A Lightweight Framework for Arbitrary Scale Super-Resolution in Ultrasound Imaging

Om Kumar *, Second Author *, *EMBS Student, IEEE*, Third Author *, *EMBS Student, IEEE*, and Debdoot Sheet, *Senior Member, IEEE*

***Abstract*— High-resolution ultrasound (US) imaging is critical in clinical diagnosis, enabling early detection of abnormalities and precise assessment of anatomical structures. While super-resolution (SR) techniques have been widely explored in medical imaging, most existing approaches are restricted to fixed or integer scaling factors. Arbitrary-scale super-resolution, especially for US images, remains largely unaddressed. This study presents a pipeline integrating a lightweight model, ElitNet, with architectural and training modifications to support arbitrary and asymmetric scaling for US images. A resize layer is introduced at the head of the network to accept user-defined scaling factors, and a two-step training strategy is employed to enhance output quality. A dedicated dataset was collected and augmented using flips and reflective padding to ensure structural consistency. Low-resolution images were synthesized using both symmetric and asymmetric scale factors. Our approach yields visually superior results and demonstrates better generalization across arbitrary scales. Quantitatively, it achieves a PSNR of 22.971 and SSIM of 0.6332, outperforming existing baselines such as ArbRCAN, SRDNet, RDUNet, and ABPN. Extensive ablation studies validate the effectiveness of the loss configuration and training strategy. This work lays foundational groundwork for adaptive, high-quality US imaging and opens opportunities for real-time, resource-efficient diagnostic applications.**

***Index Terms*— Arbitrary image SR, deep learning, single image SR, ultrasound imaging**

I. Introduction

In medical imaging, the demand for high-resolution images is paramount. Enhanced resolution improves diagnostic accuracy, aids in surgical planning, and facilitates research in disease understanding. Ultrasound (US), computed tomography (CT), and magnetic resonance imaging (MRI) rely heavily on image quality to extract meaningful clinical insights. Among these, US imaging is widely favoured for its real-time capabilities, cost-effectiveness, and non-invasiveness. However, US inherently suffers from resolution limitations due to hardware constraints, noise, and attenuation, which can obscure critical details and impact diagnostic reliability. Super-resolution (SR) techniques have emerged as a powerful tool to address these resolution limitations. Traditional SR methods [1] aim to upscale low-resolution images by fixed scaling factors such as 2x, 3x, or 4x, enhancing visual and diagnostic quality. While effective, these methods often fail to address real-world medical applications' unique and diverse resolution needs. For instance, US imaging in tissue characterization, vascular studies, or fetal assessments may require tailored resolution adjustments to visualize structures of interest optimally. Fixed scaling factors are insufficient to meet such varied demands. Arbitrary Scale Super-Resolution (ASR) [2] offers a transformative approach, enabling image enhancement at any desired scale, including fractional scales such as 1.2x, 1.4x, and 2.3x. This flexibility is particularly valuable in US imaging, where clinicians often require customized resolution enhancements for specific diagnostic tasks. For example, a fractional scaling factor might be essential to highlight delicate structures in vascular imaging or detect subtle abnormalities in soft tissues, which might not be feasible with predefined fixed scales. The ability to achieve arbitrary SR ensures that imaging systems adapt dynamically to the clinical context, improving diagnostic precision and patient outcomes. The development of ASR techniques has been closely tied to advancements in deep learning, particularly Convolutional Neural Networks (CNNs) [3]. CNNs have proven exceptionally effective in modelling spatial hierarchies and extracting complex features from medical images. These capabilities make them well-suited for arbitrary SR tasks, where the goal is to upscale images across a continuous range of scaling factors while preserving anatomical fidelity and minimizing artifacts such as blurring or aliasing. This is especially critical in US imaging, as minor distortions can significantly affect diagnostic accuracy. Despite these advancements, arbitrary scale super-resolution in US imaging presents unique challenges. US images are characterized by high variability in resolution and texture due to differences in probe settings, imaging depth, and tissue properties [4]. Designing models that efficiently learn multi-scale representations and interpolate or extrapolate them at

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment. For example, "This work was supported in part by the U.S. Department of Commerce under Grant BS123456."

The next few paragraphs should contain the authors' current affiliations, including current address and e-mail. For example, F. A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail:author@boulder.nist.gov).

S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar.colostate.edu).

T. C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba, Japan (e-mail: author@nrim.go.jp).

*These authors contributed equally to this work.

arbitrary scales without introducing artifacts is a complex task. Moreover, the computational demands of ASR must be carefully managed to ensure the technique is viable for real-time applications, which are essential for many ultrasound-based procedures. To address these challenges, this work focuses on modifying the ElitNet network [**?**], to enhance its performance for US imaging. The modifications aim to improve the network's ability to accurately handle arbitrary scaling factors by feeding images of integer and fractional scales, reducing artifacts and preserving fine anatomical details via a unique two-step training approach. We use different combinations of loss functions in both the steps. The configuration of the first step helps with retaining the crucial US data while increasing resolution arbitrarily, while the configuration of second step helps get rid of the additional artifacts introduced as part of the super-resolution step. By leveraging CNN-based architectures, the proposed approach balances computational efficiency with the need for high-quality super-resolution, making it suitable for the unique demands of US imaging. This paper explores the methodologies underpinning ASR, focusing on CNN-based approaches tailored to US imaging. The discussion encompasses theoretical frameworks and practical implementations, with a detailed evaluation of their effectiveness in handling fractional scaling factors [5]. The study also highlights the broader potential of ASR in medical imaging, underscoring its transformative impact on clinical practice. Finally, ASR's challenges and future directions are discussed, focusing on enhancing adaptability and performance in diverse medical imaging applications.

II. Literature Survey

A. Single image super resolution

Recent advancements in single-image super-resolution (SISR) have significantly leveraged convolutional neural networks (CNNs) to enhance image resolution. The pioneering SRCNN model marked a breakthrough by enabling end-to-end learning of resolution mapping, outperforming traditional methods such as sparse-coding-based super-resolution techniques [6]. Deeper CNN architectures, including VGG-inspired models, followed this, further incorporating residual learning to improve accuracy [7]. The Enhanced Deep Super-Resolution Network (EDSR) was another milestone, optimizing residual networks and eliminating unnecessary modules to achieve high-quality performance, particularly on benchmark datasets and in competitions like NTIRE 2017 [8]. Other notable methods like LapSRN [9] and DBPN [10] utilized progressive and iterative learning techniques, further improving resolution. Additionally, innovations such as sub-pixel convolution networks [11] and RCAN [12] have optimized computational efficiency in the field, enabling better performance with fewer resources. These developments in SISR have inspired parallel progress in medical imaging. For example, deep learning models in CT imaging have effectively exploited repetitive structures in medical images to reconstruct higher-resolution details, outperforming earlier approaches such as SRCNN [13]. In US imaging, where physical and hardware constraints limit resolution, CNN-based frameworks have emerged as powerful alternatives. Unsupervised super-resolution (USSR) methods have been proposed to enhance US resolution without requiring external paired datasets [14]. Similarly, hybrid strategies integrating vision-based interpolation with learning-based models have improved spatial resolution in static and dynamic US data, enabling real-time predictions [15]. These advancements underscore the potential of CNNs to overcome resolution limitations in US and other medical imaging modalities, aligning with the broader success of single-image SR methods in natural image domains. While conventional SISR approaches are typically constrained to fixed upscaling factors, recent research has emphasized arbitrary-scale super-resolution (SR) methods, which allow flexible scaling to any desired resolution. This versatility enhances precision and adaptability in clinical and research applications, particularly where diagnostic accuracy relies on capturing fine anatomical details.

B. Arbitrary image super resolution

Super-resolution (SR) techniques have shown remarkable progress in recent years, enhancing image quality across diverse application domains, including medical imaging. Within SR, single-image super-resolution (SISR) refers to reconstructing a high-resolution (HR) output from a single low-resolution (LR) input. A more recent extension of SISR is arbitrary-scale super-resolution (ArbSISR), which allows scaling by any real factor (integer or fractional) rather than being restricted to fixed scales such as 2x or 4x. This flexibility is highly valuable for medical imaging, where clinicians may require custom scaling to analyze subtle anatomical features. Despite advances in SISR for natural images, the application of ArbSISR to US remains limited due to challenges such as irregular anatomical structures, intense speckle noise, and inherently low spatial resolution. These unique characteristics demand methods capable of handling non-linear and complex transformations while preserving diagnostically relevant features. Our focus is on fractional ArbSISR, a significant but underexplored direction for US imaging.

In medical imaging, ArbSISR has begun to attract attention. For example, an implicit neural voxel function was introduced for MRI SR to enable arbitrary rescaling of volumetric data [16]. While this approach extended flexibility beyond fixed scales, it was still constrained to integer scaling and thus limited for fine adjustments required in clinical applications. Similarly, CNN-based multiscale training methods [17] offered adaptability to varying MRI resolutions, but their designs were not optimized for fractional scaling. These works highlight initial steps toward medical ArbSISR but underscore the need for greater precision and robustness in clinical modalities such as US. Insights from natural image ArbSISR provide a strong foundation for medical adaptation. Meta-SR proposed a Meta-Upscale module that dynamically generates scale-specific filters, enabling flexible scaling without retraining [18]. SRWarp introduced adaptive warping layers to accommodate spatially varying transformations, improving robustness to real-world distortions [19]. Lightweight approaches such as scale-aware feature adaptation [20] and efficient networks like OverNet

[21] and ASDN [22] refine feature extraction while reducing computational costs, making them suitable for deployment on resource-limited devices. However, most of these methods implicitly assume structured natural image content, which reduces their direct applicability to US data characterized by irregular patterns and noise. Several ArbSISR frameworks integrate image restoration techniques to improve robustness against degradations. For instance, combining blind deblurring with SR has been shown to suppress artifacts in natural images, suggesting potential utility for reducing US-specific speckle noise [23]. Similarly, CiaoSR's scale-aware non-local attention mechanism enhances feature representation across arbitrary scales [24]. In contrast, implicit neural representation models such as OPE-Upscale employ position encoding to achieve efficient scaling, providing another promising direction for US imaging [25]. Hybrid approaches like MambaSR integrate state-space modelling with Fourier Convolution Blocks, capturing spatial and frequency-domain information, which could benefit complex US datasets [26]. Fractional scaling is particularly critical in clinical US applications. Like a physician evaluating a suspicious breast tumor may need to enhance a scan by a non-integer factor to delineate lesion boundaries more clearly for biopsy planning. Existing ArbSISR models provide partial solutions. The Adaptive Implicit Deconvolution Network (AIDN) uses a Conditional Resampling Module (CRM) to achieve fractional scale adaptation [27], while A-LIIF extends local implicit functions to arbitrary-scale tasks, offering a framework that could be adapted for ultrasound-specific fractional scaling [28].

III. Methodology

A. Dataset

In a typical US process, high-frequency sound waves are transmitted into the body, and the reflected echoes are received, converted into electrical signals, and processed to form images on the machine's display. RF data carries valuable information about the intensity and frequency of these echoes and serves as the foundation for generating various imaging modes, including B-mode, Doppler, and elastography. In this study,images were acquired using a research-grade US system[1] equipped with 128 channels. For each transmitted pulse at a specific frequency f, 32 channels operated as transmitters and 64 channels as receivers in a sequential manner, repeated four times to capture the complete echo profile. The probe used for image acquisition was the $L11-5V$ linear array transducer, and scanning was performed on the CAE Blue Phantom[2] breast US model, a high-quality medical training phantom designed to replicate the acoustic and physical properties of real human breast tissue. A total of 400 US images are captured from a CAE breast phantom at a centre frequency of 5 MHz. The size of all the images (Figure 2) is 128×156 ($w \times h$). The overall data acquisition and processing pipeline is illustrated in Figure 1.

[1]https://verasonics.com/vantage-32-le/

[2]https://www.optisafe.se/blue-phantom-elastography-ultrasound-breast-phantom

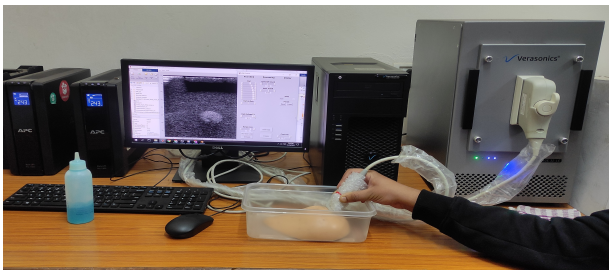
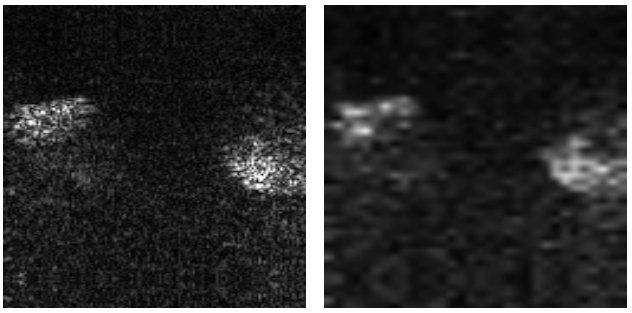


Fig. 1: B-mode image acquisition setup



(a) High resolution US image (b) Low-resolution US image

Fig. 2: The figure shows the high and low resolution of the images that we have used in the experiment

B. Network Architecture

In our proposed framework, we adopt ELiTNet [29] as the core network architecture and extend its functionality to handle arbitrary and non-symmetric scale factors along spatial dimensions. Specifically, the scale factor is allowed to differ between the horizontal and vertical axes, accommodating cases where the upscaling requirements are not uniform across height and width. We call this version of ElitNet capable of handling arbitrary scales super-resolution; SupeRELiTNet.

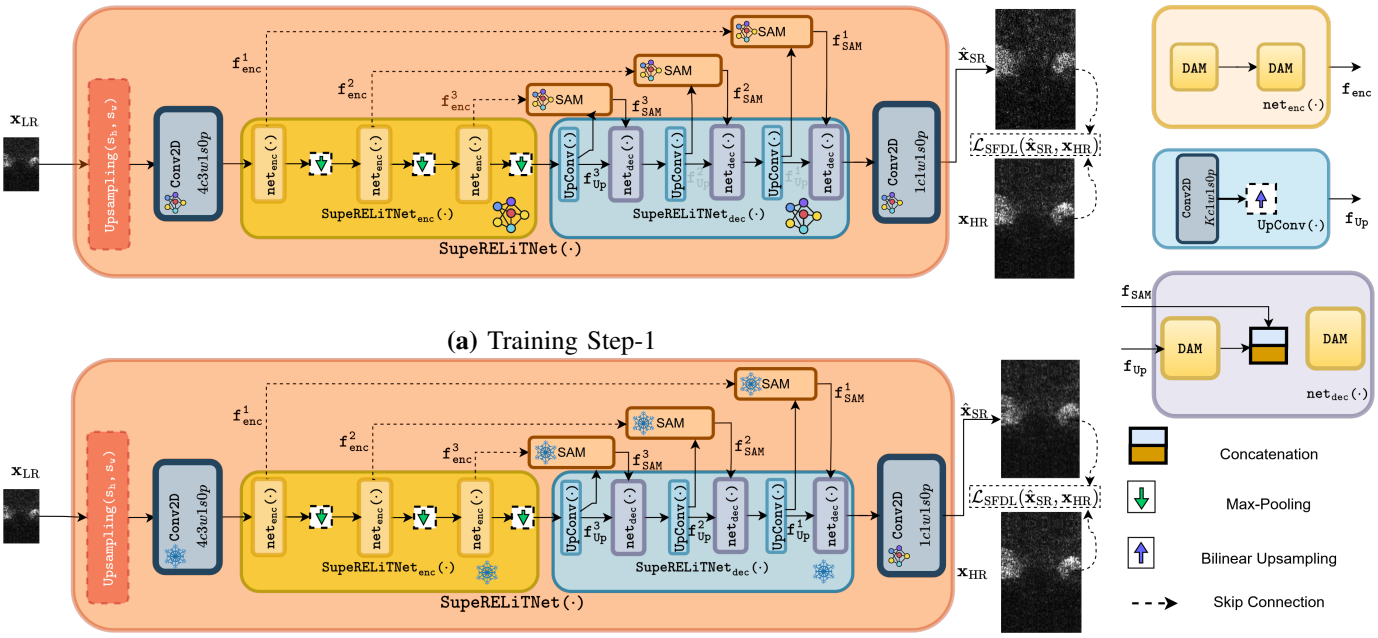
Given an image of high resolution (HR) $\mathbf{x}_{\text{HR}} \in \mathbb{R}^{H \times W \times C}$, where H and W represent spatial dimensions and C is the number of channels, a corresponding low resolution (LR) image $\mathbf{x}_{\text{LR}} \in \mathbb{R}^{\frac{H}{s_h} \times \frac{W}{s_w} \times C}$ has been produced (described in IV-A.1). Here s_h and s_w represent the scaling factor for height and width, respectively. The goal here is to produce a superresolved image $\hat{\mathbf{x}}_{\text{SR}} \in \mathbb{R}^{H \times W \times C}$, obtained by;

$$\hat{\mathbf{x}}_{\text{SR}} = \texttt{SupeRELiTNet}(\mathbf{x}_{\text{LR}}) \tag{1}$$

$$\texttt{SupeRELiTNet}(\cdot) \mapsto \texttt{Upsampling}(s_h, s_w) \rightarrow \texttt{ELiTNet}(\cdot) \tag{2}$$

which approximate the HR ground truth $\mathbf{x}_{\text{HR}}$.

By retaining ELiTNet's robust multiscale feature extraction and spatial coherence while introducing the ability to process and reconstruct images at arbitrary, nonsymmetric scale factors, the framework remains both flexible and efficient. No architectural modifications are made except for integrating a Upsampling layer that can natively support distinct scaling along the two spatial dimensions, thereby demonstrating the generalizability of SupeRELiTNet to diverse super-resolution applications.



(a) Training Step-1

(b) Training Step-2

Fig. 3: Overall training pipeline of SupeRELiTNet consisting two-step training process. denotes the learnable layer/module and denotes the frozen layers/moduls in the network. In (a) training step-1, all the layers are trainable, whereas in (b) training step-2, all the layers except the final `Conv2d` layer are frozen.

C. Multi-Objective Loss Design

In medical image super-resolution, preserving high-frequency content is critical, as it directly impacts diagnostic accuracy. To emphasize this, a two-step training paradigm has been employed and is illustrated in Figure 3. We proposed a custom loss function, the Structured Frequency Distribution Loss (SFDL), for the first-step training, which can be defined as:

$$\mathcal{L}_{\text{SFDL}}(\hat{\mathbf{x}}_{\text{SR}}, \mathbf{x}_{\text{HR}}) = \alpha \cdot \mathcal{L}_{\text{SSIM}}(\hat{\mathbf{x}}_{\text{SR}}, \mathbf{x}_{\text{HR}}) + (1-\alpha) \cdot \mathcal{L}_{\text{FDL}}(\hat{\mathbf{x}}_{\text{SR}}, \mathbf{x}_{\text{HR}}) \quad (3)$$

Where α is a weighting parameter that balances the Structural Similarity Index (SSIM) [30] loss and the Frequency Distribution Loss (FDL) [31]. The optimal value of α was determined via grid search over the interval $[0, 1]$ with a step size of 0.1, monitored using Weights and Biases[3] to ensure consistent performance across training runs. The SSIM [30] loss function is given by:

$$\mathcal{L}_{\text{SSIM}}(\hat{\mathbf{x}}_{\text{SR}}, \mathbf{x}_{\text{HR}}) = 1 - \text{SSIM}(\hat{\mathbf{x}}_{\text{SR}}, \mathbf{x}_{\text{HR}}) \quad (4)$$

Where the SSIM is defined as:

$$\text{SSIM}(\hat{\mathbf{x}}_{\text{SR}}, \mathbf{x}_{\text{HR}}) = \frac{(2\mu_{\mathbf{x}_{\text{HR}}}\mu_{\hat{\mathbf{x}}_{\text{SR}}} + C_1)(2\sigma_{\mathbf{x}_{\text{HR}}\hat{\mathbf{x}}_{\text{SR}}} + C_2)}{(\mu^2_{\mathbf{x}_{\text{HR}}} + \mu^2_{\hat{\mathbf{x}}_{\text{SR}}} + C_1)(\mu^2_{\mathbf{x}_{\text{HR}}} + \sigma^2_{\hat{\mathbf{x}}_{\text{SR}}} + C_2)} \quad (5)$$

where $\mu_{\mathbf{x}_{\text{HR}}}$ is the mean over a window in the image $\hat{\mathbf{x}}_{\text{SR}}$ and $\sigma_{\hat{\mathbf{x}}_{\text{SR}}\mathbf{x}_{\text{HR}}}$ denotes the covariance between the two images. $\mathcal{L}_{\text{FDL}}(\hat{\mathbf{x}}_{\text{SR}}, \mathbf{x}_{\text{HR}})$ in (3) refers to the Frequency Distribution Loss, adopted from [31]. FDL defines a perceptual similarity metric that compares two images based on their spatial structures and frequency content. Given two input images $\mathbf{x}_{\text{HR}}$ and $\hat{\mathbf{x}}_{\text{SR}}$, we first extract their deep feature representations using a shared backbone network $\mathcal{F}$. In our implementation, the network chosen was EfficientNet:

$$\mathbf{f}_{\mathbf{x}_{\text{HR}}} = \mathcal{F}(\mathbf{x}_{\text{HR}}), \quad \mathbf{f}_{\hat{\mathbf{x}}_{\text{SR}}} = \mathcal{F}(\hat{\mathbf{x}}_{\text{SR}}) \quad (6)$$

where $\mathbf{f}_{\mathbf{x}_{\text{HR}}}$ and $\mathbf{f}_{\hat{\mathbf{x}}_{\text{SR}}}$ denote the feature maps of $\mathbf{x}_{\text{HR}}$ and $\hat{\mathbf{x}}_{\text{SR}}$, respectively.

To analyze the structural and frequency-based differences, we transform these features into the frequency domain via the multidimensional Fast Fourier Transform (FFT):

$$\hat{\mathbf{f}}_{\mathbf{x}_{\text{HR}}} = \mathcal{F}_{\text{FFT}}(\mathbf{f}_{\mathbf{x}_{\text{HR}}}), \quad \hat{\mathbf{f}}_{\hat{\mathbf{x}}_{\text{SR}}} = \mathcal{F}_{\text{FFT}}(\mathbf{f}_{\hat{\mathbf{x}}_{\text{SR}}}) \quad (7)$$

where $\hat{\mathbf{f}}_{\mathbf{x}_{\text{HR}}}$ and $\hat{\mathbf{f}}_{\hat{\mathbf{x}}_{\text{SR}}}$ are the complex-valued frequency-domain representations.

We then decompose them into magnitude and phase components:

$$M_{\mathbf{x}_{\text{HR}}} = |\hat{\mathbf{f}}_{\mathbf{x}_{\text{HR}}}|, \ \Phi_{\mathbf{x}_{\text{HR}}} = \angle\hat{\mathbf{f}}_{\mathbf{x}_{\text{HR}}}, \ M_{\hat{\mathbf{x}}_{\text{SR}}} = |\hat{\mathbf{f}}_{\hat{\mathbf{x}}_{\text{SR}}}|, \ \Phi_{\hat{\mathbf{x}}_{\text{SR}}} = \angle\hat{\mathbf{f}}_{\hat{\mathbf{x}}_{\text{SR}}} \quad (8)$$

where $M_{\mathbf{x}_{\text{HR}}}$ and $M_{\hat{\mathbf{x}}_{\text{SR}}}$ represent the magnitudes of the frequency components, and $\Phi_{\mathbf{x}_{\text{HR}}}$ and $\Phi_{\hat{\mathbf{x}}_{\text{SR}}}$ denote the corresponding phase angles.

The mean absolute differences between the distributions of magnitude and phase components are then computed to quantify dissimilarity in both magnitude and phase:

$$s_{\text{mag}} = \frac{1}{N}\sum_{n=1}^{N} |M_{\mathbf{x}_{\text{HR}}}[n] - M_{\hat{\mathbf{x}}_{\text{SR}}}[n]| \quad (9)$$

$$s_{\text{phase}} = \frac{1}{N}\sum_{n=1}^{N} |\Phi_{\mathbf{x}_{\text{HR}}}[n] - \Phi_{\hat{\mathbf{x}}_{\text{SR}}}[n]| \quad (10)$$

[3] https://wandb.ai/

where N is the total number of frequency components, and s_{mag}, s_{phase} denote the magnitude and phase dissimilarity scores, respectively.

The final similarity score for each layer i is computed as a weighted sum of the two components:

$$s = s_{\text{mag}} + \lambda \cdot s_{\text{phase}} \tag{11}$$

where λ is a hyperparameter controlling the contribution of phase information, tuned experimentally.

The overall perceptual similarity between two inputs is obtained by averaging over all layers:

$$\text{Score}(\mathbf{x}_{\text{HR}}, \hat{\mathbf{x}}_{\text{SR}}) = \frac{1}{L} \sum_{i=1}^{L} s^{(i)} \tag{12}$$

where L denotes the total number of feature layers used in the comparison.

This combination of SSIM [30] and FDL [31] losses formed the first step of our training process. Although this combination yielded satisfactory results in preserving high-frequency content and structural similarity, it also introduced unexpected white artifacts in the output images. This is likely because the frequency domain constraints do not directly regulate pixel intensity values.

As described extensively in the Experiments Section, to address this challenge, we incorporated a second training step. In the second step, we replaced FDL with L1 loss, maintaining the same overall form of the training loss function:

$$\mathcal{L}_{\text{train}}(\hat{\mathbf{x}}_{\text{SR}}, \mathbf{x}_{\text{HR}}) = \alpha \cdot \mathcal{L}_{\text{SSIM}}(\hat{\mathbf{x}}_{\text{SR}}, \mathbf{x}_{\text{HR}}) + (1-\alpha) \cdot \mathcal{L}_{\text{L1}}(\hat{\mathbf{x}}_{\text{SR}}, \mathbf{x}_{\text{HR}}) \tag{13}$$

where the L1 loss is defined as:

$$\mathcal{L}_{\text{L1}}(\mathbf{x}_{\text{HR}}, \hat{\mathbf{x}}_{\text{SR}}) = \frac{1}{H \times W \times C} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} |\mathbf{x}_{\text{HR}_{h,w,c}} - \hat{\mathbf{x}}_{\text{SR}_{h,w,c}}| \tag{14}$$

with H, W, and C representing the image's height, width, and number of channels, respectively. Similarly to the first step, the value of α in (13) was determined by hyper-parameter tuning. For this stage, the optimal value of α was 0.5. Since the $L1$ loss severely penalizes pixel differences, it effectively reduces white artifacts while preserving the high-frequency details learned in the first training step.

IV. Experiments

A. Experimental Setting

1) Dataset Preparation: Our experiments utilized a custom US dataset comprising 400 positional images acquired from a CAE breast tumor phantom, including representative tumor samples, cysts, and non-tumor regions. To prepare the acquired US data for practical model training, a structured preprocessing pipeline was developed to enhance data quality, increase variability, and ensure consistency across all samples. To align with the architectural requirements of the model, each image was padded reflectively along the height to reach a uniform size of 128×192. Reflective padding was chosen to preserve edge features and minimize boundary artifacts, which can otherwise distort learning at image borders. We generated low-resolution (LR) training images using symmetric and asymmetric scale factors to develop a model capable of arbitrary scale super-resolution. Symmetric scaling factors ranged from 1.0 to 4.0, with a stride of 0.1, ensuring a dense range of uniform downscaling. Additionally, asymmetric scale factors were introduced by independently varying the horizontal and vertical axes with strides of 0.5, enabling the model to learn non-uniform scaling behaviors. This process allowed the model to learn robust mappings across various resolutions. Finally, a custom dataset class was implemented to handle the multiscale nature of the data. This ensured that each training batch contained images of different scale factors and their corresponding scaling information, enabling the model to handle variable input conditions effectively and making the training process more flexible and interpretable. The entire data set was then divided into training (80%), validation (10%), and testing (10%) sets to ensure balanced evaluation at different stages of model development.

2) Implementation details: The experiments were conducted on a system equipped with an Intel(R) Core i5-8600k CPU, 2x16 GB DDR4 RAM, and an NVIDIA Quadro P6000 GPU with 24GB GDDR5 memory. Details of the hyperparameters are given in Table I.

TABLE I: Hyperparameter Settings

Training Parameters	Setting Value
Batch size	16
Epoch	500
Learning rate	0.001
Optimizer	Adam
Weight decay	0.01
Scheduler	ReduceLRonPlatue
LR Reduce factor	0.8
Scheduler wait time	20

3) Evaluation Metrics: To evaluate the performance of SupeRELiTNet in reconstructing high-quality US images, we employed two standard image quality assessment metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [30]. PSNR is a widely used metric that quantifies the difference between predicted and ground truth images. It estimates how much noise or distortion is present in the reconstructed image. Higher PSNR values typically indicate better image quality and closer resemblance to the original image. SSIM [30], on the other hand, measures the perceptual similarity between two images by comparing their structural information, luminance, and contrast. Unlike PSNR, SSIM [30] is more aligned with human visual perception and better indicates how similar the reconstructed image appears to the ground truth.

V. Results

We evaluated the performance of SuperElitNET by comparing it against several existing super-resolution architectures. The combination of our novel loss functions and multi-step training strategy consistently led to superior results, as reflected in both quantitative metrics and qualitative visual comparisons. Given the limited number of models that support arbitrary-scale super-resolution, we focused our benchmark primarily on the 4× super-resolution task, a standard evaluation scale where improvements are most noticeable and widely reported. The models included in our comparison are: Enhanced

Super-Resolution Training via Mimicked Alignment for Real-World Scenes [32], ArbRCAN [33], RDUNet [34], SRDNet [35], and ABPN [36]. Among these, ArbRCAN is one of the few models capable of handling arbitrary scale factors, making it a particularly relevant baseline. While the original ArbRCAN implementation used either L1 or VGG loss during training, we re-trained it using a combination of both losses to ensure a fair comparison, aligning with our objective of preserving both perceptual quality and pixel-level accuracy.

A. Qualitative Results

To ensure a comparison across all super-resolution models, we prepared the training data according to each architecture's expected input format and preprocessing requirements described in their original implementations. This often involved modifying or rebuilding dataset pipelines to match the resolution, normalization, and augmentation strategies used in state-of-the-art configurations. Figure 4 presents the performance comparison for 4× super-resolution, where the improvements were most substantial and clearly highlight the strengths of our approach.

The figure has the qualitative results from each model architecture. One can clearly see the stark difference between the high frequency data preserved by our architecture and that of the others.

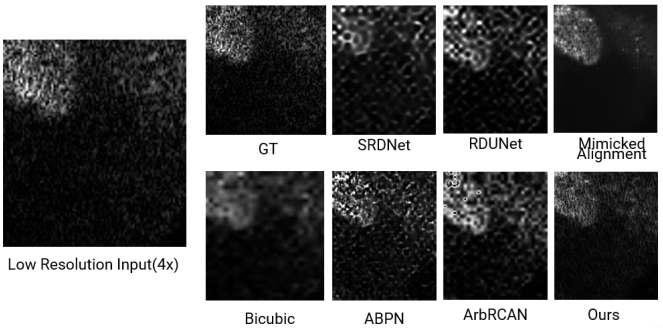


Fig. 4: Comparison of Our Proposed architecture's performance with others

TABLE II: PSNR and SSIM Comparisons

Model	PSNR	SSIM	Time (ms)	Params (M)
ArbSR	19.999	0.5867	0.255	16.93
Mim.All.	21.256	0.3682	0.157	0.57
SRDNet	17.495	0.2497	0.037	1.50
RDUNet	19.446	0.3797	1.795	166.37
ABPN	18.827	0.1402	0.100	2.51
Ours	**22.971**	**0.6332**	**0.008**	**0.05**

B. Quantitative Results

To evaluate the performance of our models, we primarily used two widely accepted image quality metrics: Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR). All evaluations were conducted at the 4× super-resolution scale, where the improvements achieved by our proposed architecture were most significant. As shown in Table II, our method consistently outperforms other state-of-the-art models across both metrics.

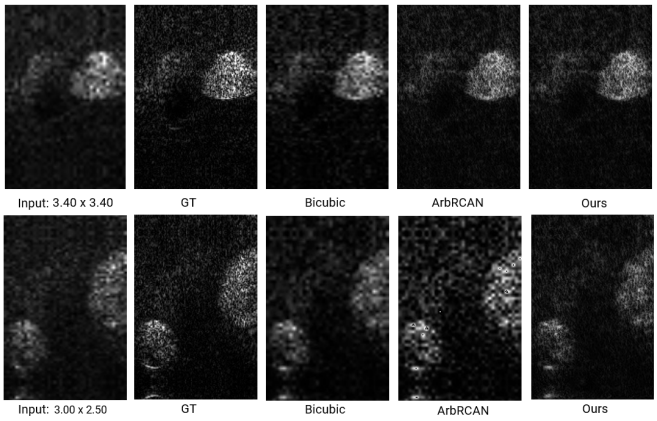


Fig. 5: Comparison with ArbRCAN highlighting the arbitrary scale component.

C. Arbitrary and non-symmetric Scale factors

To demonstrate the flexibility of our model, we evaluated its performance on arbitrary and non-symmetric super-resolution scale factors, an aspect that has received limited attention in prior work. Given the scarcity of architectures explicitly designed for arbitrary scaling, we used **ArbRCAN** [33] as the primary baseline for comparison.

Qualitative examples are provided in Figure 5, while quantitative metrics for select scale factors are shown in Table III. The results indicate that our proposed architecture consistently matches or exceeds the performance of **ArbRCAN**, highlighting its robustness and adaptability across diverse scaling scenarios.

TABLE III: Quantitative Evaluation with ArbRCAN

Scale	Ours		ArbRCAN	
	PSNR	SSIM	PSNR	SSIM
2.50 × 1.50	23.8723	0.6714	23.2296	0.7501
3.00 × 2.50	23.1230	0.6146	21.6624	0.6610
3.40 × 3.40	23.0014	0.6059	20.3507	0.6009
4.00 × 4.00	22.9710	0.6332	19.9999	0.5867
Overall	**23.2419**	**0.6408**	**21.3107**	**0.6497**

VI. Ablation Study

The primary reason behind training the model in two steps was the preservation of high frequency data and getting rid of any artifacts that would emerge as a result of the first step. Both the steps involved various loss configurations which are detailed in the following sections.

A. First Step

As mentioned in the sections above, the particular loss function combination chosen by us was arrived upon after rigorous experimentation, trial, and errors with several loss functions. This holds true for both the first and the second steps of training. Table IV highlights the loss functions that we tested in the first step.

This step was crucial in obtaining a solid frequency-retention loss function, for each step, the loss configurations were used to train the model with the hyper parameters α selected from the range [0.1,0.9] using hyperparameter tuning .

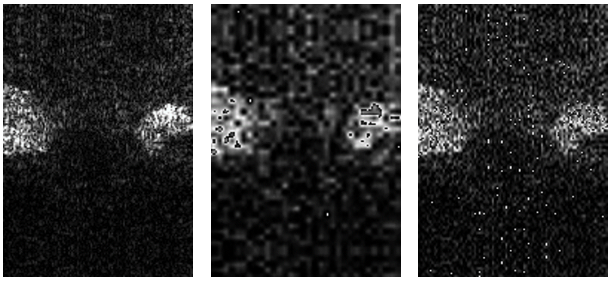
All experiments were hosted on W&B with the metrics and the training and validation loss recorded for effective monitoring.

TABLE IV: Loss Function Configs for the first step of training.

Loss Function	α	Lr.	SSIM	PSNR
SSIM + FFL	0.51	0.001	0.6771	21.457
SSIM + Laplace	0.51	0.001	0.6814	21.463
SSIM + FDL(EffNet)	0.5	0.001	0.6857	21.359
SSIM + FDL(EffNet)	0.5	1e-5	0.5768	20.424
SSIM + FDL(VGG)	0.49	0.0004	0.6343	23.500

We have already discussed the SSIM [30] loss function in the methodology section III, here we will briefly dive into the variants used along with it for the first Step. FFL or Focal Frequency Loss [37] is also meant to focus on preserving high frequency components of images and involves fast fourier transforms similar to FDL [31], however one key difference here was the FFL did not involve the extraction of features from the images using a feature extractor(Deep CNNs). Again the performance was good perceptually for smaller scale factors but were not so good for 4x SR.

The second variant tested here along with SSIM [30] was the Laplacian Filter. We used the implementation in the Kornia Computer Vision Library[4]. The Laplacian filter is primarily used for edge detection and image sharpening by highlighting regions of rapid intensity change, something we saw useful in our task. So the Laplacian loss essentially takes the Super-Resolved output and the ground truth image as inputs and applies the Laplacian filter to both of them. The mean of the absolute difference between the extracted features is returned as the laplacian loss.

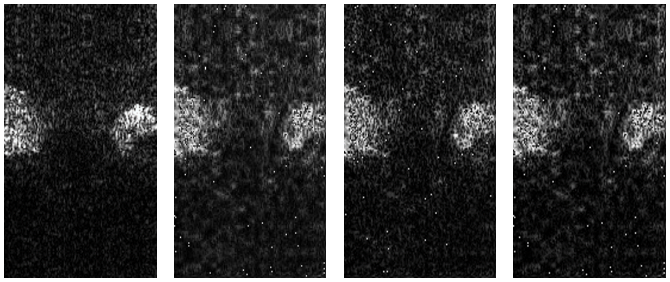


(a) GT (b) 4.0 x 4.0 (c) 3.5 x 1.5

Fig. 6: Super-resolved outputs using Laplacian loss: (a) Ground Truth (GT), (b) result at scale 4.0×4.0, and (c) result at scale 3.5×1.5. Laplacian loss helps retain edge details across different scales.

The model's performance with this loss configuration was interesting to observe. Similar to SSIM [30] + FDL [31], it seemed to preserve the high frequency details for small scales with white artifacts appearing, but the same was not true for higher scaling factors. This was something shared across the other variants as well, even for the first combination - SSIM [30] + FFL [37], SR was satisfactory for small scaling factors, barring appearance of white artifacts in the final output, but for higher scaling factors, the model seemed to hallucinate a lot, as is evident in Figure 6.

The SSIM [30] + FDL [31] loss turned out to be a very promising configuration because it seemed to preserve the overall frequency content of the images both for the small and the large scaling factors. Learning rate and the weight assigned to each loss function in the configuration also played an important role. We have already detailed FDL loss in the methodology section III, however the choice of the feature extractor was also something that we experimented for FDL loss.



(a) GT (b) EffiNet (c) EffiNet+lr(1e-5) (d) VGG

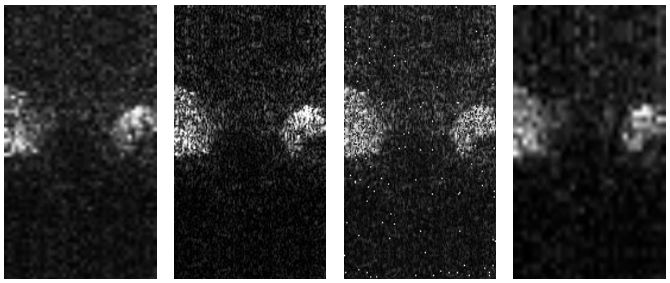
Fig. 7: Visual comparison of super-resolution outputs using (a) Ground Truth (GT), (b) EffiNet, (c) EffiNet with learning rate $1e{-}5$, and (d) VGG, evaluated under SSIM + FDL settings.

Initially, the loss function used the VGG's backbone to extract the relevant features. We changed the backbone from VGG to EfficientNet, which reduced the training time without compromising the quality of the final output, as is evident in the comparisons drawn in Figure 7.

B. Second Step

From the first step, the loss configuration SSIM [30] + FDL [31] proved to be very promising, but as is evident from Figure 8, there are a lot of white artifacts, especially over the areas indicative of tumor, which is derogatory for diagnosis purposes.

So the purpose of a second step was to train the model to get rid of the white artifacts while holding on to the high frequency data that it had learned to preserve. Therefore we decided to freeze all but the last convolutional layer of the architecture and train all over again. Instinctively, to test if freezing alone would do the job, we did not change the loss function and retrained. But it ended up degrading the output image.



(a) LR Input (b) GT (c) SR (d) Bicubic

Fig. 8: Visual results of super-resolution using SSIM + FDL (EffNet). (a) Low-resolution input, (b) Ground truth (GT), (c) Reconstructed output from our method, and (d) Bicubic interpolation.

[4]https://kornia.readthedocs.io/en/stable/filters.html#kornia.filters.Laplacian

Table V lists out the various loss configurations that we experimented with for the second step. Since the lr and α were chosen from ranges of values between 0.1 and 0.9, we closely monitored the loss curve and metrics that were logged onto W&B and we trained the models with the configurations that showed consistent decrease in validation loss, and promising metrics.

We have discussed all of the loss functions earlier except one. The first configuration in the table CHAR + FDL where CHAR stands for the Charbonnier loss function [38] [39] [40].Once again, we use the implementation from the Kornia library[5]. And that sepcific implementation essentially computes the L1-L2 loss and is computed as follows:

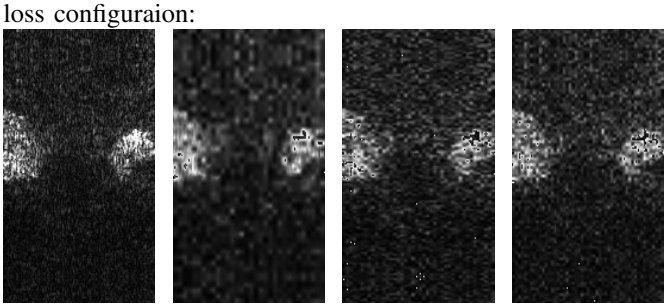
$$\mathrm{WL}(x, y) = \sqrt{(x-y)^2 + 1} - 1 \tag{15}$$

The inspiration for using this loss function came from the results with the L1 loss function, and since the Charbonnier loss incorporates both L1 and L2 loss, it seemed ideal.

TABLE V: Loss Configurations that were tried for the second step.

Loss Function	α	Lr.	SSIM	PSNR
CHAR + FDL	0.51	0.0001	0.7032	22.5530
L1 + FDL	0.51	0.001	0.6332	22.9710
L1	0.51	1e-5	0.6168	21.0691
MSE + FFL	0.6	1e-5	0.6290	21.2985
SSIM + FFL	0.51	0.001	0.6771	21.4571
SSIM + FDL(EffNet)	0.57	0.001	0.6778	18.6688
SSIM + L1	0.57	0.001	0.6771	21.4571
SSIM + Laplace	0.51	0.001	0.6393	21.2660
SSIM + Laplace	0.48	0.001	0.6112	19.8712

But despite decent metrics with Charbonnier loss, the output for higher scaling factors was not perceptually accurate, and the model hallucinated a lot of details, as depicted in Figure 9. As a result, the loss configuration of the second step was **L1 + FDL** both perceptually and metric wise, delivering a **PSNR** of **22.9710** and a **SSIM** of **0.6332**, and was used in the second step of training. Figure 10 and Figure 11 are Super Resolved Images for each loss combination for different scaling factors, one can clearly see the superior performance for L1 + FDL loss configuraion:



(a) GT (b) 4.0 x 4.0 (c) 3.5 x 1.5 (d) 2.5 x 2.0

Fig. 9: Super-resolution results using EffNet with Charbonnier + FDL loss for different scaling factors: (a) Ground Truth, (b) 4.0×4.0, (c) 3.5×1.5, (d) 2.5×2.0.

VII. Conclusion

In this study, we proposed SupeRELiTNet, a lightweight super-resolution architecture, to address the challenge of ar-

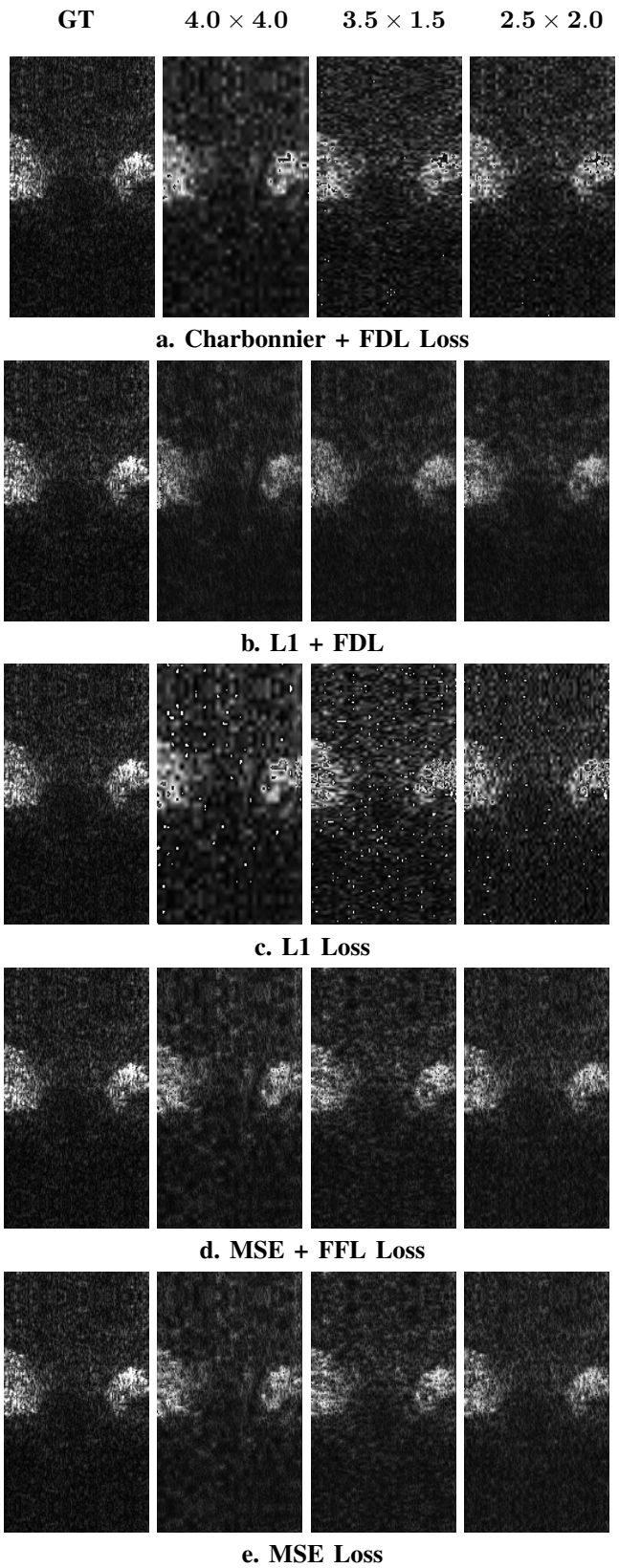


Fig. 10: Qualitative comparison of super-resolved ultrasound images generated using various loss configurations (a-e). The subfigures show: (1) Ground Truth (GT), (2) super-resolved output at scaling factor 4.0×4.0, (3) super-resolved output at 3.5×1.5, and (4) super-resolved output at 2.5×2.0

[5]https://kornia.readthedocs.io/en/latest/losses.html#kornia.losses.CharbonnierLoss

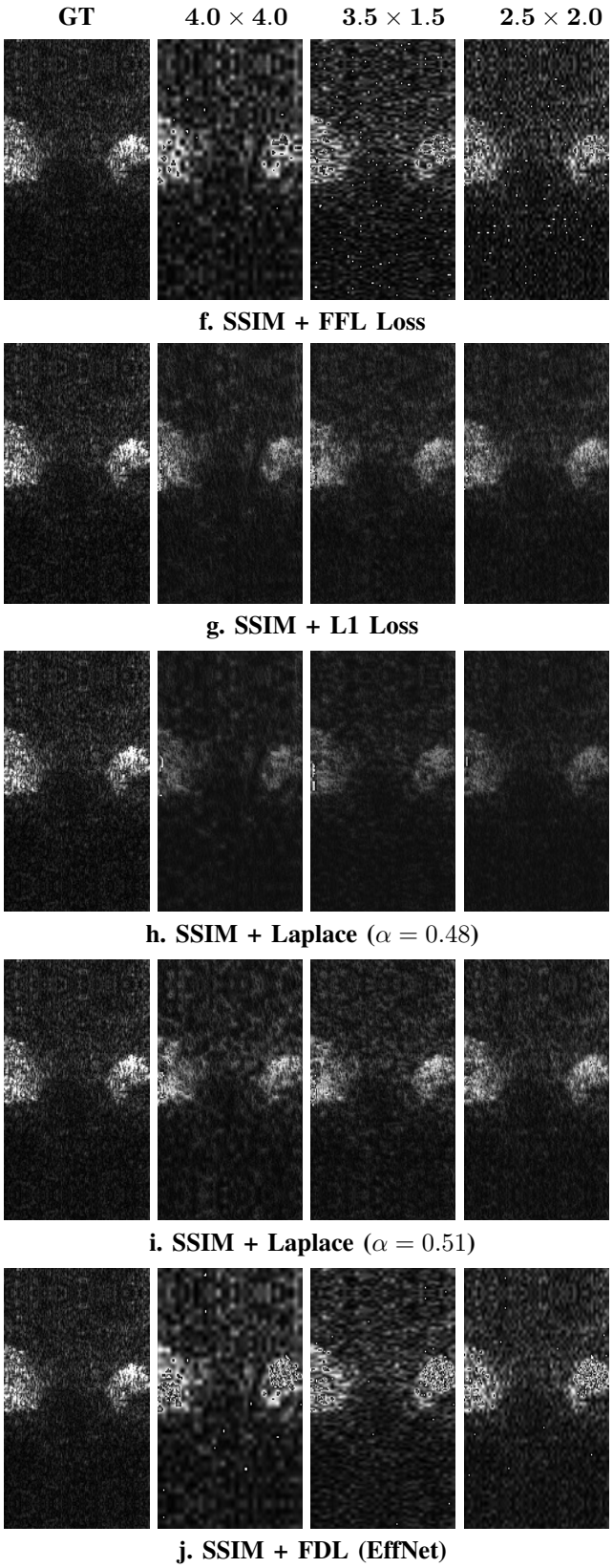


Fig. 11: Extension of Figure 10 (f-i). The subfigures show: (1) Ground Truth (GT), (2) super-resolved output at scaling factor 4.0×4.0, (3) super-resolved output at 3.5×1.5, and (4) super-resolved output at 2.5×2.0

bitrary and non-uniform scaling in US imaging, a domain with limited prior work. Our contributions include acquiring and preparing a specialized US dataset tailored for arbitrary scale super-resolution and designing a novel hybrid loss function that combines L1, SSIM, VGG perceptual loss, and frequency domain loss to improve reconstruction quality. This loss effectively balances pixel accuracy and perceptual fidelity, enhancing performance in PSNR, SSIM, and visual results across various scale factors. Quantitative and qualitative comparisons with baseline models such as ArbRCAN, SRDNet, RDUNet, and ABPN demonstrate that our approach achieves superior accuracy while maintaining computational efficiency. The lightweight nature of SupeRELiTNet enables practical deployment in low-resource settings and real-time applications. This work provides a comprehensive solution for arbitrary-scale super-resolution of US images by combining a carefully curated dataset, an efficient model architecture, and a powerful learning objective. Future work will focus on automatic scale estimation, modelling temporal consistency for US video, and clinical validation in real diagnostic workflows.

REFERENCES

[1] Karansingh Chauhan, Shail Nimish Patel, Malaram Kumhar, Jitendra Bhatia, Sudeep Tanwar, Innocent Ewean Davidson, Thokozile F. Mazibuko, and Ravi Sharma. Deep learning-based single-image super-resolution: A comprehensive review. *IEEE Access*, 11:21811–21830, 2023.

[2] Hongying Liu, Zekun Li, Fanhua Shang, Yuanyuan Liu, Liang Wan, Wei Feng, and Radu Timofte. Arbitrary-scale super-resolution via deep learning: A comprehensive survey. *Information Fusion*, 102:102015, 2024.

[3] Jianxin Wu. Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China*, 5(23):495, 2017.

[4] Johan Thijssen. Spectroscopy and image texture analysis. *Ultrasound in medicine & biology*, 26 Suppl 1:S41–4, 06 2000.

[5] Xiaohang Wang, Xuanhong Chen, Bingbing Ni, Hang Wang, Zhengyan Tong, and Yutian Liu. Deep arbitrary-scale image super-resolution via scale-equivariance pursuit. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1786–1795, June 2023.

[6] Xiancai Ji, Yao Lu, and Li Guo. Image super-resolution with deep convolutional neural network. In *2016 IEEE First International Conference on Data Science in Cyberspace (DSC)*, pages 626–630, 2016.

[7] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654, 2016.

[8] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[9] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017.

[10] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1664–1673, 2018.

[11] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[12] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018.

[13] Yunxing Gao, Hengjian Li, Jiwen Dong, and Guang Feng. A deep convolutional network for medical image super-resolution. In *2017 Chinese Automation Congress (CAC)*, pages 5310–5315, 2017.

[14] Jingfeng Lu and Wanyu Liu. Unsupervised super-resolution framework for medical ultrasound images using dilated convolutional neural networks. In *2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC)*, pages 739–744, 2018.

[15] Simone Cammarasana, Paolo Nicolardi, and Giuseppe Patane. Super-resolution of 2d ultrasound images and videos. *Medical & biological engineering & computing*, 61, 05 2023.

[16] Qing Wu, Yuwei Li, Yawen Sun, Yan Zhou, Hongjiang Wei, Jingyi Yu, and Yuyao Zhang. An arbitrary scale super-resolution approach for 3d mr images via implicit neural representation. *IEEE Journal of Biomedical and Health Informatics*, 27(2):1004–1015, 2023.

[17] Chi-Hieu Pham, Carlos Tor-Díez, Hélène Meunier, Nathalie Bednarek, Ronan Fablet, Nicolas Passat, and François Rousseau. Multiscale brain mri super-resolution using deep 3d convolutional networks. *Computerized Medical Imaging and Graphics*, 77:101647, 2019.

[18] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-sr: A magnification-arbitrary network for super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1575–1584, 2019.

[19] Sanghyun Son and Kyoung Mu Lee. Srwarp: Generalized image super-resolution under arbitrary transformation, 04 2021.

[20] Longguang Wang, Yingqian Wang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning a single network for scale-arbitrary super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4801–4810, 2021.

[21] Parichehr Behjati, Pau Rodriguez, Armin Mehri, Isabelle Hupont, Carles Fernandez Tena, and Jordi Gonzalez. Overnet: Lightweight multi-scale super-resolution with overscaling network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2694–2703, 2021.

[22] Jialiang Shen, Yucheng Wang, and Jian Zhang. Asdn: A deep convolutional network for arbitrary scale image super-resolution. *Mobile Networks and Applications*, 26(1):13–26, 2021.

[23] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Deep Plug-And-Play Super-Resolution for Arbitrary Blur Kernels . In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1671–1681, Los Alamitos, CA, USA, Jun 2019. IEEE Computer Society.

[24] Jiezhang Cao, Qin Wang, Yongqin Xian, Yawei Li, Bingbing Ni, Zhiming Pi, Kai Zhang, Yulun Zhang, Radu Timofte, and Luc Gool. Ciaosr: Continuous implicit attention-in-attention network for arbitrary-scale image super-resolution, 12 2022.

[25] Gaochao Song, Qian Sun, Luo Zhang, Ran Su, Jianfeng Shi, and Ying He. Ope-sr: Orthogonal position encoding for designing a parameter-free upsampling module in arbitrary-scale image super-resolution. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10009–10020, 2023.

[26] Jin Yan, Zongren Chen, Zhiyuan Pei, Xiaoping Lu, and Hua Zheng. Mambasr: Arbitrary-scale super-resolution integrating mamba with fast fourier convolution blocks. *Mathematics*, 12(15), 2024.

[27] Jinbo Xing, Wenbo Hu, Menghan Xia, and Tien-Tsin Wong. Scale-arbitrary invertible image downscaling. *IEEE Transactions on Image Processing*, 32:4259–4274, 2023.

[28] Hongwei Li, Tao Dai, Yiming Li, Xueyi Zou, and Shu-Tao Xia. Adaptive local implicit image function for arbitrary-scale super-resolution. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 4033–4037. IEEE, 2022.

[29] Dipayan Dewan, Anupam Borthakur, and Debdoot Sheet. Attention in a little network is all you need to go green. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–4, 2023.

[30] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[31] Zhangkai Ni, Juncheng Wu, Zian Wang, Wenhan Yang, Hanli Wang, and Lin Ma. Misalignment-robust frequency distribution loss for image transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2910–2919, 2024.

[32] Omar Elezabi, Zongwei Wu, and Radu Timofte. Enhanced super-resolution training via mimicked alignment for real-world scenes. In *Proceedings of the Asian Conference on Computer Vision*, pages 4122–4140, 2024.

[33] Longguang Wang, Yingqian Wang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning a single network for scale-arbitrary super-resolution. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4781–4790, 2021.

[34] Javier Gurrola-Ramos, Oscar Dalmau, and Teresa E Alarcón. A residual dense u-net neural network for image denoising. *IEEE Access*, 9:31742–31754, 2021.

[35] Tingting Liu, Yuan Liu, Chuncheng Zhang, Liyin Yuan, Xiubao Sui, and Qian Chen. Hyperspectral image super-resolution via dual-domain network based on hybrid convolution. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–18, 2024.

[36] Zhi-Song Liu, Li-Wen Wang, Chu-Tak Li, Wan-Chi Siu, and Yui-Lam Chan. Image super-resolution via attention based back projection networks. In *2019 IEEE/CVF international conference on computer vision workshop (ICCVW)*, pages 3517–3525. IEEE, 2019.

[37] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13919–13929, 2021.

[38] Jonathan Barron. A general and adaptive robust loss function. pages 4326–4334, 06 2019.

[39] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st International Conference on Image Processing*, volume 2, pages 168–172 vol.2, 1994.

[40] Zhengyou Zhang. Parameter estimation techniques: a tutorial with application to conic fitting. *Image and Vision Computing*, 15(1):59–76, 1997.