Supplementary Material for Overcoming Distribution Mismatch in Quantizing Image Super-Resolution Networks

Cheeun Hong[1] and Kyoung Mu Lee[1,2]

[1] Dept. of ECE & ASRI, {cheeun914, kyoungmu}@snu.ac.kr
[2] IPAI, Seoul National University

In this supplementary material, we present additional experimental results in Section S1; additional ablation study in Section S2; additional analyses in Section S3.

S1 Additional Experiments

Along with the evaluations on SR networks of scale $\times 4$ discussed in the main manuscript, we also assess our framework on networks of scale $\times 2$. As shown in Table S1, our framework outperforms existing SR quantization methods in terms of both PSNR and SSIM, demonstrating its effectiveness on scale $\times 2$ SR networks. Specifically, the PSNR gain on Set5 is 0.28 dB on EDSR, 0.74 dB on RDN, and 0.25 dB on SwinIR. These results confirm that our framework is effective for both CNN- and Transformer-based SR networks of scale 2.

Table S1: Quantitative comparisons on SR networks of scale $\times 2$

Model	Bit	Set5		Set14		B100		Urban100	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR	32	37.93	0.960	33.46	0.916	32.10	0.899	31.71	0.925
EDSR-PAMS	2	35.30	0.946	31.63	0.899	30.66	0.879	28.11	0.875
EDSR-DAQ	2	36.82	0.955	32.50	0.908	31.34	0.891	29.85	0.905
EDSR-DDTB	2	37.25	0.958	32.87	0.911	31.67	0.893	30.34	0.910
EDSR-ODM (Ours)	2	**37.53**	**0.958**	**33.06**	**0.913**	**31.81**	**0.895**	**30.81**	**0.915**
RDN	32	38.05	0.961	33.59	0.917	32.20	0.900	32.13	0.927
RDN-PAMS	2	35.45	0.946	31.67	0.899	30.69	0.879	28.14	0.874
RDN-DAQ	2	37.23	0.957	32.84	0.910	31.66	0.893	30.46	0.908
RDN-DDTB	2	36.76	0.955	32.54	0.908	31.44	0.890	29.77	0.903
RDN-ODM (Ours)	2	**37.50**	**0.958**	**33.03**	**0.913**	**31.80**	**0.895**	**30.57**	**0.913**
SwinIR	32	38.14	0.961	33.86	0.921	32.31	0.901	32.76	0.934
SwinIR-PAMS	2	35.38	0.947	31.63	0.899	30.65	0.880	28.07	0.873
SwinIR-DAQ	2	34.98	0.943	31.38	0.896	30.47	0.876	27.83	0.869
SwinIR-DDTB	2	37.17	0.957	32.78	0.911	31.42	0.888	30.24	0.908
SwinIR-ODM (Ours)	2	**37.42**	**0.958**	**33.03**	**0.913**	**31.79**	**0.895**	**30.76**	**0.914**

Furthermore, we compare our method with existing methods by training each method for 300K iterations. The results in Table S2 show that the gains achieved by our approach for 60K iterations reported in the main manuscript are maintained. Our framework still achieves more than a 0.37 dB gain over other methods for Set5 when trained for extended iterations.

Table S2: Quantitative comparisons of SR quantization methods with 300K iterations

Model	Bit	Set5		Set14		B100		Urban100	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR	32	32.10	0.894	28.58	0.781	27.56	0.736	26.04	0.785
EDSR-PAMS	2	30.05	0.852	27.17	0.744	26.70	0.705	24.09	0.707
EDSR-DAQ	2	31.11	0.874	27.98	0.765	27.14	0.720	24.96	0.745
EDSR-DDTB	2	31.19	0.878	27.97	0.767	27.14	0.723	25.01	0.749
EDSR-ODM (Ours)	2	**31.56**	**0.884**	**28.15**	**0.771**	**27.30**	**0.728**	**25.24**	**0.758**

Also, to ensure a fair comparison with DAQ [6], we follow their settings and apply our method to EDSR of 32 residual blocks with 256 channel dimensions. As shown in Table S3, our method outperforms DAQ even though DAQ employs a channel-wise quantization function, whereas our method utilizes a more efficient layer-wise function.

Table S3: Quantitative comparisons on EDSR (full) of scale $\times 4$ which consists of 32 residual blocks and 256 channel dimensions. For a fair comparison with DAQ, our model (ODM*) is trained for 300K iterations.

Model	Bit	Set5		Set14		B100		Urban100	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR-full	32	32.46	0.897	28.80	0.788	27.72	0.742	26.64	0.803
EDSR-full-DAQ	2	32.05	0.890	28.53	0.778	27.50	0.733	25.97	0.781
EDSR-full-ODM* (Ours)	2	**32.15**	**0.893**	**28.56**	**0.781**	**27.57**	**0.737**	**26.04**	**0.786**

Moreover, we compare our method with a fully-quantized SR network, EDSR-FQSR [10], in which all layers and also the skip connections are quantized. For a fair comparison, we also quantize all convolutional layers and the skip connections. The results in Table S4 show that our ODM outperforms FQSR, indicating that our approach is also effective when the network is fully quantized.

Additionally, along with SwinIR, as demonstrated in the main manuscript, we also apply our method to a more recent, large Transformer-based model, HAT [3] ($\sim$10M

Table S4: Quantitative comparisons on EDSR with fully quantized method. S.C. refers to the bit-width of skip-connections.

Scale	Model	Bit	S.C.	Set5		Set14		B100		Urban100	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
$\times 4$	EDSR	32	32	32.10	0.894	28.58	0.781	27.56	0.736	26.04	0.785
	EDSR-FQSR	4	8	30.93	0.870	27.82	0.761	27.07	0.715	24.93	0.744
	EDSR-ODM (Ours)	4	8	**31.99**	**0.890**	**28.42**	**0.777**	**27.47**	**0.733**	**25.70**	**0.775**
$\times 2$	EDSR	32	32	37.93	0.960	33.46	0.916	32.10	0.899	31.71	0.925
	EDSR-FQSR	4	8	37.04	0.951	32.84	0.908	31.67	0.889	30.65	0.911
	EDSR-ODM (Ours)	4	8	**37.86**	**0.960**	**33.42**	**0.916**	**32.08**	**0.898**	**31.71**	**0.924**

parameters). The results in Table S5 indicate that our method can be effectively applied to Transformer models.

Table S5: Quantitative comparisons on HAT of scale $\times 4$

Model	Bit	Set5		Set14		B100		Urban100	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
HAT	32	32.92	0.905	29.15	0.796	27.97	0.751	27.87	0.835
HAT-PAMS	2	30.30	0.859	27.35	0.750	26.78	0.710	24.23	0.713
HAT-DAQ	2	30.21	0.856	27.29	0.747	26.74	0.708	24.14	0.708
HAT-DDTB	2	31.23	0.878	27.96	0.766	27.16	0.724	25.08	0.751
HAT-ODM (Ours)	2	**32.06**	**0.891**	**28.56**	**0.780**	**27.53**	**0.736**	**26.10**	**0.787**

S2 Additional Ablation Study

In this section, we present an ablation study on the hyperparameters of our framework. First, we conduct an ablation study on the percentile j, which is used to initialize quantization range clipping parameters. The results in Table S6a show that when j=100, meaning the quantization range is not clipped and is determined by the maximum value, accuracy severely degrades. This hints that clipping is important for performance and that using the max function does not serve as an effective initialization policy.

Also, we analyze the impact of the gradient balance terms λ_R and λ_M, which balances the gradient of the reconstruction loss and the mismatch regularization loss, and initial learning rate β^0. The results in Tables S6b and S6c support our choice of λ_R=1, λ_M=1e-5, and β^0=1e-4.

Table S6: Ablation on hyperparameters on EDSR $\times 4$ (2-bit)

j	Set5	Set14	B100	U100
100	29.63	26.86	26.53	23.82
99	31.50	28.14	27.27	25.17
95	31.51	28.13	27.27	25.18

(a) Ablation on j

λ_M	Set5	Set14	B100	U100
2e-5	31.40	28.07	27.22	25.09
1e-5	31.50	28.14	27.27	25.17
1e-6	31.48	28.15	27.28	25.18

(b) Ablation on λ_M

β^0	Set5	Set14	B100	U100
1e-3	30.82	27.72	27.01	24.63
1e-4	31.50	28.14	27.27	25.17
1e-5	31.17	27.96	27.14	24.93

(c) Ablation on β^0

Furthermore, we compare our weight clipping correction scheme with the commonly used quantization scheme for weights, LSQ [5]. The results in Table S7 validate the effectiveness of our clipping correction approach.

Table S7: Comparison of WCC with LSQ on EDSR $\times 4$ (2-bit)

Method	Set5 (PSNR/SSIM)	Urban100 (PSNR/SSIM)
LSQ + Cooperative MR	31.26 / 0.877	25.02 / 0.746
WCC + Cooperative MR (Ours)	31.50 / 0.882	25.17 / 0.755

S3 Additional Analyses

S3.1 Complexity Analysis

We provide additional complexity analysis on SwinIR in Table S8. The results show that our method achieves superior SR performance with minimal or no computational overhead in terms of model storage size and bitOPs. BitOPs for SwinIR are calculated for processing a 64×64 input patch.

Table S8: Computational complexity comparison on SwinIR of scale $\times 4$

Model	Bit	Storage size	BitOPs	PSNR	SSIM
SwinIR	32	929.6K	5.071T	32.44	0.898
SwinIR-PAMS	2	160.2K	1.100T	29.48	0.834
SwinIR-DAQ	2	160.1K	1.176T	29.10	0.824
SwinIR-DDTB	2	160.4K	1.100T	31.01	0.873
SwinIR-ODM (Ours)	2	**160.4K**	**1.100T**	**31.44**	**0.880**

S3.2 Distribution mismatch

For the generalizability of the observed distribution mismatch problem of the main manuscript, we analyze the variance of features in SR networks across a set of images. As reported in Table S9, the classification network (ResNet20) exhibits much less image-wise and channel-wise variance compared to SR networks (EDSR, RDN). This suggests that the distribution mismatch problem is particularly severe in SR networks.

Table S9: Average variance in feature. The metrics are measured on DIV2K validation set for SR networks and ImageNet validation set for the classification network.

Task	Model	Image-wise Variance	Channel-wise Variance
Image super-resolution	EDSR ($\times 4$)	15.08	40.29
Image super-resolution	RDN ($\times 4$)	6.40	58.14
Image classification	ResNet-20	0.04	0.09

Moreover, we analyze the feature mismatch after quantization-aware training (QAT) using different methods. To track feature similarity, we compute the variance between feature distribution statistics (mean and standard deviation) within the benchmark dataset. As shown in Table S10, the feature mismatch is effectively reduced with our framework.

Table S10: Feature mismatch after quantization-aware training. The metrics are measured on DIV2K validation set for SR networks and ImageNet validation set for the classification network.

Similarity metric	Var[mean]	Var[std]	Avg. Mismatch
Before QAT	0.28	3.60	2.41e+02
After QAT w/ **ODM**	**0.25**	**1.02**	**1.37e+02**
After QAT w/ PAMS	0.29	2.34	4.04e+02
After QAT w/ DDTB	2.54	1.67	5.46e+05
After QAT w/ DAQ	2.14	6.59	6.10e+06

We provide additional analysis supporting the choice of distance from the quantization grid as a measure of distribution mismatch in Eq. (2) of the main manuscript. We note that QAT is a process searching for a quantization grid that best fits the discrepant input distributions. If the average distance of each feature from the quantization grid is small, it implies that most features are aligned with the grid, indicating a low distribution mismatch. According to Table S10, it is verified that using distance as the mismatch measure reduces the variance in feature distribution statistics across test images.

S3.3 Cooperative Regularization

In the main manuscript, we emphasized the importance of *cooperatively* using mismatch regularization and reconstruction losses. For the cooperative update, we weigh

the gradient of mismatch regularization using the cosine similarity with the gradient of the reconstruction loss. The weighing term is formulated as $0.5 \cdot (cos(\boldsymbol{v}_a, \boldsymbol{v}_b) + 1)$. Here, we present results using the weighing term as $cos(\boldsymbol{v}_a, \boldsymbol{v}_b)$ following Du *et al*. [4] and that of $u(cos(\boldsymbol{v}_a, \boldsymbol{v}_b))$ where $u(\cdot)$ is the unit step function. The results in Table S11 show that all these functions that alleviate the conflict between the two losses achieve high reconstruction accuracy, indicating that the general cooperative property is the key to performance gain.

Table S11: Different formulation for cooperative behaviour on EDSR $\times 4$ (2-bit)

Formulation	Set5 (PSNR/SSIM)	Urban100 (PSNR/SSIM)
$cos(\boldsymbol{v}_a, \boldsymbol{v}_b)$ [4]	31.46 / 0.881	25.14 / 0.755
$u(cos(\boldsymbol{v}_a, \boldsymbol{v}_b))$	31.49 / 0.883	25.15 / 0.756
$0.5 \cdot (cos(\boldsymbol{v}_a, \boldsymbol{v}_b) + 1)$ (Ours)	31.50 / 0.882	25.17 / 0.755

S3.4 More Visualizations

For better comprehension, we provide additional results of the effect of our loss after training in Figure S1. Along with the results in Figure 5 of the main manuscript, these results show that our loss term updates the activation distributions to a further quantization-friendly state while mostly preserving the high-density values of the original distribution.

Moreover, we provide visualizations of layer-wise mismatch in weights for different SR networks in Figure S2. According to the visual results, the weight distributions of different layers have a similar mean (*i.e*., near 0), but exhibit varying minimum and maximum values. This motivates us to use a layer-wise different policy for determining the weight quantization range.

S3.5 Training Time

Although our framework primarily aims to achieve an accurate quantized SR network in which the inference cost is reduced via quantization, we also provide comparisons on the training time. According to Table S12, our training scheme requires a shorter training time than DDTB and DAQ. Although our training incurs slightly more time overhead compared to PAMS, the gains in test accuracy compensate for this additional training cost.

License of the Used Assets

- DIV2K [1] dataset is publicly available for academic research purposes.
- Set5 [2], Set14 [8], BSD100 [9], Urban100 [7] datasets are made available at https://github.com/jbhuang0604/SelfExSR.

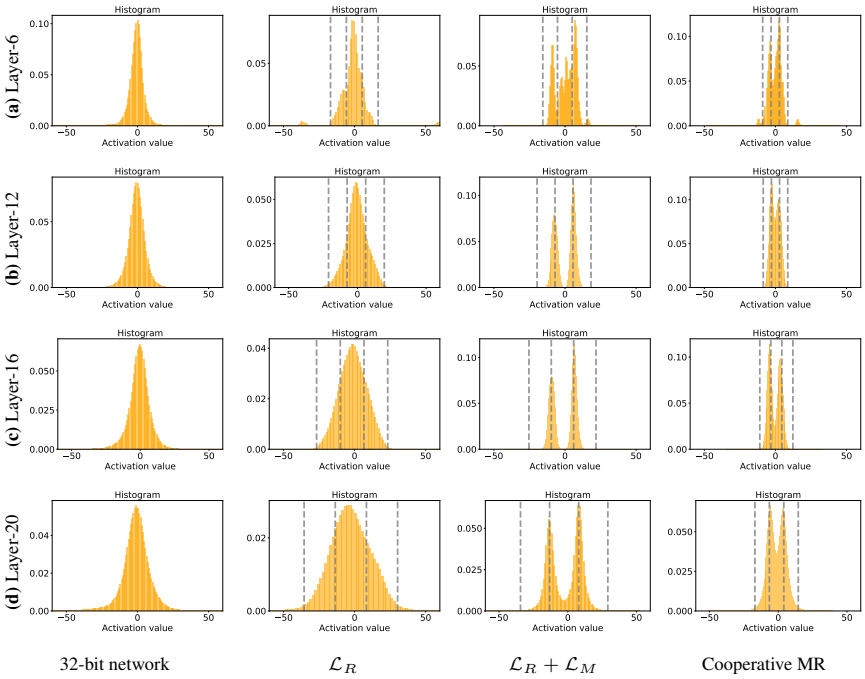

Figure S1: Distribution of activations after training and before quantization. Activations of the different convolution layers in 2-bit EDSR-ODM are visualized.

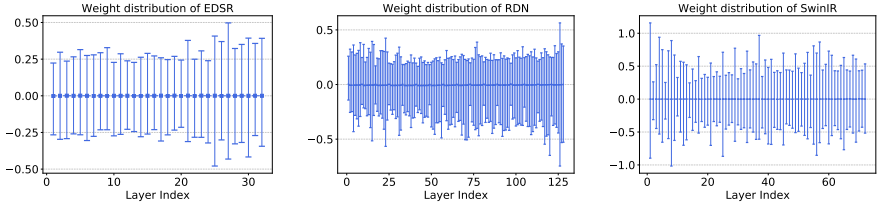

Figure S2: Layer-wise different weight distributions of SR networks. The minimum, maximum, and mean of each convolution/linear layer weight are plotted.

Table S12: Training time of QAT methods on SR networks. The training time is measured by running the experiment on a single RTX 2080Ti GPU.

Method	EDSR-PAMS	EDSR-DAQ	EDSR-DDTB	EDSR-ODM (Ours)
Time (hours)	1.5	3.7	2.5	2.4

References

1. Agustsson, E., Timofte, R.: NTIRE 2017 challenge on single image super-resolution: Dataset and study. In: CVPR Workshops (2017)
2. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.L.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: BMVC (2012)
3. Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Activating more pixels in image super-resolution transformer. In: CVPR (2023)
4. Du, Y., Czarnecki, W.M., Jayakumar, S.M., Farajtabar, M., Pascanu, R., Lakshminarayanan, B.: Adapting auxiliary losses using gradient similarity. arXiv preprint arXiv:1812.02224 (2018)
5. Esser, S.K., McKinstry, J.L., Bablani, D., Appuswamy, R., Modha, D.S.: Learned step size quantization. In: ICLR (2020)
6. Hong, C., Kim, H., Baik, S., Oh, J., Lee, K.M.: Daq: Channel-wise distribution-aware quantization for deep image super-resolution networks. In: WACV (2022)
7. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: CVPR (2015)
8. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR (2017)
9. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: ICCV (2001)
10. Wang, H., Chen, P., Zhuang, B., Shen, C.: Fully quantized image super-resolution networks. In: ACMMM (2021)