Edge accelerated reconstruction using sensitivity analysis for single-lens computational imaging - supplementary materials

Xuquan Wang,[a,b,c,†] **Tianyang Feng,**[a,b,c,†] **Yujie Xing,**[a,b,c,†] **Ziyu Zhao,**[a,b,c] **Xiong Dun,**[a,b,c,*] **Zhanshan Wang,**[a,b,c,d] **Xinbin Cheng**[a,b,c,d,*]

[a]MOE Key Laboratory of Advanced Micro-Structured Materials, Shanghai, 200092, China
[b]Institute of Precision Optical Engineering, School of Physics Science and Engineering, Tongji University, Shanghai, 200092, China
[c]Shanghai Frontiers Science Center of Digital Optics, Shanghai, 200092, China
[d]Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai, 200092, China

1 Supplementary experiments of quantization sensitivities

In this paper, we calculate the edge quantization sensitivity using a network with 50% uniform pruning, where PSNR and SSIM are weighted equally. To validate the feasibility of employing a uniformly pruned network for quantization sensitivity analysis, we conducted additional ablation experiments to evaluate the impact of pruning operations on quantization sensitivity.

The same quantitative sensitivity testing was applied to networks without pruning, with uniform pruning, and with sensitivity-aware pruning. Experimental results shown in Fig. S1 indicate that their edge quantization sensitivities are quite similar, suggesting that pruning operations have minimal impact on quantization sensitivity. This demonstrates the effectiveness of the proposed method for quantization sensitivity analysis. A fixed quantization sensitivity result can significantly simplify the compression and deployment process, thereby enhancing usage efficiency.

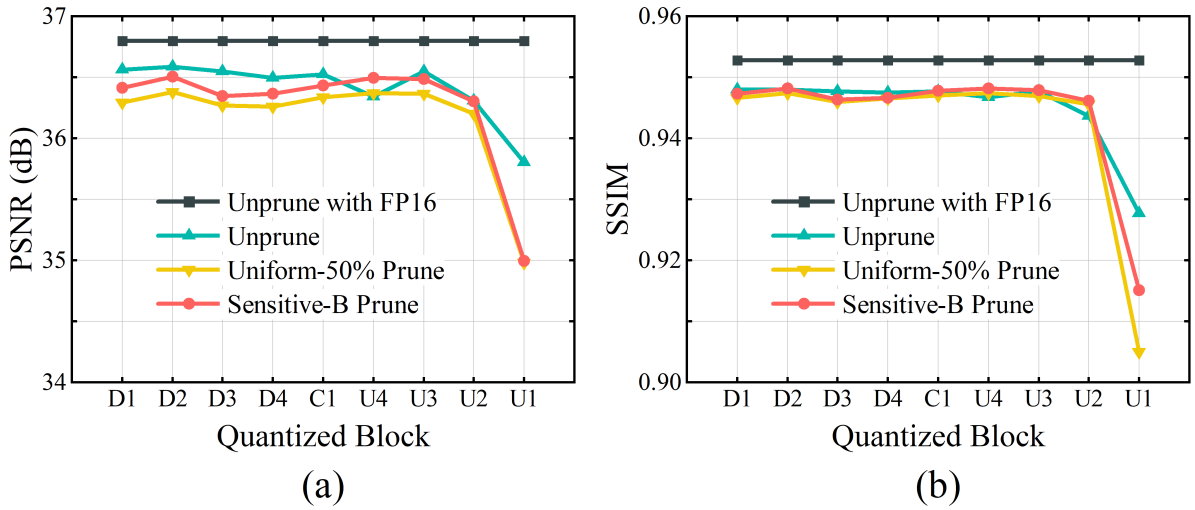

Fig. S1 Performance degradation caused by quantization for networks without pruning, with uniform pruning, and with sensitivity-aware pruning.

To explore the performance boundaries of ultra-low-bit quantization in computational imaging reconstruction, we conducted supplementary experiments using INT4 quantization. These experiments involved evaluating the quantization sensitivity of the proposed U-Net under INT4 precision and assessing the effectiveness of various 4-bit quantization strategies. It is worth noting that, since the edge chip used in our study does not support INT4 quantization, the experiments were conducted using MQBench quantization tools on a GPU platform, which is consistent with the related works.

As shown in Fig. S2, the results of the supplementary experiments indicate that the performance degradation trend introduced by block-wise 4-bit quantization is largely comparable to that of 8-bit quantization. However, the magnitude of performance degradation is significantly greater than that observed with 8-bit quantization. Specifically, after quantizing all blocks to INT4, the resulting PSNR and SSIM values are 17.15 and 0.5783, respectively, as shown in Table S1. For the mixed quantization using INT4, with FP16 applied to the last block, the resulting PSNR and SSIM values are 27.22 and 0.7401, respectively. This level of performance degradation is unacceptable for image restoration networks, even when applied only to the so-called insensitive blocks. This conclusion is more directly illustrated by the comparison shown in Fig. S3, where noticeable distortions appear in certain regions of the reconstructed images produced by the 4-bit mixed-precision quantized network.

We believe that the boundaries of model quantization are highly task-dependent—for example, recognition tasks generally exhibit greater tolerance to quantization errors compared to reconstruction tasks. At least in our study, 4-bit quantization fails to meet the minimum performance requirements of the task. This may also explain why mature, general-purpose AI edge chips have

yet to fully adopt INT4 quantization support. Given the current state of research, INT8 quantization remains a more practical and feasible choice for our application.

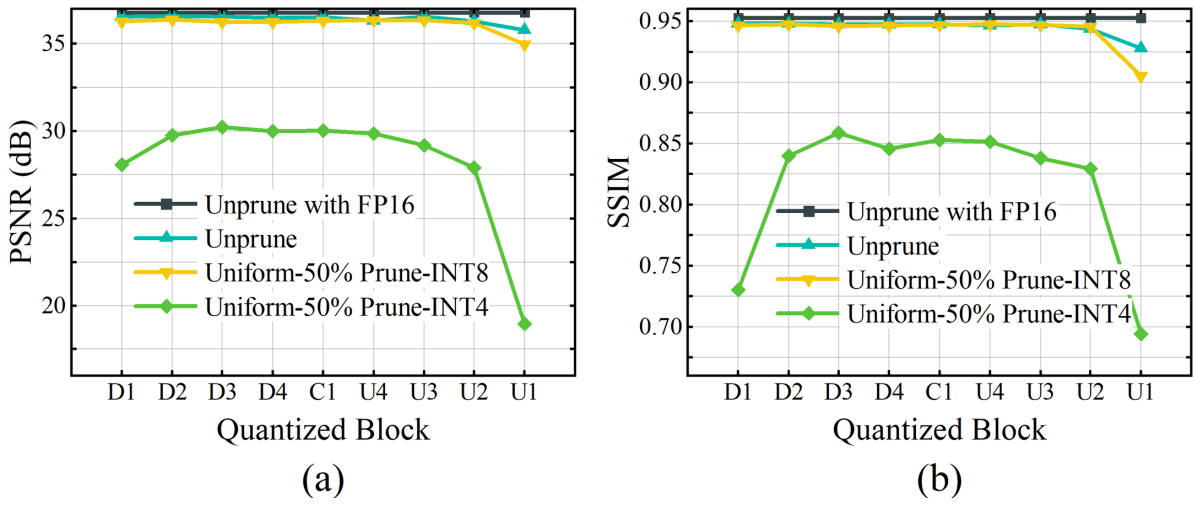

Fig. S2 Performance degradation for uniform 50% pruning network caused by INT4 quantization.

Table S1 The results of model quantization with different bit-width.

	FP16	INT8	Mixed-INT8	INT4	Mixed-INT4
PSNR	36.55	34.76	35.60	17.15	27.22
SSIM	0.9518	0.9148	0.9414	0.5783	0.7401

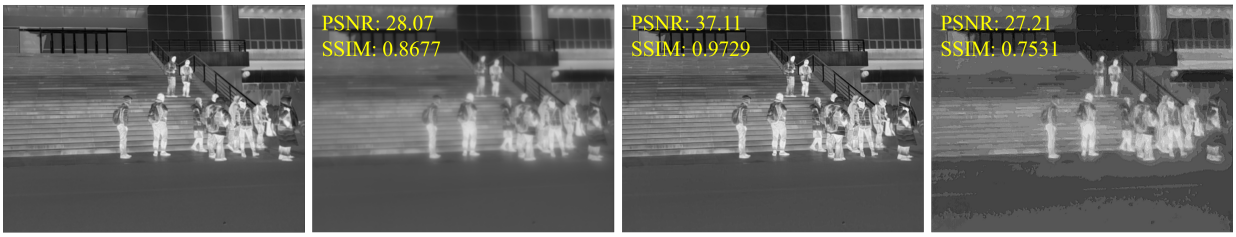

Fig. S3 Performance degradation for uniform 50% pruning network caused by INT4 quantization.

2 Supplementary explanation for hardware constraints in pruning

We provide the number of input and output channels for the models with different pruning ratios in Table 3, as shown in Table S2. The final convolution layer is not pruned. Therefore, the number of channels remains the same across all models. It is clear that the *Uniform-60%* and *Sensitive-A* models contain many layers whose channel numbers are not powers of 2 (bold numbers in the Table S2), which contradicts the hardware constraint in formula (3). As a result, these two models show poor inference speed on the RK3588. Although *Sensitive-A* demonstrates relatively good performance in image reconstruction, its poor inference speed makes it unsuitable for edge computing (see Table 3).

3 Live video of single-lens camera in outdoor experiments

To demonstrate the real-time performance of the single-lens camera at a frame rate of 25Hz, we provide the corresponding outdoor live video, including raw blurry video and clear video reconstructed in real-time on the integrated edge chip. The experiments were conducted at the Siping Road campus of Tongji University, Shanghai, China. This material validates the substantial potential

Table S2 The number of input and output channels for each layer in the models with different pruning ratios.

Block	Layer	Model					
		Unprune	Uniform-50%	Uniform-60%	Sensitive-A	Sensitive-B	Sensitive-C
D1	Conv 1	(1, 32)	(1, 16)	(1, **12**)	(1, **28**)	(1, 16)	(1, 16)
	Conv 2	(32, 32)	(16, 32)	(**12**, 32)	(**28**, 32)	(16, 32)	(16, 32)
D2	Conv 3	(32, 64)	(32, 32)	(32, **25**)	(32, **48**)	(32, 32)	(32, 32)
	Conv 4	(64, 64)	(32, 64)	(**25**, 64)	(**48**, 64)	(32, 64)	(32, 64)
D3	Conv 5	(64, 128)	(64, 64)	(64, **51**)	(64, 64)	(64, 64)	(64, 64)
	Conv 6	(128, 128)	(64, 128)	(**51**, 128)	(64, 128)	(64, 128)	(64, 128)
D4	Conv 7	(128, 256)	(128, 128)	(128, **102**)	(128, 64)	(128, 64)	(128, 64)
	Conv 8	(256, 256)	(128, 256)	(**102**, 256)	(64, 256)	(64, 256)	(64, 256)
C1	Conv 9	(256, 512)	(256, 256)	(256, **204**)	(256, 64)	(256, 128)	(256, 64)
	Conv 10	(512, 512)	(256, 512)	(**204**, 512)	(64, 512)	(128, 512)	(64, 512)
U4	ConvT 1	(512, 256)	(512, 128)	(512, **102**)	(512, 64)	(512, 64)	(512, 64)
	Conv 11	(512, 256)	(256, 128)	(**204**, **102**)	(128, 64)	(128, 64)	(128, 64)
	Conv 12	(256, 256)	(128, 256)	(**102**, 256)	(64, 256)	(64, 256)	(64, 256)
U3	ConvT 2	(256, 128)	(256, 64)	(256, **51**)	(256, 64)	(256, 64)	(256, 64)
	Conv 13	(256, 128)	(128, 64)	(**102**, 51)	(128, 64)	(128, 64)	(128, 64)
	Conv 14	(128, 128)	(64, 128)	(**51**, 128)	(64, 128)	(64, 128)	(64, 128)
U2	ConvT 3	(128, 64)	(128, 32)	(128, 25)	(128, 48)	(128, 32)	(128, 32)
	Conv 15	(128, 64)	(64, 32)	(**50**, **25**)	(**96**, **48**)	(64, 32)	(64, 32)
	Conv 16	(64, 64)	(32, 64)	(**25**, 64)	(**48**, **64**)	(32, 64)	(32, 64)
U1	ConvT 4	(64, 32)	(64, 16)	(64, **12**)	(64, **28**)	(64, 16)	(64, 16)
	Conv 17	(64, 32)	(32, 16)	(**24**, **12**)	(**56**, **28**)	(32, 16)	(32, 16)
	Conv 18	(32, 32)	(16, 32)	(**12**, 32)	(**28**, 32)	(16, 32)	(16, 32)
Final	Conv 19	(32, 1)	(32, 1)	(32, 1)	(32, 1)	(32, 1)	(32, 1)

of proposed edge acceleration strategy in practical single-lens computational imaging.

4 Ablation experiments on the performance impact of operator reconfiguration

We conducted detailed ablation experiments to evaluate the impact of operator reconfiguration. We empirically evaluated the effects on performance and inference speed resulting from the removal of LeakyReLU and the replacement of MaxPooling, both individually and in combination. The ablation experiments were conducted on unquantized networks running on a PC platform. As shown in Table S3, the experimental results demonstrate that our operator reconfiguration strategy maintains high reconstruction fidelity while effectively optimizing inference time.

Table S3 The ablation results of operator reconfiguration on imaging fidelity.

Method	PSNR	SSIM
Unprune	36.89	0.9533
Removing LeakyReLU	36.82	0.9528
Replacing MaxPooling	36.85	0.9530
Operator reconfiguration	36.80	0.9528

5 Basis for noise addition

During the experiments conducted for this study, we calibrated the noise model of the infrared detector used in our system. The calibration results indicate that the detector noise follows a Gaussian distribution with a mean of 0 and a variance of 0.0003. However, in practical imaging scenarios, the grayscale range of captured images can vary significantly due to differences in ambient illumination and target reflectivity across scenes. To meet the model's input requirements, we normalize the image grayscale values to the range [0,1] prior to restoration. This normalization process linearly amplifies the inherent noise.

Based on statistical analysis of real-world data, the amplification factor introduced by normalization typically ranges from 1 to 20. To ensure robustness under worst-case conditions, we adopted the maximum observed amplification factor of 20 in our simulations. Consequently, the noise variance used in simulation was set to: $0.0003 \times 20 = 0.006$. This approach ensures that the model is trained to handle the upper bound of noise levels that may be encountered in real-world applications.