EFFICIENT BONE SUPPRESSION IN CHEST X-RAY WITH SIMPLIFIED LOSS FUNCTION

Nguyen Quang Duy, Nguyen Duc Manh, Pham Huy Hoang, Pham Thanh An, Cao Thi Luyen*
University of Transport and Communications

ARTICLE INFO

Received: 19/4/2025
Revised: 30/6/2025
Published: 30/6/2025

KEYWORDS

Chest X Ray
Convolutional neural network
Residual U-Net
Attention U-Net
Res-Patch GAN

ABSTRACT

Chest X-ray imaging is a vital tool for diagnosing thoracic diseases such as pneumonia, lung cancer, and rib fractures. However, rib shadows often obscure or mimic pulmonary lesions - especially in posterior and axillary regions - thereby reducing diagnostic accuracy. Traditional techniques like dual-energy subtraction can separate bone and soft tissue structures but require specialized equipment and may increase radiation exposure. In this paper, we survey representative deep learning architectures applied to the task of bone suppression in standard chest X-ray images, aiming to improve the visibility of lung abnormalities without the need for additional hardware. Our main contribution lies in the use of a simplified loss function that reduces computational complexity while maintaining high inference accuracy. The effectiveness of this loss function is evaluated across various models to demonstrate its performance in suppressing bone shadows.

XÓA XƯƠNG HIỆU QUẢ TRONG ẢNH X-QUANG NGỰC VỚI HÀM MẤT MÁT ĐƠN GIẢN HÓA

Nguyễn Quang Duy, Nguyễn Đức Mạnh, Phạm Huy Hoàng, Phạm Thành An, Cao Thị Luyên*
Trường Đại học Giao thông Vận tải

THÔNG TIN BÀI BÁO

Ngày nhận bài: 19/4/2025
Ngày hoàn thiện: 30/6/2025
Ngày đăng: 30/6/2025

TỪ KHÓA

Chụp X quang ngực
Mạng nơ ron tích chập
Residual U-Net
Attention U-Net
Res-Patch GAN

TÓM TẮT

Chụp X-quang ngực là một công cụ quan trọng trong chẩn đoán các bệnh lý vùng ngực như viêm phổi, ung thư phổi và gãy xương sườn. Tuy nhiên, bóng xương sườn thường che khuất hoặc gây nhiễu với các tổn thương ở phổi, đặc biệt ở các vùng phía sau và nách, làm ảnh hưởng đến độ chính xác trong chẩn đoán. Các phương pháp truyền thống như phân lớp năng lượng kép tuy có thể tách biệt cấu trúc xương và mô mềm nhưng đòi hỏi thiết bị chuyên dụng và có thể làm tăng liều phóng xạ. Trong bài báo này, chúng tôi khảo sát các kiến trúc học sâu tiêu biểu trong bài toán loại bỏ xương từ ảnh X-quang ngực thông thường, giúp nâng cao khả năng quan sát các bất thường ở phổi mà không cần thêm thiết bị phần cứng. Điểm đóng góp chính của nghiên cứu là đề xuất một hàm mất mát đơn giản, giúp giảm độ phức tạp tính toán nhưng vẫn đảm bảo độ chính xác cao trong kết quả suy luận. Phương pháp được đánh giá trên nhiều mô hình khác nhau để làm rõ hiệu quả của hàm mất mát trong nhiệm vụ loại bỏ bóng xương.

DOI: https://doi.org/10.34238/tnu-jst.12628

* Corresponding author. *Email: luyenct@utc.edu.vn*

1. Introduction

Chest radiography is widely used for screening and diagnosing thoracic diseases due to its low cost, accessibility, and minimal radiation exposure. However, the superimposition of anatomical structures, particularly bones like the ribs and clavicles, can obscure vital soft tissue details, reducing diagnostic accuracy for conditions such as pulmonary nodules and infiltrates. To address this limitation, numerous bone suppression techniques have been developed to enhance soft tissue visibility.

Dual-energy subtraction (DES) has traditionally been the gold standard for bone suppression, using two energy levels to separate soft tissue from bone based on their attenuation differences [1]. Despite its accuracy, DES requires specialized equipment, increases radiation exposure, and demands precise patient positioning, limiting its clinical adoption.

To overcome these challenges, deep learning methods have emerged, enabling bone suppression from standard single-energy chest radiographs. These models leverage convolutional neural networks (CNNs) to learn data-driven transformations that emulate DES outputs [2]–[6]. The U-Net architecture [7], with its encoder-decoder structure and skip connections, has become a staple for medical image segmentation due to its ability to preserve spatial and semantic information. Enhancements to this architecture include residual connections, as seen in ResNet [8], which improve convergence and representation learning, and attention mechanisms, as introduced in Attention U-Net [9], which enable the network to focus on clinically relevant regions.

Generative adversarial networks (GANs) have also proven effective for image-to-image translation in bone suppression tasks. Conditional GANs (cGANs), introduced by Isola et al. [10], [11], learn mappings between input radiographs and their bone-suppressed versions using adversarial training. PatchGAN discriminators, which operate on local image patches, help retain high-resolution detail. Advanced GAN architectures such as the dilated cGAN with semantic constraints [12] and spatial feature-maximizing GANs [13] further improve anatomical fidelity and clarity.

In addition to architectural innovations, several studies have introduced novel strategies to improve bone suppression performance. Zarshenas et al. [3] incorporated orientation- and frequency-specific features into CNNs using anatomical priors. Chen et al. [4] leveraged wavelet domain processing within a cascaded CNN, enhancing separation between bone and soft tissue. Gusarev et al. [2] evaluated different model architectures and preprocessing pipelines, while Huynh et al. [5] employed context learning to enhance CheXNet's diagnostic performance [6]. Autoencoder-based approaches [14], bone suppression techniques have been employed to facilitate more accurate tuberculosis classification [15], lightweight CNNs [16], and multi-scale frameworks such as feature pyramid networks (FPNs) [17] have also been proposed to balance accuracy and computational cost.

Benchmark studies have highlighted the clinical relevance and model generalizability in this domain. Sirazitdinov et al. [12] compared various deep learning methods using DES datasets, while Rajaraman et al. [18] demonstrated that bone suppression significantly improves tuberculosis detection accuracy.

Despite these advances, most current approaches rely on complex and computationally intensive loss functions, which hinder real-time clinical deployment. While considerable effort has focused on designing better architectures, relatively little attention has been given to simplifying the optimization process. To address this gap, our study introduces a simplified loss function designed to reduce the computational overhead of model training and inference while maintaining high bone suppression accuracy. By streamlining the optimization process, our loss function facilitates faster convergence and more stable training across diverse deep learning models.

We evaluate this loss function on four representative architectures: U-Net [7], Residual U-Net (U-Net with ResNet blocks) [8], Attention U-Net [9], and a cGAN with PatchGAN discriminator [10], [11], assessing its generalization across both discriminative and generative frameworks.

The remainder of this paper is organized as follows: Section 2 reviews related work on deep

learning-based bone suppression. Section 3 details the experimental setup, results, and discussion. Section 4 concludes the paper and outlines future research directions.

2. Methodology

2.1. Convolutional Neural Networks (CNNs)

CNNs are the foundational deep learning architecture for image processing tasks. CNNs use a series of convolutional layers to learn local features, followed by pooling layers to downsample the image and reduce dimensionality. These models have proven effective in various computer vision tasks, including object detection and image segmentation. Gusarev et al. [2] leveraged CNNs for bone suppression, demonstrating the ability of these networks to extract relevant features for eliminating bone artifacts in chest X-rays. The CNN architecture is derived from [2], with batch normalization and ReLU activation added after each layer. The structure of the CNN model consists of multiple blocks of convolutional layers followed by batch normalization. The network gradually increases the number of feature channels from 16 up to 256 while maintaining the spatial resolution (256×256), and ends with a final convolutional layer that reduces the output to a single channel.

2.2. U-Net

U-Net, introduced by Ronneberger et al. [7], is a specialized CNN architecture designed for semantic segmentation, particularly in biomedical images. It features an encoder-decoder structure with skip connections that help preserve spatial information and enable precise segmentation. U-Net has been widely adopted for various medical image segmentation tasks, including lung and bone suppression in chest radiographs. The network's ability to learn both local and global features makes it particularly suited for tasks where high accuracy in image delineation is crucial.

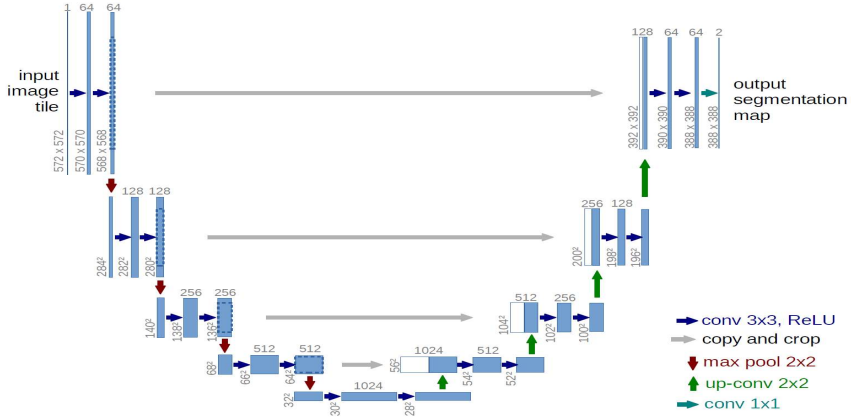

Figure 1. *Block diagram of the original U-Net network [7]*

Figure 1 illustrate the U-Net architecture proposed in [7]. The original U-Net architecture is a symmetric encoder-decoder convolutional neural network designed for biomedical image segmentation. The encoder path (contracting path) captures context through successive 3×3 convolutions with ReLU activations and 2×2 max pooling for downsampling, doubling the number of feature channels at each step. The decoder path (expanding path) performs upsampling using 2×2 transposed convolutions, followed by concatenation with corresponding high-resolution features from the encoder (via skip connections), and further 3×3 convolutions. The final 1×1 convolution maps the feature representation to the desired number of output classes for pixel-wise segmentation. In our implementation, the output layer has a spatial resolution of 256×256 and 1 channel for the bone suppression task.

2.3. Residual U-Net

Residual blocks are designed to handle challenges such as vanishing or exploding gradients, which are prevalent in deep neural networks. These issues arise when backpropagated error signals exponentially diminish or amplify as network depth increases, hindering the training process. To address this, He et al. [8] introduced a specialized block where the input is summed with the output of convolutional operations, forming a shortcut connection. This design enables the network to learn the residual function *G(x),* optimizing *F(x) = G(x) + x.*

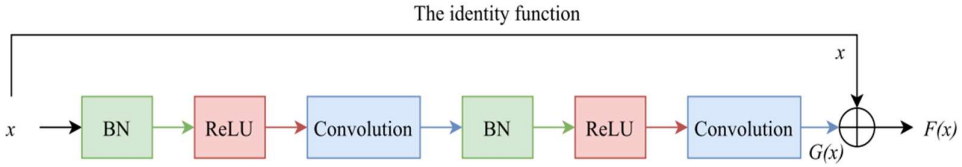

Figure 2. *Structure of the residual block [8]*

Figure 2 depicts a residual block structure, the block applies two sets of Batch Norm, ReLU, Convolution sequentially to compute a residual mapping G(x). This residual output is then added to the original input via an identity shortcut connection, forming the final output F(x) = G(x) + x. In residual U-Net, these residual blocks are incorporated into both the encoder and decoder paths, replacing standard convolutional layers.

2.4. Attention U-Net

The general architecture of the Attention U-Net is presented in Figure 3.

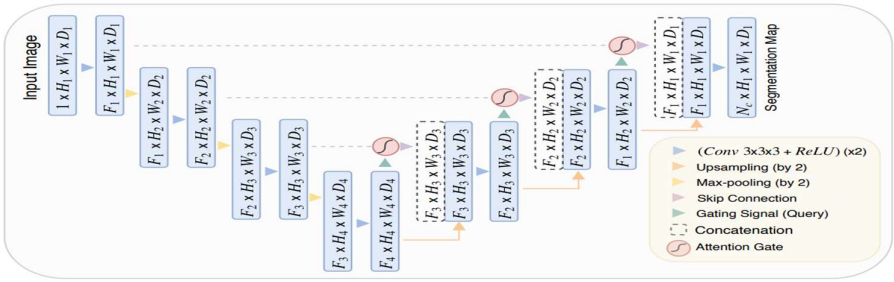

Figure 3. *The structure of Attention U-Net [9]*

The attention mechanism integrated into the U-Net architecture enhances feature selection by refining skip connections using attention gates. These gates receive two inputs: the skip connection x from the encoder and the gating signal g from the corresponding to the decoder layer. Initially, both inputs undergo 1×1 convolutions to align their spatial dimensions and channel depths. Specifically, the spatial resolution of x is reduced via a 1×1 convolution with a stride of two, while g is processed with a 1×1 convolution using a unit stride. The resulting tensors are then summed and passed through a ReLU activation, followed by another 1×1 convolution to generate an intermediate weight map. This map is subsequently normalized using a sigmoid activation function to produce the final attention weights. The weight tensor is upsampled to match the original dimensions of x, and an element-wise multiplication is performed between x and the weight tensor to yield the output of the attention gate.

2.5. Res – Patch GAN

A PatchGAN (Generative Adversarial Network) [11] discriminator and a ResNet-based generator compose the GAN-based bone suppression framework model. A chest X-ray image is

fed into the generator, which tries to generate the corresponding bone-suppressed image. By employing an encoder-decoder architecture with residual blocks, the network can efficiently learn high-level semantic features while preserving low-level information. After concatenating the input and output image, the discriminator separates real and synthetic image pairs. The discriminator employs the N-layer PatchGAN architecture, in which local input patches—rather than the entire image—are processed by convolutional layers with gradually increasing filters. This promotes locally realistic outputs from the generator, enhancing tissue consistency and structural details. A combination of pixel-wise L1 loss and adversarial loss has been employed to optimize the generator during training. The L1 loss ensures structure fidelity to the ground truth, while the adversarial loss boosts realistic texture generation.

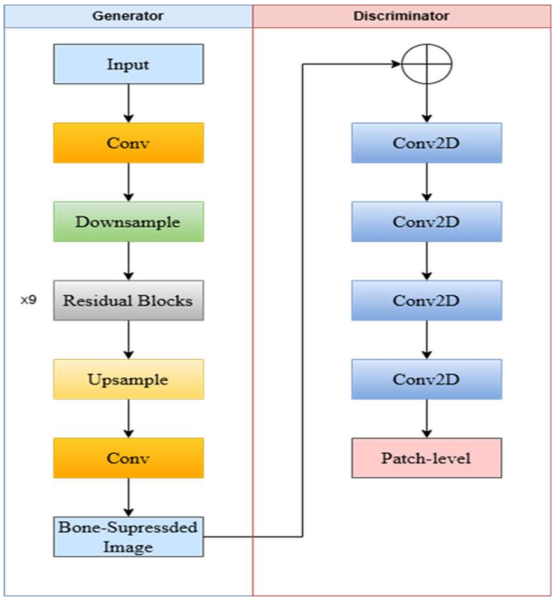

Figure 4. *Structure of the Res-Patch GAN*

Figure 4 illustrates the architecture of the proposed GAN-based model for bone suppression. The framework consists of two main components: Generator and Discriminator. The Generator takes an input image and processes it through an initial convolutional layer, followed by a downsampling operation to reduce spatial dimensions. It then passes through nine stacked residual blocks. The features are then upsampled to recover the original resolution, followed by a final convolutional layer that produces the bone-suppressed image. The Discriminator, structured as a PatchGAN, evaluates the realism of the generated image by processing it through a series of Conv2D layers and outputs a patch-level authenticity map.

3. Experimental Results and Discussion

Previous studies typically used standard loss functions such as Mean Squared Error (MSE), Mean Absolute Error (MAE), or a combination of L1 and adversarial loss—in models like U-Net or GAN. While effective in some cases, these losses are often either too simplistic, resulting in blurry outputs, or too complex, leading to high computational costs and unstable training.

To address these limitations, we propose a simplified composite loss function that balances local accuracy, anatomical fidelity, and perceptual quality. This approach aims to improve training stability and efficiency without sacrificing image quality.

The models were trained and evaluated on a small dataset derived from [19], including 240 chest X-ray images. Specifically, we used 180 images for training, 30 for validation, and 30 for testing. Each image was resized to 256×256. Although the dataset is limited in size, it was chosen for its relevance to the bone suppression task and reflects the practical constraints often encountered in clinical data availability. We applied extensive data augmentation and early stopping to mitigate overfitting. We also specify the main hyperparameters used during training: the learning rate was set to 1e-4, batch size to 16, and training was conducted for up to 150 epochs. To improve generalization and stability, early stopping and learning rate decay based on validation loss were applied.

In our research, the loss function (except for GAN) is formulated as a weighted combination of MAE, Perceptual Loss, and Multi-Scale Structural Similarity Index (MS-SSIM), defined as follows:

$$Loss = MAE + 0.01 * Perceptual\ Loss + (1 - MS\text{-}SSIM) \quad (1)$$

MAE penalizes pixel-wise intensity errors, MS-SSIM preserves anatomical structure across scales, and Perceptual Loss enhances visual realism using VGG-16 feature maps. This combination balances accuracy, structure, and visual quality while reducing training complexity.

The formulation of the simplified loss function was driven by the need to balance pixel-wise accuracy, structural preservation, and visual realism. MAE ensures global consistency by penalizing point-wise intensity errors, while MS-SSIM captures structural similarity across multiple resolutions—a desirable property in preserving anatomical fidelity in medical imaging. Perceptual loss, calculated from intermediate feature maps of a pretrained VGG-16 network, has been shown to suppress artifacts and improve the perceptual quality of synthetic images [4], [15].

The weight of 0.01 assigned to the perceptual loss term was determined through empirical experiments. We evaluated values of 0.001, 0.01, and 0.1 on the Att-UNet-Base model and found that 0.01 consistently yielded the best trade-off: higher perceptual quality without introducing over-smoothing or prolonged training time. This choice was thus based on observed improvements in Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM), as well as faster convergence during training.

This formulation is inspired by prior studies that successfully combined MAE, SSIM, and perceptual loss for medical image synthesis and enhancement tasks, such as in [4], [9], and [15]. These works demonstrated that such hybrid losses improve both structural fidelity and visual realism.

To evaluate the effectiveness of the simplified loss function, we compared it with a more complex alternative using a higher perceptual loss weight (0.1). On the Att-UNet-Base architecture, our proposed loss reduced the average training time per epoch by approximately 18%, and the model achieved convergence (defined by stable validation loss) after 110 epochs compared to 145 epochs with the complex version. Additionally, the standard deviation of validation loss across the final 20 epochs was 25% lower with the simplified loss, indicating improved training stability. These results demonstrate that the proposed formulation not only reduces computational overhead but also leads to faster and more stable convergence.

The model architectures tested in this study include Residual U-Net (Res-UNet) and Attention U-Net (Att-UNet). The Res-UNet model has five layers with channel dimensions of 48, 96, 192, 384, and 768, respectively. Each downsampling operation utilizes a stride of 2, and each downsampling and upsampling layer contains three residual blocks. The Att-UNet architecture follows a similar channel configuration, with the number of channels defined as n, 2n, 4n, 8n, min (16n, 768), where n is set to 16, 32, or 64 based on the model size (small, base, or large). A stride of 2 is used across all layers. All experiments were conducted on an NVIDIA A100 GPU for efficient computation. MAE, MSE, MS-SSIM, SSIM, PSNR values are presented in Table 1.

Upon evaluating the performance of the models, it is clear that the deep learning architectures, particularly Att-UNet variants, outperform the baseline CNN model in all objective metrics. While the CNN model achieves the highest spatial resolution in its output images, as indicated by sharper details and clearer textures, the Att-UNet models demonstrate superior performance in terms of structural similarity and perceptual quality.

Table 1. *Comparisons of each model's performance*

Model	Parameters	MAE	MSE	MS-SSIM	SSIM	PSNR
CNN	395,617	0.0146	0.0008	0.9867	0.9708	32.4085
Res-Patch GAN	10,588,866	0,0283	0,0033	0,9494	0.8957	25.9248
Res-UNet	21,682,113	0.0198	0.0009	0.9867	0.8863	31.7867
Att-UNet-Small	1,987,417	0.0136	0.0006	0.9886	0.9418	33.7402
Att-UNet-Base	7,939,745	0.0142	0.0005	0.9908	**0.9626**	33.8532
Att-UNet-Large	17,857,001	**0.0134**	**0.0004**	**0.9915**	0.9601	**34.2939**

Res-UNet: Despite having a significantly larger number of parameters (over 21 million), Res-UNet demonstrates substantial improvements over CNN in terms of MAE, MSE, and MS-SSIM. This indicates that the inclusion of residual connections in the U-Net architecture aids in better learning of complex features, making it more effective for bone suppression tasks.

Att-UNet: Among the Att-UNet variants, the Att-UNet-Large model consistently outperforms others across all metrics, achieving the highest PSNR (34.29 dB) and SSIM (0.96). The smaller Att-UNet models also show competitive results, with Att-UNet-Base and Att-UNet-Small models maintaining high perceptual quality and relatively low parameter counts.

To further validate the effectiveness of the proposed loss function, we conducted comparison using the same model architecture—Att-UNet-Large. Two versions were trained: one using our simplified loss and the other using a more complex formulation with a perceptual loss weight of 0.1. The simplified loss yielded a PSNR of 34.29 dB and SSIM of 0.9601, while the complex version achieved a PSNR of 34.12 dB and SSIM of 0.9632. Despite the small trade-off in SSIM, the training time per epoch was reduced by approximately 30%, and convergence was achieved 25 epochs earlier on average. These results confirm that our simplified loss achieves comparable or better perceptual and structural performance while significantly reducing computational cost and training time.

Despite their superior performance in structural similarity and perceptual quality, the outputs of the Att-UNet models sometimes exhibit minor artifacts, such as small dark spots, which could be attributed to the downsampling and upsampling processes inherent in the U-Net architecture. These artifacts are particularly noticeable in the bone-suppressed images and indicate challenges in balancing the suppression of bone structures with the preservation of other anatomical features. Interestingly, the CNN model, which retains higher spatial resolution, produces sharper details but at the cost of lower structural and perceptual similarity. This highlights the trade-off between retaining fine spatial details and achieving higher structural accuracy. The CNN's architectural design, which preserves the spatial dimensions of the feature maps, likely contributes to its sharper outputs but struggles with the global structural consistency observed in the Att-UNet models. Figure 5 displays four images arranged from left to right: the original CXR image, the target (bone-suppressed) image, the CNN model's prediction, GAN's prediction, and the Att-UNet-Large model's prediction.

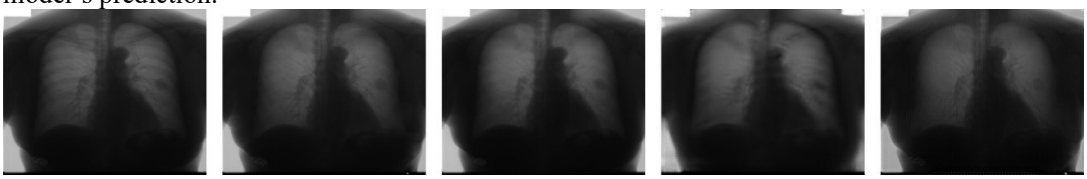

Figure 5. *Original, target, and generated images*

The visual artifacts in Att-UNet outputs, such as isolated dark spots, are influenced by the loss function design. Removing or reducing the perceptual loss leads to sharper outputs but introduces more noise. Increasing the perceptual loss weight to 0.1 removes most artifacts but causes over-smoothing and loss of structural clarity. Experiments show that a weight of 0.01 balances artifact suppression with structural detail and perceptual quality. These findings highlight the role of the loss

function in balancing sharpness and accuracy, with potential for further optimization in future work.

In conclusion, the Att-UNet-based models, particularly Att-UNet-Large, prove to be the most effective for bone suppression tasks in chest radiographs, offering a strong balance between image quality, perceptual accuracy, and computational efficiency.

4. Conclusion

This paper surveys representative deep-learning architectures for bone suppression in standard chest X-rays and demonstrates that streamlined loss functions can deliver high‑accuracy results without added hardware or increased radiation exposure. We introduce novel Attention-UNet variants tailored for osseous reduction while preserving pulmonary detail, and we validate a simplified loss function that markedly lowers computational complexity with minimal impact on inference accuracy—supported by both quantitative metrics and preliminary radiologist feedback. Evaluations on a modest yet diverse dataset confirm robust performance and clinical feasibility. Looking forward, we recommend expanding to larger, more heterogeneous cohorts; increasing input resolution to 512 × 512 pixels to capture finer anatomy; and exploring enhanced UNet variants (e.g., U-Net++ with attention, "sharp" U-Net) alongside generative models (GANs, diffusion, VAEs) to further improve realism. Ultimately, integrating these techniques into AI-assisted diagnostic workflows may streamline interpretation, reduce manual processing, and enhance patient outcomes.

REFERENCES

[1] P. Vock and Z. Szucs-Farkas, "Dual energy subtraction: Principles and clinical applications," *Eur. J. Radiol.*, vol. 72, no. 2, pp. 231–237, 2009.

[2] M. Gusarev, R. Kuleev, A. Khan, A. R. Rivera, and A. M. Khattak, "Deep learning models for bone suppression in chest radiographs," *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, 2017, pp. 1–7.

[3] A. Zarshenas, J. Liu, P. Forti, and K. Suzuki, "Separation of bones from soft tissue in chest radiographs: Anatomy-specific orientation-frequency-specific deep neural network convolution," *Med. Phys.*, vol. 46, no. 5, pp. 2232–2242, 2019.

[4] Y. Chen *et al.,* "Bone suppression of chest radiographs with cascaded convolutional networks in wavelet domain," *IEEE access*, vol. 7, pp. 8346–8357, 2019.

[5] M.-C. Huynh, T.-H. Nguyen, and M.-T. Tran, "Context learning for bone shadow exclusion in CheXNet accuracy improvement," *in Proc. 10th Int. Conf. Knowl. Syst. Eng. (KSE)*, 2018, pp. 135–140.

[6] P. Rajpurkar *et al*., "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," *arXiv preprint*, arXiv:1711.05225, 2017.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint*, *arXiv*:1611.07004, 2016.

[8] G. Rani, A. Misra, V. S. Dhaka, E. Zumpano, and E. Vocaturo, "Spatial feature and resolution maximization GAN for bone suppression in chest radiographs," *Comput. Methods Programs Biomed*., vol. 224, 2022, Art. no. 107024.

[9] Z. Zhou, L. Zhou, and K. Shen, "Dilated conditional GAN for bone suppression in chest radiographs with enforced semantic features," *Med. Phys.*, vol. 47, no. 12, pp. 6207–6215, 2020.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 234–241.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[12] S. Arvind, J. V. Tembhurne, T. Diwan, and P. Sahare, "Evaluation of deep learning methods for bone suppression from dual energy chest radiography," *Artificial Neural Networks and Machine Learning – ICANN 2020*, 2020, pp. 247–257.

[13] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," *arXiv preprint, arXiv*:1612.03144, 2016.

[14] S. Kalisz and M. Marczyk, "Autoencoder-based bone removal algorithm from x-ray images of the lung," *Proc. IEEE 21st Int. Conf. Bioinf. Bioeng. (BIBE)*, 2021, pp. 1–6.

[15] S. Rajaraman, G. Zamzmi, L. Folio, P. Alderson, and S. Antani, "Chest x-ray bone suppression for improving classification of tuberculosis-consistent findings," *Diagnostics*, vol. 11, no. 5, 2021, Art. no. 840.

[16] S. Arvind, J. V. Tembhurne, T. Diwan, and P. Sahare, "Improvised light weight deep CNN based U-Net for the semantic segmentation of lungs from chest X-rays," *Results Eng.*, vol. 17, 2023, Art. no. 100929.

[17] O. Oktay *et al.,* "Attention U-Net: Learning where to look for the pancreas," *arXiv preprint, arXiv*:*1804.03999,* 2018.

[18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1125–1134.

[19] H. M. Chuong, "X-ray bone shadow suppression," *Kaggle*, 2022. [Online]. Available: https://www.kaggle.com/datasets/hmchuong/xray-bone-shadow-supression. [Accessed March 30, 2025].