(Received 12 September 2025, Accepted 15 November 2025, Scheduled 26 November 2025)

Multi-Scale Visibility Fusion Network for Super-Resolution Near-Field Imaging in Synthetic Aperture Interferometric Radiometer

Fuxin Cai[1], Jianfei Chen[1,2,*], Ziang Zheng[1], and Leilei Liu[1]

[1] College of Electronic and Optical Engineering and the College of Flexible Electronics (Future Technology) Nanjing University of Posts and Telecommunications, Nanjing 210023, China
[2] State Key Laboratory of Millimeter Waves, Nanjing 210096, China

ABSTRACT: Synthetic Aperture Interferometric Radiometer (SAIR) has demonstrated significant potential in Earth remote sensing and radio astronomy. However, most existing imaging methods rely on single-scale visibility function, while SAIR systems typically employ sparse arrays with insufficient sampling, which results in unsatisfactory imaging quality. In this paper, we propose a novel deep learning-based imaging method that addresses this limitation by leveraging multi-scale visibility function. The multi-scale visibility fusion network (MS-VFNet) introduces cross-attention mechanisms in the visibility domain for feature fusion across different scales, fully exploiting the implicit structural information, and subsequently reconstructs the brightness temperature images through a dedicated reconstruction module. The simulation results demonstrate that the proposed MS-VFNet achieves superior reconstruction accuracy and image quality compared to state-of-the-art methods, further validating the feasibility of multi-scale fusion in SAIR super-resolution imaging.

1. INTRODUCTION

Synthetic Aperture Interferometric Radiometer (SAIR) employs interferometric measurement techniques that synthesize a large aperture virtual antenna from a sparsely distributed array of small aperture antennas. This approach overcomes the limitations imposed by the physical size of traditional radiometers. Consequently, SAIR has been widely applied in fields such as Earth remote sensing and astronomical observation. For example, Microwave Imaging Radiometer using Aperture Synthesis (MIRAS) [1], developed by the European Space Agency, has provided extensive soil moisture and sea surface salinity data for global meteorological studies. Similarly, the HY-4A remote sensing satellite [2] developed by the China Academy of Space Technology is equipped with a SAIR instrument for high-resolution observation of sea surface salinity. These applications underscore the research value of SAIR as a high-resolution remote sensing instrument.

Due to the significant advantages of SAIR over traditional radiometers, near-field SAIR shows promise for various applications. However, near-field imaging faces unique challenges including spherical wave interference and multipath effects, demanding higher algorithmic precision than traditional far-field methods. To address the near-field imaging challenges, researchers have developed several improvement strategies. Modified-FFT (MFFT) method [3], based on the imaging process of traditional far-field FFT method, incorporates phase compensation terms to correct the near-field relationship between brightness temperature images and visibility functions. However, it exhibits low description accuracy and is only applicable to rectangular visibility functions measured by conventional antenna arrays. G-matrix imaging method [4] simplifies the complex near-field SAIR imaging process into a mathematical model $V = GT$. With the help of regularization, G-matrix method can accurately reconstruct brightness temperature images and has good versatility. To efficiently solve this regularized problem, optimization algorithms such as Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) are widely employed. Nevertheless, G-matrix method still suffers from issues such as high sensitivity to parameter settings.

In recent years, deep learning has become a research hotspot in the field of SAIR imaging, providing new solutions. In far-field SAIR imaging, the application of deep learning has achieved breakthrough progress. Zhang et al. [5] first employed a data-driven approach by using a convolutional neural network as a decoding model to realize the mapping from visibility to brightness temperature images. Xiao et al. [6] adopted a modified Transformer encoder structure to extract spectral features and employed a brightness temperature image reconstruction module to recover microwave brightness temperature images. In addition, a recent study introduced a new method based on deep neural networks (DNNs) [7], which adopts a customized architecture specifically designed for the SMOS instrument. However, far-field methods cannot be directly applied to near-field SAIR. Guo et al. [8] first applied fully connected networks to near-field reconstruction, and their proposed Fully Connected Imaging Network (FCIN) method achieved better reconstruction results than traditional methods. To address the sensitivity of near-field imaging to distance, Chen et al. [9] fur-

* Corresponding author: Jianfei Chen (chenjf@njupt.edu.cn).

ther proposed a distance-adaptive network Synthetic Aperture Interferometric Radiometer Distance Adaptive Imaging Network (SAIR-DAIN), which can achieve stable high-precision imaging at different distances, with performance surpassing FCIN. Nevertheless, both traditional and deep learning-based methods are generally limited to single-scale inputs and fail to fully utilize visibility information.

Inspired by multimodal architectures, this paper proposes a multi-scale visibility fusion network (MS-VFNet), which takes visibility function at multiple scales as input and outputs brightness temperature images. The multi-scale visibility refers to the visibility functions obtained from multiple observations of the same scene, under different conditions, within the same SAIR system. Specifically, it represents the spectrum of the SAIR visibility function, at different sampling scales, by observing the target scene at varying imaging distances. Unlike previous work that performs feature fusion, MS-VFNet does not operate in the image domain. Instead, it efficiently fuses visibility feature representations at different scales using a cross-attention-based Transformer architecture [10], thereby reducing information loss. After generating high-precision visibility of the target scene, a multi-input multi-output encoder-decoder network is used for inversion. Simulation experiments demonstrate that the proposed MS-VFNet achieves superior performance in terms of reconstructed image quality, high-frequency information recovery, and high-resolution inversion.

2. RELATED WORK

The basic unit of a SAIR system is a binary interferometer. The imaging principle is illustrated in Fig. 1. The visibility function is sampled by performing complex correlation operations on the scene radiation signals that are received by two antennas, while preserving their phase information.

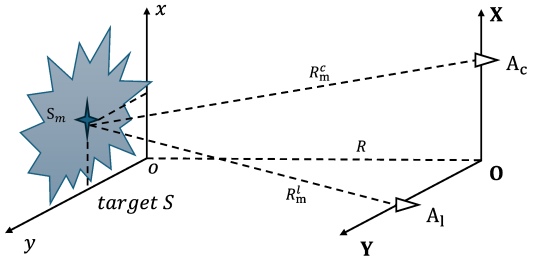

FIGURE 1. Schematic of binary coherence measurement.

Target S is located in the local coordinate system oxy, while the antenna pair (A_c, A_l) is located in the system coordinate frame OXY. The target S is decomposed into M small subcomponents. The distance between the coordinate systems OXY and oxy is denoted by R. The distances from a sub-target S_m to antennas A_c and A_l are defined as R^c_m and R^l_m, respectively. According to paper [11], the visibility function of the target scene is expressed as follows:

$$V_{c,l} = \left\langle E_c(R^c_m, t) \cdot E^*_l(R^l_m, t) \right\rangle = \sum_{m=1}^{M} T(x_m, y_m) F_c(x_m, y_m) F^*_l(x_m, y_m) e^{-jk\left(R^c_m - R^l_m\right)} \quad (1)$$

where $\langle\cdot\rangle$ denotes the time integration operation, $E_{\#}(t)$ the radiation signal received by antenna #, and (x_m, y_m) the coordinates of a point source S_m. $F_{\#}$ is the normalized radiation pattern of antenna #, and the exponential term $\exp[-jk(R^c_m - R^l_m)]$ accounts for the phase difference arising from the optical path difference ΔR between antennas A_c and A_l for source S_m, which is an important factor in SAIR. Using a Taylor approximation, the difference between R^c_m and R^l_m can be expressed as:

$$R^c_m - R^l_m \approx \frac{x_n(X_l - X_c) + y_n(Y_l - Y_c)}{R} + \frac{(X_c^2 + Y_c^2) - (X_l^2 + Y_l^2)}{2R} \quad (2)$$

In the far field, the second term is approximately zero. However, in near-field imaging, it acts as a phase modulation term, which is crucial for near-field imaging. By substituting (2) into (1), we obtain the MFFT imaging formula and approximate matrix Gap [12].

$$T^{MF}(x, y) = FT_2[e^{j\varphi(v,h)} V(v, h)] \quad (3)$$

$$G_{ap}(m, n) = F_c F^*_l e^{\frac{j\pi[2x_n(X_{mc} - X_{ml}) + 2y_n(Y_{mc} - Y_{ml}) + X^2_{ml} + Y^2_{ml} - X^2_{mc} - Y^2_{ml}]}{R\lambda}} \quad (4)$$

where T_{MF} represents the reconstructed image obtained by the MFFT method, $FT_2[\cdot]$ the two-dimensional Fourier transform, and $V(v, h)$ the visibility function with $v = k(X_l - X_c)/R$, $h = k(Y_l - Y_c)/R$, and $\varphi(v, h) = k(X_c^2 + Y_c^2 - X_l^2 - Y_l^2)/2R$. In (4), (X_{mc}, Y_{mc}) and (X_{ml}, Y_{ml}) are the coordinates of antennas (c, l), and the corresponding visibility sample is denoted as V_m. Thus, the following matrix equation can be derived.

$$V_{M\times 1} = G_{M\times N} \cdot T_{N\times 1} \quad (5)$$

where V represents the measured near-field visibility function, T the brightness temperature distribution vector, and G the imaging matrix. In the context of near-field SAIR, the accuracy of the G-matrix model is crucial for image reconstruction, as minor modeling errors can lead to severe artifacts and reconstruction distortions. To accurately reconstruct the image, we directly derive G-matrix from the original synthetic aperture imaging formula (1), with the precise G-matrix elements as follows:

$$G(m, n) = F_{mc} F^*_{ml} e^{j\pi(R^c_m - R^l_m)/\lambda} \quad (6)$$

Since solving the G-matrix is a typical ill-posed inverse problem, the reconstruction results are highly sensitive to measurement noise, and the iterative solution process requires substantial computational resources and time. To address this challenge, various optimization algorithms have been proposed to constrain the solution space by introducing regularization terms, thereby reformulating the problem into a solvable optimization task.

For example, sparse optimization methods such as FISTA [13] exploit the sparsity prior of the image T to effectively suppress noise and improve reconstruction quality.

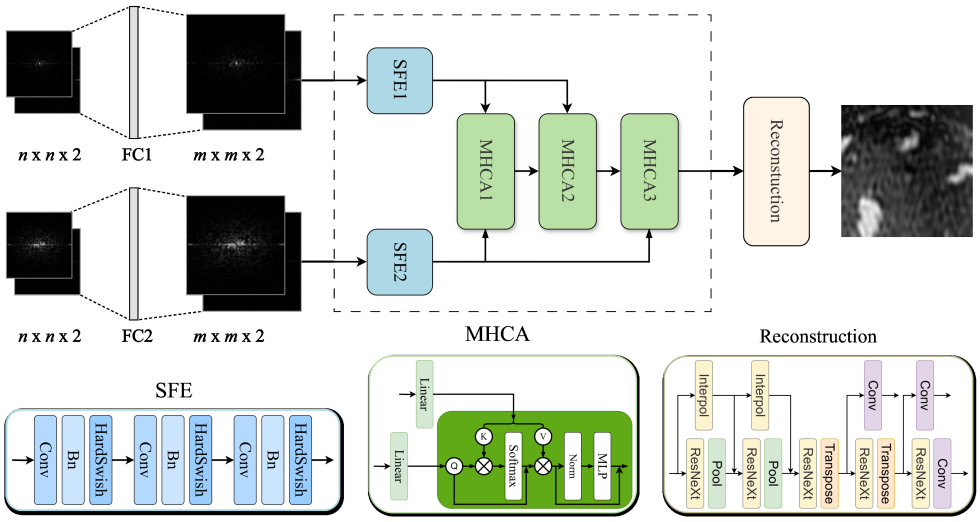

FIGURE 2. Overall network architecture.

However, these methods typically rely on handcrafted priors and still incur high computational costs, which limits their practicality in real-world applications.

In recent years, deep learning methods have been increasingly applied to near-field SAIR imaging. However, most existing methods fail to fully exploit the inherent multi-scale characteristics of the data. For instance, these methods often process visibility function at a single scale, overlooking the complementary information contained in data acquired at different distances, which inherently limits the model generalization capability and the final reconstruction accuracy.

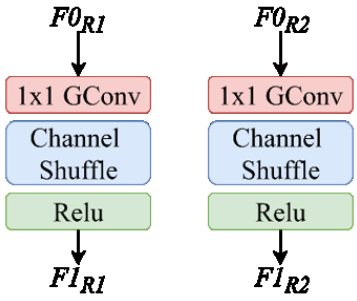

FIGURE 3. Improved fully connected layer.

3. METHODOLOGY

To address the above deficiencies and improve reconstruction quality, this paper presents a network architecture specifically designed for visibility function processing. The network progressively extracts and fuses multi-distance raw data to reconstruct high-fidelity visibility functions and finally reconstructs high-precision brightness temperature images.

3.1. Network Architecture of MS-VFNet

The proposed architecture is shown in Fig. 2, where MS-VFNet consists of three different modules.

3.1.1. Fully Connected Dimension Expansion Module

The fully connected dimension expansion module is responsible for expanding dimensions of the visibility function, i.e., performing spectral dimension expansion. For target scenes, the visibility function is continuously distributed. Therefore, visibility function distributions of various scenes can be learned by the network. The upper branch represents the original visibility function V_{R1} at imaging distance $R1$ scale, while the lower branch represents the original visibility function V_{R2} at imaging distance $R2$. Both V_{R1} and V_{R2} are complex-valued tensors of size $2 \times n \times n$, where the first dimension corresponds to the real and imaginary parts. They are respectively fed into symmetrically structured fully connected layers FC1 and FC2 for dimension expansion, yielding vectors with dimensions of $2 \times m \times m$.

In addition, multi-scale input typically generates multiple branches, where each branch requires fully connected layers with identical structures but different parameters. Since fully connected layers usually constitute the most parameter-intensive part of the network, this design results in a multiplicative increase in both the number of parameters and the computational cost during training. To address potential issues in practical deployment and considering lightweight design and computational efficiency, we replace the conventional fully connected layers with grouped 1×1 convolutions and channel shuffling [14], as shown in Fig. 3. This substitution achieves nearly the same inversion performance as the original design, while significantly reducing the parameter count and decreasing the model size by several times.

3.1.2. Multi-Scale Visibility Fusion Module

To effectively fuse visibility function from different imaging distances, we design a visibility-domain fusion network. This network structure is inspired by the attention mechanism in the multimodal image fusion network ATFuse [15], incorporating shallow feature extraction (SFE) and multi-head cross-attention (MHCA).

We employ SFE composed of multiple 1×1 convolutional blocks to further extract visibility information features, thereby reducing the loss of original visibility information. The core of the multi-scale visibility domain fusion module is MHCA. The computation formula for the cross-attention mechanism is:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

where the query vector Q, key vector K, and value vector V are generated from feature maps through corresponding linear mapping layers. d_k is the dimension of the key vector. The softmax function is used to normalize the similarity matrix into attention weights in probability form. The key difference between this cross-attention mechanism and self-attention is that Q, K, and V come from different feature sequences.

We stack three MHCA to achieve progressive fusion across multi-scale visibility features. Specifically, in MHCA1, we use the query vector from the $R1$ distance feature mapping as the query input, and the key and value vectors from the $R2$ distance feature mapping as the key and value inputs. The module generates attention weight maps by computing the similarity between Q and K, which indicate the parts that $R1$ features should focus on in the $R2$ feature space. The final output is the weight-weighted V, which fuses $R2$ information into the $R1$ perspective.

$$F1 = \text{MHCA1}(V_{R1Q}, V_{R2K}, V_{R2V}) \tag{8}$$

In MHCA2, the output $F1_Q$ from MHCA1 serves as the query input, with V_{R1K} and V_{R1V} from $R1$ as the key and value inputs. Similarly, MHCA3 uses the output $F2_Q$ from MHCA2 as the query input, with V_{R2K} and V_{R2V} from $R2$ as the key and value inputs. Finally, we obtain a high-precision visibility function with the same size as the input visibility.

$$F2 = \text{MHCA2}(F1_Q, V_{R1K}, V_{R1V}) \tag{9}$$

$$F3 = \text{MHCA3}(F2_Q, V_{R2K}, V_{R2V}) \tag{10}$$

Beyond introducing multi-scale data to increase information, another significant advantage of this module lies in its capability to perform spectral fusion in the visibility domain, which reduces information loss more effectively than image domain fusion. To emphasize this advantage, we design a Multi-Scale Image Fusion Network (MS-IFNet). Its architecture is derived primarily through the following modifications: First, the output of the fully connected dimension expansion module is reformulated into an $m \times m$ image domain representation; second, our visibility domain fusion module is replaced with ATFuse's image domain fusion module; finally, the same multi-input multi-output reconstruction module is employed. Consequently, the core difference between MS-IFNet and MS-VFNet lies solely in whether multi-scale fusion is performed in the visibility domain or image domain, as illustrated in Fig. 4. Specifically, T_{R1} and T_{R2} denote the image-domain representations transformed from V_{R1} and V_{R2}, respectively.

3.1.3. Multi-Input Multi-Output Reconstruction Module

After the fusion module, a refined visibility function of size $2 \times m \times m$ is obtained. A hybrid architecture combining

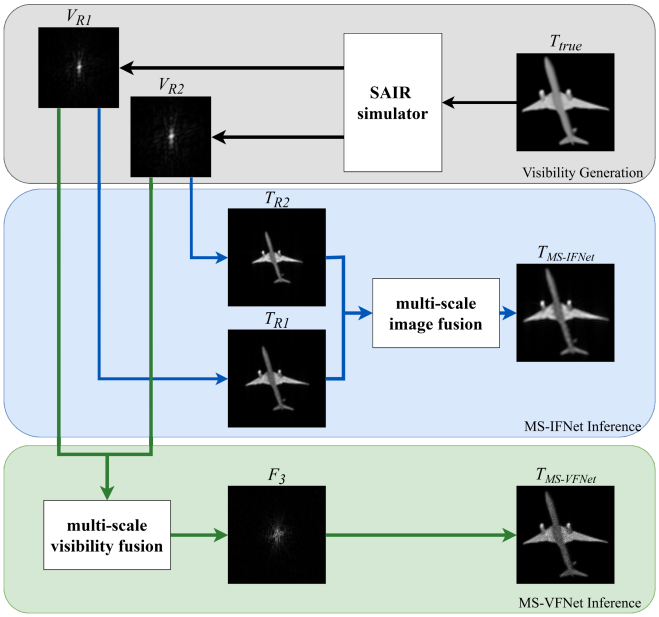

FIGURE 4. Multi-scale visibility fusion and multi-scale image fusion.

ResNeXt [16] and U-Net [17] is employed as the final module, incorporating multi-scale auxiliary inputs and outputs to facilitate hierarchical feature learning for brightness temperature image reconstruction.

Let the original input resolution be $H \times W$. After two stages of downsampling, the feature maps are reduced to $\frac{H}{2} \times \frac{W}{2}$ and $\frac{H}{4} \times \frac{W}{4}$, respectively. At each stage, the original input is resized via interpolation to match the corresponding scale, denoted as Interp $(H \times W, s)$, where s is the scale factor (e.g., $s = 2$ corresponds to $\frac{H}{2} \times \frac{W}{2}$). After encoding, the spatial resolution is reduced to $\frac{H}{4} \times \frac{W}{4}$, and the channel dimension increases to embedding_dim.

The decoder generates multi-scale outputs at $H \times W$, $\frac{H}{2} \times \frac{W}{2}$, and $\frac{H}{4} \times \frac{W}{4}$. A composite pixel-wise loss is applied to ensure both global consistency and local detail preservation during training. The optimization objective, obtained by backpropagating the loss function at each resolution, can be expressed as:

$$L_{pixel} = L_{part1} + L_{part2} + L_{part3} \tag{11}$$

where $\mathcal{L}_{\text{part1}}$, $\mathcal{L}_{\text{part2}}$, and $\mathcal{L}_{\text{part3}}$ correspond to resolutions of $H \times W$, $\frac{H}{2} \times \frac{W}{2}$, and $\frac{H}{4} \times \frac{W}{4}$, respectively. Multi-scale supervision is applied to improve both fine-grained details and overall reconstruction quality.

The network is trained using the Root Mean Square Error (RMSE) as the pixel-wise loss, defined as:

$$\text{RMSE} = \sqrt{\frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(T_{\text{Pred}}(i,j) - T_{\text{GT}}(i,j)\right)^2} \tag{12}$$

where $T_{Pred}(i,j)$ and $T_{GT}(i,j)$ denote the predicted and ground truth pixel values at position (i,j), respectively. This metric calculates the squared difference at each pixel, averages over all pixels, and then takes the square root to measure the overall reconstruction error. A lower RMSE indicates that the

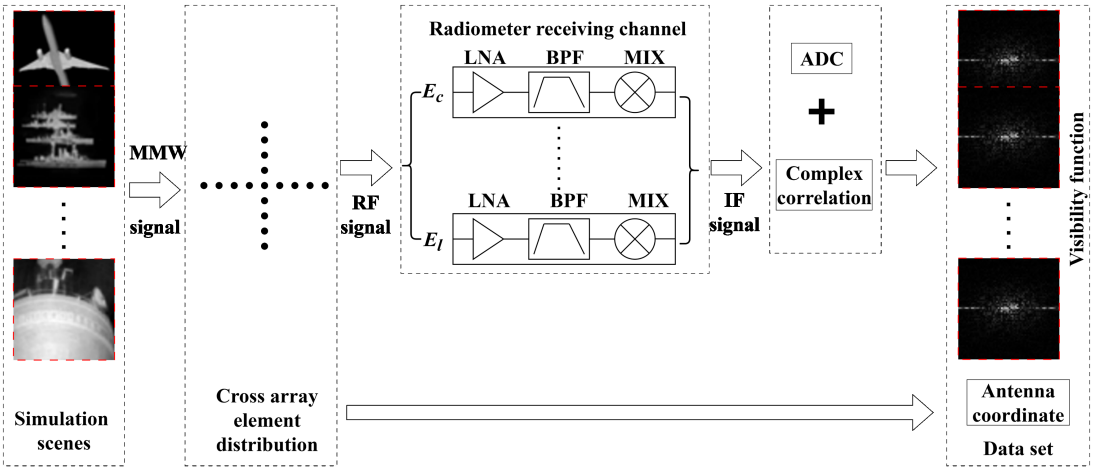

FIGURE 5. SAIR simulator.

predicted image is closer to the ground truth, reflecting higher reconstruction accuracy.

3.2. Dataset Generation and Training

Based on the imaging process of SAIR, we have established a SAIR simulation analysis model, which aims to generate visibility functions from grayscale images. The simulation parameters are shown in Table 1.

TABLE 1. Simulation parameters of SAIR.

Parameters	Values
Array Size	64×64
Wavelength	3 mm
Source spacing	1.56 cm
Antenna Spacing	1 cm
Antenna aperture	0.64 m
Imaging distance	8 m, 10 m

In this study, we utilize the SAIR simulator, as illustrated in Fig. 5, to generate the dataset. The simulator is designed to accurately reproduce the signal processing flow of a practical SAIR system. Specifically, the imaging simulator takes a 128×128 grayscale image as the input scene, and a 64×64 cross-shaped antenna array is employed to capture the millimeter-wave radiation signals from the target. The received radiation signals are integrated to obtain E_c and E_l. According to the principle of linear superposition of radiation, the radio frequency (RF) signal received at each antenna element is simulated. These signals are then processed through amplification, filtering, and mixing stages to produce the intermediate frequency (IF) signal of the radiometer. Finally, the visibility function V is obtained by an analog-to-digital converter (ADC) and the simulator outputs V together with the antenna coordinates, thus completing the entire simulation process.

The simulation procedure adopted in this study is as follows:

1. 50,000 natural scene images were selected and converted into grayscale images, with a target image resolution of 128×128.

2. Select an image T for SAIR imaging simulation, and obtain visibility functions V_{R1} and V_{R2} with the dimension of 64×64 at imaging distances $R1$ and $R2$, respectively.

3. Repeat Step 2 for each natural image to construct the complete dataset. The dataset is then randomly divided into training, validation, and testing subsets, with 80% of the data used for training, 10% for validation, and the remaining 10% for evaluating the imaging performance of the proposed method.

The network training process is shown in Fig. 6. High-resolution scenes are processed through SAIR simulation to generate visibility functions V_{R1} and V_{R2} at different distances as inputs, outputting predicted images T_{pred}. The predictions are compared with target images T_{target} to calculate loss. If the result is below a predetermined threshold, training is ter-

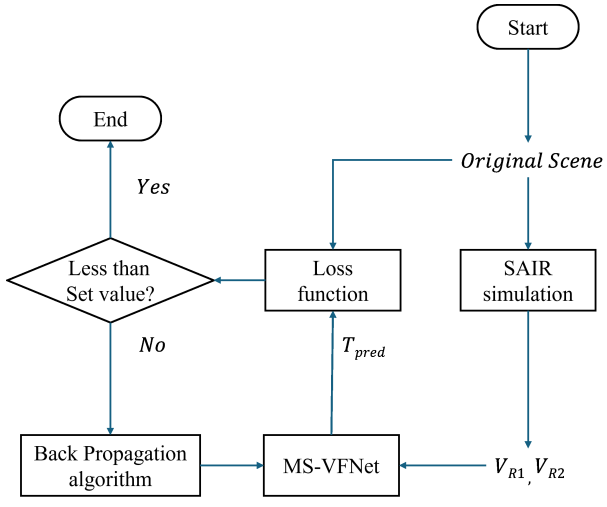

FIGURE 6. The process of network training.

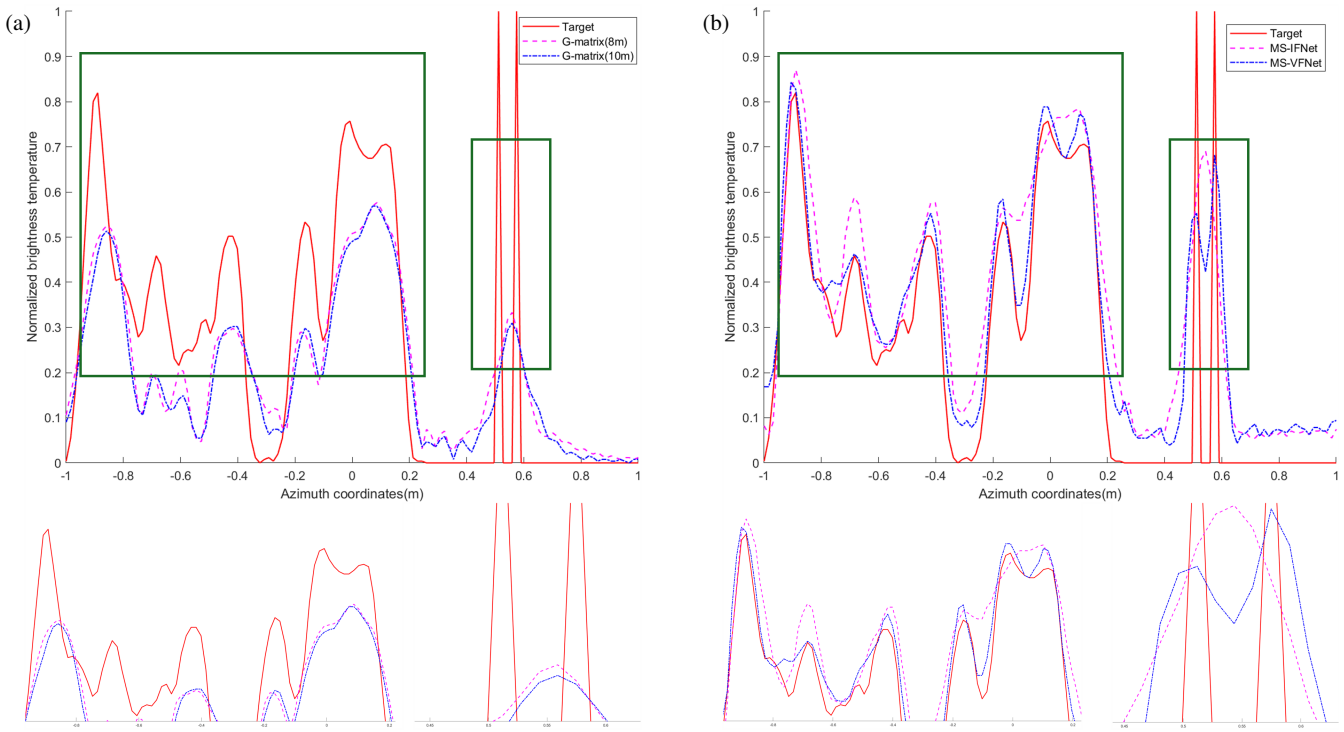

FIGURE 7. Reconstructed images of 1-D. (a) G-matrix method at imaging distances of 8 m and 10 m. (b) MS-IFNet and MS-VFNet methods.

minated. Otherwise, parameters are updated using the back-propagation.

The entire network was implemented based on the PyTorch framework. During training, the learning rate was set to 0.03 and the batch size to 64. Adam optimizer was employed, and the training process was conducted for 300 epochs. The number of attention heads was set to four.

4. SIMULATION AND RESULTS

4.1. Experimental Environment

In this section, to evaluate the effectiveness of the MS-VFNet method, we designed comparative experiments with multiple imaging methods (G-matrix method, SAIR-DAIN method, MS-IFNet method, and MS-VFNet method). The experimental platform is a PC equipped with CPU (i9-10900K) and GPU (NVIDIA RTX 2080 Ti).

4.2. 1-D Imaging Experiments

To evaluate the effectiveness of the proposed MS-VFNet method in improving imaging resolution, we conducted imaging simulation experiments on a 1-D target. The original scene consists of two parts: the left part is a 1-D simulation curve, and the right part contains two point sources with a source spacing of $1\Delta L$. Since the image-domain fusion method achieves better reconstruction than SAIR-DAIN, only the results of the MS-IFNet method are presented. Fig. 7(a) shows the reconstruction results using the G-matrix method applied separately to visibility functions at imaging distances of 8 m and 10 m. Fig. 7(b) displays the results of MS-IFNet and MS-VFNet methods, both fusing visibility function from 8 m and 10 m. For ease of comparison, the target image is shown in red, while the reconstructed images are shown in blue.

The simulation results show that the traditional G-matrix method suffers from strong over-smoothing. It fails to recover fine peak-valley features in the target curve and cannot distinguish the two right-side point sources, leading to major detail loss. As seen in Fig. 7(b), MS-IFNet and MS-VFNet yield reconstructions that better match the original curve. MS-IFNet captures most details yet still shows deviations in peak height and shape. This can be attributed to information loss and the introduction of errors during the transformation from the visibility domain to the image domain. By leveraging multi-scale visibility fusion, the proposed MS-VFNet demonstrates stronger reconstruction capabilities. As evident from the zoomed-in view, it successfully recovers multiple peak-valley structures on the left and clearly distinguishes the two point sources on the right. In summary, the MS-VFNet method outperforms both the G-matrix and MS-IFNet methods in terms of resolution enhancement.

4.3. 2-D Imaging Experiments

To evaluate the effectiveness of the proposed MS-VFNet in enhancing imaging accuracy, we conducted 2-D imaging simulation experiments in this section. Since the visibility function sampled at 8 m contains finer spatial details, the single-scale input method SAIR-DAIN utilizes only the 8 m visibility function. Fig. 8(a) and Fig. 8(e) show the original high-resolution target images. Figs. 8(b)-(d) and Figs. 8(f)–(h) present the re-

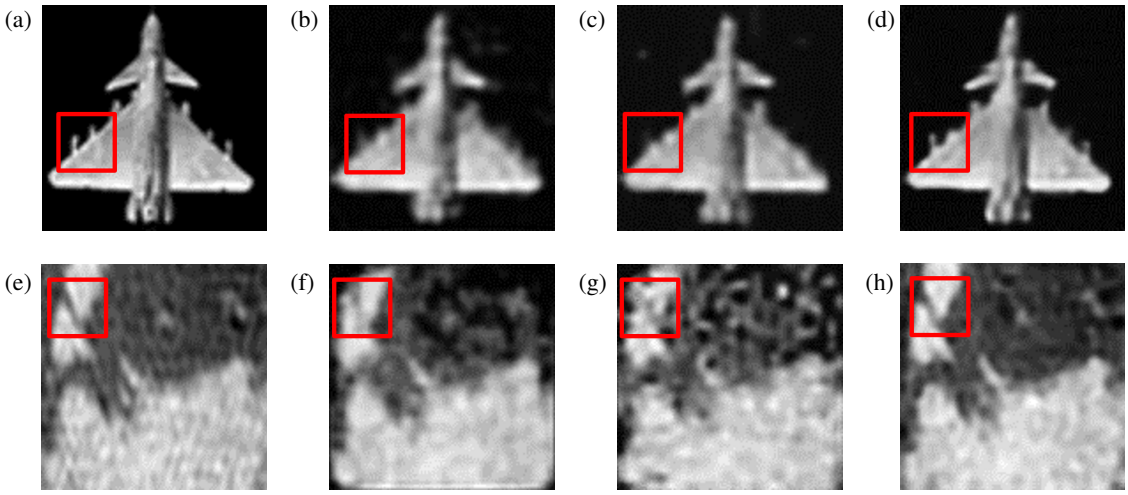

FIGURE 8. Reconstructed images of 2-D. (a) Earth Scene. (b) Image $T_{SAIR\text{-}DAIN}$. (c) Image $T_{MS\text{-}IFNet}$. (d) Image $T_{MS\text{-}VFNet}$. (e) Plane Scene. (f) Image $T_{SAIR\text{-}DAIN}$. (g) Image $T_{MS\text{-}IFNet}$. (h) Image $T_{MS\text{-}VFNet}$.

constructed images obtained by the SAIR-DAIN, MS-IFNet, and MS-VFNet methods, respectively.

The reconstructed image $T_{SAIR\text{-}DAIN}$ can preliminarily restore the target scene information but introduces background noise and loses some content. As clearly observed in Figs. 8(b) and (f), noise artifacts appear around the aircraft target, and a small portion is missing in the upper right corner of the Earth scene. For the MS-IFNet method, the reconstructed images shown in Figs. 8(c) and (g) exhibit reduced information loss through multi-scale fusion, but the noise contamination issue remains unresolved, resulting in suboptimal performance. The reconstruction results of the MS-VFNet method, as illustrated in Figs. 8(d) and (h), are closer to the original images with visually enhanced clarity, clean background, and rich image information, accurately reconstructing aircraft wings and continental fissures. This demonstrates that the MS-VFNet method can effectively fuse original multi-scale visibility function without introducing errors and recover high-frequency information in images.

To more accurately evaluate the quality of reconstructed images, we calculated two commonly used objective evaluation metrics for the reconstructed images in Fig. 8: Peak Signal-to-Noise Ratio (PSNR) and RMSE. A higher PSNR value indicates a smaller difference between the reconstructed image and original image, meaning higher similarity between them. RMSE is used to quantify the average error between predicted values and true values, reflecting the accuracy of model predictions. A smaller RMSE value indicates a smaller difference between the predicted image and target image, corresponding to higher image restoration quality. The PSNR can be derived as,

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{\mathrm{MAX}^2}{\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(T_{Pred}(i,j)-T_{GT}(i,j)\right)^2}\right) \tag{13}$$

For further objective evaluation, Table 2 lists PSNR and RMSE values for each method. The proposed MS-VFNet achieves the highest PSNR and lowest RMSE in both the aircraft and Earth scenes. This objectively demonstrates that MS-VFNet can achieve higher-precision imaging than SAIR-DAIN and MS-IFNet methods.

TABLE 2. Performance metrics for three method.

Scene	Criterion	SAIR-DAIN	MS-IFNet	MS-VFNet
Plane	RMSE	20.710	20.518	12.641
	PSNR	21.807	21.888	26.095
Earth	RMSE	30.036	27.499	14.236
	PSNR	18.578	19.344	25.063

During the imaging process, inversion time is a critical factor to consider. Although data preparation and model training usually require considerable time in the early stages, once the network training is completed, the inversion process only involves loading the trained model weights. As a result, deep learning-based inversion methods exhibit extremely fast imaging speeds. In contrast, for the traditional G-matrix method, the imaging process requires first constructing the G-matrix, followed by a large number of iterative computations during the inversion stage. Typically, hundreds of iterations are needed to obtain the final result, making this approach highly time-consuming.

Table 3 lists in detail the inversion time required by the three methods, all tested under the same hardware platform and runtime environment. For the software environment, G-matrix method was implemented in MATLAB R2024b, while SAIR-DAIN, MS-IFNet, and MS-VFNet methods were implemented in PyCharm 2024. It can be clearly observed that the

TABLE 3. Image reconstruction time comparison.

Scene	G-matrix	SAIR-DAIN	MS-IFNet	MS-VFNet
Earth	27.5 s	1.5 s	1.1 s	1.1 s

G-matrix method takes a significantly longer time, whereas the deep learning-based methods are much faster. The difference between MS-VFNet and MS-IFNet is relatively small, mainly because the fully connected layers, which dominate the network's parameter size and computational cost, are similar in both networks, and the other modules also show no substantial differences. Compared with the SAIR-DAIN method, our proposed approach achieves a smaller number of parameters and faster inversion speed owing to its lightweight network design. These results demonstrate that the proposed MS-VFNet method possesses excellent real-time performance and can effectively meet the imaging speed requirements of practical applications.

5. CONCLUSION

In near-field SAIR imaging, existing methods using original single-scale visibility result in suboptimal imaging accuracy. To address this problem, we propose a novel deep learning network MS-VFNet in this paper. We design a cross-attention Transformer architecture to extract global features from multi-scale visibility functions and reduce information loss, thereby obtaining high-precision original visibility function for inversion imaging. Simulation results show that the reconstructed image quality of the proposed MS-VFNet method is better than the existing near-field SAIR brightness temperature image reconstruction methods (G-matrix method, SAIR-DAIN method, and MS-IFNet method). In future work, incorporating additional multi-scale visibility function from more imaging distances may further enhance SAIR imaging performance.

ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 62371255 and Grant 61601237, in part by the State Key Laboratory of Millimeter-Waves under Grant K202526.

REFERENCES

[1] Kerr, Y. H., A. Al-Yaari, N. Rodriguez-Fernandez, M. Parrens, B. Molero, D. Leroux, S. Bircher, A. Mahmoodi, A. Mialon, P. Richaume, S. Delwart, A. A. Bitar, T. Pellarin, R. Bindlish, T. J. Jackson, C. Rüdiger, P. Waldteufel, S. Mecklenburg, and J.-P. Wigneron, "Overview of SMOS performance in terms of global soil moisture monitoring after six years in operation," *Remote Sensing of Environment*, Vol. 180, 40–63, 2016.

[2] Liu, Y., D. Zhu, P. Liu, R. Yun, T. Wang, and X. Zang, "The HY-4A scatterometer: A novel active system for measuring surface roughness for accurate sea-surface brightness temperature correction," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 63, 1–16, 2025.

[3] Chen, J., J. Guo, S. Zhang, and X. Zhu, "Blind restoration method for near-field millimeter-wave SAIR," in *2018 43rd International Conference on Infrared, Millimeter, and Terahertz Waves (IRMMW-THz)*, 1–2, Nagoya, Japan, 2018.

[4] Du, S., F. Zhao, J. He, X. Chen, and L. Jiang, "A high-resolution image reconstruction technology based on G-matrix and fourier transform," in *2022 3rd China International SAR Symposium (CISS)*, 1–7, Shanghai, China, 2022.

[5] Zhang, Y., Y. Ren, W. Miao, Z. Lin, H. Gao, and S. Shi, "Microwave SAIR imaging approach based on deep convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 57, No. 12, 10 376–10 389, 2019.

[6] Xiao, C., H. Dou, H. Li, R. Jin, R. Zhai, W. Wang, R. Lv, and Y. Li, "Image reconstruction of synthetic aperture radiometer by transformer," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 61, 1–15, 2023.

[7] Faucheron, R., E. Anterrieu, L. Yu, A. Khazaal, and N. J. Rodríguez-Fernández, "Deep-learning-based approach in imaging radiometry by aperture synthesis: An alias-free method," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 17, 6693–6711, 2024.

[8] Guo, Z., J. Chen, and S. Zhang, "Fully connected imaging network for near-field synthetic aperture interferometric radiometer," *IEICE Transactions on Information*, Vol. E105-D, No. 5, 1120–1124, 2022.

[9] Chen, J., J. Zhou, and R. Peng, "Distance adaptive imaging algorithm for synthetic aperture interferometric radiometer in near-field," *IEEE Geoscience and Remote Sensing Letters*, Vol. 20, 1–5, 2023.

[10] Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv:1706.03762v7*, 2017.

[11] Yao, X., C. Zheng, J. Zhang, B. Yang, A. Hu, and J. Miao, "Near field image reconstruction algorithm for passive millimeter-wave imager BHU-2D-U," *Progress In Electromagnetics Research C*, Vol. 45, 57–72, 2013.

[12] Chen, J., X. Zhu, S. Zhang, and X. Chen, "General G-matrix imaging method for near-field millimeter-wave SAIR with any arrays," in *2018 IEEE MTT-S International Wireless Symposium (IWS)*, 1–3, Chengdu, China, 2018.

[13] Li, S., M. Amin, G. Zhao, and H. Sun, "Radar imaging by sparse optimization incorporating MRF clustering prior," *IEEE Geoscience and Remote Sensing Letters*, Vol. 17, No. 7, 1139–1143, 2020.

[14] Zhang, X., X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6848–6856, Salt Lake City, UT, USA, 2018.

[15] Jian, L., S. Xiong, H. Yan, X. Niu, S. Wu, and D. Zhang, "Rethinking cross-attention for infrared and visible image fusion," *arXiv:2401.11675v1*, 2024.

[16] Xie, S., R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1492–1500, Honolulu, HI, USA, Jul. 2017.

[17] Ronneberger, O., P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention — MICCAI 2015*, 234–241, Springer, Cham, 2015.