RALSTM: Autoregressive Multi-Phase Contrast-Enhanced CT Synthesis via Spatio-Temporal Priors

Jiajian Xie[1], Wenfeng Xu[1], Cong Lai[2], Cheng Liu[2], Gansen Zhao[1]* and Kewei Xu[2]*
[1]*School of Computer Science, South China Normal University, Guangzhou, China*
[2]*Department of Urology, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China*
Email: xiejiajian@m.scnu.edu.cn, xuwenfeng@m.scnu.edu.cn, laic@mail2.sysu.edu.cn,
liuch278@mail.sysu.edu.cn, gzhao@m.scnu.edu.cn, xukewei@mail.sysu.edu.cn

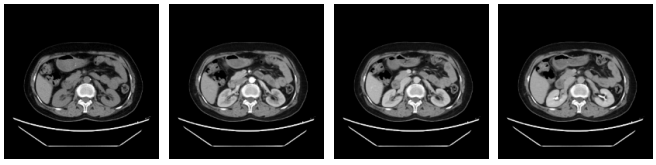
***Abstract*—To extend the benefits of dynamic lesion characterization to patients with contrast agent contraindications, synthesizing multi-phase contrast-enhanced computed tomography (CECT) from non-contrast CT (NCCT) has emerged as a vital tool for personalized radiotherapy planning. However, existing methods often struggle to fully utilize the anatomical context priors of NCCT and the temporal dependencies between different phases. This results in anatomical distortions across different phases and a significant increase in training iterations and parameter count, limiting their practical application. To overcome these challenges, this paper introduces a novel framework named RALSTM for the autoregressive and sequential synthesis of multi-phase CECT from abdominal NCCT. The framework leverages a Residual ConvLSTM module to model encoded features and capture the temporal transitions across phases. Additionally, a Region-Aware Gating Unit is used to localize contrast-enhancing regions and impose spatial constraints on the decoded features. Experimental results on both internal and external datasets demonstrate that this method achieves superior metrics and delivers outstanding synthesis quality.**

***Index Terms*—autoregressive, synthesis, prior knowledge, generative adversarial network**

I. INTRODUCTION

In radiotherapy planning, delineation of organs-at-risk typically relies on both contrast-enhanced computed tomography (CECT) images and non-contrast CT (NCCT) images. CECT, in particular, substantially improves the conspicuity of pathological regions after intravenous iodinated contrast administration, facilitating accurate lesion detection and delineation [1]. As illustrated in Fig. 1, NCCT is acquired prior to contrast injection, whereas arterial, venous, and delayed phases are sequentially obtained post-injection; the temporal comparison of the same anatomical region across these phases enables clinicians to dynamically evaluate lesions [2]. Nevertheless, a subset of patients cannot receive contrast media because of allergies, renal insufficiency, or comorbidities. For these contraindicated or high-risk individuals, synthesizing multi-phase CECT from NCCT represents a promising alternative [3].

In recent years, deep learning-based image synthesis has witnessed rapid advancement. Generative Adversarial Networks (GANs) [4] and their adaptive variants have been extensively applied to medical image synthesis. More recently, Denoising Diffusion Probabilistic Models (DDPMs) [5], and their latent-space variants [6], have emerged as a compelling alternative due to their ability to generate high-quality and diverse samples. Current medical image synthesis methodologies can be broadly categorized by their architectural paradigms: GAN-based methods, which drive the generated distribution toward the target through a minimax game between a generator and a discriminator [7]–[11], and diffusion-based methods, which approximate the target distribution via iterative or latent denoising processes [12]–[16].



(a) NCCT (b) arterial (c) portal venous (d) delayed

Fig. 1. CT images of different phases.

Despite the efficacy of these methods in general medical imaging, their application to multi-phase CECT synthesis remains to be fully explored. Spatially, while CECT enhances soft-tissue contrast, it maintains anatomical structures consistent with NCCT. The single-step mapping of GANs effectively preserves the semantic information of source images; conversely, diffusion models—which sample directly from noise—often fail to fully exploit the spatial priors of NCCT and suffer from the heavy computational overhead inherent in likelihood-based models [3], [17]. Furthermore, existing synthesis approaches typically neglect the intrinsic temporal correlations among sequentially acquired CECT phases, often relying on phase-by-phase training and generation. Such high complexity and model scales limit their utility in clinical practice.

To overcome these challenges, this paper proposes an autoregressive network for abdominal multi-phase CECT synthesis. As illustrated in Fig. 2, the framework first extracts semantic features synchronously across each downsampling level via

* G. Zhao and K. Xu are all co-corresponding authors.

an encoder. A Gated Unit decoder module then constrains the feature space by predicting soft masks, ensuring anatomical consistency and guiding the model's spatial awareness of contrast-enhanced regions. In the bottleneck layer, a Residual ConvLSTM module treats the three enhancement phases as sequential time steps to model long-term dependencies in high-level semantic features, thereby learning the temporal evolution of contrast enhancement across different phases. Additionally, a scheduled sampling strategy [18] is introduced to suppress error accumulation during the autoregressive sequential synthesis.

The main contributions of this work are summarized as follows:

- This work proposes RALSTM, an autoregressive framework specifically designed for abdominal multi-phase CECT synthesis. By enabling the synthesis of multi-phase CECT from a NCCT scan in a single training pass, RALSTM significantly reduces training iterations and parameter count. Comprehensive experiments on both internal and public datasets show that RALSTM generates high-fidelity images and generalizes robustly across diverse clinical scenarios.
- This work introduces a Region-Aware Gated Unit (RAGU) that explicitly guides the model to perceive contrast-enhancement regions. By predicting soft masks to constrain the feature space, this module ensures the preservation of anatomical structure consistency while enhancing the spatial awareness of the model.
- The proposed framework incorporates a specialized Residual ConvLSTM module to capture temporal dependencies across various CECT phases. This module serves as a novel solution to the challenge of modeling long-term contrast-agent evolution, facilitating the generation of sequential phases with superior temporal coherence.

II. RELATED WORK

A. CT Synthesis

Current CT synthesis methodologies can be broadly categorized into GAN-based and diffusion-based paradigms. For instance, Luo et al. [10] proposed a mask-guided dual-network GAN for MRI-to-CT synthesis, which first predicts a skeletal mask to guide parallel branches in synthesizing bone and soft tissue components separately. Chen et al. [11] incorporated the clinical concept of subtraction images into the NCCT-to-CECT task. Their approach utilizes a mask predictor integrated into normalization layers to highlight correlations within contrast-enhanced regions decomposed via sliding windows, while employing self-similarity maps and structural invariance losses to ensure anatomical consistency. Regarding diffusion models, Feng et al. [16] introduced a contrast-mask-prior-guided denoising process. By utilizing a thresholding method to balance vessel enhancement and noise suppression, they applied a mask-weighted loss to focus the model on vascular regions, thereby improving the fidelity of anatomical features during the conversion from NCCT to angiography.

Similarly, this work leverages subtraction-based knowledge; however, the gating mechanism in RAGU adaptively predicts contrast-enhanced regions to generate multi-scale soft masks. These masks are supervised via a dedicated loss function to impose spatial feature constraints, enabling the model to maintain anatomical integrity while focusing on enhancement regions without requiring additional pre-processing.

B. Sequential Modeling

Sequential models, such as LSTM [19], Transformer [20], and Mamba State Space Model (SSM) [21], have demonstrated exceptional capabilities in modeling long-range dependencies and integrating global contextual information. Recent studies have successfully combined these models with U-Net-like encoder-decoder architectures for medical image synthesis [22]. For example, ResViT [23] was proposed to synthesize missing sequences in multi-contrast MRI and MRI-to-CT datasets by unifying the local precision of CNNs with the global context sensitivity of Vision Transformers [24]. To capture contextual features efficiently without excessive complexity, Atli et al. [25] introduced a Mamba-based approaches that extracts task-relevant information across spatial, frequency, and channel dimensions in a dual-domain parallel fashion. Recognizing that medical image synthesis encompasses temporal sequences beyond purely spatial contexts, Chung et al. [26] utilized a ConvLSTM [27] within a 3D U-Net [28] backbone. Through a two-stage training strategy and foreground-aware cropping, it iteratively refines outputs across time steps to synthesize delayed breast MRI phases from early enhancement phases.

Inspired by these advancements, this study integrates both temporal and spatial priors for multi-phase CECT synthesis. This work employ a specialized module to model temporal contexts based on the chronological acquisition order, while adaptively introducing spatial constraints at the decoder stage via a region-aware module.

III. METHODS

As illustrated in Fig. 2, the proposed RALSTM follows a U-shaped encoder-decoder architecture. The encoding path consists of standard convolutional layers designed to extract high-level semantic features from the images. At the central bottleneck layer, a Residual ConvLSTM module explicitly models the temporal evolution across different phases, enabling the autoregressive generation of subsequent multi-phase CECT images from an input NCCT. The decoder is composed of multi-scale Region-Aware Gated Units (RAGU). Each RAGU generates a soft mask, which is utilized to compute an alignment loss against the ground-truth subtraction images. This mechanism imposes constraints on the decoder-side features to maintain strict anatomical consistency. During the training phase, a Scheduled Sampling strategy is adopted to balance convergence stability with inference robustness.

A. Encoder

The proposed encoder is composed of a stack of convolutional layers that extract hierarchical semantic representations,

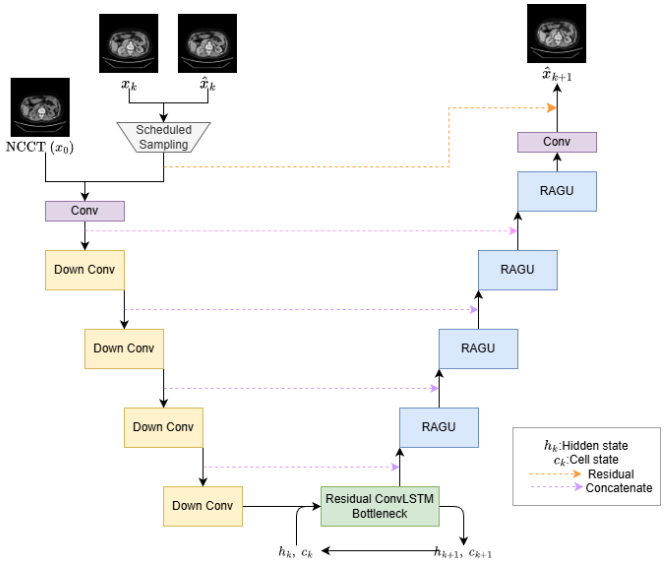


Fig. 2. Overall framework of the proposed method.

which support high-level feature modeling at the bottleneck and detailed structure recovery in the decoder. The overall framework performs multi-phase CECT synthesis in an autoregressive manner. Accordingly, the encoder is applied sequentially across four anatomical phases, denoted as $\mathcal{S} = \{x_0, x_1, x_2, x_3\}$. Specifically, each time step $k \in \{0, 1, 2\}$ corresponds to the transition from NCCT (x_0) to the arterial (x_1), portal venous (x_2), and delayed (x_3) phases, respectively.

In this autoregressive framework, the input I_k to the model at time step k is constructed by concatenating the reference NCCT image and the preceding phase's information:

$$I_k = [x_0 \oplus \tilde{x}_k] \tag{1}$$

where $\oplus$ denotes channel-wise concatenation, x_0 represents the fixed NCCT image, and $\tilde{x}_k$ is the input derived from the previous step. For the initial step ($k = 0$), $\tilde{x}_0$ is defined as x_0.

To improve the robustness of autoregressive multi-phase synthesis, the work employ Scheduled Sampling as a curriculum training strategy. Autoregressive generation inherently suffers from error accumulation, where minor prediction deviations at early phases can be amplified when propagated to subsequent phases, especially at early iterations when the predicted outputs are still unreliable. Scheduled sampling overcomes this trade-off by gradually transitioning the conditioning signal from ground truth to model-generated inputs.

During training, the model dynamically selects between the ground-truth image x_k and the predicted output $\hat{x}_k$ from step $k-1$ based on a probability ϵ. The effective input component $\tilde{x}_k$ is formulated as:

$$\tilde{x}_k = \mathbb{I} \cdot x_k + (1 - \mathbb{I}) \cdot \hat{x}_k, \quad \mathbb{I} \sim \text{Bernoulli}(\epsilon) \tag{2}$$

where the sampling probability ϵ follows an inverse sigmoid decay schedule:

$$\epsilon_i = \frac{\mu}{\mu + \exp(i/\mu)} \tag{3}$$

where i denotes the iteration index and μ is a hyperparameter controlling the decay rate.

The encoder architecture follows a symmetrical U-Net contraction path. Let $l \in \{1, 2, \ldots, L\}$ denote the index of the encoding stages, the hierarchical feature extraction process is defined as:

$$F_l = \mathcal{E}_l(F_{l-1}), \quad F_l \in \mathbb{R}^{C_l \times H_l \times W_l} \tag{4}$$

where $\mathcal{E}_l(\cdot)$ represents the l-th encoding operator comprising convolutional layers and a downsampling module. The spatial resolution at each stage is given by $H_l = H/2^l$ and $W_l = W/2^l$, where $H \times W$ denotes the initial input resolution.

These multi-scale representations $\{F_l\}_{l=1}^{L}$ provide rich semantic contexts, with the most compressed features at the bottleneck being fed into the Residual ConvLSTM module to model the spatiotemporal transitions between enhancement phases.

B. Residual ConvLSTM Bottleneck

Multi-phase CECT acquisition exhibits a natural temporal progression of contrast-agent uptake and washout, inducing strong phase-to-phase dependencies beyond a set of independent image-to-image mappings.

A Residual ConvLSTM (ResConvLSTM) is proposed at the bottleneck for explicit temporal modeling, enabling the capture of spatiotemporal transitions across contrast-enhancement regions. ConvLSTM is prioritized over standard recurrent units as its convolutional gating mechanism preserves spatial topology, which is essential for maintaining anatomical integrity. Furthermore, situating the recurrent module at the bottleneck facilitates efficient temporal reasoning within a compact yet semantically dense latent space, significantly reducing the computational cost compared to modeling dynamics at high-resolution scales.

As illustrated in Fig. 3, for each time step k, the module integrates the current spatial feature F_k with the previous hidden state h_k and cell state c_k to model the transition of enhancement patterns. The internal gated mechanism is defined as follows:

$$\begin{aligned}
i_{k+1} &= \sigma(W_{xi} * F_k + W_{hi} * h_k + b_i) \\
f_{k+1} &= \sigma(W_{xf} * F_k + W_{hf} * h_k + b_f) \\
o_{k+1} &= \sigma(W_{xo} * F_k + W_{ho} * h_k + b_o) \\
g_{k+1} &= \tanh(W_{xc} * F_k + W_{hc} * h_k + b_c) \\
c_{k+1} &= f_{k+1} \odot c_k + i_{k+1} \odot g_{k+1} \\
h'_{k+1} &= o_{k+1} \odot \tanh(c_{k+1})
\end{aligned} \tag{5}$$

where $*$ denotes the convolution operator, $\odot$ represents the Hadamard product, and σ is the sigmoid activation function. The variables i_k, f_k, o_k, and g_k represent the input gate, forget gate, output gate, and the candidate cell state, respectively.

To enhance the representational capacity and stabilize recurrent feature propagation, the intermediate hidden state h'_k undergoes a sequential dual-refinement process. Specifically, a Channel Compression (CC) module is first employed to

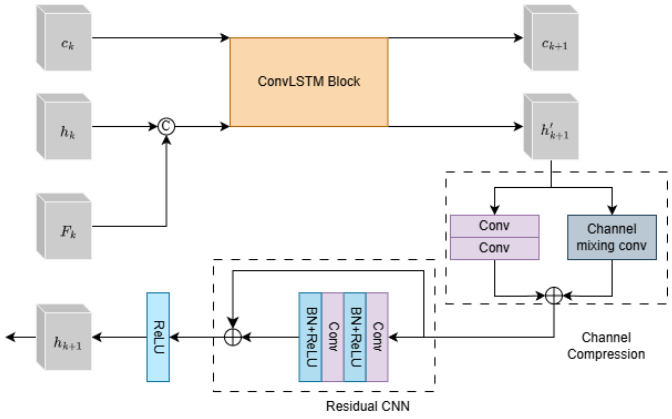


Fig. 3. Architecture of Residual ConvLSTM module.

distill salient features by reducing channel redundancy. Subsequently, a Residual CNN (ResCNN) block [29] is adopted to compensate for potential over-smoothing introduced by recurrent aggregation, while reinforcing local structural details via residual feature learning:

$$h_{k+1} = \text{ResCNN}(\text{CC}(h'_{k+1})) \tag{6}$$

This architecture achieves a joint spatial-temporal calibration, yielding a robust latent representation. The refined hidden state h_{k+1} is fed into the decoder to facilitate the synthesis of the CECT image for the current phase. Simultaneously, both h_{k+1} and the updated cell state c_{k+1} serve as the recurrent hidden representations, which are propagated to the subsequent time step $k+1$ to maintain temporal consistency throughout the autoregressive process.

C. Region-Aware Gated Units Decoder

Although multi-phase CECT synthesis aims to predict the spatiotemporal distribution of contrast-agent uptake, the underlying anatomical structures should remain consistent with the input NCCT. However, conventional decoders often modify non-enhancing tissues indiscriminately, leading to undesired structural deformation and reduced anatomical fidelity. To explicitly constrain the synthesis process to clinically meaningful enhancement regions, this work design a multi-scale Region-Aware Gated Unit (RAGU) decoder inspired by the Convolutional Block Attention Module (CBAM) [30]. The proposed RAGU selectively modulates decoder features by emphasizing enhancement-related responses while suppressing unnecessary changes in structurally stable regions.

As shown in Fig. 4, the RAGU at stage l adaptively integrates the upsampled feature map D_{l+1} from the preceding deeper layer with the high-resolution spatial features F_l from the encoder. Specifically, the concatenated representation $\hat{D}_l = [D_{l+1} \oplus F_l]$ is first refined via a channel attention module to generate a contrast-aware feature D_l^c. Subsequently, a soft anatomical mask M_l is extracted through spatial feature aggregation:

$$M_l = \sigma\left(\mathcal{C}_{7\times7}\left([\text{AvgPool}(D_l^c) \oplus \text{MaxPool}(D_l^c)]\right)\right) \tag{7}$$

where $\mathcal{C}_{7\times7}$ denotes a 7×7 convolution. To facilitate information flow while suppressing redundant modifications, the final output is formulated as a gated transformation:

$$D_l = M_l \odot D_l^c \tag{8}$$

The predicted mask M_l is explicitly supervised by the subtraction map between the target phase and the input NCCT. This supervision provides a clinically interpretable spatial prior that highlights regions with significant contrast uptake, thereby guiding the decoder to focus on enhancement-relevant structures and preventing erroneous alterations in non-enhancing tissues.

D. Loss Function

The training objective consists of four terms: adversarial loss $\mathcal{L}_{adv}$, pixel-wise reconstruction loss $\mathcal{L}_1$, mask alignment loss $\mathcal{L}_{mask}$, and perceptual loss $\mathcal{L}_{perc}$.

This work adopt a conditional PatchGAN [31] discriminator shared across all phases. For each phase k, the adversarial loss is:

$$\mathcal{L}_{adv} = \mathbb{E}_{x_k}[\log D(x_0, x_k)] + \mathbb{E}_{\hat{x}_k}[\log(1 - D(x_0, \hat{x}_k))] \tag{9}$$

where $\mathbb{E}$ denotes expectation, and D denotes the discriminator.

Pixel-level fidelity is enforced by:

$$\mathcal{L}_1 = \|\hat{x}_k - x_k\|_1 \tag{10}$$

To constrain structural changes to enhancement-related regions, this work supervise the RAGU mask M_l using a normalized subtraction prior $G_k = \text{Norm}(|x_k - x_0|)$ with BCE loss:

$$\mathcal{L}_{mask} = -\frac{1}{N}\sum_{i=1}^{N}\Big[G_{k,i}\log M_{l,i} + (1 - G_{k,i})\log(1 - M_{l,i})\Big] \tag{11}$$

This work further adopt LPIPS [32] for perceptual consistency:

$$\mathcal{L}_{perc} = \sum_j \frac{1}{H_j W_j}\|\phi_j(\hat{x}_k) - \phi_j(x_k)\|_2^2 \tag{12}$$

where $\phi_j(\cdot)$ denotes the feature map from the j-th layer of a pre-trained network [33].

The overall objective is:

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{adv} + \lambda_2\mathcal{L}_1 + \lambda_3\mathcal{L}_{mask} + \lambda_4\mathcal{L}_{perc} \tag{13}$$

where λ_1, λ_2, λ_3 and λ_4 are hyperparameters that regulate the relative contribution of each loss component.

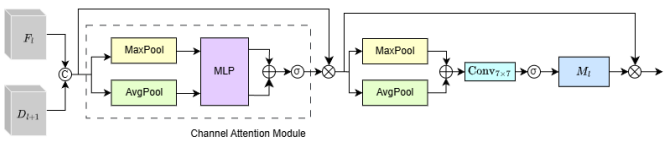


Fig. 4. Structure of RAGU.

IV. EXPERIMENTS

A. Datasets

1) Internal Dataset: To evaluate the effectiveness of the proposed model for multi-phase CECT synthesis, a clinical dataset was collected from the collaborating hospital. This dataset comprises sequential four-phase CT scans from 363 patients, including NCCT (N), arterial (A), portal venous (V), and delayed (D) phases. The cohort was partitioned at the patient level: 290 patients (13,920 slices) were allocated for training, 36 patients (1,728 slices) for validation, and 37 patients (1,776 slices) for testing. To eliminate spatial misalignments between phases caused by patient motion or respiration, all contrast-enhanced volumes were registered to their corresponding NCCT templates using the Advanced Normalization Tools (ANTs) [34]. Post-registration, the CT volumes were normalized based on clinical windowing settings (Hounsfield Units) and resized to a 512×512 resolution to meet the input requirements of the deep learning model.

2) External Test Set: This work utilized an open-access dataset released by Ristea et al. [35], which is specifically designed for NCCT-to-CECT translation. This dataset contains triple-phase lung CT scans (N, A, and V phases). This work selected a subset of 52 patients (2,496 slices) as an independent external testing cohort. The same preprocessing pipeline—including registration and intensity mapping—was applied to ensure consistency with the internal dataset.

B. Implementation Details

The proposed framework was implemented using PyTorch and CUDA, with experiments conducted on a computing server equipped with an NVIDIA A800 GPU (80 GB memory). For the optimization of the generator and discriminator, this work utilized the Adam optimizer with momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The model was trained for 100 epochs at a constant learning rate of 0.0002. Considering the total training duration, the parameter μ in (3) was set to 15. The hyperparameters for the loss functions in (13) were empirically determined as $\lambda_1 = 1.0$, $\lambda_2 = 1.0$, $\lambda_3 = 1.0$, and $\lambda_4 = 0.5$ to balance the adversarial, pixel-wise, mask alignment, and perceptual constraints. To quantitatively assess the synthesis performance, this work employed four widely recognized metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Mean Absolute Error (MAE).

C. Experimental Results

1) Comparison Settings: To demonstrate the superiority of the proposed framework, RALSTM was compared with several existing methods, including pix2pix [31], CycleGAN [36], DINO [37], F-LSeSim [38], CyTran [35], MSA-GAN [39] across two datasets. All methods were run with default hyperparameters and identical data splits to ensure fair comparison.

2) Synthesis Results Comparison of Different Models: Unlike this proposed framework, all comparative methods were trained independently three times to generate the three specific contrast-enhanced phases. Tables I and II summarize the quantitative evaluations on both internal and external datasets, indicating that RALSTM consistently outperforms all competing methods. Qualitative results on the internal dataset, as shown in Fig. 5, highlight the method's effectiveness in preserving anatomical details and accurately modeling the temporal dynamics of contrast enhancement. Moreover, visual comparisons on the external test set in Fig. 6 further validate the superior performance and robust generalization capability of this proposed framework.

TABLE I
QUANTITATIVE COMPARISON ON INTERNAL TEST SET. THE BEST THREE RESULTS ARE HIGHLIGHTED IN **BOLD**, UNDERLINED AND *italic*.

Methods	N-to-A			N-to-V			N-to-D		
	PSNR↑	SSIM↑	MAE↓	PSNR↑	SSIM↑	MAE↓	PSNR↑	SSIM↑	MAE↓
pix2pix [31]	21.6260	0.8394	0.0074	20.9755	0.8330	0.0088	21.2384	0.8293	*0.0082*
CycleGAN [36]	22.2472	*0.8510*	0.0067	*21.1971*	**0.8440**	*0.0086*	21.2120	0.8405	0.0084
DINO [37]	*22.1533*	**0.8523**	0.0078	22.0005	*0.8417*	0.0082	22.0854	0.8370	0.0079
F-LSeSim [38]	22.1349	0.8504	*0.0071*	20.7641	0.8348	0.0101	*21.7666*	*0.8374*	0.0083
CyTran [35]	21.0513	0.8454	0.0093	20.7736	0.8410	0.0093	21.2880	0.8350	0.0083
MSA-GAN [39]	21.7069	0.8120	0.0088	20.7004	0.8103	0.0100	20.7602	0.8085	0.0096
RALSTM	**22.7249**	0.8514	**0.0065**	**22.1055**	0.8421	**0.0079**	**22.2192**	**0.8408**	**0.0076**

TABLE II
QUANTITATIVE COMPARISON ON EXTERNAL TEST SET. THE BEST THREE RESULTS ARE HIGHLIGHTED IN **BOLD**, UNDERLINED AND *italic*.

Methods	N-to-A			N-to-V		
	PSNR↑	SSIM↑	MAE↓	PSNR↑	SSIM↑	MAE↓
pix2pix [31]	20.1398	0.7549	0.0120	19.9228	0.7427	0.0132
CycleGAN [36]	20.5755	0.7718	0.0165	20.4111	0.7583	**0.0111**
DINO [37]	*20.2749*	**0.7769**	0.0110	20.1513	*0.7555*	*0.0118*
F-LSeSim [38]	20.0840	0.7708	*0.0115*	19.3774	0.7532	0.0135
CyTran [35]	19.9611	*0.7733*	0.0121	*20.3706*	**0.7639**	*0.0118*
MSA-GAN [39]	19.5233	0.7353	0.0133	18.1565	0.7102	0.0188
RALSTM	**20.9359**	0.7754	**0.0105**	**20.5242**	0.7509	0.0115

Table III summarizes the computational resource consumption in terms of parameter size, FLOPs, memory usage, and training efficiency. For other methods, the reported metrics represent the cumulative requirements of three independent models trained specifically for each CECT phase. In contrast, the proposed RALSTM encapsulates the multi-phase synthesis task within a unified framework, achieving a competitive parameter size of 38.72 M, which ranks third among all compared methods. Notably, RALSTM exhibits superior training efficiency with a Time/Epoch of 0.76 hours, outperforming most decoupled multi-stage frameworks (e.g., MSA-GAN and F-LSeSim). Although the autoregressive nature of the module introduces a slight increase in computational latency compared to simple single-step models like pix2pix, the overall resource demand remains well within the acceptable range for practical clinical workflows, enabling high-fidelity sequence generation without excessive hardware overhead.

3) Ablation Study: As shown in Table IV, this work validated the contribution of each component. Architecturally, removing the RAGU Decoder or Internal Residuals degrades performance, verifying their roles in anatomical preservation and feature propagation. Notably, eliminating the Res-ConvLSTM Bottleneck causes the most significant drop, confirming that modeling temporal dependencies is critical. Regarding training strategies, the absence of Mask Super-

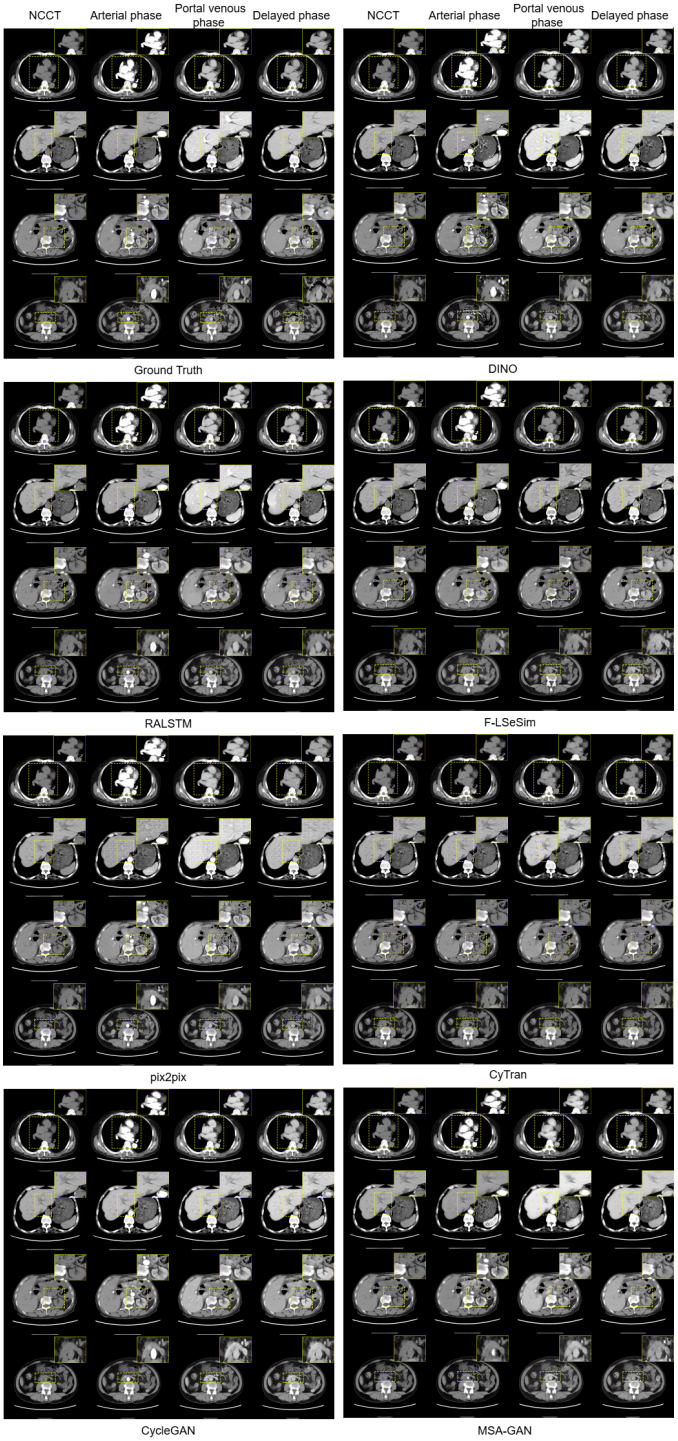


Fig. 5. Qualitative results on the internal dataset. The magnified patches highlight the contrast-enhanced regions of the synthetic CT images.

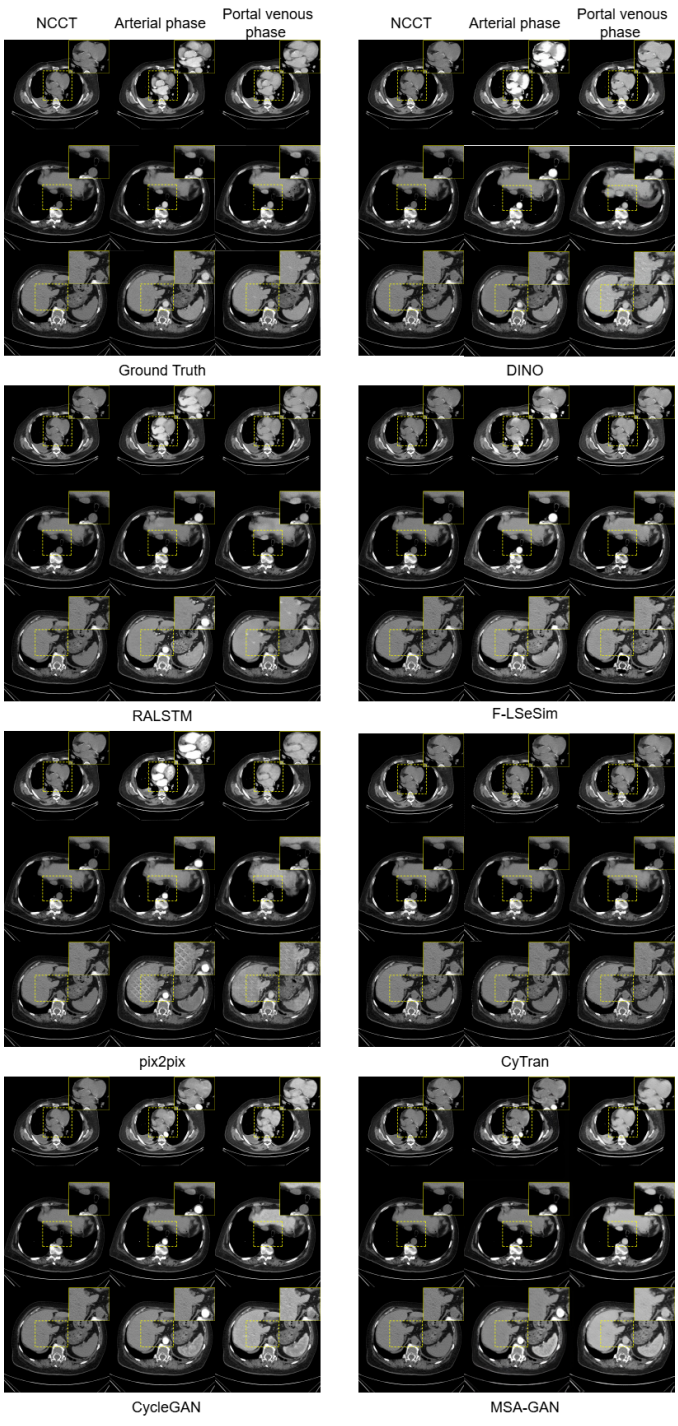


Fig. 6. Qualitative results on the external dataset. The magnified patches highlight the contrast-enhanced regions of the synthetic CT images.

TABLE III

COMPUTATIONAL RESOURCE DEMANDS OF EXISTING METHODS AND THE PROPOSED METHOD. THE TOP THREE RESULTS ARE HIGHLIGHTED IN **BOLD**, UNDERLINE, AND *italic*.

Methods	Params (M)	FLOPs (G)	Memory (MB)	Time/Epoch (h)
pix2pix [31]	163.22	*214.20*	**1088.64**	**0.20**
CycleGAN [36]	68.20	672.51	4555.26	4.53
DINO [37]	146.32	127.10	*1881.21*	*0.78*
F-LSeSim [38]	74.10	759.71	6751.05	2.24
CyTran [35]	**7.15**	**25.16**	1772.10	1.44
MSA-GAN [39]	18.24	262.05	6519.90	16.45
RALSTM	*38.72*	287.98	3290.36	0.76

TABLE IV

ABLATION STUDY OF KEY COMPONENTS AND STRATEGIES. THE RESULT IS THE AVERAGE OF THE THREE PHASES, AND THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Model Variants	PSNR ↑	SSIM ↑	MAE ↓
w/o RAGU Decoder	21.6717	0.8364	0.0085
w/o Residual ConvLSTM Bottleneck	20.7312	0.8279	0.0093
w/o Internal Residual Refinement	22.2459	0.8387	0.0081
w/o Mask Supervision ($\mathcal{L}_{mask} = 0$)	22.1872	0.8371	0.0079
w/o Scheduled Sampling	21.8518	0.8373	0.0084
RALSTM (Proposed)	**22.3499**	**0.8448**	**0.0073**

vision impairs structural fidelity, while removing Scheduled Sampling (i.e., reverting to teacher forcing) exacerbates exposure bias, reducing long-term accuracy. Consequently, the full RALSTM yields the best results across all metrics.

4) Visual Analysis: To intuitively assess the interpretability of the proposed Region-Aware Gated Units (RAGU), this work visualize the intermediate attention maps in Fig. 7. The ground truth subtraction image is derived by calculating the maximum pixel-wise difference across the three real contrast phases to represent the ideal enhancement regions. The learned soft mask is the average of the masks generated by the RAGU module across the three synthesized phases. As observed, the learned mask exhibits a high correlation with the ground truth subtraction image, accurately highlighting contrast-enhanced regions (e.g., vessels and organs) while suppressing the background. This confirms that the model successfully captures the anatomical prior of contrast agent uptake.

Furthermore, to evaluate the distributional alignment between the synthesized and real images, this work conducted a t-SNE visualization [40], as shown in Fig. 8. To ensure the feature representation is domain-specific, this work employed a ResNet50 [29] backbone pre-trained on the internal dataset to extract high-level semantic features. The visualization demonstrates that the feature clusters of the synthetic images closely overlap with those of the real images, indicating that the proposed RALSTM effectively models the complex manifold of multi-phase CECT data and generates images with realistic high-level features.

V. CONCLUSION

This paper presented RALSTM, a novel autoregressive framework designed for the sequential synthesis of abdominal

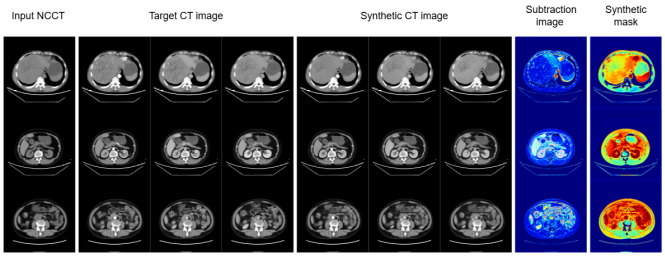


Fig. 7. Visualization of the learned attention masks. The Ground Truth Subtraction represents the maximum intensity difference across real phases, while the Learned Soft Mask is the average output of the RAGU module.

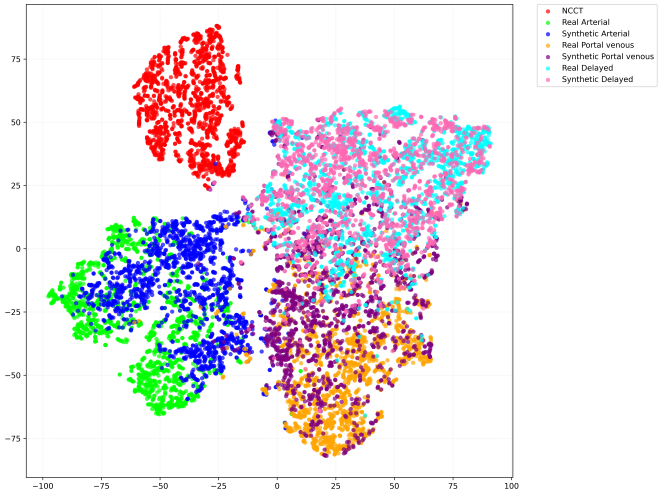


Fig. 8. The t-SNE feature space visualization of real and synthetic images. High-level features were extracted using a ResNet50 pre-trained on the internal dataset.

multi-phase CECT images from a single NCCT scan. To overcome the limitations of existing methods in anatomical structure preservation and temporal modeling, this work effectively integrates clinical domain knowledge into the deep learning architecture. Specifically, this work utilize clinical subtraction maps as explicit spatial priors to guide the Region-Aware Gated Unit, ensuring precise localization of contrast-enhancement regions. Additionally, the inherent sequential acquisition order of CECT phases serves as strong temporal priors, which are modeled by the Residual ConvLSTM to capture the dynamic evolution of contrast agent uptake. This synergistic design enables the framework to autoregressively generate high-fidelity multi-phase sequences. Extensive experiments on internal and external datasets demonstrate that RALSTM outperforms leading methods in terms of both quantitative metrics and visual quality, showcasing its promising application in personalized radiotherapy planning and clinical diagnosis.

ACKNOWLEDGMENT

This work was supported by the Guangdong Basic and Applied Basic Research Foundation (Guangdong-Hong Kong-Macao National Center for Applied Mathematics: Re-

search on the Interpretability of Medical Artificial Intelligence Technologies), National Social Science Foundation of China (19ZDA041), National Natural Science Foundation of China (82072841), Natural Science Foundation of Guangdong Province (2021A1515010199), Guangdong S&T Program (2023B1111030006).

REFERENCES

[1] L. Zhong, R. Xiao, H. Shu, K. Zheng, X. Li, Y. Wu, J. Ma, Q. Feng, and W. Yang, "Ncct-to-cect synthesis with contrast-enhanced knowledge and anatomical perception for multi-organ segmentation in non-contrast ct images," *Medical Image Analysis*, vol. 100, p. 103397, 2025.

[2] K. T. Bae, "Intravenous contrast medium administration and scan timing at ct: considerations and approaches," *Radiology*, vol. 256, no. 1, pp. 32–61, 2010.

[3] T. A. Rose Jr and J. W. Choi, "Intravenous imaging contrast media complications: the basics that every clinician needs to know," *The American journal of medicine*, vol. 128, no. 9, pp. 943–949, 2015.

[4] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[5] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[7] Y. Chen, H. Lin, W. Zhang, W. Chen, Z. Zhou, A. A. Heidari, H. Chen, and G. Xu, "Icycle-gan: Improved cycle generative adversarial networks for liver medical image generation," *Biomedical Signal Processing and Control*, vol. 92, p. 106100, 2024.

[8] F. Wang, Q. Zhang, Q. Zhao, M. Wang, and F. Sun, "Unsupervised image-to-image translation with multiscale attention generative adversarial network," *Applied Intelligence*, vol. 54, no. 8, pp. 6558–6578, 2024.

[9] X. Liu, B. Zeng, S. Gao, S. Li, Y. Feng, H. Li, B. Liu, J. Liu, and B. Zhang, "Ladiffgan: Training gans with diffusion supervision in latent spaces," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1115–1125.

[10] Y. Luo, S. Zhang, J. Ling, Z. Lin, Z. Wang, and S. Yao, "Mask-guided generative adversarial network for mri-based ct synthesis," *Knowledge-Based Systems*, vol. 295, p. 111799, 2024.

[11] W. Chen, W. Zhao, Z. Chen, T. Liu, L. Liu, J. Liu, and Y. Yuan, "Mask-aware transformer with structure invariant loss for ct translation," *Medical Image Analysis*, vol. 96, p. 103205, 2024.

[12] Y. Zhou, T. Chen, J. Hou, H. Xie, N. C. Dvornek, S. K. Zhou, D. L. Wilson, J. S. Duncan, C. Liu, and B. Zhou, "Cascaded multi-path shortcut diffusion model for medical image translation," *Medical Image Analysis*, vol. 98, p. 103300, 2024.

[13] J. Kim and H. Park, "Adaptive latent diffusion model for 3d medical image to image translation: Multi-modal magnetic resonance imaging study," in *Proceedings of the IEEE/CVF Winter conference on applications of computer Vision*, 2024, pp. 7604–7613.

[14] X. Xiao, Q. V. Hu, and G. Wang, "Pyramid hierarchical masked diffusion model for imaging synthesis," in *2025 International Joint Conference on Neural Networks (IJCNN)*, 2025, pp. 1–8.

[15] H. Jiang, M. Imran, T. Zhang, Y. Zhou, M. Liang, K. Gong, and W. Shao, "Fast-ddpm: Fast denoising diffusion probabilistic models for medical image-to-image generation," *IEEE Journal of Biomedical and Health Informatics*, 2025.

[16] L. Feng, L. Guo, W. He, X. Ye, Y. Pan, S. Gao, and R. Zhang, "Diffcta: Diffusion model-based non-contrast ct angiography with contrast-enhanced mask guidance," *Biomedical Signal Processing and Control*, vol. 112, p. 108645, 2026.

[17] S. Dayarathna, K. T. Islam, S. Uribe, G. Yang, M. Hayat, and Z. Chen, "Deep learning based synthesis of mri, ct and pet: Review and analysis," *Medical image analysis*, vol. 92, p. 103046, 2024.

[18] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," *Advances in neural information processing systems*, vol. 28, 2015.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[21] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *First conference on language modeling*, 2024.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[23] O. Dalmaz, M. Yurt, and T. Çukur, "Resvit: Residual vision transformers for multimodal medical image synthesis," *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2598–2614, 2022.

[24] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[25] O. F. Atli, B. Kabas, F. Arslan, A. C. Demirtas, M. Yurt, O. Dalmaz, and T. Cukur, "I2i-mamba: Multi-modal medical image synthesis via selective state space modeling," *arXiv preprint arXiv:2405.14022*, 2024.

[26] W. Chung, J. Kang, G. E. Park, S. H. Kim, and Y. Nam, "Synthesizing delayed-phase contrast-enhanced breast mr images from early-phase images using an iterative deep network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2025, pp. 583–592.

[27] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

[28] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 424–432.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[30] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[31] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[32] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[34] B. B. Avants, N. Tustison, G. Song *et al.*, "Advanced normalization tools (ants)," *Insight j*, vol. 2, no. 365, pp. 1–35, 2009.

[35] N.-C. Ristea, A.-I. Miron, O. Savencu, M.-I. Georgescu, N. Verga, F. S. Khan, and R. T. Ionescu, "Cytran: A cycle-consistent transformer with multi-level consistency for non-contrast to contrast ct translation," *Neurocomputing*, vol. 538, p. 126211, 2023.

[36] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[37] K. Vougioukas, S. Petridis, and M. Pantic, "Dino: A conditional energy-based gan for domain translation," *arXiv preprint arXiv:2102.09281*, 2021.

[38] C. Zheng, T.-J. Cham, and J. Cai, "The spatially-correlative loss for various image translation tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 407–16 417.

[39] F. Wang, Q. Zhang, Q. Zhao, M. Wang, and F. Sun, "Unsupervised image-to-image translation with multiscale attention generative adversarial network," *Applied Intelligence*, vol. 54, no. 8, pp. 6558–6578, 2024.

[40] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.