Supplementary Material: QuantSR: Accurate Low-bit Quantization for Efficient Image Super-Resolution

Anonymous Author(s)
Affiliation
Address
email

1 More Quantitative Results

In the supplementary materials, we provide additional detailed results of our QuantSR to demonstrate its comprehensive advantages in accuracy and efficiency.

1.1 Implementations

For the implementation of various methods, we have strived to adhere as closely as possible to the results and official code provided in their respective papers. Specifically, when using the same settings as QuantSR, we directly utilized the reported results from the paper [4, 2]. When different settings were employed, we aligned our model and training setup with the official code ([5] for PAMS and [2] for CADyQ), trained the model accordingly, and tested the obtained results.

1.2 Accuracy Results

In terms of accuracy, we present the results of the QuantSR approach. For each QuantSR variant (including QuantSR-C and QuantSR-T with different bit widths), four weight-sharing variants with varying numbers of blocks are included to achieve flexible inference with resource adaptation while breaking the accuracy upper limit. For QuantSR-C, variants with 32, 16, and 8 blocks are simultaneously trained, and for fairness, a 16-block variant equivalent to full precision is used for comparison in the paper. As for QuantSR-T, lighter variants with 4, 2, and 1 block(s) are used since SwinIR typically involves higher computational complexity, and smaller variants are advantageous for practical edge applications. From the complete results in Tab. 1, it can be observed that the largest variants almost match the full precision performance, and some 4-bit results even surpass it. This demonstrates the powerful potential of our proposed QuantSR approach in unleashing the accuracy of quantized SR networks. Furthermore, even considering the smallest variants, the accuracy remains at a reasonable level, albeit with the inference efficiency of the model pushed to the extreme.

1.3 Efficiency Results

In terms of efficiency, we present comprehensive results in Tab. 2, including QuantSR-T. To obtain more realistic estimates, we compute the FLOPs and storage savings of the quantization portion as $\frac{b}{32}$, where b represents the quantization bit-width. We found that due to the combined improvements in operators and architecture, QuantSR exhibits significant reductions in computation and FLOPs, particularly for its transformer version. This implies that it holds great potential for edge applications of transformer-based SR networks.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Method	Scale	#Bit (w/a)	#Blk	Set5		Set14		B100		Urban100		Manga109	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	×2	-/-	-	33.66	0.9299	30.24	0.8688	29.56	0.8431	26.88	0.8403	30.80	0.9339
SRResNet [3]	×2	32/32	16	38.00	0.9605	33.59	0.9171	32.19	0.8997	32.11	0.9282	38.56	0.9770
SwinIR_S [6]	×2	32/32	4	38.14	0.9611	33.86	0.9206	32.31	0.9012	32.76	0.9340	39.12	0.9783
QuantSR-C	×2	4/4	32	38.04	0.9606	33.66	0.9185	32.22	0.9002	32.19	0.9296	38.90	0.9776
			16	37.97	0.9603	33.53	0.9176	32.15	0.8993	31.91	0.9268	38.67	0.9772
			8	37.80	0.9597	33.35	0.9158	32.04	0.8979	31.46	0.9221	38.25	0.9762
QuantSR-T	×2	4/4	4	38.10	0.9604	33.65	0.9186	32.21	0.8998	32.20	0.9295	38.85	0.9774
			2	37.93	0.9602	33.51	0.9173	32.14	0.8991	31.88	0.9262	38.55	0.9768
			1	37.80	0.9596	33.31	0.9155	32.01	0.8973	31.39	0.9207	38.19	0.9759
QuantSR-C	×2	2/2	32	37.70	0.9594	33.21	0.9148	31.96	0.8970	31.11	0.9189	37.93	0.9754
			16	37.57	0.9589	33.09	0.9136	31.84	0.8954	30.77	0.9149	37.60	0.9745
			8	37.32	0.9579	32.88	0.9114	31.68	0.8930	30.29	0.9087	37.01	0.9729
QuantSR-T	×2	2/2	4	37.55	0.9587	33.12	0.9143	31.89	0.8958	30.96	0.9172	37.61	0.9745
			2	37.44	0.9583	33.02	0.9134	31.83	0.8952	30.82	0.9156	37.40	0.9740
			1	37.33	0.9576	32.89	0.9119	31.71	0.8933	30.48	0.9112	37.04	0.9730
Bicubic	×4	-/-	-	28.42	0.8104	26.00	0.7027	25.96	0.6675	23.14	0.6577	24.89	0.7866
SRResNet [3]	×4	32/32	16	32.16	0.8951	28.60	0.7822	27.58	0.7364	26.11	0.7870	30.46	0.9089
SwinIR_S [6]	×4	32/32	4	32.44	0.8976	28.77	0.7858	27.69	0.7406	26.47	0.7980	30.92	0.9151
QuantSR-C	×4	4/4	32	32.17	0.8943	28.60	0.7821	27.59	0.7368	26.12	0.7885	30.52	0.9082
			16	32.00	0.8924	28.50	0.7799	27.52	0.7342	25.88	0.7807	30.15	0.9040
			8	31.75	0.8894	28.35	0.7763	27.42	0.7307	25.59	0.7700	29.73	0.8985
QuantSR-T	×4	4/4	4	32.18	0.8941	28.63	0.7822	27.59	0.7367	26.11	0.7871	30.49	0.9087
			2	32.02	0.8922	28.52	0.7795	27.52	0.7343	25.89	0.7797	30.20	0.9051
			1	31.79	0.8891	28.36	0.7757	28.37	0.7301	27.41	0.7687	29.69	0.8982
QuantSR-C	×4	2/2	32	31.47	0.8849	28.19	0.7725	27.31	0.7277	25.29	0.7604	29.16	0.8897
			16	31.30	0.8819	28.08	0.7694	27.23	0.7246	25.13	0.7537	28.81	0.8844
			8	31.04	0.8771	27.87	0.7643	27.11	0.7202	24.85	0.7423	28.21	0.8743
QuantSR-T	×4	2/2	4	31.53	0.8845	28.16	0.7715	27.28	0.7274	25.26	0.7609	29.06	0.8898
			2	31.45	0.8832	28.11	0.7703	27.25	0.7261	25.18	0.7575	28.89	0.8871
			1	31.26	0.8801	27.97	0.7665	27.15	0.7223	24.99	0.7490	28.52	0.8801

Table 1: Full quantitative results of QuantSR. SRResNet and SwinIR-S are used as full-precision backbones. 'w/a' denotes the weight/activation bits. Results for variants with the same number of blocks as their full-precision counterparts are colored with red.

Method	#Bit (w/a)	#Blk	Params (K) (↓ Ratio)	Ops (G) (↓ Ratio)	Urban100	
					PSNR	SSIM
SRResNet	32/32	16	1,367 (0%)	90.1 (0%)	32.16	0.8951
QuantSR-C	4/4	32	451 (↓ 67.0%)	29.9 (↓ 66.9%)	32.17	0.8943
		16	303 (↓ 77.8%)	20.2 (↓ 77.5%)	32.00	0.8924
		8	230 (↓ 83.1%)	15.4 (↓ 82.9%)	31.75	0.8894
QuantSR-C	2/2	32	170 (↓ 87.6%)	11.5 (↓ 87.2%)	31.48	0.8849
		16	161 (↓ 88.2%)	10.9 (↓ 87.9%)	31.30	0.8819
		8	156 (↓ 88.6%)	10.6 (↓ 88.3%)	31.04	0.8771
SwinIR_S	32/32	4	930 (0%)	56.47 (0%)	32.44	0.8976
QuantSR-T	4/4	4	154 (↓ 83.37%)	9.39 (↓ 83.37%)	32.18	0.8941
		2	98.8 (↓ 89.38%)	6.03 (↓ 89.32%)	32.02	0.8922
		1	71.0 (↓ 92.36%)	4.34 (↓ 92.31%)	31.97	0.8891
QuantSR-T	2/2	4	50.2 (↓ 94.60%)	3.08 (↓ 94.55%)	31.53	0.8845
		2	46.8 (↓ 94.97%)	2.87 (↓ 94.92%)	31.45	0.8832
		1	45.1 (↓ 95.15%)	2.77 (↓ 95.09%)	31.26	0.8801

Table 2: Compression ratio of 2-bit and 4-bit SRResNet and SwinIR_S (×4), and their input sizes are 3×256×256 for calculating Ops.

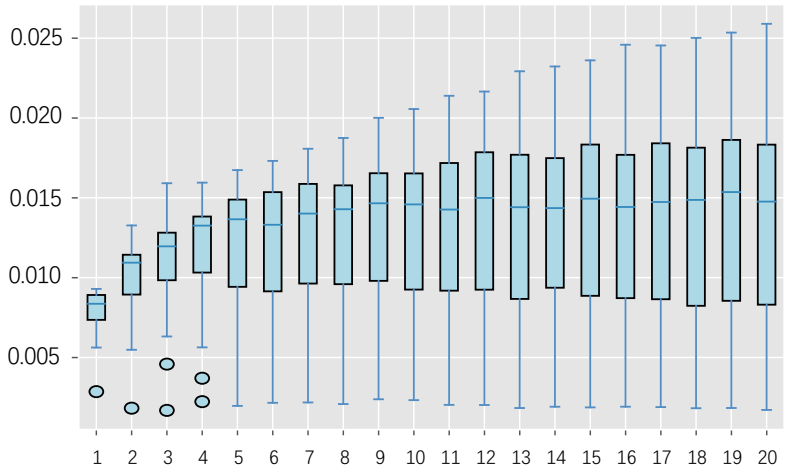


Figure 1: $\hat{v}_b$ for weight quantizer. We use a box plot to count the change of the distribution of the learnable parameter $\hat{v}_b$ in the entire QuantSR network with the training epoch, where for each epoch, we show the maximum value, minimum value, and variance of the parameter.

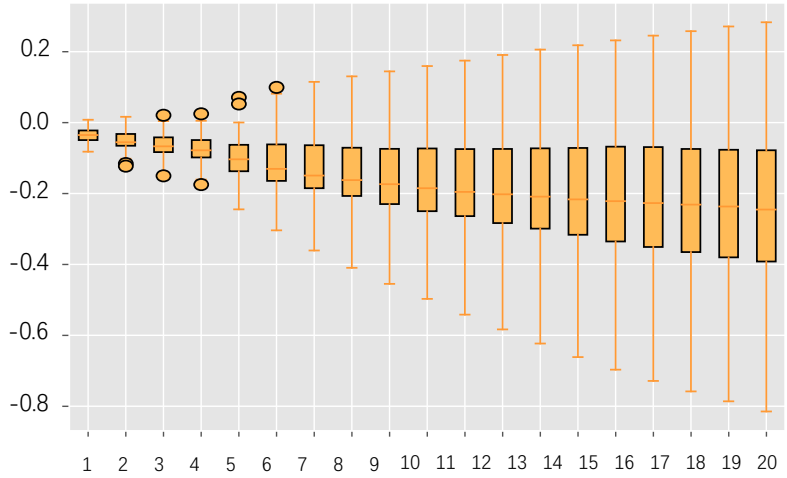


Figure 2: $\hat{\tau}$ for activation quantizer. The form of statistics is the same as Figure 1.

2 More Visualization Results

2.1 Visualization Analysis

Figures 1 and 2 present statistics on the learnable parameters of the proposed quantizer. For the quantized step parameters v_b and $\hat{\tau}$, the statistics indicate that the range of these types of learnable parameters in the network generally increases with training time. Since the form of the proposed quantizer depends on these two learnable redistribution parameters, this phenomenon suggests an increasing diversity of quantizers in our QuantSR. This increase in diversity signifies that, under a limited fixed bit-width, we effectively utilize the diversity provided by quantization to significantly recover the forward representation capacity of the quantized SR model.

Figure 3 illustrates the impact of the $\phi(\cdot)$ function on a single backward propagation and a single sample. In the proposed quantizer, the transformation function $\phi(\cdot)$ is embedded within each quantization interval, causing no effect during the forward propagation but guiding parameter updates to better reflect the behavior of the quantizer during backward. As shown in the figure, our $\phi(\cdot)$ function influences parameter updates within each propagation, and this influence is generally present but not significantly pronounced at the individual update level. This allows for more stable updates of

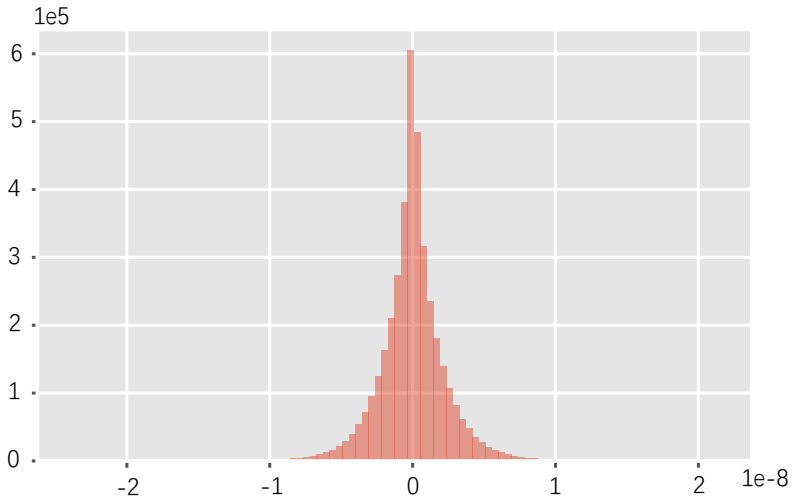


Figure 3: Gradient effect of $\phi(\cdot)$. In the 4-bit QuantSR-C, we present the gradient effect caused by $\phi(\cdot)$ in the weight quantizer in a random quantized volume set. After obtaining the gradient before and after the derivation of the function, we make a pixel-wise difference between them and then statistic the value. The resulting statistic can be defined as the effect of this function on the gradient.

the quantized SR network, gradually accumulating the impact throughout the process to accurately reflect the quantizer's influence in the final well-trained network.

In Fig. 4, we further visualize the output of the high-level feature extractor of our quantized SR model. Compared to the existing DoReFa method, our QuantSR-C and QuantSR-T exhibit clearer and richer details at the same bit-width, indicating that QuantSR has achieved better representation capacity.

2.2 Visual Results

We utilize SRResNet [3] as the underlying architecture for CNN-based image super-resolution (SR) networks. Our QuantSR-C approach is compared with DoReFa [4, 7], PAMS [4], and CADyQ [1] using 2-bit and 4-bit precision. For Transformer-based image SR networks, we adopt the lightweight SwinIR_S [6] as the backbone. Given the increased difficulty in Transformer binarization and our observations from CNN-based methods, we exclusively apply our proposed techniques to quantize SwinIR_S. Additionally, we present the results of our quantized Transformer baseline, QuantSR-T.

We present additional visual results in Figs. 5 and 6 for 4-bit setting and Figs. 7 and 8 for the more challenging 2-bit setting. These figures reveal that our proposed QuantSR achieves comparable or superior performance compared to other methods in most instances. Particularly in more challenging scenarios, our QuantSR outperforms other approaches, delivering the highest quality reconstructions. Comparing QuantSR to its corresponding full-precision model SwinIR_S, we observe minimal discrepancies between them. These findings further validate the efficacy of our proposed techniques QuantSR-C and QuantSR-T.

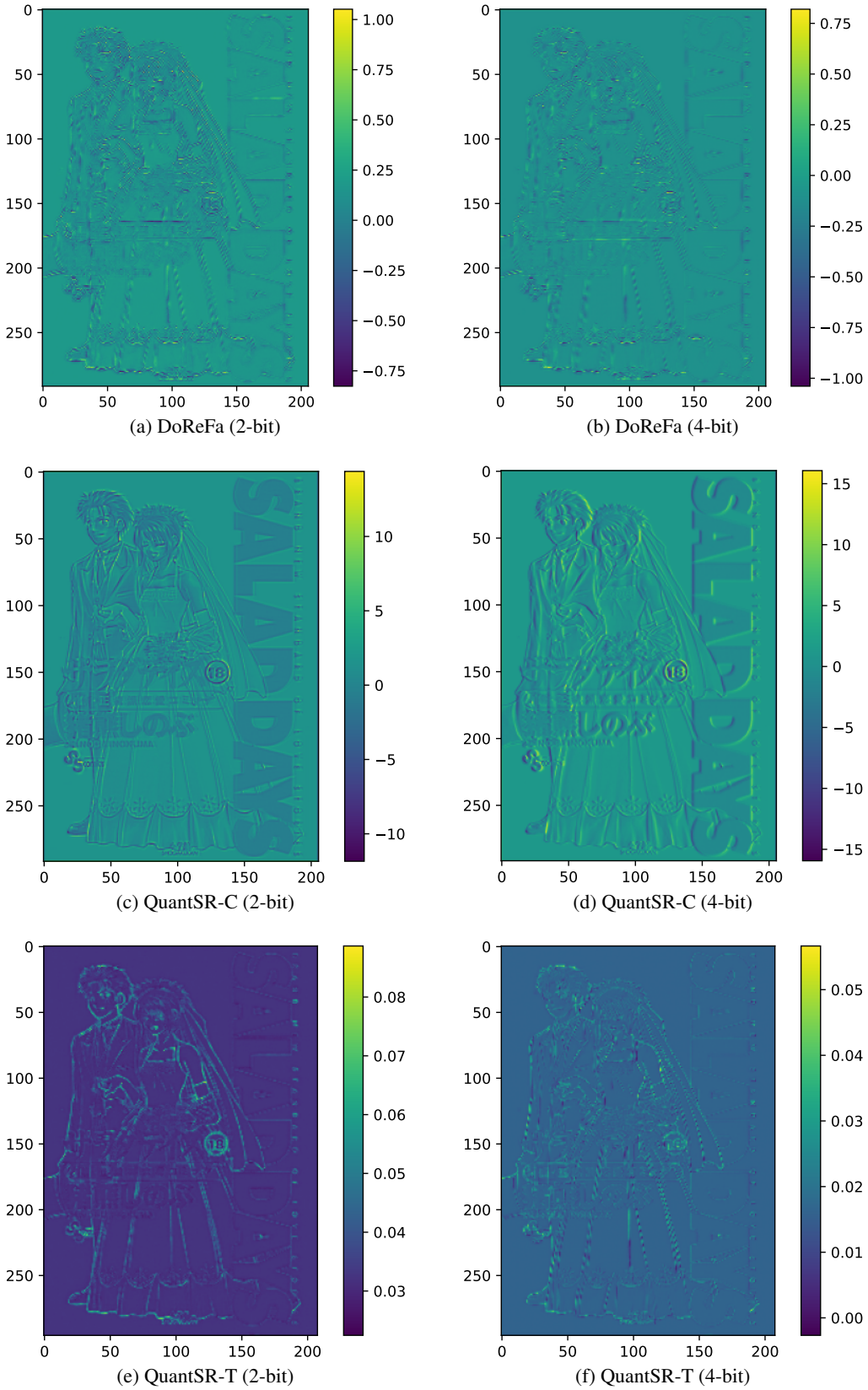
3 Checklist Explanations

3.1 Code

We have provided code to reproduce the results in this work.

3.2 Limitations

We would provide more analyses about the limitations of our method. (1) Although we have further narrowed the performance gap between the low-bit quantized model and its full-precision one, the very low-bit (*e.g.*, 2-bit) quantized model would suffer from obvious performance drops. (2) Currently, we only investigate low-bit quantization for image super-resolution. It is better to generalize the quantized networks for other image restoration applications (*e.g.*, image denoising and deblurring).



(a) DoReFa (2-bit) (b) DoReFa (4-bit)

(c) QuantSR-C (2-bit) (d) QuantSR-C (4-bit)

(e) QuantSR-T (2-bit) (f) QuantSR-T (4-bit)

Figure 4: Feature visualization for quantized SR networks. We visualize 2- and 4-bit DoReFa, QuantSR-C, and QuantSR-T features (Test sample: SaladDays_vol18 in Manga109 dataset).

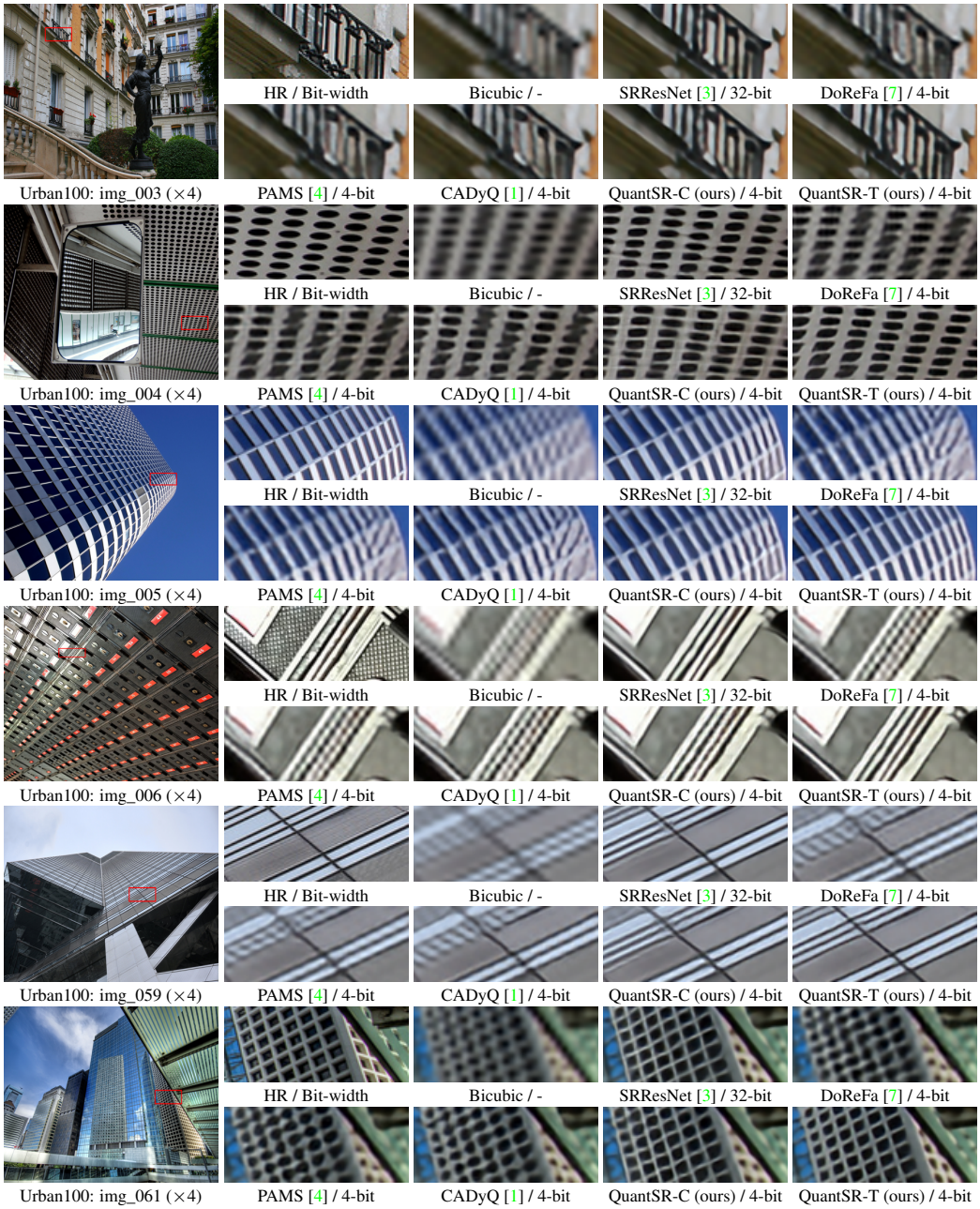


Figure 5: Visual comparison (×4) with lightweight SR in terms of 4-bit.

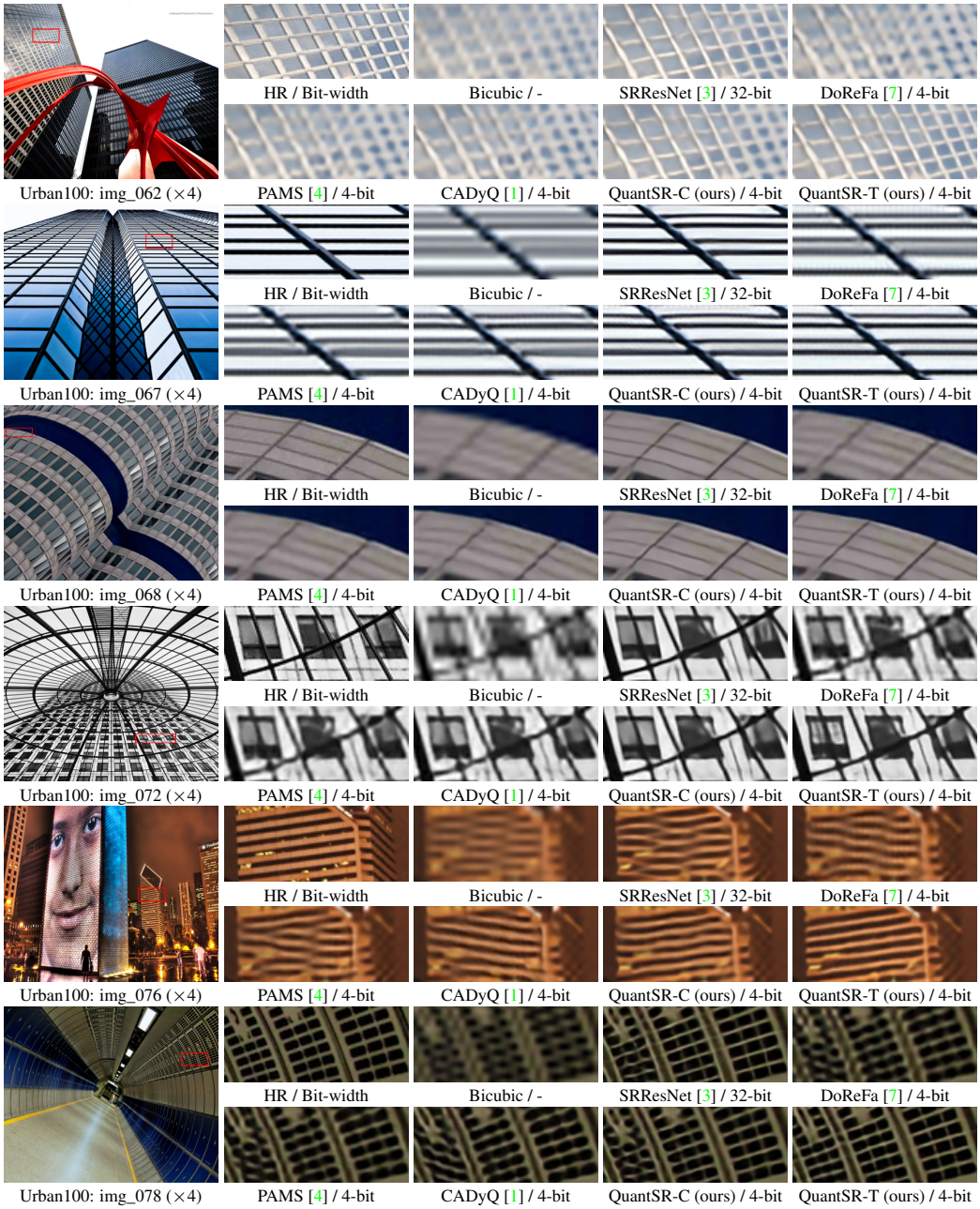


Figure 6: Visual comparison (×4) with lightweight SR in terms of 4-bit.

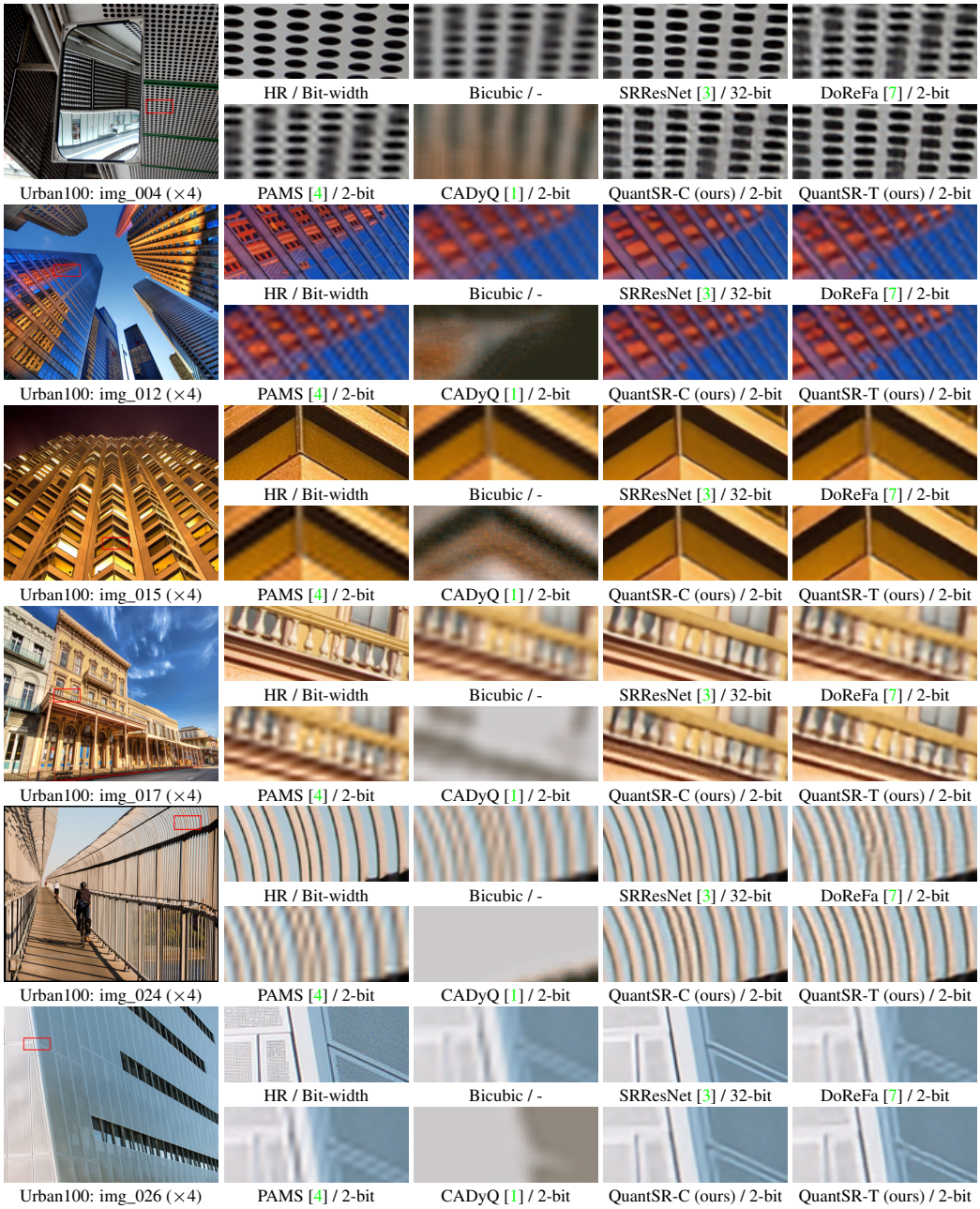


Figure 7: Visual comparison (×4) with lightweight SR in terms of 2-bit.

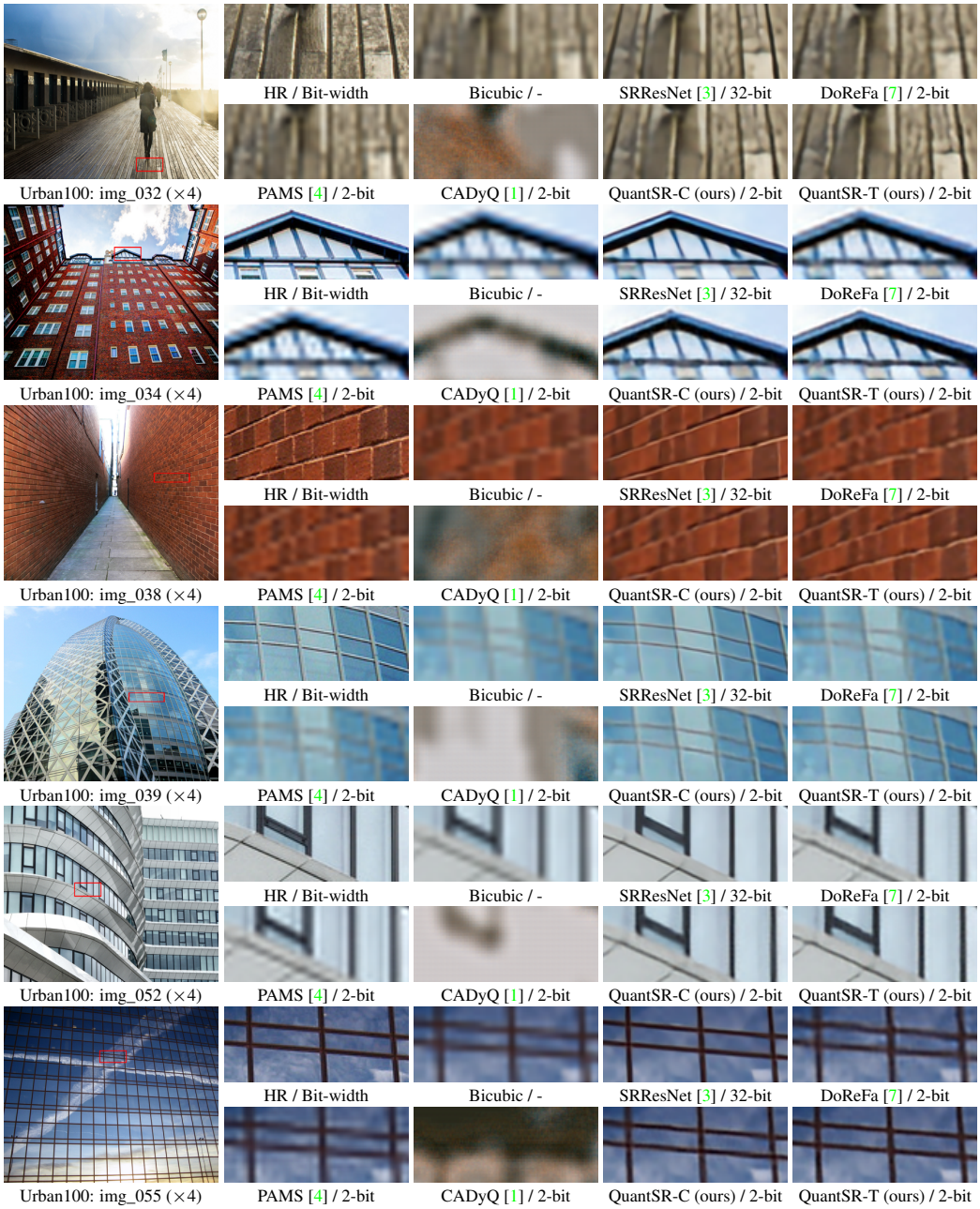


Figure 8: Visual comparison (×4) with lightweight SR in terms of 2-bit.

References

[1] Cheeun Hong, Sungyong Baik, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Cadyq: Content-aware dynamic quantization for image super-resolution. In *ECCV*, 2022. 4, 6, 7, 8, 9

[2] Cheeun Hong, Sungyong Baik, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Cadyq, 05 2023. 1

[3] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 2, 4, 6, 7, 8, 9

[4] Huixia Li, Chenqian Yan, Shaohui Lin, Xiawu Zheng, Baochang Zhang, Fan Yang, and Rongrong Ji. Pams: Quantized super-resolution via parameterized max scale. In *ECCV*, 2020. 1, 4, 6, 7, 8, 9

[5] Huixia Li, Chenqian Yan, Shaohui Lin, Xiawu Zheng, Baochang Zhang, Fan Yang, and Rongrong Ji. Pams, 05 2023. 1

[6] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCVW*, 2021. 2, 4

[7] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv preprint arXiv:1606.06160*, 2016. 4, 6, 7, 8, 9