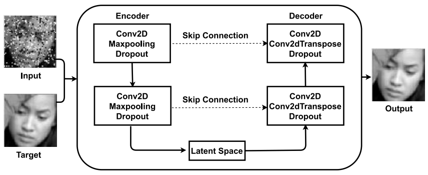

Fig. 1. UNET architecture.

B. Attention UNET

UNET is quite effective, but it uses a lot of GPU memory and wastes resources on pointless activations. Attention UNET addresses these problems by emphasizing only the pertinent activations during training. This cuts down on the amount of time lost on unimportant activations and improves the method's generalizations. Fig. 2 shows a structure of the Attention UNET [13] model's architecture.

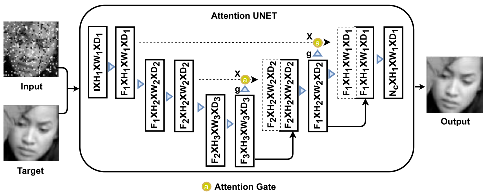

Fig. 2. Attention UNET architecture.

In Fig. 2, $'F'$ corresponds to the number of output channels, $'D'$ corresponds to the number of input channels, and $'H'$ and $'W'$ represent the length and depth of the convolution kernel.

In the Attention UNET model, the skip connection is augmented by an attention gate that requires two inputs, $'x'$ and $'g'$. The gating signal $'g'$ corresponds to the next lower layer in the network and has a superior feature representation as it originates from a deeper area of the network. On the other hand, the input $'x'$ is a skip connection that contains more precise spatial information because it comes from the earlier layers. Consequently, only relevant information is allowed to flow through the skip connection.

In Fig. 3, attention gate is shown in more detail.

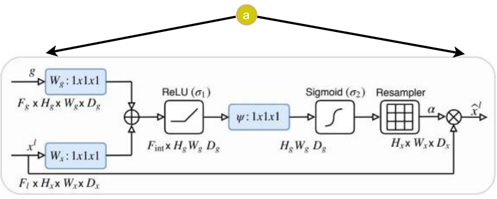
a
g
$W_g: 1x1x1$
$F_g \times H_g \times W_g \times D_g$
x^l
$W_x: 1x1x1$
$F_l \times H_x \times W_x \times D_x$
ReLU (σ_1)
$F_{int} \times H_g W_g D_g$
$\psi: 1x1x1$
Sigmoid (σ_2)
$H_g W_g D_g$
Resampler
$H_x \times W_x \times D_x$
α
$\hat{x}^l$

Fig. 3. Attention gate in-depth.

The input $'x'$ originates from the upper layer, resulting in a higher-dimensional representation, while $'g'$ originates from the lower layer, resulting in a lower-dimensional representation. To ensure consistent dimensions, appropriate convolution operation is applied, followed by concatenation. Subsequently, a rectified linear unit (ReLU) activation is applied, and the resulting tensor is passed through a convolutional layer with a filter count of 1. This yields a 1-depth vector, which represents the weight of the input. To ensure interpretability, the ReLU output is further processed using a sigmoid activation function. Finally, the obtained weight vector is upsampled to match the size of $'x'$, and element-wise multiplication is performed with $'x'$. This scaling operation adjusts the vector based on its relevance.

C. Residual Attention UNET

The flow of data in conventional feedforward neural networks occurs sequentially, where the output of one layer serves as the input for the next layer. Residual [14] connections offer an additional pathway for data to reach later segments of the neural network, bypassing certain layers.

In the case of feedforward neural networks, training a deep network can be challenging because of issues like vanishing gradients and exploding gradients. However, the use of residual connections in a neural network has been demonstrated to lead to much easier convergence during training, even with networks containing hundreds of layers.

979-8-3503-0219-6/23/$31.00 ©2023 IEEE

Fig. 4 shows the architecture of residual blocks.

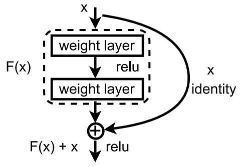

Fig. 4. Residual block.

where, $'x'$ represents input, $'F(x)'$ represents the mapping function then $F(x)+x$ is the output gained from the network. Let's assume that the desired output is denoted by function $'H(x)'$ then, $H(x) \coloneqq F(x)+x$. As the model becomes more larger there will be gradient disappearance in the network resulting $F(x)=0$, then the function $H(x)=x$ serves as an identity mapping that eliminates convolution layers and reduces the depth of a network, while still maintaining accuracy.

Assume $'R(x)'$ is the residual then,

$$R(x)=H(x)-x \qquad (2)$$

Rearranging, we get,

$$H(x)=R(x)+x \qquad (3)$$

The primary objective of the residual block is to learn the actual output, $'H(x)'$. It can be observed from the Fig. 4 that due to the presence of an identity connection from $'x'$, the layers are, in fact, attempting to learn the residual component, $'R(x)'$. Consequently, the block is referred to as a Residual Block.

An example of residual attention UNET architecture is shown in Fig. 5.

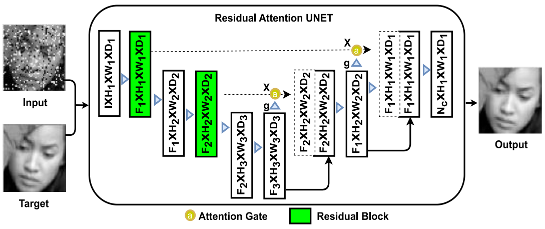

Fig. 5. Residual attention UNET.

V. Experiment Setup

In this experiment, the FER2013 dataset [15] is employed, which comprises 35887 grayscale images with a resolution of 48×48 pixels. For training and validation of the deep learning model, a total of 28709 images (80%) and 7078 images (19.72%), respectively, are used. Additionally, 100 images are kept for testing purposes to compare the performance of the conventional filter and deep learning algorithms. While the validation data improves the model during training, the testing data confirms its effectiveness.

To produce noisy data, several combinations of Gaussian noise with a mean of 0 and standard deviation of 1, and salt and pepper noise are utilized. These combinations are divided into three categories: Noise Added without Overlapping (Top & Down (G: 30 & S: 4%) and Side by Side (G: 30 & S: 4%)), Noise Added with Overlapping (Full Overlap (G: 30 & S: 4%), Full Overlap (G: 20 & S: 4%) and Full Overlap (G: 30 & S: 1.8%)), and Noise Added with Partial Overlapping (Partial Overlap (G: 30 & S: 4%) and Complex Overlap (G: 30 & S: 4%)) .Fig. 8 provides a visualization of these examples. The $'G'$ value represents the Gaussian Noise Factor, which regulates the amount of Gaussian noise added to the image. As the $'G'$ value exceeds 0, the image becomes more corrupted. $'S'$ represents Salt & Pepper, which indicates the percentage of image pixels that will be replaced with noise on a scale of [0, 1]. As the value approaches 1, image blurriness increases. The noise factors of the two noises are adjusted to achieve identical PSNR values when comparing the image with its noised and noise-free versions.

The first five mixed noises can be easily understood by examining the images and their corresponding descriptions. For Partial Overlap, the two noises overlap each other at the area that is six pixels away from the center of the image on both sides horizontally. It constitutes 25% of the total area of the image. For Complex Overlap, Salt & Pepper noise is added to half of the image horizontally on the left side, and Gaussian noise is added to half of the image vertically on the bottom part. The noise overlaps at the lower left side, and the upper right side of the image is noise-free. It occupies 75% of the overall image area.

To evaluate the performance, two metrics, Peak signal-to-noise ratio (PSNR) [16] and Structural Similarity Index (SSIM) [17], are utilized.

Mathematically, PSNR can be represented as,

$$PSNR(dB)=10log_{10}\left[MAX_{Signal}^{2}/MSE\right] \qquad (4)$$

where $'MSE'$ represents mean square error of all the pixels in the images and $'MAX'$ represents the maximum value of pixel.

Mathematically, SSIM can be represented as,

$$S(x,y)=l(x,y)\cdot c(x,y)\cdot s(x,y) \qquad (5)$$

where $'l'$ represents luminance, $'c'$ represents contrast and s represents structural similarities between two images $'x'$ and $'y'$.

For UNET schemes, this experiment uses UNET architecture based on this paper [18].

The diagram of proposed Attention UNET architecture used in this research is shown in Fig. 6 and the diagram of proposed Residual Attention UNET architecture is shown in Fig. 7.

In this experiment, Traditional Filtering schemes were employed using specific parameters: Median Filter, Gaussian Filter with a standard deviation (σ) of 1, and Bilateral Filter with $\sigma_1=1$ and $\sigma_2=0.3$. All filters have a size of 3×3. Additionally, BM3D [19] with $\sigma=30$ was utilized. For Deep Learning schemes, UNET-based models were trained using specific hyperparameters, including the Adam optimizer, 10 epochs, a batch size of 64, a learning rate of 0.001, and Mean Squared Error (MSE) loss function.

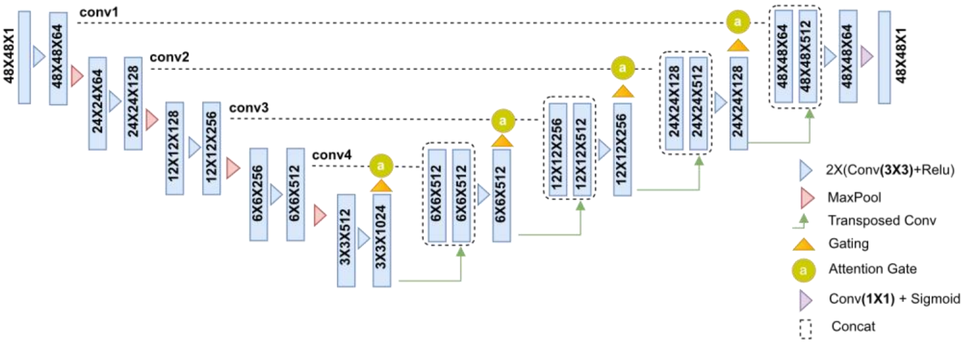

Fig. 6. Architecture of proposed attention UNET.

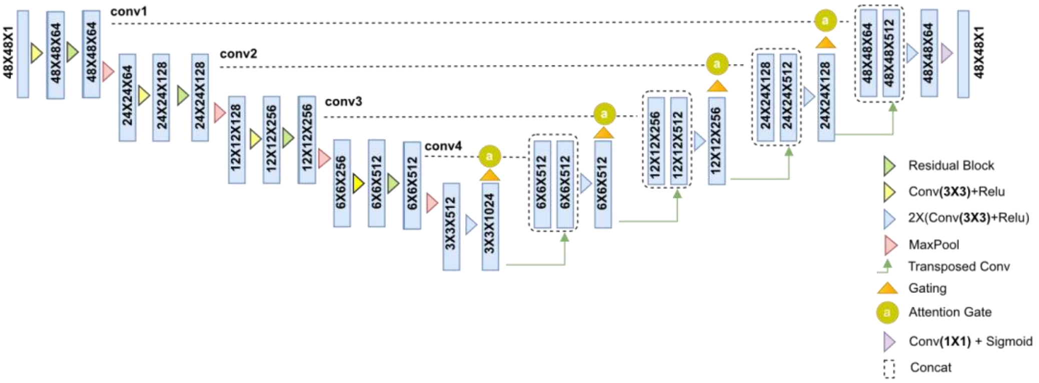

Fig. 7. Architecture of proposed residual attention UNET.

VI. RESULTS AND DISCUSSION

Tables I and II display the outcomes of comparing an original image that is free of noise with images that have been corrupted by various mixtures of Gaussian and Salt & Pepper noise. In addition, the tables showcase the results of image denoising for these corrupted images through the use of various techniques such as Gaussian Filter, Median Filter, Bilateral Filter, BM3D, UNET, Attention UNET and Residual Attention UNET. The metric values obtained from the denoising process are presented in the Tables I and II.

TABLE I. PSNR GAINED BEFORE AND AFTER DENOISING

Seven Styles of Noise Mixture (Gaussian + S&P)	**Noisy Image**	**Gaussian**	**Median**	**Bilateral**	**BM3D [19]**	**UNET [18]**	**Proposed Attention UNET**	**Proposed Residual Attention UNET**
1. Top & Down (G:30 & S:4%)	18.8399	24.1452	24.8698	21.6051	24.7184	29.7757	30.0286	**30.0373**
2. Side by Side (G:30 & S:4%)	18.8188	24.1774	24.8911	21.6085	24.7012	29.4725	29.8376	**29.8799**
3. Full Overlap (G:30 & S:4%)	15.9222	22.3650	22.8052	18.7671	21.2442	26.7138	26.7744	**26.8218**
4. Full Overlap (G:20 & S:4%)	17.4327	23.1841	24.5596	19.8160	22.9010	28.1612	28.8056	**28.8319**
5. Full Overlap (G:30 & S:1.8%)	17.1746	23.4601	23.0945	20.6884	23.9008	26.7099	26.8523	**26.8722**
6. Partial Overlap (G:30 & S:4%)	17.8299	23.6261	24.2922	20.6333	23.7847	28.3086	28.8061	**28.8660**
7. Complex Overlap (G:30 & S:4%)	18.8702	24.2588	24.8064	21.7276	24.2208	29.0547	29.5508	**29.5980**

TABLE II. SSIM GAINED BEFORE AND AFTER DENOISING

Seven Styles of Noise Mixture (Gaussian + S&P)	**Noisy Image**	**Gaussian**	**Median**	**Bilateral**	**BM3D [19]**	**UNET [18]**	**Proposed Attention UNET**	**Proposed Residual Attention UNET**
1. Top & Down (G:30 & S:4%)	0.6529	0.8463	0.8644	0.7678	0.8678	0.9550	**0.9584**	0.9578
2. Side by Side (G:30 & S:4%)	0.6525	0.8467	0.8664	0.7658	0.8638	0.9545	0.9562	**0.9580**
3. Full Overlap (G:30 & S:4%)	0.4958	0.7797	0.7862	0.6356	0.7516	0.9105	0.9127	**0.9150**
4. Full Overlap (G:20 & S:4%)	0.5824	0.8152	0.8508	0.6929	0.8114	0.9386	0.9433	**0.9444**
5. Full Overlap (G:30 & S:1.8%)	0.5506	0.8142	0.7947	0.7133	0.8395	0.9099	0.9145	**0.9149**
6. Partial Overlap (G:30 & S:4%)	0.6081	0.8266	0.8448	0.7255	0.8394	0.9369	0.9442	**0.9448**
7. Complex Overlap (G:30 & S:4%)	0.6949	0.8546	0.8647	0.7893	0.8567	0.9499	0.9546	**0.9555**

The results of denoising noise from images indicate that the residual attention UNET algorithm is more effective than other algorithms in nearly all cases. In Fig. 8, examples of images that were denoised using different filters can be seen, including Gaussian, median, bilateral, BM3D, UNET, attention UNET, and residual attention UNET. Traditional filters do not improve the quality of images, and as the noise becomes more complex, their performance worsens. However, deep learning-based filters, such as residual attention UNET, can produce high-quality denoised images. In fact, images denoised using residual attention UNET exhibit slightly greater clarity and detail than other UNET variants, which is clearly visible in Fig. 9.

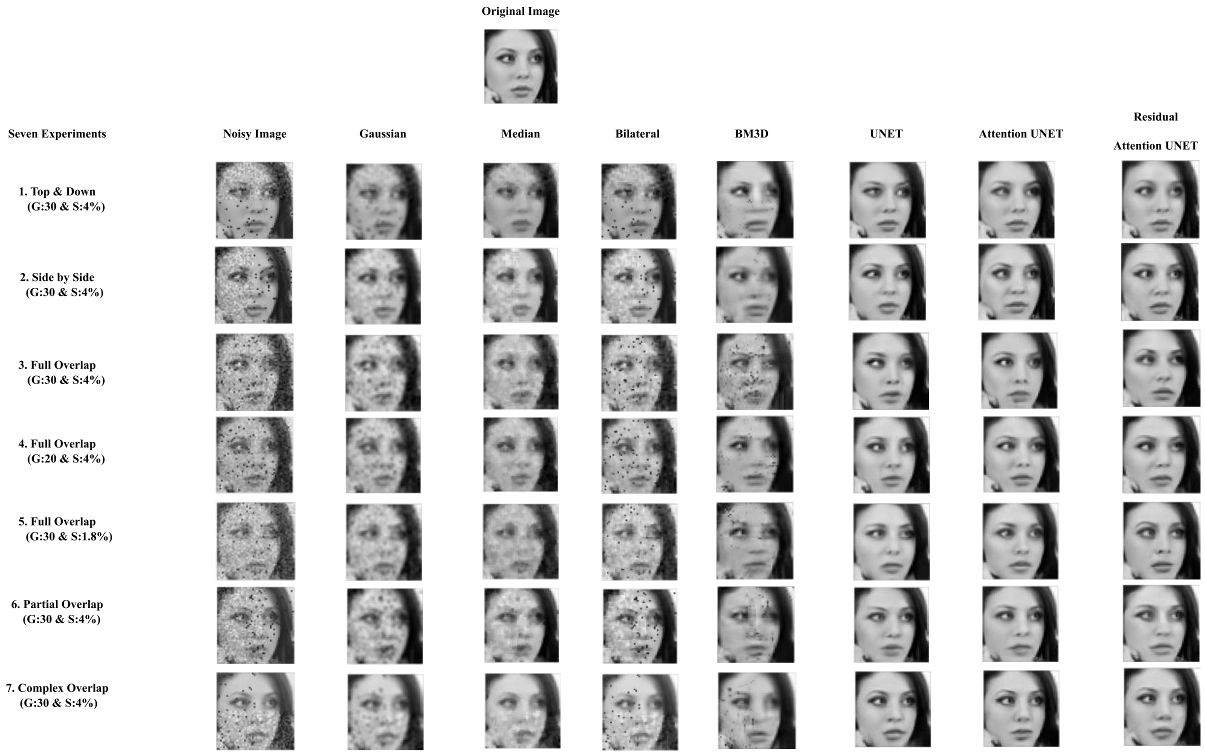
Original Image
Seven Experiments
Noisy Image
Gaussian
Median
Bilateral
BM3D
UNET
Attention UNET
Residual Attention UNET
1. Top & Down (G:30 & S:4%)
2. Side by Side (G:30 & S:4%)
3. Full Overlap (G:30 & S:4%)
4. Full Overlap (G:20 & S:4%)
5. Full Overlap (G:30 & S:1.8%)
6. Partial Overlap (G:30 & S:4%)
7. Complex Overlap (G:30 & S:4%)

Fig. 8. Visualization of image denoising on images corrupted by the mixture of the Gaussian and salt & pepper noise.

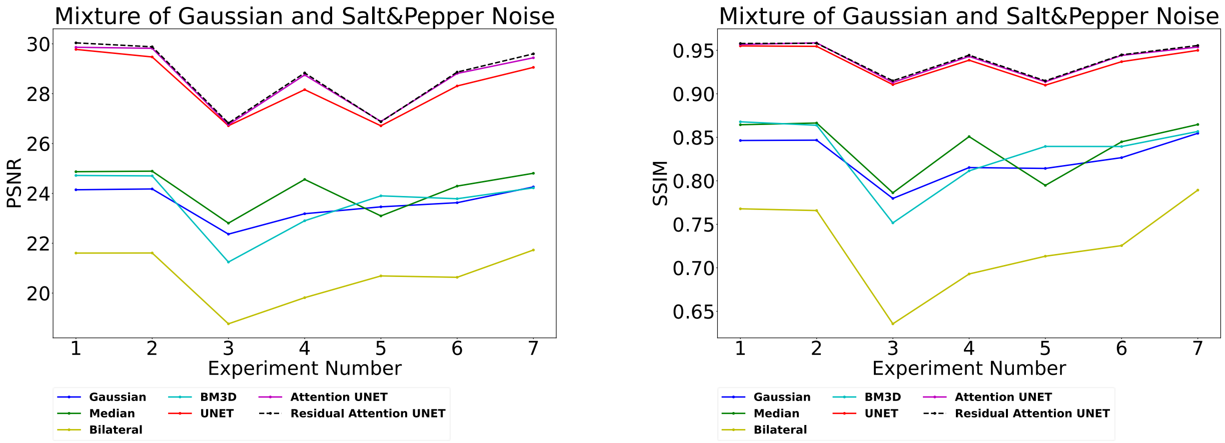

Fig. 9. Visualization of image denoising on images corrupted by the mixture of the Gaussian and salt & pepper noise. Experiment Number 1. Top & Down (G:30 & S:4%), 2. Side by Side (G:30 & S:4%)), 3. Full Overlap (G:30 & S:4%), 4. Full Overlap (G:20 & S:4%), 5. Full Overlap (G:30 & S:1.8%), 6. Partial Overlap (G:30 & S:4%), and 7. Complex Overlap (G:30 & S:4%).

VII. CONCLUSION

This research presents a comparative analysis of several UNET-based algorithms designed to remove mixed noise from facial images. Additionally, conventional filters, including Gaussian, Median, Bilateral, and BM3D, are also evaluated. The findings indicate that traditional filters are inadequate for mixed noise denoising, whereas UNET-based architectures achieve facial images with high PSNR and SSIM values. Moreover, the denoised images generated using proposed residual attention UNET are characterized by slightly greater clarity and detail than other UNET variants.

In order to enhance the denoising performance, our plan is to explore a more advanced deep learning architecture.

ACKNOWLEDGMENT

This research is financially supported by Sirindhorn International Institute of Technology (SIIT), Thammasat University (TU).

REFERENCES

[1] M. Wang, S. Zheng, X. Li, and X. Qin, "A new image denoising method based on Gaussian filter," in Proceedings - 2014 International Conference on Information Science, Electronics and Electrical Engineering, ISEEE 2014, 2014, vol. 1. doi: 10.1109/InfoSEEE.2014.6948089.

[2] X. Li, J. Ji, J. Li, S. He, and Q. Zhou, "Research on image denoising based on median filter," in IMCEC 2021 - IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference, 2021. doi: 10.1109/IMCEC51613.2021.9482247.

[3] D. Bhonsle, V. Chandra, and G. R. Sinha, "Medical image denoising using bilateral filter," International Journal of Image, Graphics and Signal Processing, vol. 4, no. 6, 2012, doi: 10.5815/ijigsp.2012.06.06.

[4] I. H. Sarker, "Deep Learning: A comprehensive overview on techniques, taxonomy, applications and research directions," SN Computer Science, vol. 2, no. 6. 2021. doi: 10.1007/s42979-021-00815-1.

[5] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, 2017, vol. 2017-December.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[7] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," Insights into Imaging, vol. 9, no. 4. 2018. doi: 10.1007/s13244-018-0639-9.

[8] J. Zhai, S. Zhang, J. Chen, and Q. He, "Autoencoder and its various variants," in Proceedings - 2018 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2018, 2019. doi: 10.1109/SMC.2018.00080.

[9] S. Ghose, N. Singh, and P. Singh, "Image denoising using deep learning: Convolutional neural network," in Proceedings of the Confluence 2020 - 10th International Conference on Cloud Computing, Data Science and Engineering, 2020. doi: 10.1109/Confluence47617.2020.9057895.

[10] J. Gurrola-Ramos, O. Dalmau, and T. E. Alarcón, "A residual dense U-Net neural network for image denoising," IEEE Access, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3061062.

[11] A. Limshuebchuey, R. Duangsoithong, and M. Saejia, "Comparison of image denoising using traditional filter and deep learning methods," in 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, ECTI-CON 2020, 2020. doi: 10.1109/ECTI-CON49241.2020.9158242.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2015, vol. 9351. doi: 10.1007/978-3-319-24574-4_28.

[13] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," arXiv.org, 2018, https://arxiv.org/abs/1804.03999.

[14] H. Zhang, Q. Lian, J. Zhao, Y. Wang, Y. Yang, and S. Feng, "RatUNet: Residual U-Net based on attention mechanism for image denoising," PeerJ Computer Science, vol. 8, p. e970, May 2022, doi: https://doi.org/10.7717/peerj-cs.970.

[15] "FER-2013 | Kaggle." https://www.kaggle.com/datasets/msambare/fer2013 (accessed Feb. 26, 2023).

[16] J. Korhonen and J. You, "Peak signal-to-noise ratio revisited: Is simple beautiful?," in 2012 4th International Workshop on Quality of Multimedia Experience, QoMEX 2012, 2012. doi: 10.1109/QoMEX.2012.6263880.

[17] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error measurement to structural similarity," IEEE transactions on image processing, vol. 13, no. 4, 2004.

[18] M. Tripathi, "Facial image denoising using autoencoder and UNET," Heritage and Sustainable Development, vol. 3, no. 2, pp. 89-96, Oct. 2021.

[19] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising with block-matching and 3D filtering," in Image Processing: Algorithms and Systems, Neural Networks, and Machine Learning, 2006, vol. 6064. doi: 10. 1117/12.643267.