UDiNet: A dilated U-net for improving OCR performance

Mahyar Fardinfar, Pouya Rashidikia, Mohammad Reza Rezaie, Mina Zolfy Lighvan
Department of Computer Engineering,
Electrical and Computer Engineering Faculty,
University of Tabriz
Tabriz, Iran
mahyarfardinfar@gmail.com, pouyarashidikia102@gmail.com,
mohammadreza.rezaei@sharif.edu, mzolfy@tabrizu.ac.ir

Abstract

Image denoising is a critical task in the field of computer vision. This paper introduces the UDiNet architecture, a dilated variant of the U-Net, specifically designed to address image denoising challenges. We present a novel dataset comprising book sheet images to rigorously evaluate the performance of the proposed method. Experimental results demonstrate that UDiNet significantly enhances the performance of established Optical Character Recognition (OCR) systems, such as Tesseract and Genome. The model effectively mitigates severe noise while preserving essential structural details of English characters. This capability positions UDiNet as a valuable preprocessing technique for various applications, including classification, detection, and OCR tasks. To promote further research in this domain, we have made the code and trained models publicly accessible.

Keywords
Image noise reduction, Image denoising, U-net architecture, dilated convolution, OCR

1. Introduction

Image noise reduction is yet an open problem in computer vision that focuses on emboldening the quality and detail of images, to make them more suitable for subsequent tasks such as classification, detection, segmentation, generation, etc. However, challenges in image denoising include accurately modeling complex noises relative to the environment, preserving important image details while reducing noise, and developing methods that can effectively eliminate or suppress diverse noise groups (Ilesanmi and Ilesanmi 2021).

Convolutional neural networks (CNNs) are a specific type of deep learning model that aims to process and extract features of matrices (Taye 2023). The convolutional layers typically combine with pooling layers, normalizer layers, dropout layers, fully connected layers, etc. Convolutional layers apply a set of learnable filters to the input image, where each filter extracts a specific feature like edges, shapes, or textures. By sliding the filters across the image and computing dot products, convolutional layers produce feature maps that capture the responses of the filters at different spatial locations. This allows CNNs to efficiently learn features independent of their position on the image grid (Dhillon and Verma 2020).

Dilated convolutions are a variation of regular convolutional operations that create spaces or holes within the filter. This design allows the filter to encompass a larger receptive field, thereby facilitating the capture of multi-scale contextual information without increasing the number of parameters (Wang et al. 2019). In a dilated convolution, the filter is applied over an area larger than its spatial dimensions by selectively skipping input values, with the spacing defined by the dilation rate. By stacking dilated convolutions with varying dilation rates, the network can effectively aggregate features from multiple scales, which is particularly advantageous for dense prediction tasks such as image denoising.

Optical Character Recognition (OCR) is a crucial technology in the field of computer vision and document analysis, enabling the conversion of text within images into machine-readable formats. This process involves complex algorithms that detect and recognize individual characters and words from various types of documents, including scanned papers, photographs, and digitally created images. In our research, we have leveraged OCR as a benchmark

task to evaluate the effectiveness of our image denoising model. By comparing the OCR accuracy on original noisy images against the results obtained from images processed through our denoising algorithm, we can quantitatively assess the impact of our noise reduction techniques (Hegghammer 2022). This approach not only demonstrates the practical applicability of our model but also provides a tangible measure of improvement in a real-world scenario, where enhanced image quality directly correlates with increased OCR precision and reliability.
The code for this research would be published online in order to help enhance future works.

2. Literature Review

Numerous CNN-based methods have been explored for image denoising, addressing various noise sources such as Gaussian, impulse, salt & pepper, and speckle noise, with extensive studies categorizing and analyzing these techniques and their evaluation on popular datasets (Ghose, Singh, and Singh 2020). Recent advancements in image denoising have led to the development of deep learning-based methods like Att-ResUNet, which combines attention mechanisms and residual UNet networks to effectively capture and remove noise while preserving fine details in complex denoising tasks (Ding et al. 2024).

Recent studies on discriminative model learning, such as DnCNN, have shown significant improvements in image denoising by leveraging residual learning and batch normalization to handle various noise levels and enhance denoising performance (Zhang et al. 2017). FFDNet, a fast and flexible denoising convolutional neural network with a tunable noise level map, addresses the limitations of existing discriminative denoisers by effectively handling a wide range of noise levels and spatially variant noise, achieving a good trade-off between inference speed and denoising performance (Zhang, Zuo, and Zhang 2017).

In the realm of self-supervised image denoising, methods like Neighbor2Neighbor have demonstrated effective denoising performance using only noisy images, overcoming the limitations of requiring large amounts of noisy-clean image pairs for supervision (Huang et al. 2021). Additionally, Noise2Void (N2V) extends the concept of Noise2Noise (N2N) by enabling training without the need for noisy image pairs or clean target images, making it particularly useful for applications like biomedical imaging where acquiring training targets is challenging (Krull, Buchholz, and Jug 2019).

Multi-scale Dilated Convolution (MsDC) networks use dilated filters to aggregate multi-scale contextual information without reducing the receptive field, effectively addressing the limitations of conventional methods and enhancing image denoising performance (Jia et al. 2022). Enhanced convolutional neural denoising network (ECNDNet) leverages residual learning, batch normalization, and dilated convolutions to address training difficulties, performance saturation, and computational cost, demonstrating superior performance over state-of-the-art methods in image denoising (Tian et al. 2019). The multiscale dilated residual network (MDRNet) employs dilated convolutions and a hybrid dilation rate pattern to enhance image quality, achieving competitive denoising performance with fewer parameters compared to existing convolutional network-based methods (Li et al. 2020).

Recent advancements in image restoration have seen the emergence of Swin Transformer-based models, such as SUNet, which integrate Swin Transformer layers into the UNet architecture, surpassing traditional CNN-based methods in performance for high-level vision tasks (Fan, Liu, and Liu 2022). Combining adaptive weighted median filtering, VGG-16, and U-Net networks, recent methods have shown significant improvements in PSNR and SSIM values for real image denoising, effectively retaining more image detail information (Li, Wang, and Tao 2023).

3. Proposed Method

3.1 UDiNet Architecture

The U-Net architecture consists of a contracting path (encoder) and an expansive path (decoder). The encoder progressively reduces spatial resolution while increasing feature map depth through a series of convolutional layers, leaky ReLU activations, and max pooling operations. This process doubles the number of feature channels at each downsampling step, enabling the capture of increasingly abstract representations. The bottleneck comprises three convolutional layers that produce a highly compressed, information-rich feature map. Importantly, in our modified architecture, parallel to the first convolutional layer in each block, there is a dilated convolutional layer. This not only helps the filters to capture features in a wider feature area but also helps the model to learn the general features of characters in the presence of noises (Ronneberger, Fischer, and Brox 2015).

The expansive path, or decoder, reconstructs the compressed feature map back to the original input size through a series of upsampling operations. Each upsampling step is followed by a convolutional layer that halves the number of feature channels and incorporates skip connections from the corresponding layers in the contracting path. These skip connections are critical as they concatenate the feature maps from the encoder to the decoder, preserving spatial information that may have been lost during the downsampling process and enhancing the network's ability to generate precise segmentations and a gradient vanish or explosion. The upsampling is typically performed using transposed convolutions, which increase the spatial dimensions of the feature maps. But, due to the checker-boards effect we have used convolutional layers. Each upsampling step is followed by a series of convolutional layers with leaky ReLU activations to refine the feature maps. The final layer of the U-Net employs a 1x1 convolution to map each feature vector to the desired pixel value; in our case it is an integer ranging from 0 for black to 255 for white.

Our model begins with five parallel streams of dilated and standard convolutional layers. This is followed by two classic U-Net encoder blocks that leverage the architecture's strength in preserving spatial information and 4 layers of convolution as a bottleneck. The decoder consists of seven classic U-Net decoder layers, which progressively upsample and refine the feature maps. Notably, skip connections are applied only to the U-Net blocks, facilitating the flow of fine-grained spatial information from the encoder to the decoder (Fig. 1).

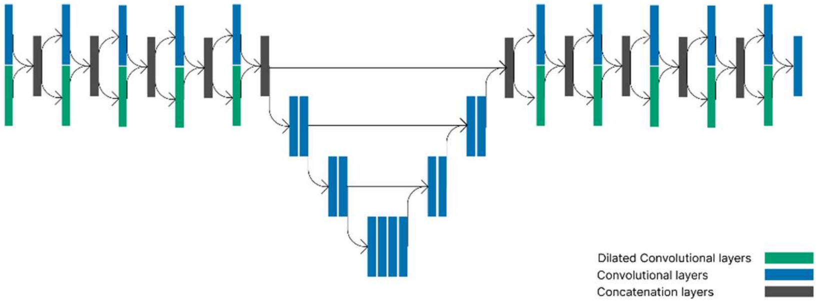


Figure 1. The proposed UDiNet architecture, which is a combination of U-net at the middle part and our hybrid approach at both sides. The blue and green blocks are representative of classic and dilated convolution layers while gray blocks are used as the concatenation layers.

3.2 Loss function

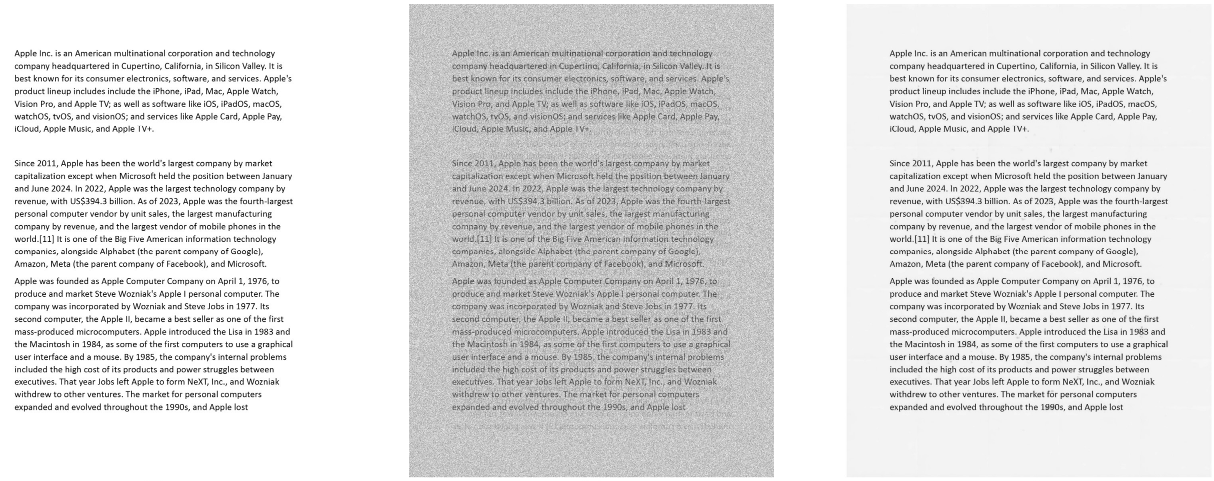
The Elastic Loss Function, also known as Elastic Net Regularization, is a hybrid approach that combines the penalties of both Ridge (L2) and Lasso (L1) regression methods. The Elastic Net loss function is defined as:

$$L(y,\hat{y}) = \frac{\alpha}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| + \frac{\beta}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i|^2$$

Here, ɑ is a mixing parameter that balances the contributions of the L1 and L2 penalties. When ɑ is set to zero, the Elastic Net reduces to Ridge regression, and when ɑ is set to one, it simplifies to Lasso regression. This flexibility allows the Elastic Net to benefit from the strengths of both methods: the L2 penalty helps in faster optimization for bigger loss values due to its bigger derivative value the bigger it gets, while the L1 penalty induces sparsity by driving some coefficients to zero. By tuning the parameters ɑ and β , one can achieve a balance that optimally fits the specific characteristics of the data set, leading to improved model performance and generalization.

4. Data Collection

For this study, we compiled a dataset by sourcing a diverse array of PDFs available online, from which we manually selected usable pages to create a collection of 1,500 book sheets. To enhance the dataset's applicability for model training, we introduced synthetic noise to the images, including normal noise, flip effects, and intensity noise, while ensuring that each noisy image was paired with its corresponding clean ground truth (a sample is provided in Fig. 2). This approach not only simulates the real-world imperfections often encountered in document images but also provides a robust foundation for evaluating the performance of our models in handling various noise conditions.



Apple Inc. is an American multinational corporation and technology company headquartered in Cupertino, California, in Silicon Valley. It is best known for its consumer electronics, software, and services. Apple's product lineup includes include the iPhone, iPad, Mac, Apple Watch, Vision Pro, and Apple TV; as well as software like iOS, iPadOS, macOS, watchOS, tvOS, and visionOS; and services like Apple Card, Apple Pay, iCloud, Apple Music, and Apple TV+.

Since 2011, Apple has been the world's largest company by market capitalization except when Microsoft held the position between January and June 2024. In 2022, Apple was the largest technology company by revenue, with US$394.3 billion. As of 2023, Apple was the fourth-largest personal computer vendor by unit sales, the largest manufacturing company by revenue, and the largest vendor of mobile phones in the world.[11] It is one of the Big Five American information technology companies, alongside Alphabet (the parent company of Google), Amazon, Meta (the parent company of Facebook), and Microsoft.

Apple was founded as Apple Computer Company on April 1, 1976, to produce and market Steve Wozniak's Apple I personal computer. The company was incorporated by Wozniak and Steve Jobs in 1977. Its second computer, the Apple II, became a best seller as one of the first mass-produced microcomputers. Apple introduced the Lisa in 1983 and the Macintosh in 1984, as some of the first computers to use a graphical user interface and a mouse. By 1985, the company's internal problems included the high cost of its products and power struggles between executives. That year Jobs left Apple to form NeXT, Inc., and Wozniak withdrew to other ventures. The market for personal computers expanded and evolved throughout the 1990s, and Apple lost

(a)

Apple Inc. is an American multinational corporation and technology company headquartered in Cupertino, California, in Silicon Valley. It is best known for its consumer electronics, software, and services. Apple's product lineup includes include the iPhone, iPad, Mac, Apple Watch, Vision Pro, and Apple TV; as well as software like iOS, iPadOS, macOS, watchOS, tvOS, and visionOS; and services like Apple Card, Apple Pay, iCloud, Apple Music, and Apple TV+.

Since 2011, Apple has been the world's largest company by market capitalization except when Microsoft held the position between January and June 2024. In 2022, Apple was the largest technology company by revenue, with US$394.3 billion. As of 2023, Apple was the fourth-largest personal computer vendor by unit sales, the largest manufacturing company by revenue, and the largest vendor of mobile phones in the world.[11] It is one of the Big Five American information technology companies, alongside Alphabet (the parent company of Google), Amazon, Meta (the parent company of Facebook), and Microsoft.

Apple was founded as Apple Computer Company on April 1, 1976, to produce and market Steve Wozniak's Apple I personal computer. The company was incorporated by Wozniak and Steve Jobs in 1977. Its second computer, the Apple II, became a best seller as one of the first mass-produced microcomputers. Apple introduced the Lisa in 1983 and the Macintosh in 1984, as some of the first computers to use a graphical user interface and a mouse. By 1985, the company's internal problems included the high cost of its products and power struggles between executives. That year Jobs left Apple to form NeXT, Inc., and Wozniak withdrew to other ventures. The market for personal computers expanded and evolved throughout the 1990s, and Apple lost

(b)

Apple Inc. is an American multinational corporation and technology company headquartered in Cupertino, California, in Silicon Valley. It is best known for its consumer electronics, software, and services. Apple's product lineup includes include the iPhone, iPad, Mac, Apple Watch, Vision Pro, and Apple TV; as well as software like iOS, iPadOS, macOS, watchOS, tvOS, and visionOS; and services like Apple Card, Apple Pay, iCloud, Apple Music, and Apple TV+.

Since 2011, Apple has been the world's largest company by market capitalization except when Microsoft held the position between January and June 2024. In 2022, Apple was the largest technology company by revenue, with US$394.3 billion. As of 2023, Apple was the fourth-largest personal computer vendor by unit sales, the largest manufacturing company by revenue, and the largest vendor of mobile phones in the world.[11] It is one of the Big Five American information technology companies, alongside Alphabet (the parent company of Google), Amazon, Meta (the parent company of Facebook), and Microsoft.

Apple was founded as Apple Computer Company on April 1, 1976, to produce and market Steve Wozniak's Apple I personal computer. The company was incorporated by Wozniak and Steve Jobs in 1977. Its second computer, the Apple II, became a best seller as one of the first mass-produced microcomputers. Apple introduced the Lisa in 1983 and the Macintosh in 1984, as some of the first computers to use a graphical user interface and a mouse. By 1985, the company's internal problems included the high cost of its products and power struggles between executives. That year Jobs left Apple to form NeXT, Inc., and Wozniak withdrew to other ventures. The market for personal computers expanded and evolved throughout the 1990s, and Apple lost

(c)

Figure 2. (a) Grand Truth (b) Synthetic noise (c) UDiNet's output

5. Results and Discussion

Edit distance (Haldar and Mukhopadhyay 2011) was used as the metric to measure the impact of UDiNet on the performance of the Tesseract OCR model. Tessdata-best was chosen as the OCR model which is the most accurate LSTM-based model offered by Tesseract. Uniform and flip noise was then added to a handful of text images and fed into the UDiNet model. noisy images and their denoised versions were given to the OCR model, in order to extract their text and calculate the average edit distance with regards to the original text. The outcomes demonstrate that Tesseract's performance has improved remarkably after de-noising. Different page segmentation modes and OCR engine modes are offered by Tesseract. The LSTM OCR engine mode was used to extract text. Promising results were achieved among different modes, which are demonstrated in Fig. 3.

Based on Fig. 3, Tesseract has trouble extracting text after about a 20% noise ratio, as shown in the above chart, even with various page segmentation modes. Moreover, Tesseract loses its ability to recognize word blocks at a noise ratio of 30%. This isn't the case with images that have been denoised, which continues to operate well even at high noise ratios.

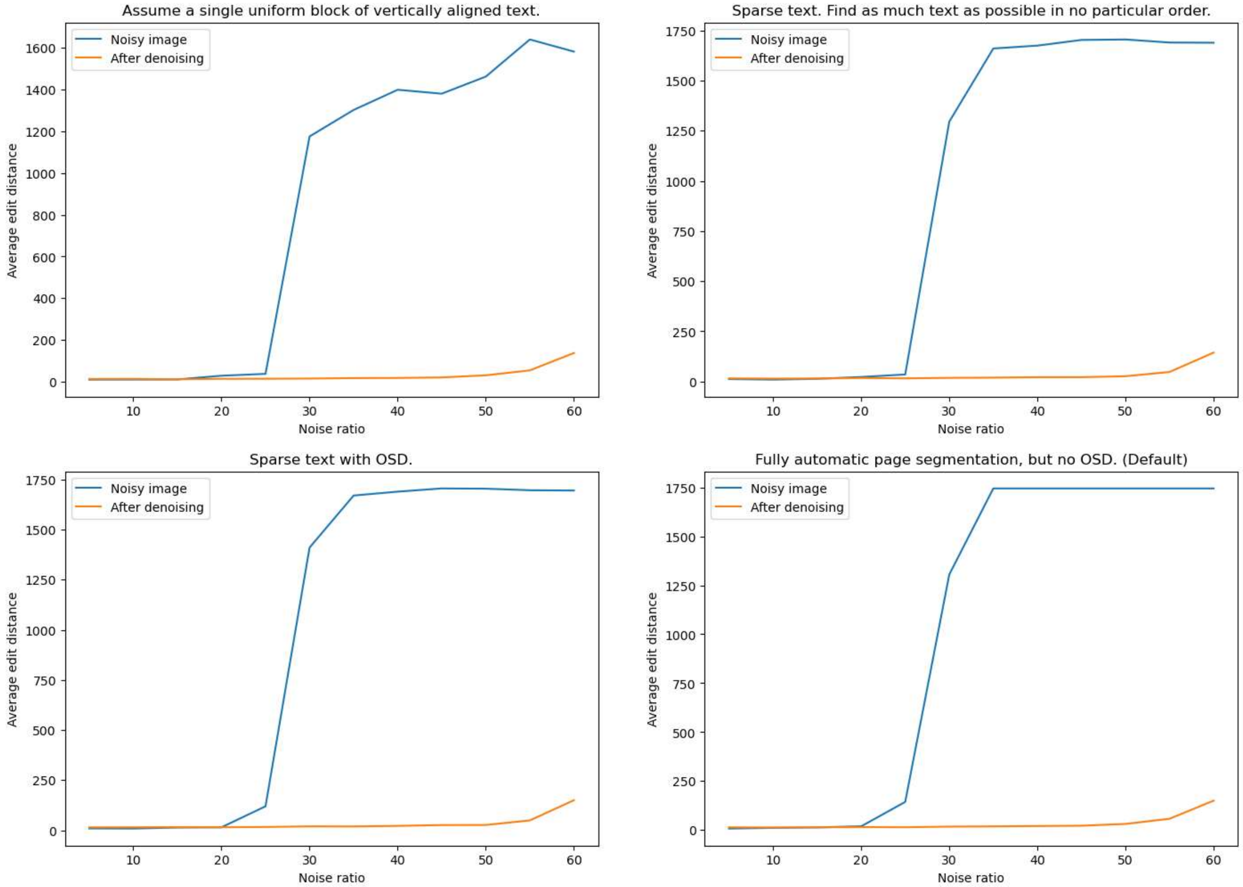


Figure 3. Performance comparison of various Tesseract detection modes on various noise intensities and noise free data

6. Conclusion

In conclusion, this study addresses the ongoing challenges in image noise reduction, a critical problem in computer vision that impacts various downstream tasks. By leveraging the power of Convolutional Neural Networks (CNNs) and incorporating advanced techniques such as dilated convolutions, we have developed a novel approach to tackle the complexities of image denoising. The use of dilated convolutions allows our model to capture multi-scale contextual information efficiently, enhancing its ability to handle complex noise patterns without significantly increasing computational overhead. This research not only contributes to the field of image denoising for OCR but also has potential implications for improving the performance of subsequent computer vision tasks.

References

Dhillon, Anamika, and Gyanendra K. Verma. 2020. 'Convolutional neural network: a review of models, methodologies and applications to object detection', *Progress in Artificial Intelligence*, 9: 85-112.

Ding, Shifei, Qidong Wang, Lili Guo, Jian Zhang, and Ling Ding. 2024. 'A novel image denoising algorithm combining attention mechanism and residual UNet network', *Knowledge and Information Systems*, 66: 581-611.

Fan, Chi-Mao, Tsung-Jung Liu, and Kuan-Hsien Liu. 2022. 'SUNet: Swin Transformer UNet for Image Denoising', *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*: 2333-37.

Ghose, S., N. Singh, and P. Singh. 2020. "Image Denoising using Deep Learning: Convolutional Neural Network." In *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 511-17.

Haldar, Rishin, and Debajyoti Mukhopadhyay. 2011. 'Levenshtein Distance Technique in Dictionary Lookup Methods: An Improved Approach', *Computing Research Repository - CORR*.

Hegghammer, Thomas. 2022. 'OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment', *Journal of Computational Social Science*, 5: 861-82.

Huang, T., S. Li, X. Jia, H. Lu, and J. Liu. 2021. "Neighbor2Neighbor: Self-Supervised Denoising from Single Noisy Images." In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14776-85.

Ilesanmi, Ademola E., and Taiwo O. Ilesanmi. 2021. 'Methods for image denoising using convolutional neural network: a review', *Complex & Intelligent Systems*, 7: 2179-98.

Jia, Xinlei, Yali Peng, Jun Li, Yunhong Xin, Bao Ge, and Shigang Liu. 2022. 'Pyramid dilated convolutional neural network for image denoising', *Journal of Electronic Imaging*, 31.

Krull, A., T. O. Buchholz, and F. Jug. 2019. "Noise2Void - Learning Denoising From Single Noisy Images." In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2124-32.

Li, Dongjie, Huaian Chen, Guoqiang Jin, Yi Jin, Changan Zhu, and Enhong Chen. 2020. 'A multiscale dilated residual network for image denoising', *Multimedia Tools and Applications*, 79: 34443-58.

Li, Xianli, Xueshan Wang, and Yitian Tao. 2023. "Image Denoising Based on Median Filter and UNet." In *2023 IEEE 3rd International Conference on Data Science and Computer Application (ICDSCA)*, 8-14. IEEE.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation." In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, edited by Nassir Navab, Joachim Hornegger, William M. Wells and Alejandro F. Frangi, 234-41. Cham: Springer International Publishing.

Taye, Mohammad M. 2023. "Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions." In *Computation*.

Tian, Chunwei, Yong xu, Lunke Fei, Junqian Wang, Wen Jie, and Nan Luo. 2019. 'Enhanced CNN for image denoising', *CAAI Transactions on Intelligence Technology*, 4.

Wang, Yanjie, Guodong Wang, Chenglizhao Chen, and Zhenkuan Pan. 2019. 'Multi-scale dilated convolution of convolutional neural network for image denoising', *Multimedia Tools and Applications*, 78: 19945-60.

Zhang, Kai, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. 2017. 'Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising', *Trans. Img. Proc.*, 26: 3142–55.

Zhang, Kai, Wangmeng Zuo, and Lei Zhang. 2017. 'FFDNet: Toward a Fast and Flexible Solution for CNN based Image Denoising', *IEEE Transactions on Image Processing*, PP.