IEEE
ICASSP 2024 KOREA

RESIDUAL DENSE SWIN TRANSFORMER FOR CONTINUOUS DEPTH-INDEPENDENT ULTRASOUND IMAGING

Jintong Hu, Hui Che, Zishuo Li, Wenming Yang

Tsinghua Shenzhen International Graduate School, Tsinghua University

Problem Statement

Ultrasound imaging serves as a pivotal tool in medical diagnostics for its non-invasive nature and real-time imaging capabilities, allowing visualization of superficial and deep structures. However, adjusting the imaging depth presents challenges that impact image quality and field-of-view, making it difficult to achieve depth-continuous imaging.

Goal

Achieving depth-independent imaging by post-processing algorithms.

Motivation

Specific shortcoming:

1. Ultrasound imaging depth adjustment compromises temporal resolution and image quality due to echo reception time limitations and signal interference.
2. Traditional zoom-in operation in ultrasonic devices using interpolation result in loss of detail and artifacts.

To address these limitations, we introduce the Residual Dense Swin Transformer Network (RDSTN), which integrates a linear embedding layer, a Residual Dense Shifted-window Transformer (RDST) encoder, and an Multi-Layer Perceptron (MLP).

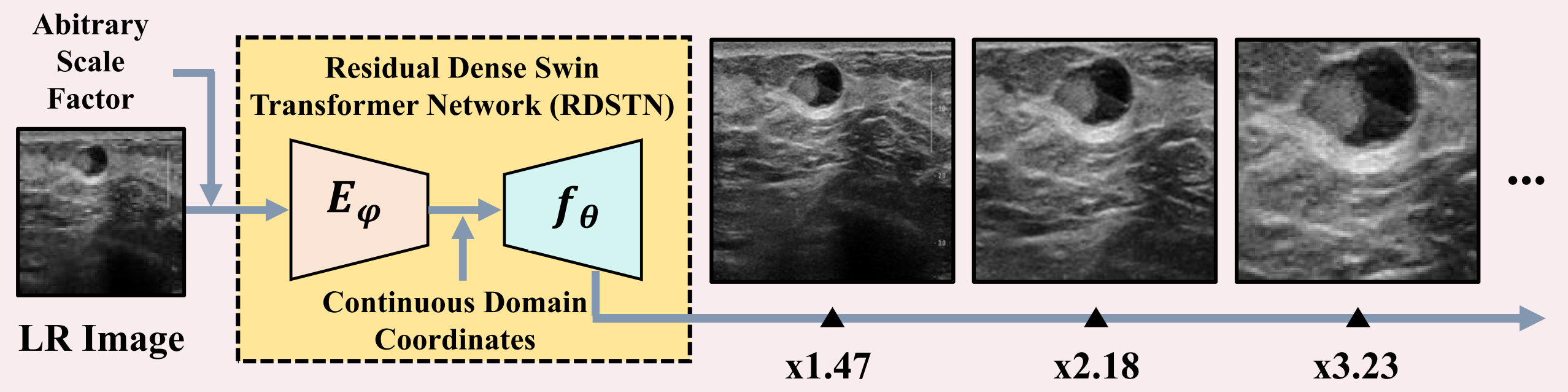
Abitrary Scale Factor
LR Image
Residual Dense Swin Transformer Network (RDSTN)
E_φ
f_θ
Continuous Domain Coordinates
x1.47
x2.18
x3.23
...

Fig. 1. **An example of SR images of abitrary scales generated by RDSTN**. RDSTN can achieve super-resolution of arbitrary scale using only single model.

Proposed RDSTN Model

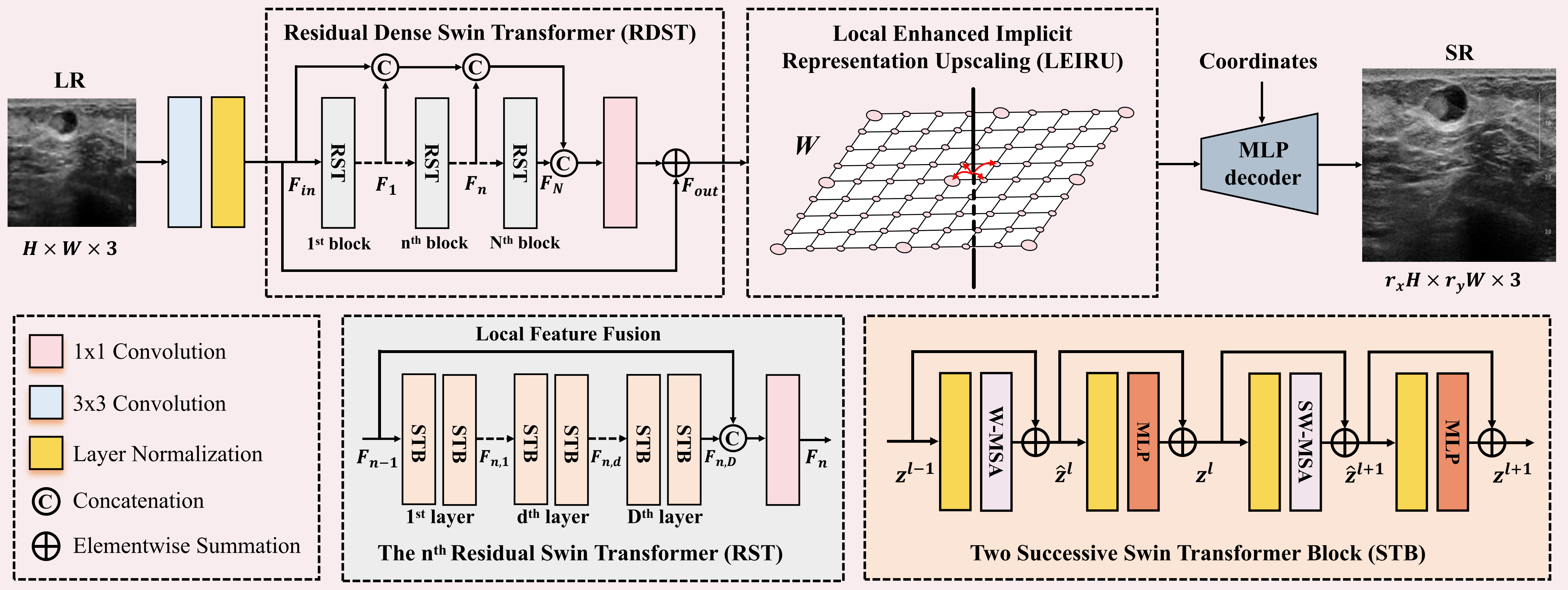

Fig. 2. **The main pipeline of our proposed RDSTN**. RDSTN consists of an encoder which extracts non-locality and a decoder that performs local enhancement. The combination of locality and non-locality improves the representation and performance of the model.

Residual Dense Swin Transformer Encoder:

The RDST encoder, based on the Swin Transformer block, maintains a fixed resolution and number of channels at each stage's output, ensuring a one-to-one correspondence between pixels in LR images and their respective latent codes. A key innovation in the RDST encoder is the fusion of local and global features, enabling the model to retain essential contextual information throughout its processing stages.

Global feature fusion: $F_{out} = F_{in} + Conv_{1\times1}([F_{in}, F_1, \ldots, F_N])$ (1)

Local feature fusion: $F_n = Conv_{1\times1}([F_{n-1}, STB^D(F_{n-1})])$ (2)

Local Enhanced Implicit Representation Upscaling:

The LEIRU decoder, operates on the coordinates of the target high-resolution image, aligning each coordinate with its nearest latent code. The RGB values for each coordinate xq are defined by: $RGB(x_q) = MLP([C(x_q), x_q - x^*)])$ (3)

where $C(x_q)$ is the nearest latent code to x_q, and x^* is the coordinate of $C(x_q)$. The relative distance between the coordinate and its nearest latent code serves as a measure of feature similarity.

Local ensemble operation: $\text{LEIRU}(x_q) = \sum_{x_i \in grid} w_{x_i} RGB(x_i)$ (4)

where w_{x_i} represents the weights for each coordinate x_i. The non-local encoder's design plays a crucial role in infusing non-locality into the local decoder, optimizing the model's performance.

Experiments and Results

Table 1. **Quantitative comparison in terms of PSNR(dB)**. The evaluation is performed on the BUSI testing set. The models are trained with continuous scale sampled from U(1, 4). Best result of each scale is in **bold**.

Methods	Num. of Parameters	In-distribution							Out-distribution		
		×1.6	×1.7	×1.8	×1.9	×2	×3	×4	×6	×8	×10
Bicubic	–	40.21	39.36	38.88	38.21	38.68	33.17	30.40	26.88	24.86	23.64
EDSR-LIIF [16]	496.4K	43.92	43.06	42.26	41.50	40.80	35.80	32.87	29.42	27.34	26.04
RDN-LIIF [16]	5.8M	44.71	43.81	43.03	42.28	41.57	**36.36**	**33.22**	29.58	27.46	26.12
Unet [20]	31.4M	42.39	41.71	41.05	40.42	39.83	35.24	32.55	29.20	27.16	25.91
Resnet50 [21]	4.1M	42.86	42.07	41.35	40.62	39.95	35.17	32.46	29.12	27.11	25.87
RDSTN (ours)	3.2M	**44.78**	**43.89**	**43.10**	**42.35**	**41.62**	36.34	33.20	**29.64**	**27.54**	**26.18**

Table 2. **Ablation study of RDSTN on Local Feature Fusion (LFF) and Global Feature Fusion (GFF)**. The evaluation is performed on the BUSI testing set to assess the performance of these strategies. The best result of each scale is in **bold**.

Model Settings	Module		In-distribution							Out-distribution		
	LFF	GFF	×1.2	×1.4	×1.6	×1.8	×2	×3	×4	×6	×8	×10
S_1	✗	✗	48.65	46.17	44.36	42.69	41.21	36.04	33.01	29.51	27.42	26.07
S_2	✗	✓	48.71	46.23	44.42	42.73	41.27	36.07	33.03	29.54	27.46	26.11
S_3	✓	✗	48.89	46.40	44.61	42.94	41.46	36.23	33.13	29.59	27.50	26.14
S_4	✓	✓	**49.27**	**46.62**	**44.78**	**43.10**	**41.62**	**36.34**	**33.20**	**29.64**	**27.54**	**26.18**

Table 3. **Generalization test of RDSTN on out-distribution dataset**. The evaluation is performed on the USenhance dataset. The best result is in **bold**.

Method	scale				
	×1.6	×1.7	×1.8	×1.9	×2
train: BUSI [23], test: MICCAI USenhance breast					
Bicubic	34.28	33.52	31.55	31.23	31.66
EDSR-LIIF	35.30	34.55	33.63	33.07	32.42
RDN-LIIF	35.12	34.41	33.56	33.13	32.47
RDSTN (ours)	**35.35**	**34.62**	**33.74**	**33.23**	**32.59**
train: BUSI [23], test: MICCAI USenhance thyroid					
Bicubic	38.17	37.14	34.91	34.26	34.81
EDSR-LIIF	39.87	38.73	37.72	36.86	36.07
RDN-LIIF	39.91	38.77	37.76	36.90	36.10
RDSTN (ours)	**39.99**	**38.81**	**37.83**	**36.96**	**36.14**
train: BUSI [23], test: MICCAI USenhance carotid					
Bicubic	38.11	37.14	34.95	34.37	34.84
EDSR-LIIF	40.05	38.96	37.90	37.07	36.21
RDN-LIIF	40.08	38.98	37.89	37.11	36.28
RDSTN (ours)	**40.16**	**39.07**	**37.99**	**37.17**	**36.30**

Conclusion

- Our advanced RDSTN effectively tackles the challenges associated with long-range modeling and non-local feature extraction in arbitrary-scale super-resolution. It streamlines the delicate balance between image quality and field-of-view, showcasing enhanced noise suppression capabilities.
- Testing reveals that RDSTN performs competitively in both metrics and visual quality compared to other methods, yet utilizes fewer parameters.
- Through RDSTN, we can adeptly navigate continuous imaging at suitable depth thresholds.