M2Trans: Multi-Modal Regularized Coarse-to-Fine Transformer for Ultrasound Image Super-Resolution

Zhangkai Ni, *Member, IEEE*, Runyu Xiao, Wenhan Yang, *Member, IEEE*,
Hanli Wang, *Senior Member, IEEE*, Zhihua Wang, *Member, IEEE*, Lihua Xiang, and Liping Sun

***Abstract*—Ultrasound image super-resolution (SR) aims to transform low-resolution images into high-resolution ones, thereby restoring intricate details crucial for improved diagnostic accuracy. However, prevailing methods relying solely on image modality guidance and pixel-wise loss functions struggle to capture the distinct characteristics of medical images, such as unique texture patterns and specific colors harboring critical diagnostic information. To overcome these challenges, this paper introduces the Multi-Modal Regularized Coarse-to-fine Transformer (M2Trans) for Ultrasound Image SR. By integrating the text modality, we establish joint image-text guidance during training, leveraging the medical CLIP model to incorporate richer priors from text descriptions into the SR optimization process, enhancing detail, structure, and semantic recovery. Furthermore, we propose a novel coarse-to-fine transformer comprising multiple branches infused with self-attention and frequency transforms to efficiently capture signal dependencies across different scales. Extensive experimental results demonstrate significant improvements over state-of-the-art methods on benchmark datasets, including CCA-US, US-CASE, and our newly created dataset MMUS1K, with a minimum improvement of 0.17dB, 0.30dB, and 0.28dB in terms of PSNR.**

***Index Terms*—Ultrasound image super-resolution, multi-modal, CLIP, transformer.**

I. Introduction

Medical imaging encompasses various modalities, including magnetic resonance imaging (MRI), X-ray, computed tomography (CT), and ultrasound imaging, each serving vital roles in clinical practice. Among these modalities, ultrasound imaging stands out as a critical tool in modern medicine due to its non-invasiveness, radiation-free nature, high repeatability, and cost-effectiveness [1]. Despite these advantages, ultrasound images often suffer from low resolution, attributed to challenges such as equipment limitations and acoustic wave diffraction. Consequently, there arises a pressing need to enhance ultrasound image resolution through super-resolution techniques, essential for facilitating downstream tasks and enabling accurate diagnoses by medical professionals.

Single-image super-resolution (SISR) is a technique aimed at enhancing the resolution of a single low-resolution image by reconstructing its high-resolution counterpart. Learning-based methods that use data-driven paradigms to learn complex super-resolution mappings, including Convolutional Neural Networks (CNN), Generative Adversarial Networks (GAN) [2], [3], and Transformers [4], have been widely used to address the challenges faced by conventional SISR methods. CNN-based methods, often using mean square error (MSE) as a loss function, have shown effectiveness in both natural and medical image super-resolution. For instance, Lim et al. [5] developed an enhanced deep SR network for nature images, and Umehara et al. [6] applied CNNs to chest CT super-resolution, surpassing traditional methods and demonstrating the effectiveness of CNN for low-level medical image tasks. However, CNN-based methods often require deeper networks to capture larger receptive fields, leading to blurred images due to their focus on low-frequency information and struggle with high-frequency details. To address

Received 7 December 2023; revised 23 May 2024 and 16 June 2024; accepted 29 August 2024. Date of publication 3 September 2024; date of current version 7 May 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62201387 and Grant 82202153, in part by the Shanghai Pujiang Program under Grant 22PJ1413300, and in part by the Fundamental Research Funds for the Central Universities. *(Corresponding authors: Lihua Xiang; Hanli Wang.)*

Zhangkai Ni, Runyu Xiao, and Hanli Wang are with the Department of Computer Science and Technology and the Key Laboratory of Embedded System and Service Computing (Ministry of Education), Tongji University, Shanghai 200092, China (e-mail: zkni@tongji.edu.cn; runyu_xiao@tongji.edu.cn; hanliwang@tongji.edu.cn).

Wenhan Yang is with Pengcheng Laboratory, Shenzhen 518066, China (e-mail: yangwh@pcl.ac.cn).

Zhihua Wang is with the Guangdong Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, Shenzhen 518115, China (e-mail: zhihua.wang@my.cityu.edu.hk).

Lihua Xiang and Liping Sun are with the Department of Medical Ultrasound, Center of Tumor Minimally Invasive Treatment, Shanghai Tenth People's Hospital, Ultrasound Research and Education Institute, Clinical Research Center for Interventional Medicine, School of Medicine, Tongji University, Shanghai 200070, China, and also with the Shanghai Engineering Research Center of Ultrasound Diagnosis and Treatment, Shanghai 200072, China (e-mail: xianglihua1121@163.com; sunliping_s@126.com).

Our code and dataset will be available at: https://github.com/eezkni/M2Trans

Digital Object Identifier 10.1109/JBHI.2024.3454068

this limitation, Ledig et al. [2] introduced the SRGAN, which uses adversarial learning to generate photo-realistic images. Choi et al. [7] further adapted SRGAN for B-mode ultrasound images, reducing aliasing artifacts and improving image quality. Nevertheless, the learning process of the GAN-based models [8] is inherently unstable and may easily produce artifacts that are detrimental to visual perception. Transformer-based models have gained popularity for their ability to capture long-term dependencies through self-attention mechanisms, particularly in low-level vision tasks like super-resolution. However, these models rely solely on image modalities during learning process, which is insufficient to produce semantic and meaningful information.

In recent years, significant advancements have been made in large foundation models designed for natural language processing (e.g., GPT [9] and BERT [10]), image-text semantic alignment (e.g., CLIP [11]), segmentation (e.g., SAM [12]), and more. These models, benefiting from rich prior knowledge, diverse data sources, and extensive model parameters, have shown outstanding performance across various high-level vision tasks. However, in the field of low-level vision, particularly in medical image SR, the exploration of effectively leveraging large foundation models to extract valuable information as training constraints remains relatively unexplored, presenting a promising avenue for further research.

Based on the considerations above, we propose leveraging prior knowledge extracted from the text modality via a pre-trained CLIP model to establish joint image-text modality guidance for training an ultrasound image SR model. Specifically, we incorporate a domain-specific medical CLIP model, *i.e.* MedCLIP [13], into the loss to ensure our model effectively identifies critical semantic and quality-related information about ultrasound images. Moreover, we introduce a novel coarse-to-fine transformer aimed at preserving high-frequency details and capturing the intrinsic signal dependency across various scales in medical images. This transformer incorporates multiple branches infused with self-attention and frequency transformations, facilitating efficient information capture within and among different scales. In summary, our main contributions are as follows:

- We propose a joint image-text modality ultrasound image SR model that leverages the rich prior knowledge of MedCLIP to guide the training of the proposed SR model. To the best of our knowledge, this study is the first exploration of utilizing multi-modal prior knowledge for ultrasound image SR.
- We design a coarse-to-fine transformer with multiple branches infused with self-attention and frequency transformations. It efficiently captures information within and among different scales, enabling the SR model to learn the intrinsic signal dependencies across various scales from the joint image-text modality guidance.
- We create the first image-text paired SR dataset to facilitate training and evaluation in this task. It contains 1023 annotated ultrasound images from multiple organs, including the bladder, gallbladder, thyroid, and more. Experimental results on existing benchmark datasets and our created dataset demonstrate the superior performance of our proposed SR model for ultrasound images.

II. Related Works

A. Single Image Super-Resolution

Image super-resolution (SR) aims to produce images with sharp, clear edges and texture details while mitigating visual artifacts from low-resolution (LR) sources. Dong et al. introduced the first CNN-based approach, SRCNN [14], pioneering the application of CNNs to SR tasks. This groundbreaking work spurred the development of numerous CNN-based methods [7]. Subsequent approaches deepened the network architecture to enhance texture details and improve SR performance [15]. However, despite surpassing traditional methods by directly mapping LR to SR with MSE supervision, these CNN-based methods still suffered from a limited receptive field, often resulting in image blur.

To address the limitations of CNN-based methods in generating image details, the development of GAN-based SR methods has gained prominence. Ledig et al. introduced SRGAN [2], employing GAN to produce SR images with rich details and textures. Subsequently, GAN-based SR networks have been continually developed [1], incorporating various improvements such as adding additional discriminators in the feature domain to mitigate meaningless artifacts [16] and designing losses more consistent with the human visual system [17]. Moreover, the application of transformers in computer vision has gained traction in recent years. With the blessing of high-performance attention mechanism [4], IPT [18] and SwinIR [19] have successfully been applied in low-level vision fields such as denoising, deraining and SR. In general, transformer-based SR methods dominate the SR field, however, the substantial computational cost of transformer-based methods remains unsolved.

B. Medical Image Super-Resolution

High-resolution medical images are valuable for medical diagnosis and downstream tasks, which has prompted the development of SR methods tailored to medical images. Pham et al. [20] first modified SRCNN to reconstruct brain MR images and later Qiu et al. [21] introduced a progressive U-Net residual network for CT images. However, these CNN-based methods are often insufficient to generate finer texture details, especially when the upscaling factor is large. Similarly to developments in natural image SR, GAN-based techniques [22] have been widely used in medical image SR to reconstruct clear details. Wu et al. [23] successfully applied GAN to ultrasound images, while almost at the same time, Bell et al. [24] extended the method to unsupervised field and achieved superior results. Unlike CNN-based methods, GAN-based approaches prioritize high-frequency information and effectively reconstruct texture details, resulting in clearer images. However, these methods tend to produce pseudo structures and textures that are very fatal for medical diagnosis.

Benefiting from the ability to capture long-term dependencies, transformer-based models have also found extensive applications in the field of medical imaging. Liu et al. [25] utilized residual learning with a memory mechanism to store and update details by estimating blur kernels to obtain more fine-grained information for ultrasound image blind SR. Puttagunta et al. [26] extended the use of Swin transformer [27] to chest X-ray and skin lesion images, resulting in improved PSNR and SSIM indicators compared to other methods. Recently, Georgescu et al. [28] introduced a multi-modal multi-head convolutional attention mechanism for medical image SR, using multi-modal low-resolution inputs to achieve significant performance improvements across various datasets. Transformer-based methods have become mainstream in medical image SR, but transformer module typically demands high computational cost and memory consumption. In this work, we aim to design an efficient transformer-based medical image SR model optimized with multi-modal regularization.

C. Contrastive Language-Image Pre-Training Models

By computing the cosine similarity of images and text features, the Contrastive Language-Image Pre-training Model (CLIP) [11] understands and bridges the gap between two modalities. CLIP demonstrates remarkable performance across various downstream tasks such as object detection [29], video text retrieval [30], Embodied AI [31] and etc. Furthermore, pre-trained CLIP models have found applications in the medical imaging domain. Lin et al. [32] expanded CLIP to medical image understanding by scaling data volume to 1.6 million image-caption pairs. Meanwhile, Liu et al. [33] proposed a CLIP-driven universal model to address the limitations of previous models that focused on segmenting specific organs or tumors while overlooking the semantics of anatomical structures. Besides, Wang et al. [13] developed MedCLIP with multi-modal contrastive learning to improve the capabilities of medical image analysis.

The extensive prior knowledge of CLIP has proven invaluable for various tasks, such as matching loss [34] guiding conditional NeRF, consistency, and similarity losses [35] constraining essence transfer. In addition to these high-level tasks, priors of CLIP can also be effectively applied to various image processing tasks such as image enhancement [36], denoising [37], 3D shape generation [38] and etc. Although CLIP has seen extensive use in a wide range of tasks, it has not yet been explored for ultrasound image super-resolution tasks. This study introduces a semantic loss based on CLIP to leverage its prior knowledge and utilizes it to guide the generation of image super-resolution. This approach is motivated by the observation that CLIP can effectively assess both the quality and abstract perceptions of images [39].

III. Multi-Modal Regularized Coarse-to-Fine Transformer (M2Trans)

In this section, we first present the motivation and overall framework of the proposed M2Trans in Section III-A. Next, we delve into the details of the proposed coarse-to-fine transformer module (CFTM) and the image-text semantic loss based on CLIP in Section III-B and III-C, respectively.

A. Overview

1) Motivation: Given a low-resolution ultrasound image I_{lr}, we aim to address the following two issues:

- *Limited modality leading to insufficient recovery:* The mapping from LR to HR is ill-posed, but previous methods that rely solely on image modality fail to effectively capture and recover the semantic information in ultrasound images, which have unique texture patterns and colors crucial for diagnosis.
- *Single-scale leading to under-detailed recovery:* Most methods with cascaded structures often use only single-scale information to reconstruct SR images. This can result in under-detailed ultrasound SR images due to neglecting dependencies across different scales.

To address these issues, we propose M2Trans that progressive refines features via a coarse-to-fine transformer regularized by a joint **image-text semantic loss**. This enables us to fully exploit cues from different scales and modalities to capture **multi-granularity dependencies** and reveal more meaningful semantic information about ultrasound images.

2) Framework: In this section, we will briefly introduce the overall framework of the proposed M2Trans in Fig. 1. Given a low-resolution (LR) image $\mathrm{I}_{lr} \in \mathbb{R}^{3\times H\times W}$, the objective of M2Trans is to obtain the super-resolution (SR) image $\mathrm{I}_{sr} \in \mathbb{R}^{3\times sH\times sW}$, where H and W denote the height and width of the LR image, and s represents the scaling factor. Specifically, the LR image is first projected into the feature space through a shallow feature extraction module $\mathrm{M}_{SF}(\cdot)$, which consists of a convolution layer and can be expressed as:

$$\mathrm{F}_{sf} = \mathrm{M}_{SF}(\mathrm{I}_{lr}). \tag{1}$$

The shallow feature F_{sf} of the LR image is progressively refined by the proposed coarse-to-fine transformer modules to reconstruct the high-resolution (HR) feature F_{df}, which can be formulated as:

$$\mathrm{F}_{df} = \mathrm{M}_{DF}(\mathrm{F}_{sf}), \tag{2}$$

where $\mathrm{M}_{DF}(\cdot)$ consists of stacked CFTMs. The t-th CFTM takes F_t as input and outputs the refined feature F_{t+1}, which can be expressed as:

$$\mathrm{F}_{t+1} = \mathrm{CFTM}_t(\mathrm{F}_t), \quad t = 0, 1, \ldots, n, \tag{3}$$

where F_t is the output of the previous CFTM, F_0 is initialized as the shallow feature F_{sf}, and n is the number of CFTMs.

Finally, the learned deep feature F_{df} and shallow feature F_{sf} are combined as inputs to the HR image reconstruction module M_{RC}, which reconstructs the final SR image. The overall process can be represented as follows:

$$\mathrm{I}_{sr} = \mathrm{M}_{RC}(\mathrm{F}_{df} + \mathrm{F}_{sf}), \tag{4}$$

where M_{RC} consists of a series of components, including a convolutional layer, a pixelshuffle layer, a GeLU layer, and a convolutional layer. This network can extract hierarchical

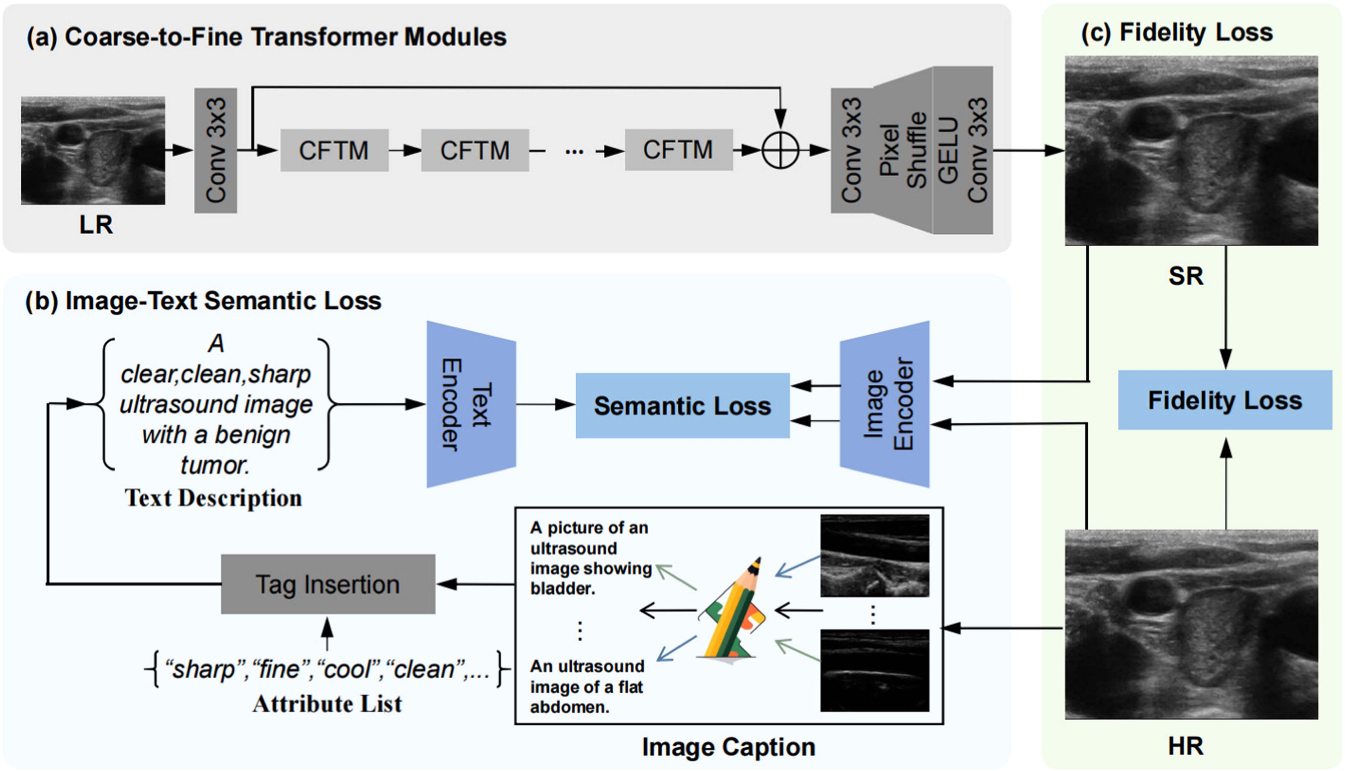

(a) Coarse-to-Fine Transformer Modules
LR
Conv 3x3
CFTM
CFTM
CFTM
Conv 3x3
Pixel Shuffle
GELU
Conv 3x3
(b) Image-Text Semantic Loss
A clear,clean,sharp ultrasound image with a benign tumor.
Text Description
Text Encoder
Semantic Loss
Image Encoder
Tag Insertion
{"sharp","fine","cool","clean",...}
Attribute List
A picture of an ultrasound image showing bladder.
An ultrasound image of a flat abdomen.
Image Caption
(c) Fidelity Loss
SR
Fidelity Loss
HR

Fig. 1. The framework of the proposed M2Trans for ultrasound image SR consists of three main components: (a) The LR image is initially projected into the feature space through a convolution operation, and then undergoes progressive refinement by a series of **coarse-to-fine** transformer modules (**CFTM**), and is finally reconstructed into an SR image. (b) The **image-text semantic loss** encourages M2Trans to produce high-quality SR images with improved understanding and meaning. (c) The **fidelity loss** ensures the reconstructed image maintains a high level of visual fidelity.

features from local to global by the coarse-to-fine mechanism, leading to high-quality reconstruction results with finer details and clearer structures.

B. Coarse-to-Fine Transformer Module

As illustrated in Fig. 2, we introduce a coarse-to-fine transformer module (CFTM) comprising multiple branches infused with self-attention and frequency transforms, effectively capturing information within and among different scales. Specifically, given the input feature $\mathrm{F}_t \in \mathbb{R}^{C\times H\times W}$, the CFTM initially splits it into four groups based on the channel dimension to derive the inputs of multiple branches:

$$[\mathrm{X}^0, \mathrm{X}^1, \mathrm{X}^2, \mathrm{X}^3] = \theta(\mathrm{F}_t), \tag{5}$$

where $\theta(\cdot)$ represents the split operation by channel dimension and $\mathrm{X}^i \in \mathbb{R}^{\frac{C}{4}\times H\times W}$ is the features after splitting, where i ranges from 0 to 3. Inspired by the lossless reversible property of discrete wavelet transform (DWT [40]), we construct a coarse-to-fine self-attention mechanism that combines multiple reversible down/up-sampling with self-attention. For feature X^i, we use DWT to perform i times of reversible down-sampling, followed by self-attention, and finally use the corresponding i times inverse DWT (IDWT) to upsample it to the same dimensions as X^i. This module aggregates information from diverse scales, enabling the modeling of global, regional, and local features, thereby enhancing the effectiveness of image reconstruction. It can be formulated as:

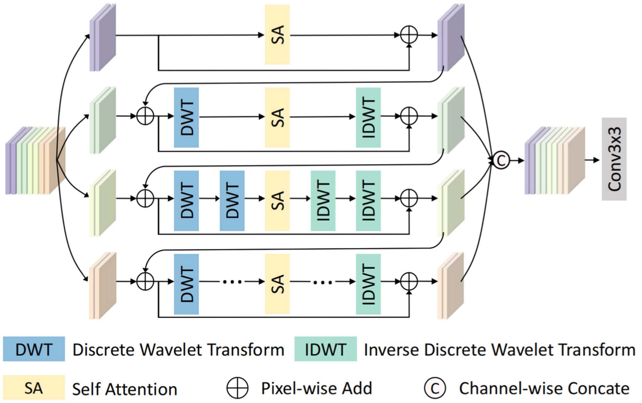


Fig. 2. Detailed structure of the proposed CFTM, including three parts: (a) The input feature is grouped according to the channel dimensions; (b) Different groups perform self-attention (SA) operations at different scales by wavelet transforms, and residual operations are used in adjacent groups to enhance the interaction between groups.; (c) The processed features of each group are concatenated and fused through a convolutional layer to obtain the final output feature.

$$\mathrm{H}^i = \phi_j(\mathrm{X}^i), \quad j = 0, 1, 2, \ldots, \tag{6}$$

where $\phi_j(\cdot)$ represents a series of DWT modules and j is the number of the DWT module. Note that we do not conduct a triple downsample of the 3-th branch, as this would result in an extensive increase in module parameters but only a slight improvement in performance. Then the downsampled features are fed into the SA module to extract deep features:

$$\bar{\mathrm{H}}^i = \mathrm{SA}(\mathrm{H}^i), \tag{7}$$

where $\mathrm{SA}(\cdot)$ denotes the self-attention module. Afterward, the features which are generated by the SA module are transformed back to the original dimensions using the corresponding number of IDWT modules:

$$\hat{\mathrm{X}}^i = \mathrm{X}^i + \psi_j(\bar{\mathrm{H}}^i), \quad j = 0, 1, 2, \ldots, \tag{8}$$

where $\psi_j(\cdot)$ means stacked IDWT modules and j is the corresponding times. In order to speed up the convergence of the model, the residual operation is adopted for each branch.

Due to the benefits of wavelet theory, this approach enables the preservation of original information for self-attention while achieving multi-scale receptive fields. Additionally, we compute self-attention in a cascaded manner to facilitate the exchange of information between different branches. This involves adding the output of each branch to the subsequent branch to refine the feature representations progressively:

$$\bar{\mathrm{X}}^{i+1} = \mathrm{X}^{i+1} + \hat{\mathrm{X}}^i, \quad i = 0, 1, 2, \tag{9}$$

where $\bar{\mathrm{X}}^{i+1}$ is updated as the sum of the i-th branch output $\hat{\mathrm{X}}^i$ and the $(i+1)$-th branch input X^{i+1}. The updated $\bar{\mathrm{X}}^{i+1}$ serves as the new input feature for the $(i+1)$-th branch, substituting X^i in (6) and (8). Cascading branches can enhance interactions between feature groups which leading to coarse-to-fine feature aggregation.

Finally, the concatenated features are fed into the feedforward network that is achieved by a 3×3 convolution operation to fuse features, which can be represented as:

$$\mathrm{F}_{t+1} = \mathrm{Conv}(\Theta[\hat{\mathrm{X}}^0, \hat{\mathrm{X}}^1, \hat{\mathrm{X}}^2, \hat{\mathrm{X}}^3]), \tag{10}$$

where $\mathrm{Conv}(\cdot)$ represents convolution operation and $\Theta(\cdot)$ denotes the concatenate operation.

C. Image-Text Semantic Loss

Inspired by [39], the CLIP model demonstrates an exceptional ability to build semantic relationships between texts and visual entities, reflecting both the quality and abstract understanding of an image. Given that the medical image SR task aims to assist doctors in making more accurate diagnoses from images, the loss we designed is more consistent with the doctors' judgment at the semantic level, aiming for a clean and clear ultrasound image. Thus, we optimize the proposed M2Trans under the multi-modality regularization framework based on the CLIP model. Compared with previous methods, our proposed image-text semantic loss encourages M2Trans to produce high-quality SR ultrasound images in terms of style, understanding, and meaning, rather than merely optimizing fidelity. Since our task is to super-resolve ultrasound images, our proposed image-text semantic loss is based on the pre-trained MedCLIP [13], a variant of CLIP trained on a large dataset of medical images.

The proposed image-text semantic loss function is calculated based on the image-text triple $(\mathrm{I}_{hr}, \mathrm{I}_{sr}, \mathrm{T}_p)$, where the dimension of I_{hr} and I_{sr} are both $\mathbb{R}^{3\times W\times H}$, and T_p is the pre-annotated text description associated with I_{hr}. To be specific, we use the image encoder and text encoder of MedCLIP to obtain the features of the image-text triple respectively:

$$\begin{aligned} \mathrm{v}_{hr} &= \mathrm{E}_i(\mathrm{I}_{hr}) \\ \mathrm{v}_{sr} &= \mathrm{E}_i(\mathrm{I}_{sr}) \\ \mathrm{v}_p &= \mathrm{E}_t(\mathrm{T}_p), \end{aligned} \tag{11}$$

where $\mathrm{E}_i(\cdot)$ represents the image encoder and E_t denotes the text encoder. v_{hr}, v_{sr}, and v_p are the HR image feature, SR image feature, and text description feature, respectively. Similarity scores between the HR image feature v_{hr} and the text feature v_p, as well as between the SR image feature v_{sr} and the text feature v_p, are obtained by calculating the cosine similarity of $(\mathrm{v}_{hr}, \mathrm{v}_p)$ and $(\mathrm{v}_{sr}, \mathrm{v}_p)$ as follows:

$$\begin{aligned} \mathrm{s}_{hr} &= \cos(\mathrm{v}_{hr}, \mathrm{v}_p) = \frac{\mathrm{v}_{hr}\cdot \mathrm{v}_p}{||\mathrm{v}_{hr}||\,||\mathrm{v}_p||} \\ \mathrm{s}_{sr} &= \cos(\mathrm{v}_{sr}, \mathrm{v}_p) = \frac{\mathrm{v}_{sr}\cdot \mathrm{v}_p}{||\mathrm{v}_{sr}||\,||\mathrm{v}_p||}, \end{aligned} \tag{12}$$

where $\cos(\cdot)$ represents the calculation of cosine similarity. The image-text semantic loss is defined as the $\mathcal{L}_1$ norm between s_{hr} and s_{sr}, expressed as:

$$\mathcal{L}_{clip} = ||\mathrm{s}_{sr} - \mathrm{s}_{hr}||. \tag{13}$$

Different from the proposed M2Trans that can accept images of arbitrary resolution, MedCLIP requires the resolution of the input images to be 224 × 224. However, directly resizing input images to meet the requirement of MedCLIP may result in the degradation of image quality, potentially leading to incorrect results [41]. To address this challenge, we implement additional measures to ensure that MedCLIP effectively captures the information contained in the input images. Specifically, for each pair of input images $(\mathrm{I}_{hr}, \mathrm{I}_{sr})$, where I_{sr} is the SR result of the I_{lr} and I_{hr} is the corresponding HR image, we generate specified K patches $\{\mathrm{p}_k\}_{k=1}^{K}$ for each image. It is important to note that the first patch p_1 captures information about the entire image through resizing operations, while the remaining K-1 patches obtain partial information about the image through random cropping. The cropping operations on I_{hr} and I_{sr} are consistent to ensure that the input to the image encoder undergoes cropping in the same region, thus eliminating irrelevant errors. The final image-text semantic loss is computed as the weighted sum of $\mathcal{L}_{clip}$ of these image patches, expressed as:

$$\mathcal{L}_{sem} = \sum_{i=1}^{N}\sum_{j=1}^{K} \delta^{i,j} \cdot ||\mathrm{s}_{sr}^{i,j} - \mathrm{s}_{hr}^{i,j}||, \tag{14}$$

where N is the number of training images in a batch. $\delta^{i,j}$ represents the weight factor of the j-th patch of the i-th image, which is empirically set to $\frac{1}{K}$ in this work.

IV. Muti-Modal Ultrasound Image 1K Dataset

The existing ultrasound image datasets are insufficient in volume to fulfill the demands of deep model learning. Additionally, many of these datasets are primarily annotated for specific organ parts to facilitate downstream tasks like lesion identification or tumor segmentation, rather than focusing on the core objective of ultrasound image SR. To address this gap, we create a new multi-modal ultrasound image dataset (MMUS1K) containing 1023 images of various organ parts, each paired with a corresponding text description. This study is approved by the institutional review board of Shanghai Tenth People's Hospital of Tongji University (Shanghai, China; NO.SHSY-IEC-5.0/23K115/P01). We curate a substantial collection of ultrasound images from Shanghai Tenth People's Hospital, covering various organ parts such as the bladder, gallbladder, thyroid, kidney, and more. However, these data contain unnecessary information such as watermarks, blur, and noise. To address this, we use an open-source data labeling tool, LabelImg, to crop the pertinent content of the ultrasound images for desensitization. Finally, we collecte a total of 1023 ultrasound images after excluding data that is either too small or redundant, ensuring that the minimum height and width of these images are no less than 448 and 600, respectively.

As described in Section III-C, we utilize the pretrained image encoder and text encoder from MedCLIP to extract features for both images and texts in order to compute the image-text semantic loss. However, since the collected ultrasound images lack corresponding diagnostic reports, we employ a large language model (LLM) to annotate these images with text descriptions. To achieve this, we utilize the CaptionAnything tool [42], a versatile multi-modal image processing tool that combines the capabilities of Segment-Anything [12] and ChatGPT to generate textual descriptions for the ultrasound images. The overall annotation process can be summarized as follows:

$$(\mathrm{T}_p, \mathrm{R}, \mathrm{e}) = \mathrm{CAT}(\mathrm{I}_{hr}, \mathrm{V}_c, \mathrm{L}_c), \tag{15}$$

where $\mathrm{CAT}(\cdot)$ represents the CaptionAnything model, R and e are the segment region and the corresponding score of the text description, respectively. V_c and L_c are the visual controls and language controls decided by user interest. Specifically, the meaningful positioning of information in the HR ultrasound image is determined through a segmentation model $\mathrm{SAM}(\cdot)$ set by visual control, which can be expressed as:

$$\mathrm{R} = \mathrm{SAM}(\mathrm{I}_{hr}, \mathrm{V}_c). \tag{16}$$

Then, the region position R along with the HR ultrasound image are jointly input to a pre-trained image captioner $\mathrm{ICT}(\cdot)$ to generate a raw description T_c of the interest areas as:

$$\mathrm{T}_c = \mathrm{ICT}(\mathrm{I}_{hr}, \mathrm{R}). \tag{17}$$

Later, to generate text descriptions more suitable for the needs of ultrasound image SR, the coarse descriptions are refined using a text refiner $\mathrm{TRE}(\cdot)$ with the guidance of language controls to generate finer text descriptions:

$$\mathrm{T}_p = \mathrm{TRE}(\mathrm{T}_c, \mathrm{L}_c). \tag{18}$$

Finally, a cosine similarity score is calculated between the HR ultrasound image and the generated text description to evaluate the quality of the description:

$$\mathrm{e} = \cos(\mathrm{I}_{hr}, \mathrm{T}_p). \tag{19}$$

We repeat the above steps multiple times for each image, resulting in a series of text descriptions for each image with corresponding confidence estimates.

Since these descriptions may contain interference and errors, and the focus is on the overall quality information of the entire image rather than specific regions, we select the sentence with the highest score and the largest region as the ultrasound image description. We select three types of sentences in all as templates, as follows:

- a picture of a $\{attr\}$ ultrasound image,
- a $\{attr\}$ image of an ultrasound,
- a $\{attr\}$ ultrasound image.

where $\{attr\}$ represents the attribute words that express the quality information of the images.

To ensure that the image-text semantic loss captures the overall quality and semantic information of the image, we employ a straightforward sentence template to direct the model's attention to the global understanding of the image. According to the findings mentioned in [46], the CLIP model can discern the quality information of the image based on the added attribute words. We list a series of attribute words to describe the image quality. They include:

- nice, sharp, fine, clear, clean, cool, great.

Finally, we randomly add from 1 to 7 of the above 7 attribute words to each image in the dataset based on the selected sentence templates to generate the final text descriptions of the ultrasound images.

V. Experiments and Analysis

A. Experimental Setups

1) Datasets: To evaluate the performance and generalization capability of our model, we train the M2Trans on our created MMUS1K and evaluate it on three datasets: CCA-US[1] and US-CASE,[2] and our MMUS1K dataset. The first two datasets are widely used public datasets. The CCA-US dataset comprises 84 ultrasound images of the carotid artery, while the US-CASE dataset consists of over 7000 ultrasound images covering various anatomical regions such as the liver, heart, and mediastinum, among others. From the US-CASE dataset, we select images referenced in papers [1], [25], excluding 14 images that are too small or lack meaningful semantic information, resulting in a final testing set of 111 images.

2) Training Objective: The M2Trans is optimized with a combination of the proposed image-text semantic loss and fidelity loss. The fidelity loss is defined as the $\mathcal{L}_1$ norm between the SR image and the HR image. The training loss of our model is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{fid} + \beta\mathcal{L}_{sem}, \tag{20}$$

[1] [Online]. Available: http://splab.cz/en/download/databaze/ultrasound

[2] [Online]. Available: http://www.ultrasoundcases.info/Cases-Home.aspx

TABLE I
RESULTS OF OUR PROPOSED M2TRANS COMPARED WITH SOTA METHODS IN PSNR AND SSIM

Method	Scale	Params (M)	FLOPs (G)		PSNR ↑			SSIM ↑	
				CCA-US	US-CASE	MMUS1K	CCA-US	US-CASE	MMUS1K
Bicubic	× 2	-	-	37.41	37.59	34.91	0.9600	0.9570	0.9429
EDSR [5]		14.17	915.26	39.63	38.25	36.17	0.9676	0.9580	0.9478
SwinIR [19]		11.75	858.24	37.00	36.54	33.99	0.9567	0.9492	0.9312
ELAN [43]		3.56	131.54	37.44	36.54	34.73	0.9618	0.9499	0.9351
ESRT [44]		4.89	186.25	38.60	38.02	36.42	0.9679	0.9603	0.9493
HAT [45]		20.62	1659.52	37.32	37.57	35.55	0.9601	0.9531	0.9411
M2Trans		3.61	20.23	**40.37**	**39.79**	**37.74**	**0.9736**	**0.9682**	**0.9589**
Bicubic	× 3	-	-	32.99	32.69	30.90	0.9103	0.8871	0.8591
EDSR [5]		14.17	365.33	35.36	33.29	32.20	0.9344	0.8976	0.8821
SwinIR [19]		11.94	369.61	32.24	30.89	29.72	0.9048	0.8662	0.8501
ELAN [43]		3.58	58.79	35.25	33.54	31.93	0.9332	0.8989	0.8753
ESRT [44]		4.98	82.78	34.50	32.99	31.76	0.9226	0.8841	0.8693
HAT [45]		20.81	714.70	33.46	32.66	31.39	0.9129	0.8791	0.8607
M2Trans		3.63	9.48	**35.87**	**34.42**	**32.86**	**0.9360**	**0.9094**	**0.8945**
Bicubic	× 4	-	-	30.12	28.90	28.24	0.8432	0.7892	0.7817
EDSR [5]		14.17	230.34	32.55	30.82	30.04	0.8902	0.8497	0.8326
SwinIR [19]		11.90	214.56	29.72	28.50	27.66	0.8397	0.7834	0.7758
ELAN [43]		3.61	33.07	31.72	31.02	30.40	0.8845	0.8464	0.8309
ESRT [44]		4.96	46.56	31.42	30.84	30.25	0.8778	0.8374	0.8235
HAT [45]		20.77	414.88	29.23	28.72	28.08	0.8378	0.7812	0.7582
M2Trans		3.63	5.87	**32.72**	**31.32**	**30.68**	**0.8977**	**0.8516**	**0.8392**

For ease of identification, the best- and second-best-performing are highlighted in bold and underlined, respectively.

where β is the weighting factor controlling the relative importance of the image-text semantic loss.

3) Evaluation Metrics: For quantitative comparison, we select classic metrics including PSNR and SSIM [47], along with FSIM [48] and GMSD [49], to assess image quality. PSNR and SSIM are commonly used indicators for assessing fidelity and structural similarity, while FSIM and GMSD focus more on perceptual quality. It is important to note that higher values of PSNR, SSIM, and FSIM indicate better performance, while lower values of GMSD indicate better performance.

4) Training Details: Settings of M2Trans: We randomly crop HR images to generate 384×384 patches for training the network. Subsequently, using bicubic downsampling at scales 2, 3, and 4, we obtain LR patches of sizes 192×192, 128×128, and 96×96, respectively. Data augmentation techniques include horizontal flipping, vertical flipping, and 90-degree rotation for each image. Our network starts with an initial learning rate of $1e-4$, gradually decreasing to $1e-6$ using the cosine decay strategy. Training lasts for 200 epochs (a total of 5×10^5 iterations) across different scales. Our model is implemented under the PyTorch architecture and trained on a single NVIDIA GeForce RTX 4090. The hyperparameter β in (20) is set to 0.01 empirically. The number of CTFM modules is fixed at 8, the input batch size N is set to 2 for fair comparison with different baselines, and we utilize the ADAM optimizer with default settings. Additionally, the number of cropped patches for semantic loss, K, is set to 3 empirically.

Settings of Baselines: We conduct a comparative analysis of the proposed M2Trans with classic methods like bicubic interpolation and state-of-the-art learning-based approaches. Given that there are few open source codes for ultrasound image SR, we select several well-known natural image region methods and made some adjustments to suit the ultrasound image SR task, including EDSR [5], SwinIR [19], ELAN [43], ESRT [44] and HAT [45], which are all transformer-based SR models similar to ours. In our comparisons, we aim to preserve the original configurations of these methods as much as possible to ensure fairness and consistency with our settings and dataset. For EDSR, we set the batch size to 2 while keeping other parameters unchanged. For the SwinIR model, we also modify the parameters to keep it as consistent as possible with the configuration of our model. In the case of ELAN, we set the batch size to 2 and adjust the patch size to 384, while the channel number is set to 150 to match the scale of our model. For ESRT, we only modify the total number of iterations while keeping other parameters intact. For the HAT model, the batch size is set to 1, while keeping other parameters as consistent as possible with that of our model.

B. Quantitative Results

The results in Tables I and II demonstrate that our proposed method consistently outperforms other approaches across various upscaling factors and evaluation metrics. When the amplification scale is 2, our proposed M2Trans surpasses other methods in all aspects of different metrics. The PSNR exceeds the second best with a margin of 0.74, 1.54, and 1.32 while SSIM grows 0.0057, 0.0079, and 0.0096 on different datasets. When the amplification scale becomes 3 and 4, our proposed M2Trans also shows consistent superiority over compared baselines. Higher FSIM and lower GMSD represent a clearer structure and detailed texture. The only method with a similar number of parameters, ELAN, performs worse than our approach and demands a high computational load, resulting in excessively high FLOPs (Floating Point Operations). In summary, our proposed M2Trans achieves a consistent high performance in quantitative comparisons on different datasets at different amplification scales,

TABLE II
Results of Our Proposed M2Trans Compared With SOTA Methods in FSIM and GMSD

Method	Scale	Params (M)	FLOPs (G)	FSIM ↑			GMSD ↓		
				CCA-US	US-CASE	MMUS1K	CCA-US	US-CASE	MMUS1K
Bicubic	× 2	-	-	0.9455	0.9495	0.9575	0.0443	0.0456	0.0593
EDSR [5]		14.17	915.26	0.9660	0.9690	0.9832	0.0319	0.0387	0.0350
SwinIR [19]		11.75	858.24	0.9421	0.9414	0.9396	0.0586	0.0670	0.0876
ELAN [43]		3.56	131.54	0.9524	0.9564	0.9714	0.0432	0.0478	0.0500
ESRT [44]		4.89	186.25	0.9488	0.9505	0.9626	0.0503	0.0555	0.0600
HAT [45]		20.62	1659.52	0.9465	0.9484	0.9616	0.0495	0.0566	0.0603
M2Trans		3.61	20.23	**0.9733**	**0.9784**	**0.9849**	**0.0166**	**0.0185**	**0.0302**
Bicubic	× 3	-	-	0.9105	0.9091	0.9274	0.0734	0.0889	0.0978
EDSR [5]		14.17	365.33	0.9219	0.9168	0.9394	0.0691	0.0834	0.0839
SwinIR [19]		11.94	369.61	0.8840	0.8790	0.9141	0.1067	0.1174	0.1081
ELAN [43]		3.58	58.79	0.9300	0.9284	0.9429	0.0549	0.0698	0.0800
ESRT [44]		4.98	82.78	0.8989	0.8952	0.9207	0.0894	0.1019	0.1026
HAT [45]		20.81	714.70	0.8850	0.8813	0.9061	0.1019	0.1141	0.1151
M2Trans		3.63	9.48	**0.9352**	**0.9335**	**0.9526**	**0.0491**	**0.0645**	**0.0705**
Bicubic	× 4	-	-	0.8752	0.8672	0.8963	0.1122	0.1314	0.1348
EDSR [5]		14.17	230.34	0.8920	0.8799	0.9102	0.0965	0.1153	0.1167
SwinIR [19]		11.90	214.56	0.8120	0.8234	0.8820	0.1620	0.1513	0.1341
ELAN [43]		3.61	33.07	0.8747	0.8780	0.9104	0.1135	0.1158	0.1156
ESRT [44]		4.96	46.56	0.8316	0.8244	0.8574	0.1448	0.1513	0.1523
HAT [45]		20.77	414.88	0.8351	0.8301	0.8625	0.1356	0.1448	0.1478
M2Trans		3.63	5.87	**0.8966**	**0.8901**	**0.9213**	**0.0873**	**0.1060**	**0.1046**

For ease of identification, the best- and second-best-performing are highlighted in bold and underlined, respectively.

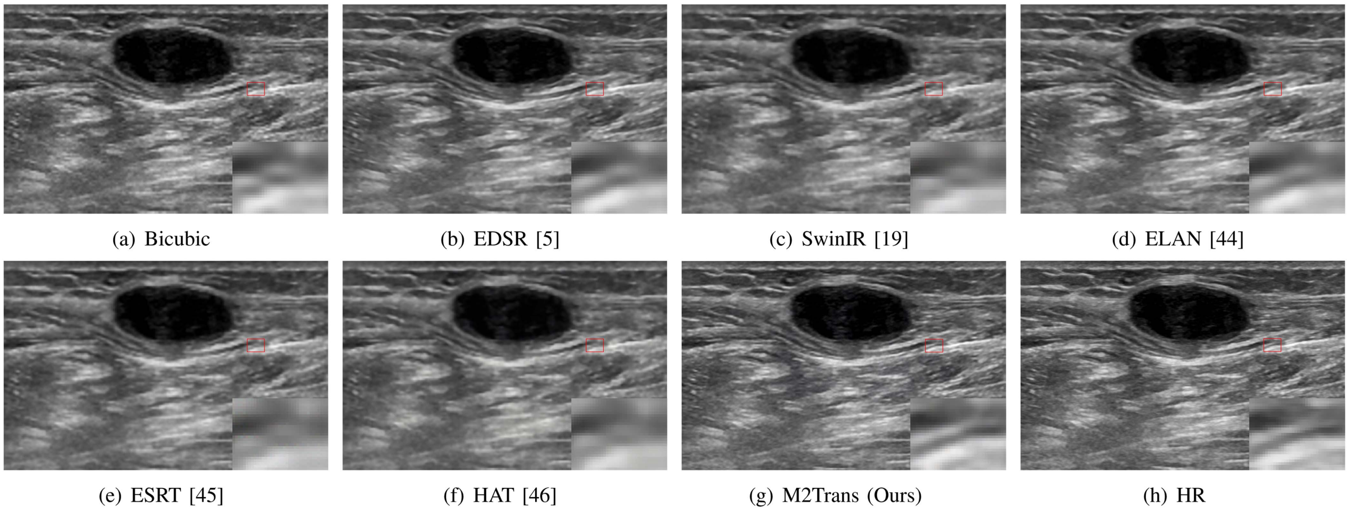

(a) Bicubic (b) EDSR [5] (c) SwinIR [19] (d) ELAN [44]
(e) ESRT [45] (f) HAT [46] (g) M2Trans (Ours) (h) HR

Fig. 3. The visual comparisons of ×4 SR results of different methods for ultrasound images from the MMUS1K dataset.

which fully proves the effectiveness and superiority of our method.

C. Qualitative Results

Three sets of images with an upscaling factor of 4 are selected from different testing datasets to showcase the effectiveness of our model for SR tasks. Firstly, we compare the results of different methods on images from the US-CASE dataset with an upscaling factor of 4. As illustrated in Fig. 4, the SR result of our method closely resembles the HR image compared to other methods. It maintains sharper details in color transition regions and preserves a greater amount of fine-grained information. When compared to the SwinIR and HAT methods, our approach exhibits less brightness deviation from the ground truth and faithfully reconstructs the super-resolved results. Similar comparisons are conducted between the MMUS1K and CCA-US datasets, as depicted in Figs. 3 and 5. Consistently, the results produced by our method retain clearer edge information and excel in constructing detailed textures, despite the challenges posed by 4× upscaling, which inherently entails more significant information loss and ambiguity in the SR results.

Therefore, when compared to other methods, our approach generates results that are more visually close to the ground truth image. By effectively utilizing prior information for image quality guidance, it achieves clearer images with richer detail

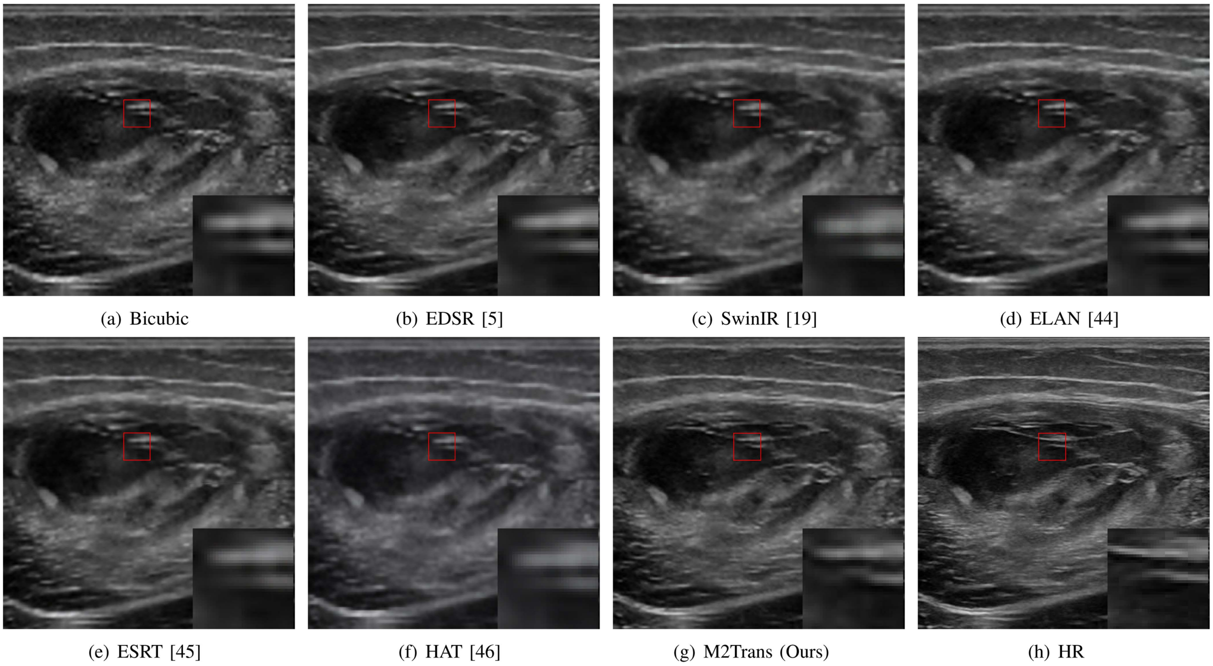


Fig. 4. The visual comparisons of ×4 SR results of different methods for ultrasound images from the US-CASE dataset.

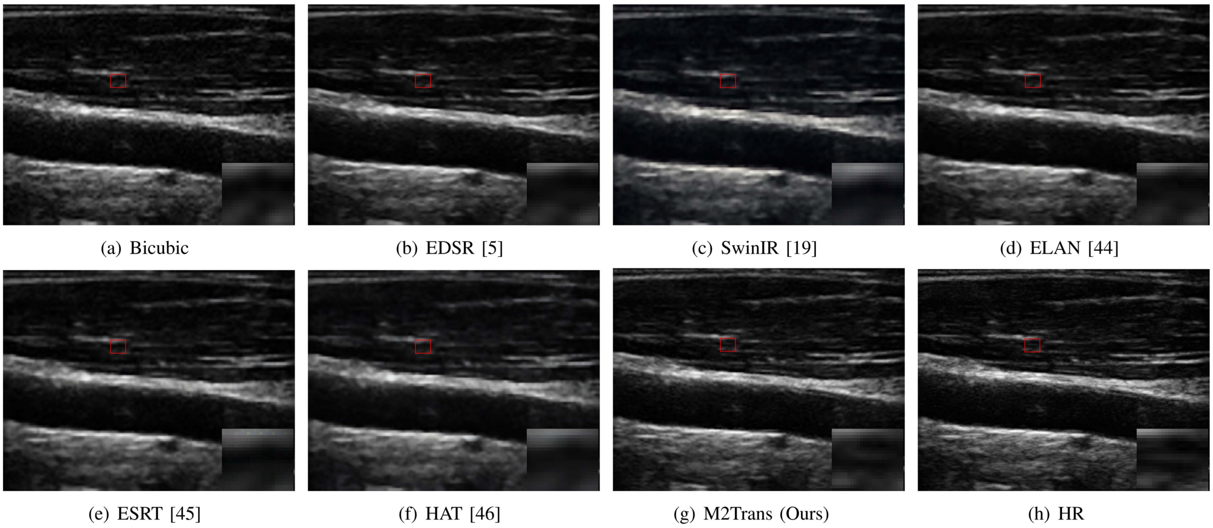


Fig. 5. The visual comparisons of ×4 SR results of different methods for ultrasound images from the CCA-US dataset.

information. In summary, guided by prior knowledge from MedCLIP, our proposed M2Trans excels in reconstructing super-resolution ultrasound images with the assistance of image-text semantic loss. This methodology enhances brightness and color consistency with HR images while preserving richer texture details as well.

D. Ablation Study

This subsection conducts extensive ablation studies to evaluate the impact of each component in the proposed M2Trans, including CFTM, image-text semantic loss, hyper-parameter of β, and patch number K.

1) Effectiveness of CFTM: We conduct ablation experiments on the proposed CFTM with a variant model. Version 1 removes the DWT and IDWT modules. To maintain consistent channel dimensions when passing through the SA module, we adjusted the input channel dimension to four times the original one, *i.e.*, $D = 256$, to match the dimensions of the last two layers passing through the SA module, for a fair comparison.

From the experimental results in Table III, it is observed that the DWT structure can significantly improve the performance

TABLE III
COMPARISONS OF PROPOSED M2TRANS VARIANTS ON VARIOUS DATASETS WITH DIFFERENT SCALES

	Scale	DWT	$\mathcal{L}_{sem}$	PSNR ↑			SSIM ↑		
				CCA-US	US-CASE	MMUS1K	CCA-US	US-CASE	MMUS1K
Version 1	× 2		✓	38.84	38.69	36.64	0.9708	0.9644	0.9539
Version 2		✓		39.14	38.88	37.18	0.9687	0.9643	0.9560
Full		✓	✓	**40.37**	**39.79**	**37.74**	**0.9736**	**0.9682**	**0.9589**
Version 1	× 3		✓	34.47	33.72	32.60	0.9287	0.8977	0.8889
Version 2		✓		35.42	33.99	32.68	0.9325	0.8981	0.8897
Full		✓	✓	**35.87**	**34.42**	**32.86**	**0.9360**	**0.9094**	**0.8945**
Version 1	× 4		✓	31.75	30.91	30.32	0.8869	0.8422	0.8283
Version 2		✓		32.38	30.87	30.18	0.8917	0.8389	0.8217
Full		✓	✓	**32.72**	**31.32**	**30.68**	**0.8977**	**0.8516**	**0.8392**

TABLE IV
PARAMETERS AND FLOPS OF MODELS WITH OR WITHOUT DWT AND IDWT IN CFTM

	Params (M)			FLOPs (G)		
	×2	×3	×4	×2	×3	×4
Version 1	5.41	5.74	5.67	199.62	94.69	60.41
Full	3.61	3.63	3.63	20.23	9.48	5.87

Version 1 represents the model without DWT and IDWT in cftm, while full represents the proposed M2Trans.

of the model. Since the DWT module effectively reduces the spatial dimension, it is akin to performing dimensional reduction before conducting self-attention operations on the features. Therefore, as shown in Table IV, the number of parameters in the model decrease by 33% compared to the model without the DWT module, and the FLOPs are only 10.13% of the original. Nevertheless, the model still achieves excellent results in PSNR and SSIM evaluation measures.

2) Effectiveness of Image-Text Semantic Loss: We conduct ablation experiments on the proposed image-text semantic loss with a variant model. Version 2 only uses $\mathcal{L}_{fid}$ loss without incorporating image-text semantic loss. From the results in Table III, we can observe that using the image-text semantic loss results in a significant performance improvement. For example, with a scale factor of 2, there is a notable increase from 0.56 to 1.23 in PSNR and from 0.0029 to 0.0049 in SSIM. Similarly, with a scale factor of 3, the increases in PSNR on the three datasets are 0.45, 0.43, and 0.18, respectively. Besides, we also utilize MDA (Manifold Discovery and Analysis [50]) tools to demonstrate how our network aligns the text and imaging modalities at different layers by the proposed Image-Text Semantic Loss. The results, shown in Fig. 6, indicate that our model with Image-Text Semantic Loss (the first row) exhibits smoother and more continuous color transitions compared to the model without Image-Text Semantic Loss (the second row).

3) Effectiveness of Hyper-Parameter β: To determine the optimal weighting factor β of our proposed semantic loss, we conduct ablation experiments on the hyper-parameter β, as shown in Table V. From Table V, it is evident that when β is set to 0.01, most experimental results achieve the best or second performance. Therefore, we finally select 0.01 as the hyper-parameter β for our method.

TABLE V
DIFFERENT SETTINGS OF HYPER-PARAMETER β

β	Scale	CCA-US PSNR↑	CCA-US SSIM↑	US-CASE PSNR↑	US-CASE SSIM↑	MMUS1K PSNR↑	MMUS1K SSIM↑
0.001	×2	40.03	**0.9736**	39.51	0.9676	37.05	0.9575
	×3	35.51	0.9341	34.03	0.9023	32.53	0.8882
	×4	32.54	0.8910	31.40	0.8525	30.62	0.8366
0.005	×2	40.19	0.9733	39.66	0.9675	37.57	**0.9589**
	×3	35.53	0.9338	33.69	0.8974	31.79	0.8773
	×4	**32.76**	0.8945	**31.45**	0.8500	**30.68**	0.8347
0.01	×2	**40.37**	**0.9736**	**39.79**	**0.9682**	**37.74**	**0.9589**
	×3	**35.87**	**0.9360**	**34.42**	**0.9094**	**32.86**	**0.8945**
	×4	32.72	**0.8977**	31.32	0.8516	**30.68**	**0.8392**
0.1	×2	40.19	0.9727	39.26	0.9661	36.95	0.9561
	×3	35.32	0.9279	34.01	0.8990	32.79	0.8925
	×4	32.25	0.8878	31.31	0.8527	30.44	0.8303
1.0	×2	40.10	0.9720	39.72	0.9679	37.49	0.9586
	×3	34.88	0.9238	34.12	0.9016	32.60	0.8910
	×4	32.25	0.8868	31.42	**0.8528**	30.60	0.8366

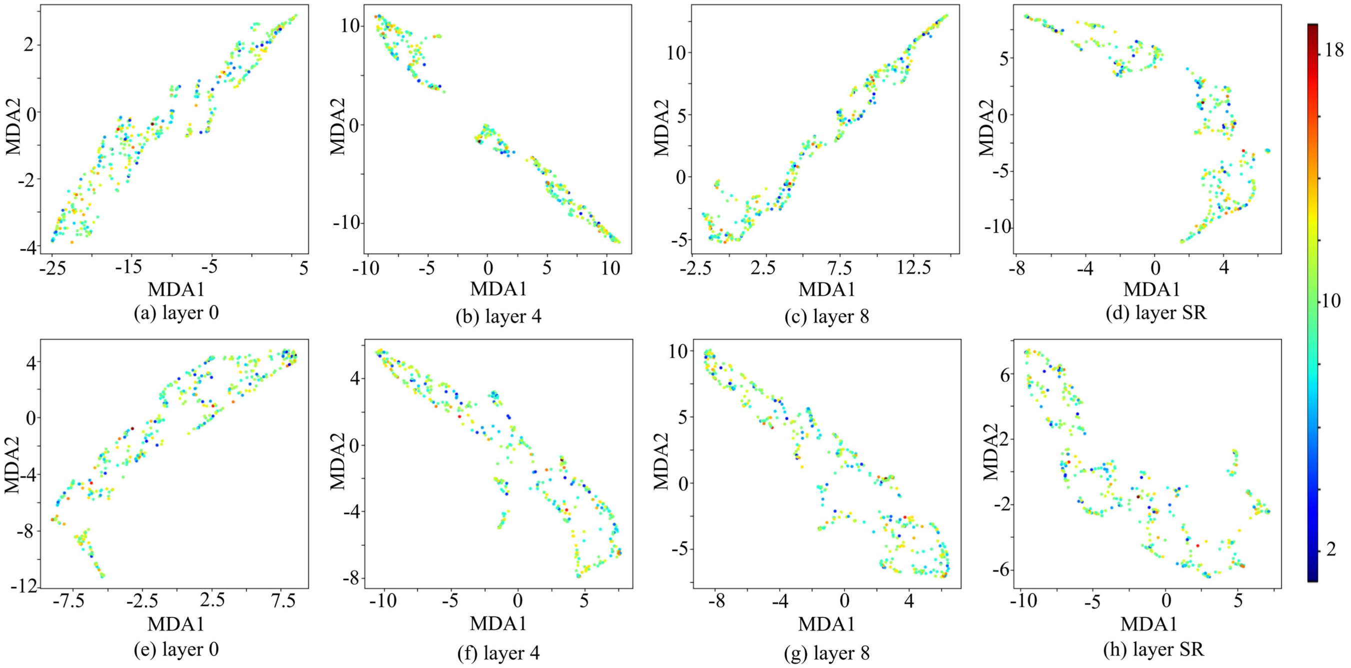
For ease of identification, the best- and second-best-performing are highlighted in bold and underlined, respectively.

TABLE VI
DIFFERENT SETTINGS OF PATCH NUMBER K OF SCALE 2

K	CCA-US PSNR↑	CCA-US SSIM↑	US-CASE PSNR↑	US-CASE SSIM↑	MMUS1K PSNR↑	MMUS1K SSIM↑
1	40.31	**0.9737**	39.58	0.9674	37.46	0.9579
2	39.87	0.9729	39.71	0.9668	37.63	**0.9589**
3	**40.37**	0.9736	**39.79**	**0.9682**	**37.74**	**0.9589**
4	39.75	0.9722	39.57	0.9676	37.35	0.9574

For ease of identification, the best- and second-best-performing are highlighted in bold and underlined, respectively.

4) Effectiveness of Patch Number K: Our initial intuition in adopting patches to calculate image-text semantic loss is to meet the requirement of the input size of MediCLIP. We conduct ablation experiments on the K with various values. From the results in Table VI, one can observe that most experimental results achieve the best or second performance when the patch number equals 3. Therefore, we finally set 3 as the default value of image patch K for our proposed method.



(a) layer 0 (b) layer 4 (c) layer 8 (d) layer SR

(e) layer 0 (f) layer 4 (g) layer 8 (h) layer SR

Fig. 6. MDA visualization of the proposed M2Trans features for super-resolution tasks after network training. The first row illustrates the feature distribution with image-text semantic loss, while the second row illustrates the feature distribution without image-text semantic loss.

VI. Conclusion

In this paper, we propose a novel ultrasound image SR method, *i.e.* M2Trans, that explores the potential of leveraging prior knowledge for ultrasound image SR tasks. By utilizing the MedCLIP to build the joint image-text guidance as the semantic loss, our method successfully achieves ultrasound SR results with richer semantic information and detailed textures. In the M2Trans model, we design a novel coarse-to-fine transformer comprising multiple branches infused with self-attention and frequency transforms to efficiently capture signal dependencies across different scales. Extensive experiments on three datasets demonstrate the superiority of our method, outperforming existing state-of-the-art methods in terms of quantitative and qualitative evaluations. Furthermore, the ablation studies show the effectiveness of each component as well as the rationality of the architecture design and hyper-parameter choices. Our future work will focus on leveraging prior knowledge to better guide the super-resolution (SR) of ultrasound images at larger amplification scales. Additionally, collecting actual clinical electronic health record (EHR) information to enhance the model's applicability and explainability in clinical settings, ensuring it is more effective and relevant for practical use.

References

[1] H. Liu, J. Y. Liu, S. D. Hou, T. Tao, and J. G. Han, "Perception consistency ultrasound image super-resolution via self-supervised CycleGAN," *Neural Comput. Appl.*, vol. 35, pp. 12331–12341, 2021.

[2] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4681–4690.

[3] Z. K. Ni, W. H. Yang, S. Q. Wang, L. Ma, and S. Kwong, "Towards unsupervised deep image enhancement with generative adversarial network," *IEEE Trans. Image Process.*, vol. 29, pp. 9140–9151, 2020.

[4] A. Vaswani et al., "Attention is all you need," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 1–11.

[5] B. Lim, S. Son, H. Kim, S. Nah, and M. L. Kyoung, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 136–144.

[6] K. Umehara, J. Ota, and T. Ishida, "Application of super-resolution convolutional neural network for enhancing image resolution in chest CT," *J. Digit. Imag.*, vol. 31, pp. 441–450, 2018.

[7] W. Choi, M. Kim, J. HakLee, J. Kim, and J. BeomRa, "Deep CNN-based ultrasound super-resolution for high-speed high-resolution B-mode imaging," in *2018 IEEE Int. Ultrasonics Symp.*, 2018, pp. 1–4.

[8] Z. K. Ni, W. H. Yang, H. L. Wang, S. Q. Wang, L. Ma, and S. Kwong, "Cycle-interactive generative adversarial network for robust unsupervised low-light enhancement," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 1484–1492.

[9] T. Brown et al., "Language models are few-shot learners," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 1877–1901.

[10] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Ch. Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[11] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[12] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4015–4026.

[13] Z. F. Wang, Z. B. Wu, D. Agarwal, and J. M. Sun, "MedCLIP: Contrastive learning from unpaired medical images and text," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2022, pp. 3876–3887.

[14] C. Dong, C. C. Loy, K. M. He, and X. O. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[15] Y. D. Liang, J. J. Wang, S. P. Zhou, Y. H. Gong, and N. N. Zheng, "Incorporating image priors with deep convolutional neural networks for image super-resolution," *Neurocomputing*, vol. 194, pp. 340–347, 2016.

[16] S. J. Park, H. Son, S. Cho, K. S. Hong, and S. Lee, "SRFeat: Single image super-resolution with feature discrimination," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 439–455.

[17] X. N. Zhu, L. Zhang, L. J. Zhang, X. Liu, Y. Shen, and S. J. Zhao, "GAN-based image super-resolution with a novel quality loss," *Math. Problems Eng.*, vol. 2020, 2020, Art. no. 5217429.

[18] H. T. Chen et al., "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12299–12310.

[19] J. Y. Liang, J. Z. Cao, G. L. Sun, K. Zhang, V. G. Luc, and R. Timofte, "SwinIR: Image restoration using swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1833–1844.

[20] C. H. Pham, A. Ducournau, R. Fablet, and F. Rousseau, "Brain MRI super-resolution using deep 3D convolutional networks," in *2017 IEEE 14th Int. Symp. Biomed. Imag.*, 2017, pp. 197–200.

[21] D. F. Qiu, Y. H. Cheng, and X. S. Wang, "Progressive U-net residual network for computed tomography images super-resolution in the screening of COVID-19," *J. Radiat. Res. Appl. Sci.*, vol. 14, no. 1, pp. 369–379, 2021.

[22] Z. Ni, W. Yang, S. Wang, L. Ma, and S. Kwong, "Unpaired image enhancement with quality-attention generative adversarial network," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1697–1705.

[23] Y. Y. Wu, F. Yang, J. Huang, and Y. Q. Liu, "Super-resolution construction of intravascular ultrasound images using generative adversarial networks," *J. Southern Med. Univ.*, vol. 39, no. 1, pp. 82–87, 2019.

[24] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-GAN," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 1–10.

[25] H. Liu, J. Y. Liu, F. Chen, and C. F. Shan, "Progressive residual learning with memory upgrade for ultrasound image blind super-resolution," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 9, pp. 4390–4401, Sep. 2022.

[26] M. Puttagunta and R. Subban, "SwinIR transformer applied for medical image super-resolution," *Procedia Comput. Sci.*, vol. 204, pp. 907–913, 2022.

[27] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[28] M. I. Georgescu et al., "Multimodal multi-head convolutional attention with various kernel sizes for medical image super-resolution," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 2194–2204.

[29] S. Y. Zhao et al., "Exploiting unlabeled data with vision and language models for object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 159–175.

[30] H. S. Luo et al., "CLIP4clip: An empirical study of CLIP for end to end video clip retrieval and captioning," *Neurocomputing*, vol. 508, pp. 293–304, 2022.

[31] A. Khandelwal, L. Weihs, R. Mottaghi, and A. Kembhavi, "Simple but effective: CLIP embeddings for embodied AI," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14809–14818.

[32] W. X. Lin et al., "PMC-CLIP: Contrastive language-image pre-training using biomedical documents," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2023, pp. 525–536.

[33] J. Liu et al., "CLIP-driven universal model for organ segmentation and tumor detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 21152–21164.

[34] C. Wang, M. L. Chai, M. M. He, D. D. Chen, and J. Liao, "CLIP-NeRF: Text-and-image driven manipulation of neural radiance fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 3825–3834.

[35] H. Chefer, S. Benaim, R. Paiss, and L. Wolf, "Image-based CLIP-guided essence transfer," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 695–711.

[36] Z. X. Liang, C. Y. Li, S. C. Zhou, R. C. Feng, and C. C. Loy, "Iterative prompt learning for unsupervised backlit image enhancement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 8094–8103

[37] J. N. Pinkney and C. Li, "Clip2latent: Text driven sampling of a pre-trained StyleGAN using denoising diffusion and CLIP," in *Proc. Brit. Mach. Vis. Conf.*, 2022, pp. 1–12.

[38] A. Sanghi et al., "CLIP-sculptor: Zero-shot generation of high-fidelity and diverse shapes from natural language," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 18339–18348.

[39] J. Y. Wang, K. C. Chan, and C. C. Loy, "Exploring CLIP for assessing the look and feel of images," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, 2023, pp. 2555–2563.

[40] M. J. Shensa, "The discrete wavelet transform: Wedding the a trous and Mallat algorithms," *IEEE Trans. Signal Process.*, vol. 40, no. 10, pp. 2464–2482, Oct. 1992.

[41] M. J. Saikia, S. Kuanar, D. Mahapatra, and S. Faghani, "Multi-modal ensemble deep learning in head and neck cancer HPV sub-typing," *Bioengineering*, vol. 11, no. 1, 2023, Art. no. 13.

[42] T. Wang et al., "Caption anything: Interactive image description with diverse multimodal controls," 2023, *arXiv:2305.02677*.

[43] X. D. Zhang, H. Zeng, S. Guo, and L. Zhang, "Efficient long-range attention network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 649–667.

[44] Z. S. Lu, H. Liu, J. C. Li, and L. L. Zhang, "Efficient transformer for single image super-resolution," 2021, *arXiv:2108.11084*.

[45] X. Y. Chen, X. T. Wang, J. T. Zhou, Y. Qiao, and C. Dong, "Activating more pixels in image super-resolution transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 22367–22377.

[46] K. Y. Zhou, J. K. Yang, C. C. Loy, and Z. W. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, 2022.

[47] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[48] L. Zhang, L. Zhang, X. Q. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[49] W. F. Xue, L. Zhang, X. Q. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2014.

[50] M. T. Islam et al., "Revealing hidden patterns in deep neural network feature space continuum via manifold learning," *Nature Commun.*, vol. 14, no. 1, 2023, Art. no. 8506.