

Local Implicit Wavelet Transformer for Arbitrary-Scale Super-Resolution

Minghong Duan^{1,2}

22111010025@m.fudan.edu.cn

Linhao Qu^{1,2}

lhqu20@fudan.edu.cn

Shaolei Liu^{1,2}

19111010029@fudan.edu.cn

Manning Wang^{1,2*}

mnwang@fudan.edu.cn

¹ Digital Medical Research Center

Fudan University

Shanghai, China

² Shanghai Key Lab of Medical

Image Computing and Computer

Assisted Intervention

1 Haar Wavelet Transform

The Haar wavelet transform is a simple and computationally efficient method for decomposing input signals into low-frequency and high-frequency sub-bands, widely employed in the field of computer vision [9, 8, 10, 14, 16, 19]. In this paper, we utilize the Haar wavelet transform to perform the discrete wavelet transform (DWT) on the features Z obtained from the encoder. The Haar wavelet transform typically involves processing the input signal with high-pass filter H^T and low-pass filter L^T to obtain different sub-bands. Specifically, the low-pass and high-pass filters are:

$$\mathbf{L}^T = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \end{bmatrix}, \quad \mathbf{H}^T = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \end{bmatrix} \quad (1)$$

Similarly, the filters of the Haar wavelet transform consist of four 2×2 kernels, including LL^T , LH^T , HL^T and HH^T . In this paper, we use LL^T to process the feature Z to obtain the low-frequency component LL , and respectively use LH^T , HL^T and HH^T to process the feature Z to obtain the high-frequency components LH , HL and HH . Following the wavelet transform, the low-frequency component exhibits smooth surface and texture information, while the high-frequency components capture more complex texture details. We denote the low-frequency component LL as F^L and concatenate the high-frequency components LH , HL , and HH along the channel axis, represented as F^H .

2 Analysis on the training strategy

We analyzed the impact of training strategies on the DIV2K dataset [11]. As shown in Table 1, training within a small scale sampling range of $s \sim U(1, 4)$ enables the model to achieve good

*Corresponding Author.

© 2024. The copyright of this document resides with its authors.

It may be distributed unchanged freely in print or electronic forms.

performance at small upscaling scales but sacrifices reconstruction accuracy at higher scaling factors. Expanding the scale sampling range to $s \sim U(1, 8)$ during training can enhance the model’s reconstruction performance at larger upscaling scales but decrease performance at smaller upscaling scales. Our proposed curriculum learning strategy gradually expands the sampling range during training. Although the performance at scaling factors of $\times 2$ and $\times 3$ is not as good as training within the small-scale sampling range of $s \sim U(1, 4)$, overall, it achieves the best balance across different scaling factors. Our training approach ensures effective reconstruction at large sampling scales and achieves the most optimal or suboptimal results across various scaling factors. To validate the generality of the proposed curriculum learning training strategy, we applied the same training setup to LIIF [8] and LTE [9]. As shown in Table 2, we observed performance improvements, indicating the effectiveness and generalizability of our training strategy.

Training strategy	$\times 2$	$\times 3$	$\times 4$	$\times 6$	$\times 8$	$\times 12$	$\times 24$	$\times 27$	$\times 30$
Curriculum learning strategy	34.79	31.12	29.15	26.91	25.55	23.86	21.3	20.92	20.6
Training with $U(1, 4)$	34.84	31.13	29.14	26.89	25.52	23.83	21.27	20.89	20.57
Training with $U(1, 8)$	34.78	31.11	29.14	26.91	25.55	23.86	21.31	20.92	20.6

Table 1: The average PSNR (dB) of different training strategies on the DIV2K validation set [4].

Training strategy	Method	$\times 2$	$\times 3$	$\times 4$	$\times 6$	$\times 12$	$\times 18$
Original	EDSR-LIIF [8]	34.67	30.96	29	26.75	23.71	22.17
	EDSR-LTE [9]	34.72	31.02	29.04	26.81	23.78	22.23
Curriculum	EDSR-LIIF [8]	34.65(+0.02)	30.99(+0.03)	29.05(+0.05)	26.81(+0.06)	23.77(+0.06)	22.22(+0.05)
	EDSR-LTE [9]	34.7(+0.02)	31.03(+0.01)	29.07(+0.03)	26.85(+0.04)	23.82(+0.04)	22.28(+0.05)

Table 2: The average PSNR (dB) of LIIF [8] and LTE [9] training with and without curriculum learning on the DIV2K validation set [4].

3 More evaluation metrics

We employ two metrics, SSIM and LPIPS, to further demonstrate the effectiveness of LIWT compared to other arbitrary-scale SR methods. We compare the performance of LIWT, LTE [9], LIIF [8], and MetaSR [6] using SwinIR as the encoder on Set14 [14] and Urban100 [10] datasets. Typically, higher SSIM and lower LPIPS correspond to better performance. From the results in Table 3, it can be observed that except for $\times 2$ scaling, our method achieves the highest SSIM and the lowest LPIPS. This indicates that our approach can recover more structural information and has better perceptual quality.

Dataset	Method	$\times 2$		$\times 3$		$\times 4$		$\times 6$		$\times 8$	
		SSIM \uparrow	LPIPS \downarrow	SSIM \uparrow	LPIPS \downarrow	SSIM \uparrow	LPIPS \downarrow	SSIM \uparrow	LPIPS \downarrow	SSIM \uparrow	LPIPS \downarrow
Set14 [14]	Meta-SR [6]	0.923	0.134	0.850	0.227	0.791	0.291	0.704	0.382	0.648	0.446
	LIIF [8]	0.923	0.134	0.851	0.227	0.792	0.293	0.707	0.380	0.652	0.443
	LTE [9]	0.924	0.133	0.852	0.224	0.794	0.292	0.709	0.377	0.655	0.440
	LIWT(Ours)	0.924	0.132	0.853	0.223	0.795	0.289	0.712	0.373	0.657	0.436
Urban100 [10]	Meta-SR [6]	0.939	0.102	0.873	0.186	0.810	0.251	0.709	0.347	0.638	0.415
	LIIF [8]	0.939	0.102	0.876	0.188	0.817	0.258	0.719	0.349	0.650	0.415
	LTE [9]	0.941	0.100	0.877	0.186	0.820	0.254	0.722	0.343	0.653	0.408
	LIWT(Ours)	0.941	0.099	0.878	0.183	0.821	0.250	0.726	0.336	0.657	0.401

Table 3: Comparison of more evaluation metrics (SSIM \uparrow and LPIPS \downarrow) on Set14 [14] and Urban100 [10].

4 Comparison with DWT-based SR methods

We compare LIWT with other DWT-based SR methods [8, 9, 10, 11, 16, 19] using PSNR and SSIM metrics on Set14 [11] and Urban100 [9], where LIWT utilizes SwinIR [10] as the encoder. As shown in Table 4, our LIWT achieves the best results at various scaling factors.

Dataset	Method	×2		×3		×4	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Set14 [11]	MWCNN [8]	33.71	0.918	30.14	0.841	28.58	0.788
	DWSR [9]	33.07	0.911	29.83	0.831	28.04	0.767
	WRAN [10]	34.21	0.922	30.71	0.852	28.60	0.786
	WDRN [11]	33.90	0.921	30.50	0.845	28.75	0.786
	JWSGN [16]	34.17	0.923	-	-	28.96	0.789
	WaveMixSR [19]	31.27	0.904	28.77	0.841	26.25	0.751
	LIWT(Ours)	34.31	0.924	30.86	0.853	29.05	0.795
Urban100 [9]	MWCNN [8]	32.36	0.931	28.19	0.852	26.37	0.789
	DWSR [9]	30.46	0.916	-	-	25.26	0.755
	WRAN [10]	33.47	0.940	28.99	0.869	26.74	0.803
	WDRN [11]	32.64	0.937	28.59	0.862	26.41	0.797
	JWSGN [16]	33.17	0.938	-	-	26.82	0.807
	WaveMixSR [19]	29.14	0.908	25.82	0.819	23.57	0.730
	LIWT(Ours)	33.52	0.941	29.46	0.878	27.30	0.821

Table 4: Comparison of PSNR↑ and SSIM↑ for different DWT-based methods on Set14 [11] and Urban100 [9].

5 Comparison of different arbitrary-scale SR methods at non-integer scale

To further assess the advantages of our method over other arbitrary-scale SR methods, we present comparative results of PSNR and SSIM metrics at non-integer scaling factors on Set14 [11] and Urban100 [9]. As shown in Table 5, Our LIWT achieves optimal results at various scaling factors.

Dataset	Method	×2.2		×2.5		×3.3		×3.5		×4.4		×5.5		×6.6		×7.7	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Set14 [11]	Meta-SR [1]	31.84	0.899	31.48	0.883	29.14	0.825	29.26	0.813	28.18	0.768	26.92	0.720	26.14	0.684	25.32	0.654
	LIIF [1]	31.91	0.899	31.54	0.884	29.44	0.825	29.29	0.814	28.25	0.770	27.02	0.722	26.25	0.688	25.41	0.658
	LTE [1]	31.93	0.900	31.55	0.884	29.48	0.826	29.30	0.815	28.29	0.771	27.07	0.724	26.30	0.689	25.50	0.661
	LIWT(Ours)	31.95	0.900	31.60	0.885	29.49	0.827	29.34	0.816	28.31	0.772	27.08	0.725	26.33	0.691	25.50	0.662
Urban100 [9]	Meta-SR [1]	31.72	0.923	28.37	0.884	28.07	0.851	27.45	0.836	25.96	0.785	24.71	0.730	23.76	0.684	23.01	0.645
	LIIF [1]	31.78	0.923	28.48	0.885	28.26	0.854	27.67	0.840	26.24	0.792	24.96	0.739	23.97	0.694	23.18	0.656
	LTE [1]	31.92	0.925	28.51	0.887	28.35	0.856	27.76	0.842	26.33	0.796	25.04	0.742	24.04	0.698	23.21	0.659
	LIWT(Ours)	31.95	0.925	28.51	0.886	28.41	0.857	27.81	0.843	26.39	0.797	25.13	0.746	24.12	0.701	23.29	0.662

Table 5: Comparison of different arbitrary-scale SR methods at non-integer scale on Set14 [11] and Urban100 [9] (PSNR (dB)).

6 The analysis of large-scale SR

We analyzed the advantages of our LIWT on the benchmark dataset for large-scale SR scenarios. We define scales larger than ×6 as large-scale. As shown in Table 6, our method achieves optimal results for large-scale at ×6, ×8, and ×12. As shown in Figure 4, LIWT with RDN [18] as the encoder can even outperform other methods with SwinIR [10] as the encoder at ×8 SR on Set5 [2].

Methods	Set5 [0]			Set14 [0]			B100 [0]			Urban100 [0]		
	$\times 6$	$\times 8$	$\times 12$	$\times 6$	$\times 8$	$\times 12$	$\times 6$	$\times 8$	$\times 12$	$\times 6$	$\times 8$	$\times 12$
RDN-Meta-SR [0]	29.04	26.96	-	26.51	24.97	-	25.90	24.83	-	23.99	22.59	-
RDN-LIIF [0]	29.15	27.14	24.86	26.64	25.15	23.24	25.98	24.91	23.57	24.20	22.79	21.15
RDN-UltraSR [0]	29.33	27.24	24.81	26.69	25.25	23.32	26.01	24.96	23.59	24.30	22.87	21.20
RDN-IPE [0]	29.25	27.22	-	26.58	25.09	-	26.00	24.93	-	24.26	22.87	-
RDN-LTE [0]	29.32	27.26	24.79	26.71	25.16	23.31	26.01	24.95	23.6	24.28	22.88	21.22
RDN-LIWT (Ours)	29.45	27.38	25.00	26.80	25.30	23.36	26.05	24.99	23.63	24.41	23.01	21.33
SwinIR-Meta-SR [0]	29.09	27.02	24.82	26.58	25.09	23.33	25.94	24.86	23.59	24.16	22.75	21.31
SwinIR-LIIF [0]	29.46	27.36	24.99	26.82	25.34	23.39	26.07	25.01	23.64	24.59	23.14	21.43
SwinIR-LTE [0]	29.50	27.35	25.07	26.86	25.42	23.44	26.09	25.03	23.66	24.62	23.17	21.50
SwinIR-LIWT (Ours)	29.60	27.51	25.07	26.90	25.43	23.48	26.13	25.06	23.69	24.71	23.24	21.57

Table 6: Comparison for scales at $\times 2$, $\times 3$, and $\times 4$ on benchmark datasets (PSNR (dB)).

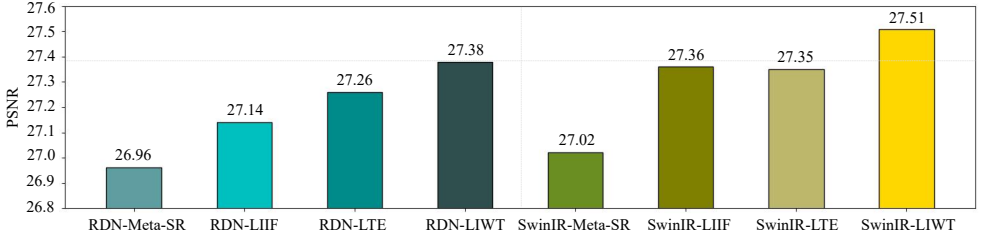


Figure 1: Comparison of $\times 8$ SR task on Set5 [0] (PSNR (dB)).

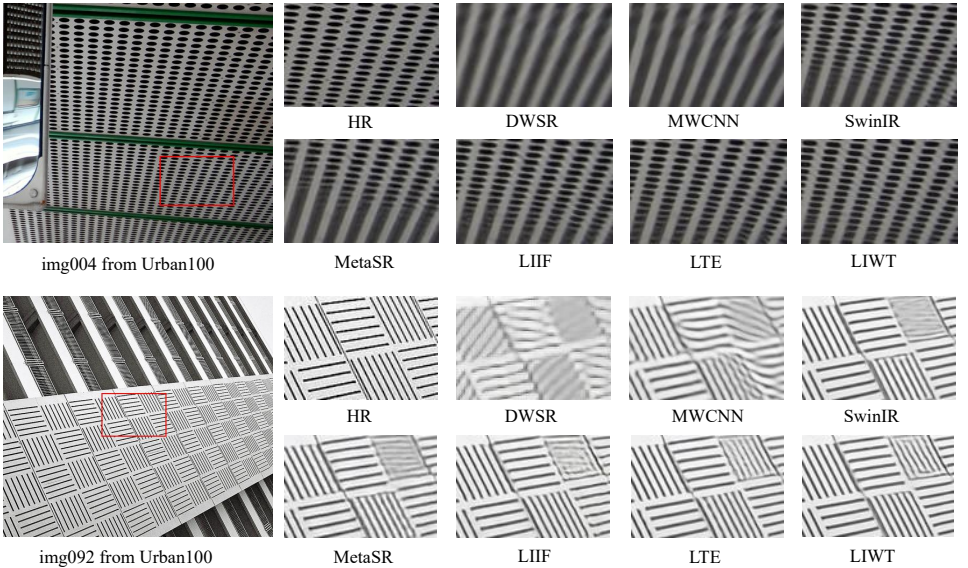


Figure 2: Visual comparisons for $\times 4$ SR on Urban100 [0]. Zoom in for best view.

References

- [1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017.

- [2] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012.
- [3] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8628–8638, 2021.
- [4] Tiantong Guo, Hojjat Seyed Mousavi, Tiep Huu Vu, and Vishal Monga. Deep wavelet prediction for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 104–113, 2017.
- [5] Wei-Yen Hsu and Pei-Wen Jian. Wavelet detail perception network for single image super-resolution. *Pattern Recognition Letters*, 166:16–23, 2023.
- [6] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-sr: A magnification-arbitrary network for super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1575–1584, 2019.
- [7] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015.
- [8] Pranav Jeevan, Akella Srinidhi, Pasunuri Prathiba, and Amit Sethi. Wavemixsr: Resource-efficient neural network for image super-resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5884–5892, 2024.
- [9] Jaewon Lee and Kyong Hwan Jin. Local texture estimator for implicit representation function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1929–1938, 2022.
- [10] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.
- [11] Pengju Liu, Hongzhi Zhang, Wei Lian, and Wangmeng Zuo. Multi-level wavelet convolutional neural networks. *IEEE Access*, 7:74973–74985, 2019.
- [12] Ying-Tian Liu, Yuan-Chen Guo, and Song-Hai Zhang. Enhancing multi-scale implicit learning in image super-resolution with integrated positional encoding. *arXiv preprint arXiv:2112.05756*, 2021.
- [13] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423. IEEE, 2001.
- [14] Jingwei Xin, Jie Li, Xinrui Jiang, Nannan Wang, Heng Huang, and Xinbo Gao. Wavelet-based dual recursive network for image super-resolution. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):707–720, 2020.

- [15] Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Ultrasr: Spatial encoding is a missing key for implicit image function-based arbitrary-scale super-resolution. *arXiv preprint arXiv:2103.12716*, 2021.
- [16] Shengke Xue, Wenyuan Qiu, Fan Liu, and Xinyu Jin. Wavelet-based residual attention network for image super-resolution. *Neurocomputing*, 382:116–126, 2020.
- [17] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*, pages 711–730. Springer, 2012.
- [18] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018.
- [19] Wenbin Zou, Liang Chen, Yi Wu, Yunchen Zhang, Yuxiang Xu, and Jun Shao. Joint wavelet sub-bands guided network for single image super-resolution. *IEEE Transactions on Multimedia*, 2022.