(Received 13 July 2024, Accepted 16 October 2024, Scheduled 6 November 2024)

The Dual-Modality Fusion Imaging Method for EMT and UTT Based on DSCTFusion-ECA

Jinxun Le[1], Ronghua Zhang[1,*], Wenying Fu[1], Shuqing Jia[1], Xuefeng Bai[2], and Boyang Li[2]

[1]*School of Artificial Intelligence, Tiangong University, Tianjin 300387, China*
[2]*School of Control Science and Engineering, Tiangong University, Tianjin 300387, China*

ABSTRACT: Dual-modality tomography integrates two different imaging technologies, allowing for the acquisition of more comprehensive sensing data. By combining information from both modalities, the accuracy of final imaging results is enhanced. However, due to the use of different physical sensitive field backgrounds by different measurement modalities, integrating information from different modalities with differing dimensions presents a challenge. To address this issue, a supervised DSCTFusion-ECA deep learning method is proposed. This method consists of four modules: initial imaging, feature extraction, feature fusion, and image reconstruction. In the feature extraction module, dense connections are utilized first to extract shallow cross-modal features, then two dual-branch feature extraction networks are utilized to separately capture modality-specific low-frequency global features and high-frequency local features for both modalities. The performance and robustness of multi-modality tomography can be effectively improved through the extraction of more comprehensive features. In the feature fusion module, Efficient Channel Attention is employed to capture channel dependencies and generate attention weights. The modal complementarity and the representation ability of key features have been enhanced, while avoiding information redundancy, thereby improving the discriminative power of the features. Simulation results show that the proposed network can fully extract and fuse features from EMT and UTT modalities, demonstrating strong robustness and generalization. Compared to the widely used U-Net network in tomography, DSCTFusion-ECA yields better reconstruction results.

1. INTRODUCTION

Tomographic imaging is a functional imaging technique that has been extensively researched and applied in both industrial and medical fields. Ultrasonic Transmission Tomography (UTT) [1] and Electromagnetic Tomography (EMT) [2] have gained significant attention and widespread application over the past few decades due to their advantages of low cost and radiation-free imaging [3–5].

Traditional single-modality tomographic imaging techniques are often limited in their applications due to the inherent sensitivity principles. However, multi-modality imaging technology, which combines the advantages of two or more imaging modalities, has the potential to overcome these limitations and effectively improve imaging accuracy [6, 7]. Multi-modality tomographic imaging reconstruction methods primarily consist of two approaches: incorporating prior information from one modality and fusing images from two modalities. Incorporating prior information from one modality involves utilizing the reconstruction result of one modality as prior knowledge for the reconstruction of another modality. For instance, Steiner et al. used prior information about the medium distribution obtained from URT to constrain the EIT image reconstruction process, improving the reconstruction accuracy of small targets [8]. Jiang et al. designed a capacitive-coupled EIT/UT dual-modality imaging system, comparing and analyzing the characteristics, correlations, and complementarities of the two imaging modalities to enhance image reconstruction quality [9]. Xu et al. used bubble distribution information obtained from UT measurements as prior information for ERT reconstruction, achieving complementary spatial resolution of dual-modality sensitive field distributions [10]. Liang et al. used UT positional measurements as prior information to guide the free interface reconstruction of EIT, employing constrained least squares for the fusion of different modality information [11]. Fusion of two modality images involves combining the reconstruction results from both modalities. For instance, Zhang et al. used digital image fusion methods to combine ECT and CT reconstructed images, enhancing the image reconstruction accuracy of gas/liquid two-phase flow cross-sectional distribution [12]. Pusppanathan et al. used convolution back projection to reconstruct images separately from ECT and UT measurements and proposed a fuzzy logic pixel fusion algorithm to merge the reconstructed images from ECT and UT [13]. Yue et al. utilized fuzzy clustering methods to determine the fuzzy membership of image pixels in the reconstructed results from ERT and ECT, thereby calculating the grayscale value of each pixel for image reconstruction [14].

The two aforementioned methods can indeed enhance the quality of reconstruction to a certain extent, but they each have their respective limitations. In the approach where one modality serves as prior information, the effectiveness of the final imaging is constrained by the quality of the constructed prior information. In the method of fusing images from two modalities, the two single-modal reconstruction results are directly

* Corresponding author: Ronghua Zhang (zhangronghua@tiangong.edu.cn).

fused. It makes the fusion results constrained by the single-modal reconstruction results and the fusion strategy.

In recent years, the advancement of deep learning has led to an increasing application of deep learning algorithms in tomographic imaging. Deep learning is capable of capturing complex nonlinear relationships, enabling it to handle intricate imaging conditions and irregular data for improved image reconstruction quality. Malikov et al. used U-Net architecture and transfer learning to enhance ultrasound tomographic imaging results of containment liner plates [15]. Shi et al. proposed a convolutional neural network segmentation method to ensure the accuracy of ultrasound velocity inversion, followed by using the velocity corrected version of the Kirchhoff migration method to reconstruct skeletal images, achieving good results [16]. Xiao et al. introduced two deep learning algorithms, SSAE+RBF and optimized fully connected networks, to enhance the imaging quality of EMT [17]. Zhang et al. first reconstructed initial conductivity and permeability images using the Landweber algorithm based on the aforementioned measurements, then input these initial images into an improved DeepLabv3 network for image segmentation, resulting in high-quality conductivity and permeability distribution reconstruction images [18].

In order to enable comprehensive fusion of EMT and UTT modality data, a data-driven deep learning network named DSCTFusion-ECA is proposed for dual-modality fusion imaging. This network consists of two initial imaging modules, a feature extraction module, a feature fusion module, and an image reconstruction module. In the initial imaging module, a fully connected network is utilized to map modality data into complex nonlinear relationships for target reconstruction. The aim is that measurement information of different dimensions is transformed into pixel information in the same dimensional image space to address heterogeneity between different modalities. In the feature extraction module, characteristics such as shape, position, and size are meticulously extracted from pixel information and then integrated in the feature fusion module. Finally, the integrated features are employed in the image reconstruction module to reconstruct the image. The main contributions of this approach can be summarized as follows:

(1) A deep learning network for dual-modality fusion imaging of EMT and UTT is proposed. The inputs of the network are EMT boundary voltage measurements and UTT boundary acoustic time measurements, and the output is the predicted medium distribution.

(2) During the feature extraction process, both common features between modalities and private features specific to each modality are considered. In extracting modality-specific features, dual branches are used to extract high-frequency local features and low-frequency global features respectively for each modality, ensuring more comprehensive feature extraction by the network.

(3) In the feature fusion process, high-frequency local features and low-frequency global features of EMT and UTT modalities are first summed correspondingly. Then, Efficient Channel Attention (ECA) is applied to assign different weights to the features, highlighting important features while suppressing irrelevant or minor features. Therefore, the network focuses more on critical information.

2. FUSION METHOD OF EMT AND UTT

2.1. Principles of EMT and UTT

Assuming that the electromagnetic fields in EMT follow quasi-static field laws and that the medium is isotropic, substituting the magnetic vector potential into Maxwell's equations allows derivation of the mathematical model for the EMT system [19]:

$$\nabla^2 A = j\omega\mu\sigma \tag{1}$$

where A represents the magnetic vector potential, ω the angular frequency, μ the magnetic permeability, and σ the electrical conductivity.

From Equation (1), it is evident that the magnetic vector potential is influenced by the distribution of electrical conductivity and permeability in the imaging region. The magnetic vector potential A and magnetic induction B satisfy:

$$\nabla A = B \tag{2}$$

Based on the known magnetic vector potential, the induced voltage in the detection coil can be obtained using the following formula:

$$\mu = -\frac{d\psi}{dt} = -n \cdot \frac{d(B \cdot S_d)}{dt} = -n \cdot \frac{d(A \cdot l)}{dt} \tag{3}$$

where μ is the induced voltage, ψ the magnetic flux linkage through the coil, n the number of turns of the detection coil, S_d the area of the coil, and l the axial length of the coil.

From Equations (1) to (3), it is evident that different distributions of electrical conductivity in the sample will result in varying magnetic fields in the imaging region, thereby yielding different induced voltage data. The EMT modality measurement consists of an array of 8 circularly distributed sensors, as shown in Figure 1(a). The data collection in EMT involves single-coil current excitation and multiple-coil voltage measurement. Specifically, one coil is selected as the excitation coil during each measurement cycle, while the remaining 7 coils serve as detection coils to measure induced voltages. After each coil has been used as an excitation coil, a total of 56 measurements are collected for one sample of EMT modality data.

The mathematical model of UTT is determined by the acoustic wave propagation model, which includes both wave acoustics and ray acoustics models. The wave acoustics model considers various interactions between acoustic waves and the medium, providing a more accurate description of acoustic wave propagation in the medium. However, in practical applications, the wave acoustics model is often computationally intensive and sensitive to noise. The ray acoustics model simplifies the modeling process of acoustic wave-medium interactions, making it more feasible computationally. Additionally, in the absence of diffraction effects, the ray acoustics model can be considered an approximation and simplification of the wave acoustics model [20]. In the ray acoustics model, the relationship between the speed of sound along the propagation path and

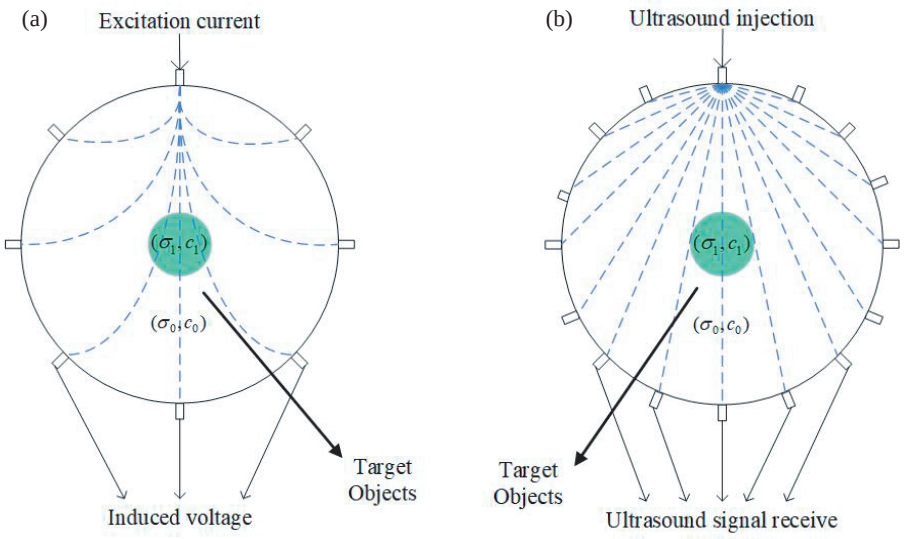

FIGURE 1. Dual-modality simulation model. (a) EMT. (b) UTT.

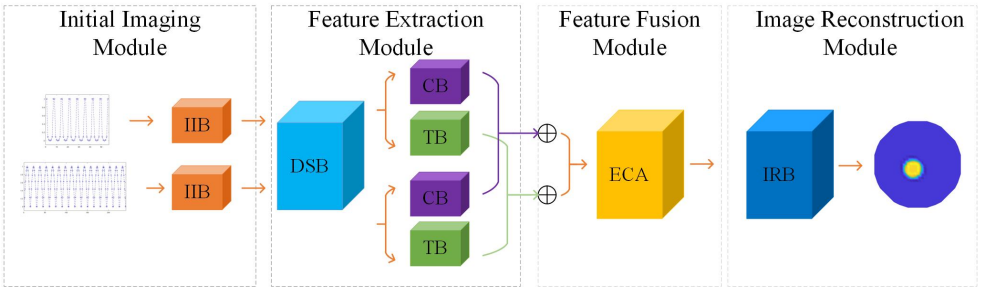

FIGURE 2. The structure of DSCTFusion-ECA.

the time of flight (TOF) can be associated through the following mapping:

$$T = \int_{ray} \frac{1}{c(x)} dx = \int_{ray} s(x) dx \tag{4}$$

where T represents TOF, $c(x)$ the speed of sound distribution, $s(x) = \frac{1}{c(x)}$ rthe slowness distribution, and ray the propagation path.

The measurement of the UTT modality consists of an array of 16 sensors distributed in a circular pattern, as shown in Figure 1(b). The data collection in UTT involves single ultrasound transducer emission and multiple ultrasound transducer reception. Specifically, one ultrasound transducer is selected as the emitter during each measurement cycle, while the remaining 15 transducers serve as receivers to collect acoustic timing data. Once each of the 16 transducers has been used as an emitter, a total of 240 measurements are collected for one sample of UTT modality data.

2.2. Network Structure

When describing the modal information on the distribution of the medium in the same measurement area, there may be redundant, conflicting, or complementary information. Additionally, both modalities exhibit characteristics such as heterogeneity, high dimensionality, and differing dimensions. Effectively and reasonably utilizing this information is a core challenge in the process of dual-modality tomographic imaging reconstruction.

In order to comprehensively and effectively integrate the measurement information from EMT and UTT, a network named DSCTFusion-ECA is proposed for dual-modality tomographic imaging reconstruction of EMT and UTT. The network consists of initial imaging, feature extraction, feature fusion, and image reconstruction modules. The structure of the network is shown in Figure 2.

2.2.1. Initial Imaging Module

The input to the initial imaging module is normalized measurement data, including 56 EMT boundary voltage measurements and 240 UTT boundary acoustic time measurements. These inputs are processed through two Initial Imaging Blocks (IIBs), each dedicated to learning the nonlinear mapping between boundary voltage and acoustic time measurements within the medium's distribution. Each IIB consists of a 4-layer fully connected neural network with 128, 256, 512, and 1024 neurons per layer respectively (as shown in Figure 3). The data from the final layer is reshaped to create an initial image of size 32×32 pixels.

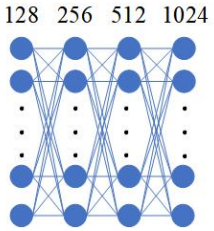

FIGURE 3. The structure of IIB.

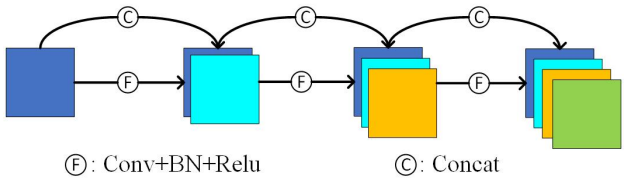

FIGURE 4. The structure of DSB.

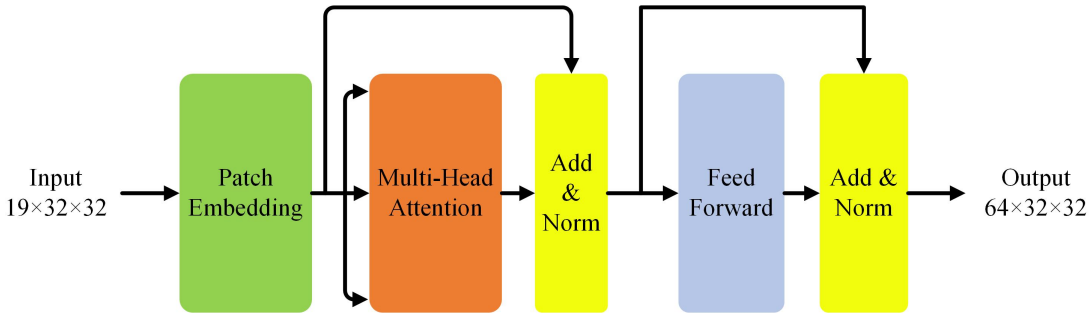

FIGURE 5. The structure of TB.

2.2.2. Feature Extraction Module

The feature extraction module consists of three parts: a Dense Block (DSB) based on Dense Connecting [21], a Transformer Block (TB) based on Transformer Encoder Block [22], and a CNN Block (CB) based on Invertible Neural Networks Block (INNB) [23, 24].

The DSB is a shared feature extractor. Its structure is shown in Figure 4. By dense connection mechanism, the DSB effectively retains original information during the feature extraction process, thereby enhancing the network's information flow and feature reuse. The DSB can increase feature maps to improve the network's representation and learning capabilities. Therefore, the DSB is used to extract shallow features of cross-modality. The entire DSB can be represented as:

$$x_l = F\left([x_0, x_1, x_2, \ldots, x_{l-1}]\right) \tag{5}$$

where $x_0, x_1, x_2, \ldots, x_l$ represent the concatenated output feature maps of network layers $0, 1, 2, \ldots, l$, and F is the nonlinear transformation function composed of Convolution, Batch Normalization (BN), and ReLU layers.

The Transformer Block (TB) is a private feature extractor designed to extract modality-specific low-frequency global features from the shallow information of cross-modality. The structure is shown in Figure 5. This module takes the output of the DSB as its input. First, Patch Embedding is used to divide the input image into fixed-size patches and convert each patch into a fixed-dimensional vector representation. Next, Multi-Head Attention mechanism is used to compute attention weights. By self-attention mechanism, the model can capture global relationships in sequential data regardless of distance. Self-attention mechanism enables the model to dynamically allocate attention between different positions, thereby enhancing its ability to understand long-range dependencies. The formula for the self-attention mechanism is as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimension of the key vectors.

Subsequently, the output passes through a residual connection and layer normalization, followed by a feedforward neural network layer, and another residual connection and normalization layer to obtain the final output.

The CNN Block (CB) is also a private feature extractor to extract modality-specific high-frequency local features from the shallow cross-modality information. The CB consists of a Patch Embedding and an Invertible Neural Network Block (INNB), where the Patch Embedding is the same as in the TB. To capture edge and texture information in the features as effectively as possible, the INNB with affine coupling layers is employed. The INNB generates input and output features mutually, allowing for better preservation of input information and can be considered a lossless feature extraction block. The structure is shown in Figure 6. Each reversible layer transformation is:

$$I_{k+1}[c+1, C] = I_k[c+1, C] + F_1(I_k[1, c]) \tag{7}$$

$$I_{k+1}[1, c] = I_k[c+1, C] \odot \exp(F_2(I_{k+1}[c+1, C])) + F_3(I_{k+1}[c+1, C]) \tag{8}$$

$$I_{k+1} = CAT\{(I_{k+1}[1, c] + I_{k+1}[c+1, C])\} \tag{9}$$

where $\odot$ denotes the Hadamard product; $I_k[1, c]$ are the inputs to the k-th reversible layer for the first 1 to c channels; C is the

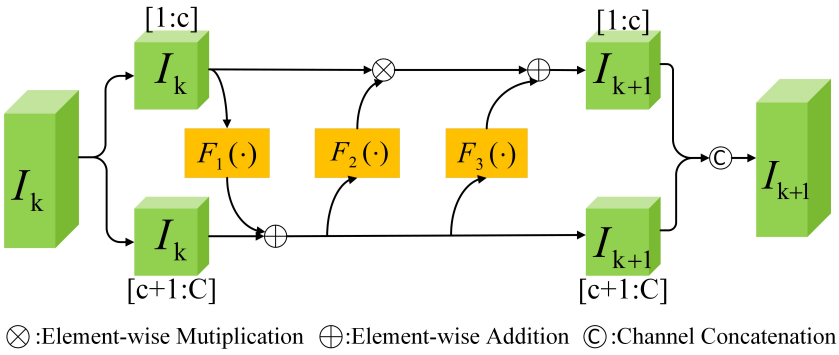

FIGURE 6. The structure of INNB.

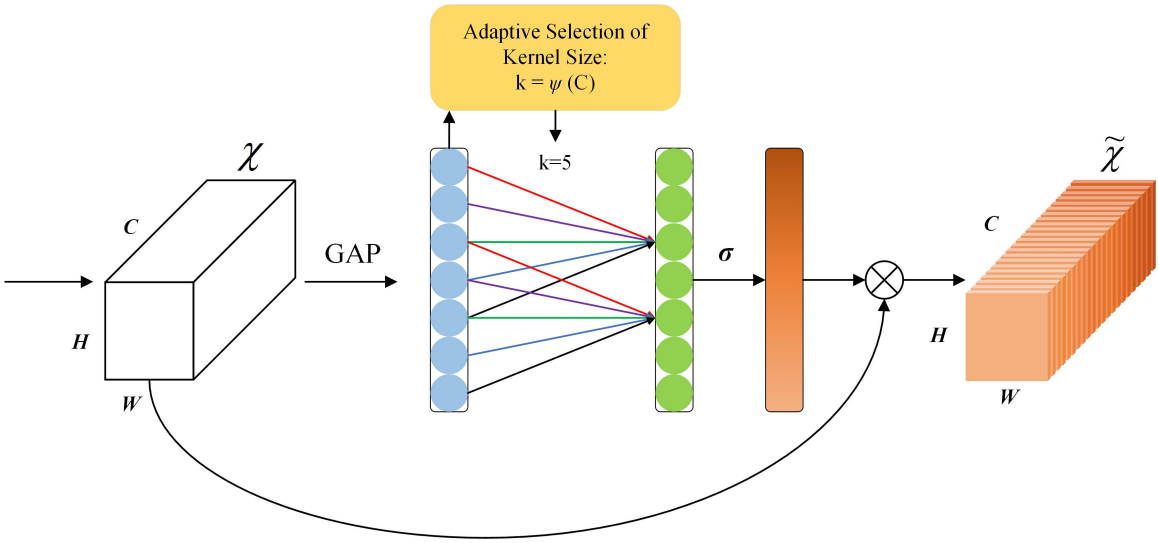

FIGURE 7. The structure of ECA.

total number of channels; $CAT(\cdot)$ is the channel concatenation operation; and $F_i(i = 1, 2, 3)$ are arbitrary mapping functions. In this paper, $F_1 = F_2 = F_3$, the mapping functions used can be represented as:

$$G(x) = ReLU(Conv_{1\times1}(x)) \tag{10}$$

$$H(x) = ReLU(Conv_{3\times3}(G(x))) \tag{11}$$

$$F(x) = ReLU(Conv_{1\times1}(H(x))) \tag{12}$$

where x denotes the input; $G(x)$ and $H(x)$ are the output of the intermediate layer; and $F(x)$ is the final output. Here, $G(x)$ represents the up sampling operation, $H(x)$ the feature extraction operation, and $F(x)$ the down sampling operation. The convolutional kernel matrices used in $Conv_{1\times1}$ (which has a kernel size of 1×1 for channel-wise transformations) and $Conv_{3\times3}$ (which has a kernel size of 3×3 for channel-wise transformations) are non-singular to ensure that the transformation is invertible. By designing invertible convolution kernels, the network can maintain computational efficiency and stability during both forward and backward propagations. Using the channel fusion property of 1×1 convolutions and the spatial feature extraction capability of 3×3 convolutions, the network effectively extracts and fuses detailed features during the up sampling and down sampling processes.

2.2.3. Feature Fusion Module

In feature fusion module, element-wise addition is used to merge high-frequency and low-frequency features of the two modalities. The fused high-frequency and low-frequency features are then concatenated and passed into the Efficient Channel Attention (ECA) module. As shown in Figure 7, the input feature tensor of size $H \times W \times C$ is processed through a Global Average Pooling (GAP) layer, which compresses the spatial dimensions (height and width) into a single value, forming a channel vector of size $1 \times 1 \times C$. A one-dimensional convolution is then applied, where the kernel size k is typically chosen adaptively to capture local relationships across channels. The output is then activated using a sigmoid function, generating weights for each channel. These weights ω are used to perform channel-wise reweighting of the original input feature map, producing an enhanced feature map of the same size $H \times W \times C$ [25].

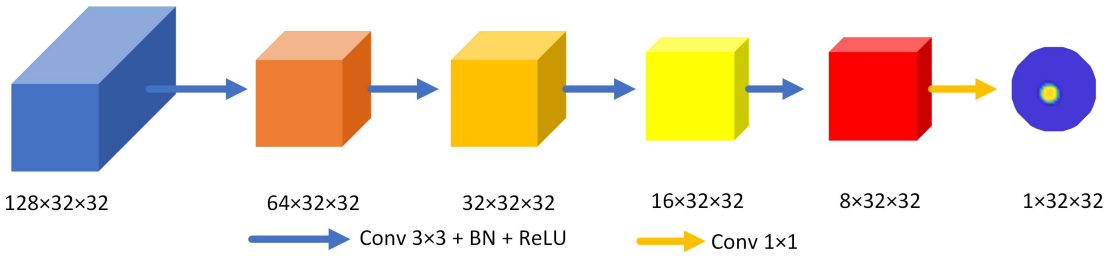

FIGURE 8. The structure of image reconstruction module.

ECA captures the inter-channel dependencies through adaptive one-dimensional convolution, generating channel attention weights and reweighting the feature maps accordingly. This process significantly enhances the complementarity between modalities and the representation of crucial features while avoiding information redundancy, thereby improving the discriminative power of the features. High-frequency and low-frequency features complement each other in terms of information content. By utilizing the attention mechanism, ECA can more effectively integrate information from these two modalities, improving the overall quality of feature representation.

2.2.4. Image Reconstruction Module

After passing through the ECA module, a feature map of size $128 \times 32 \times 32$ is obtained. The image reconstruction module is responsible for merging these feature maps to generate the final reconstructed image, as shown in Figure 8. In the image reconstruction module, the input feature maps go through four image reconstruction blocks, which integrate and reconstruct the extracted information from the feature maps, gradually restoring the feature representation of the input image. Each reconstruction block includes a 3×3 convolutional layer, a BN layer, and a ReLU activation function. Finally, a 1×1 convolutional layer is used to reduce the number of channels to 1, producing the reconstructed image.

2.3. Loss Function

Loss function plays a critical role in the training of deep learning models. The aim is to quantify the discrepancy between the model's predicted output and actual target. It steers the adjustment of model parameters, ultimately enhancing the model's performance.

Due to the complexity of the DSCTFusion-ECA architecture, when the loss function back propagates to the initial imaging layer, the gradient tends to be small. To avoid the issue of vanishing gradients, auxiliary loss functions are introduced after the initial imaging layer. These auxiliary losses provide supervision for the initial imaging layer, ensuring effective gradient propagation and mitigating the vanishing gradient problem. This ensures that the network can effectively update the parameters of the early initial imaging layers. Auxiliary loss functions not only help stabilize the network training process but also enhance the feature extraction capabilities of the intermediate layers [26].

The final total loss function of the network includes EMT initial imaging loss, UTT initial imaging loss, and loss from the final layer's output. This can be expressed as follows:

$$L = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{13}$$

$$L_{all} = \alpha \times L_{out} + \beta \times L_{emt} + \gamma \times L_{utt} \tag{14}$$

where L is the Mean Squared Error, y the true distribution of the medium, $\hat{y}$ the predicted distribution by the network, L_{all} the network's total loss, L_{out} the loss value of the final output layer, L_{emt} the loss of the EMT modality's initial imaging layer, and L_{utt} the loss of the UTT modality's initial imaging layer. The terms α, β, and γ are the weight factors that respectively influence the final imaging speed and accuracy, the EMT initial imaging speed and accuracy, and the UTT initial imaging speed and accuracy, with $\alpha + \beta + \gamma = 1$.

3. ESTABLISHMENT OF DATA SET

A data set for an 8-coil EMT simulation model and a 16-transducer UTT simulation model was established with the help of MATLAB and COMSOL Multiphysics software. In EMT simulation model, the background conductivity is 0.06 S/m, and the conductivity of the target object is 5.998×10^7 S/m. The radius of the imaging cross-section is 50 mm. The adaptive triangular mesh finite element method provided by COMSOL was used to solve the EMT forward problem, as shown in Figure 9(a). Each coil is sequentially used as the excitation coil, with the remaining coils acting as measurement coils to obtain the measurement voltage. When all coils have been

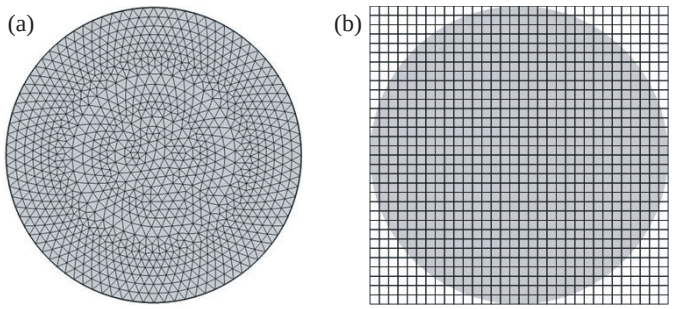

FIGURE 9. Mesh of simulation model. (a) Mesh of the forward problem. (b) Mesh of the reverse problem.

excited, a boundary measurement voltage containing 56 measurements is obtained. The imaging cross-section is divided into 32×32 pixels, retaining only the pixels within the circular imaging cross-section, resulting in 812 effective pixels, as shown in Figure 9(b). In the UTT simulation model, the background sound speed is 1450 m/s, and the sound speed of the test object is 3750 m/s. Similar to the EMT simulation model, each transducer in the UTT simulation model is sequentially used as the excitation source, with the remaining transducers acting as receivers. This process results in 240 boundary acoustic time measurements. The distribution of the test objects in the UTT simulation model is consistent with the EMT simulation model, and the imaging cross-section also contains 812 effective pixels.

The data set consists of 1–3 circular test objects with randomized positions and radii. Figure 10 illustrates some typical distributions of these test objects. The non-intersecting test objects result in 4000 simulated samples. The data set is divided into a training set (80%), a validation set (10%), and a test set (10%). The training set is utilized for network training, the validation set for adjusting training hyper parameters, and the test set for evaluating the model's generalization ability. Importantly, none of these sets overlap or contain noise.

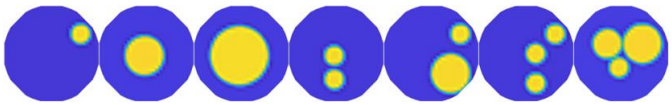

FIGURE 10. Typical sample distributions in database.

Each sample in the data set includes three vectors U_n, T_n, and R_n (where $n = 1, 2, 3, \ldots, N$), with N representing the total number of samples. U_n is obtained by normalizing the difference between the boundary measurement vectors with and without the test object in the EMT simulation model. T_n is obtained by normalizing the difference between the boundary measurement vectors with and without the test object in the UTT simulation model. R_n is the label vector of the true medium distribution, which is used as the label for the deep learning model to supervise and constrain the training process of the deep learning model.

4. SIMULATION RESULTS AND DISCUSSION

4.1. Evaluation Metrics

In order to quantitatively assess the image reconstruction quality of different samples, two commonly utilized metrics in the field of image reconstruction are employed: Relative Error (RE) and Correlation Coefficient (CC).

RE is defined as the relative error between the reconstructed distribution $\hat{y}$ by the network and the true distribution y. The equivalent formula for RE is as follows:

$$RE = \frac{\|y - \hat{y}\|_2^2}{\|y\|_2^2} \tag{15}$$

CC represents the similarity between the reconstructed distribution $\hat{y}$ and the true distribution y, defined as follows:

$$CC = \frac{\sum_{i=1}^{N} \left(\hat{y}_i - \overline{\hat{y}}\right)\left(y_i - \overline{y}\right)}{\sum_{i=1}^{N} \left(\hat{y}_i - \overline{\hat{y}}\right)^2 \sum_{i=1}^{N} \left(y_i - \overline{y}\right)^2} \tag{16}$$

where N represents the number of effective pixels.

4.2. Results of Test Set With Noise

To assess the robustness of DSCTFusion-ECA, Gaussian noise with signal-to-noise ratio (SNR) ranging from 20 to 60 dB is added to the test set. Imaging of the test set is performed under different SNR levels using a trained network for imaging with samples of varying sizes, positions, and numbers. The results are illustrated in Figure 11.

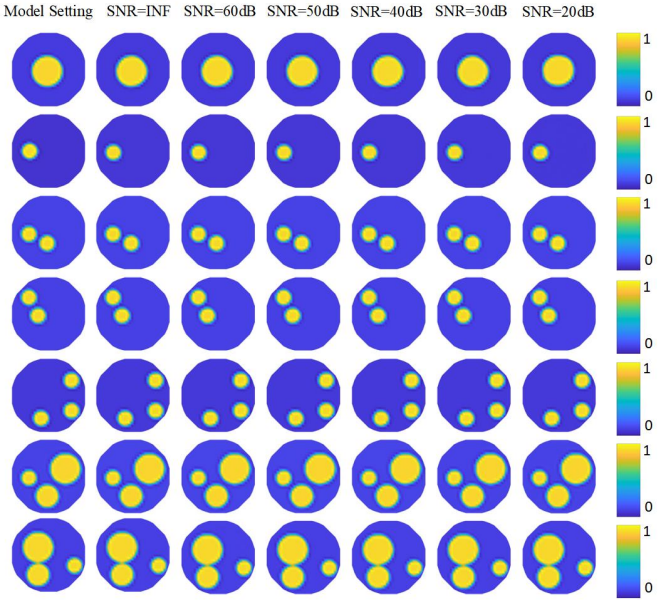

FIGURE 11. Imaging of test data at different SNR level.

The first column shows the simulated true distribution of the samples, while columns 2 to 7 display reconstructed images under different levels of noise. Despite the fluctuations in noise levels, the predicted results of the network consistently and accurately reflect the distribution of the objects being studied, demonstrating the robustness of the DSCTFusion-ECA network.

In order to quantitatively analyze the robustness of the DSCTFusion-ECA network, RE and CC as mentioned in Section 4.1 were used to assess the reconstruction results of the test data under different noise levels. As shown in Figure 12, with increasing noise levels, the average RE continues to rise while the average CC decreases. Nevertheless, the average RE remains at a relatively low level, and the average CC maintains a high level. This indicates that the network's predicted results demonstrate minimal relative error and high correlation with the true distribution, consistent with the reconstruction results shown in Figure 11. Through quantitative analysis using RE and CC, it is further confirmed that the DSCTFusion-ECA network exhibits robustness.

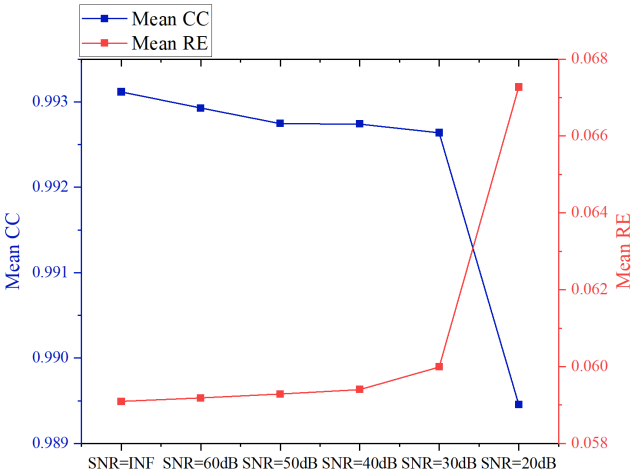

FIGURE 12. Average RE and CC of the test data with different SNR level.

4.3. Results of a New Test Set

To evaluate the generalization capability of DSCTFusion-ECA, simulations are performed using a new test set composed of completely new samples that are distinct from those in the training set and test set. These new samples consist of elliptical shapes and four circular samples, which assess DSCTFusion-ECA's ability to generalize across different shapes and quantities. Partial reconstruction results are depicted in Figure 13. From the elliptical samples, it is evident that despite the absence of elliptical data in the data set, DSCTFusion-ECA effectively reconstructs the positions, sizes, and shapes of the objects being studied. However, due to the lack of edge information for ellipses in the dataset, DSCTFusion-ECA is unable to accurately reconstruct the edges of elliptical samples. As for the four circular samples, DSCTFusion-ECA demonstrates effective reconstruction of their positions, sizes, and shapes.

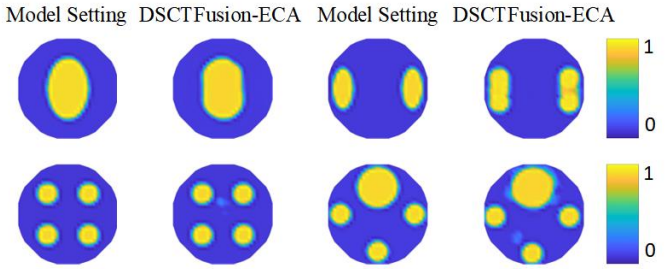

FIGURE 13. Imaging results of new test set.

DSCTFusion-ECA effectively extracts and integrates multimodal information, but it is limited in its ability to reconstruct a wider variety of object shapes due to constraints within the data set. While DSCTFusion-ECA demonstrates strong performance in generalization, its capacity to reconstruct the edges of more diverse object shapes is hindered by limitations in the data set.

4.4. Results of Different Image Reconstruction Algorithms

U-Net, a deep convolutional neural network, is widely employed in various fields due to its exceptional performance. By utilizing an encoder-decoder structure and skip connections, U-Net effectively incorporates contextual information in images for precise image segmentation. To assess the reconstruction performance and dual-modality fusion effect of DSCTFusion-ECA, the reconstruction results of U-Net and DSCTFusion-ECA were compared under EMT modality, UTT modality, and dual-modality fusion. In U-Net, an initial imaging module similar to DSCTFusion-ECA was incorporated. Under the dual-modality setting in U-Net, features were extracted from the initial imaging of the two modalities through down sampling and were fused during the up sampling reconstruction process via skip connections. When up sampling, the two obtained down sampling results are concatenated and then up sampling. Additionally, each layer of up sampling uses skip connections to concatenate the results of the corresponding layers from the two down sampling stages with the current layer's results.

The reconstruction results of different algorithms are shown in Figure 14. The first column in the figure depicts the true distribution map of the medium, while columns 2–6 show the results of U-Net under EMT modality, DSCTFusion-ECA under EMT modality, U-Net under UTT modality, DSCTFusion-ECA under UTT modality, dual-modality fusion of U-Net, and dual-modality fusion of DSCTFusion-ECA, respectively. The CC and RE of different reconstruction algorithms are illustrated in Figures 15 and 16. Across six medium distributions, dual-modality DSCTFusion-ECA consistently achieves the highest CC and lowest RE.

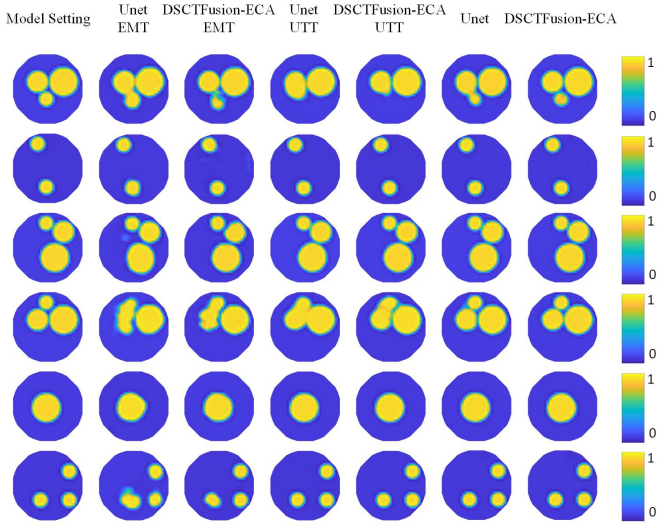

FIGURE 14. Imaging results of different algorithms.

For simpler medium distributions, such as one or two target objects, all six methods effectively reconstruct the distribution of the target objects. However, when there are many target objects, and the target objects are close to each other, the other five algorithms may produce unclear reconstructed contours. However, DSCTFusion-ECA can well reconstruct the position size and edge contour of the target objects, which can also be

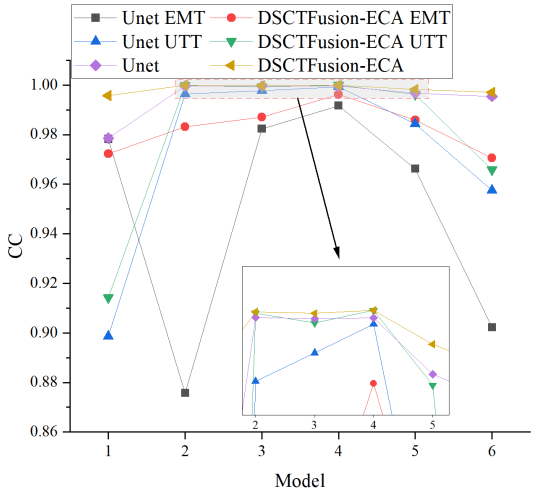

FIGURE 15. CC of different algorithms.

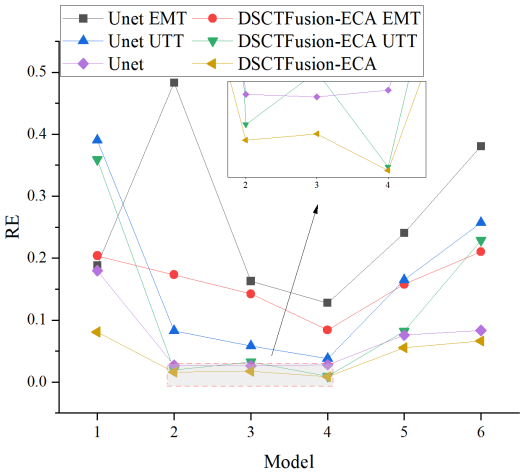

FIGURE 16. RE of different algorithms.

seen from RE and CC. Ultimately, DSCTFusion-ECA excels due to its ability to extract both global low-frequency features and local high-frequency features, effectively integrating them through channel attention mechanisms, resulting in reconstructions closely resembling the true distribution.

Moreover, the imaging results of UTT are generally better than those of EMT. This is primarily because UTT has 240 measurement data, while EMT only has 56, providing more known information, which leads to better imaging. Additionally, the physical model of UTT is simpler than that of EMT; UTT can be approximated as a hard field, allowing the network to better capture the mapping relationship to the results.

5. CONCLUSIONS

A supervised deep learning network tailored for dual-modality fusion imaging in EMT and UTT modalities named DSCTFusion-ECA is proposed in this paper. The network architecture comprises an initial imaging module, a feature extraction module, a feature fusion module, and an image reconstruction module. Initially, fully connected layers are used for the initial image reconstruction of the boundary measurement information from the EMT modality and UTT modality, respectively. The feature extraction module employs DSB to extract shallow features across modalities, utilizing dense connections to retain original information effectively. This enhances information flow and feature reuse, bolstering the network's representational and learning capabilities. Subsequently, TB and CB modules extract modality-specific low-frequency global and high-frequency local features. TB employs self-attention mechanisms to capture global dependencies dynamically, facilitating better understanding of long-range relationships across data. In CB, the use of INNB with affine coupling layers ensures lossless feature extraction, preserving input information effectively. TB and CB outputs undergo addition, followed by ECA to assign varying attention scores across channels, enabling the network to discern feature importance effectively. Finally, a fully convolutional network reconstructs images using these attention-scored features. The results demonstrate that DSCTFusion-ECA has strong robustness and generalization capabilities, and superior performance over U-Net.

In conclusion, DSCTFusion-ECA has strong feasibility in EMT and UTT dual-modality fusion imaging. However, due to the limitations of the data set, reconstructing the edges of non-circular objects is quite challenging. Enriching the edge information in the data set is the direction for future work.

REFERENCES

[1] Rahim, R. A., M. H. F. Rahiman, K. S. Chan, and S. W. Nawawi, "Non-invasive imaging of liquid/gas flow using ultrasonic transmission-mode tomography," *Sensors and Actuators A: Physical*, Vol. 135, No. 2, 337–345, 2007.

[2] Peyton, A. J., M. S. Beck, A. R. Borges, J. E. D. Oliveira, G. M. Lyon, Z. Z. Yu, M. W. Brown, and J. Ferrerra, "Development of electromagnetic tomography (EMT) for industrial applications. Part 1: Sensor design and instrumentation," in *1st World Congress on Industrial Process Tomography*, Vol. 3, 306–312, 1999.

[3] Hayashi, S., S. Yoshimoto, and A. Yamamoto, "Noncontact 2D temperature imaging of metallic foils using electromagnetic tomography," *IEEE Sensors Journal*, Vol. 23, No. 16, 17 942–17 950, 2023.

[4] Rahiman, M. H. F., R. A. Rahim, M. H. F. Rahiman, and M. Tajjudin, "Ultrasonic transmission-mode tomography imaging for liquid/gas two-phase flow," *IEEE Sensors Journal*, Vol. 6, No. 6, 1706–1715, 2006.

[5] Zhang, R., H. Fang, Q. Zhang, J. Wang, D. Zhang, J. Cheng, and W. Yin, "In situ damage monitoring of CFRPS by electromagnetic tomography with the compatible multitemplate supervised descent method," *IEEE Transactions on Instrumentation and Measurement*, Vol. 72, 1–12, 2023.

[6] Zhang, W., C. Tan, and F. Dong, "Dual-modality tomography by ERT and UTT projection sorting algorithm," *IEEE Sensors Journal*, Vol. 20, No. 10, 5415–5423, 2020.

[7] Liang, G., V. Kolehmainen, M. Vauhkonen, and F. Dong, "Structural similarity driven joint reconstruction of conductivity and

sound speed in EIT/UTT dual-modality tomography," *Inverse Problems*, Vol. 39, No. 10, 105010, 2023.

[8] Steiner, G., M. Soleimani, and D. Watzenig, "A bio-electromechanical imaging technique with combined electrical impedance and ultrasound tomography," *Physiological Measurement*, Vol. 29, No. 6, S63, 2008.

[9] Jiang, Y., M. Soleimani, and B. Wang, "Contactless electrical impedance and ultrasonic tomography: Correlation, comparison and complementarity study," *Measurement Science and Technology*, Vol. 30, No. 11, 114001, 2019.

[10] Xu, C., F. Dong, and Z. Zhang, "Dual-modality data acquisition system based on CPCI industrial computer," in *2012 IEEE International Conference on Imaging Systems and Techniques Proceedings*, 567–572, Manchester, UK, Jul. 2012.

[11] Liang, G., S. Ren, and F. Dong, "Ultrasound guided electrical impedance tomography for 2D free-interface reconstruction," *Measurement Science and Technology*, Vol. 28, No. 7, 074003, 2017.

[12] Zhang, R., Q. Wang, H. Wang, M. Zhang, and H. Li, "Data fusion in dual-mode tomography for imaging oil-gas two-phase flow," *Flow Measurement and Instrumentation*, Vol. 37, 1–11, 2014.

[13] Pusppanathan, J., R. A. Rahim, F. A. Phang, E. J. Mohamad, N. M. N. Ayob, M. H. F. Rahiman, and C. K. Seong, "Single-plane dual-modality tomography for multiphase flow imaging by integrating electrical capacitance and ultrasonic sensors," *IEEE Sensors Journal*, Vol. 17, No. 19, 6368–6377, 2017.

[14] Yue, S., T. Wu, J. Pan, and H. Wang, "Fuzzy clustering based ET image fusion," *Information Fusion*, Vol. 14, No. 4, 487–497, 2013.

[15] Malikov, A. K. U., M. F. F. Cuenca, B. Kim, Y. Cho, and Y. H. Kim, "Ultrasonic tomography imaging enhancement approach based on deep convolutional neural networks," *Journal of Visualization*, Vol. 26, No. 5, 1067–1083, 2023.

[16] Shi, Q., T. Zhou, Y. Li, C. Liu, and D. Ta, "Deep learning for TOF extraction in bone ultrasound tomography," *IEEE Transactions on Computational Imaging*, Vol. 8, 1063–1073, 2022.

[17] Xiao, J., Z. Liu, P. Zhao, Y. Li, and J. Huo, "Deep learning image reconstruction simulation for electromagnetic tomography," *IEEE Sensors Journal*, Vol. 18, No. 8, 3290–3298, 2018.

[18] Zhang, W., Z. Zhu, and Y. Geng, "Simultaneous conductivity and permeability reconstructions for electromagnetic tomography using deep learning," *IEEE Transactions on Instrumentation and Measurement*, Vol. 72, 1–11, 2023.

[19] Guo, Q., J. Ye, C. Wang, and X. Liang, "An ill-conditioned optimization method and relaxation strategy of landweber for EMT system based on TMR," *IEEE Transactions on Instrumentation and Measurement*, Vol. 70, 1–9, 2020.

[20] Gan, W. S., *Acoustical Imaging: Techniques and Applications for Engineers*, John Wiley & Sons, 2012.

[21] Huang, G., Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708, 2017.

[22] Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017.

[23] Dinh, L., J. Sohl-Dickstein, and S. Bengio, "Density estimation using Real NVP," *ArXiv Preprint ArXiv:1605.08803*, 2016.

[24] Zhou, M., X. Fu, J. Huang, F. Zhao, A. Liu, and R. Wang, "Effective pan-sharpening with transformer and invertible neural network," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 60, 1–14, 2021.

[25] Wang, Q., B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11 534–11 542, 2020.

[26] Lee, C.-Y., S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artificial Intelligence and Statistics*, 562–570, 2015.