TinyLUT: Tiny Look-Up Table for Efficient Image Restoration at the Edge

Huanan Li[1], Juntao Guan[1], Rui Lai[1], Sijun Ma[1], Lin Gu[2], Zhangming Zhu[1]

[1]Xidian University, Xi'an, China [2]RIKEN AIP,Tokyo103-0027,Japan

西安电子科技大学
XIDIAN UNIVERSITY

CONTENTS

01

Introduction

Introduction

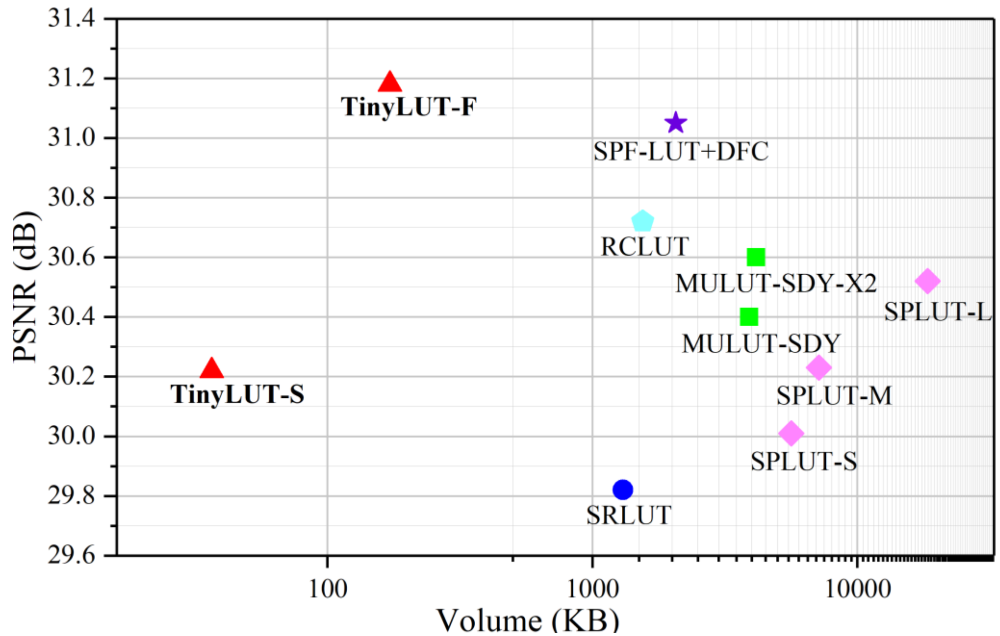
Video

Camera

Movie

Game

Introduction



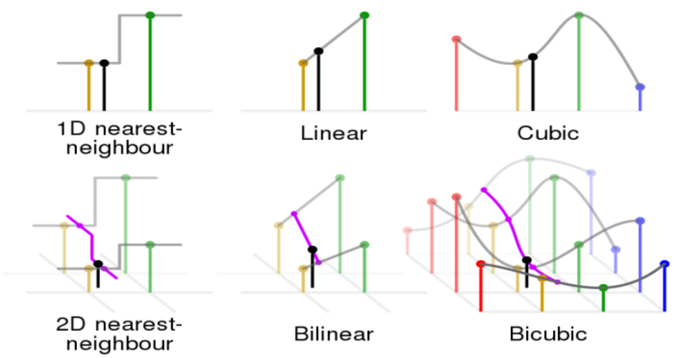
- The size of LUT exhibits exponential growth with the convolution kernel size, creating a storage bottleneck for its broader application on edge devices.
- We address the **storage explosion** challenge to promote the capacity of mapping the complex CNN models by LUT.
- The storage requirement of TinyLUT is around **4.1%** of MuLUT-SDY-X2 and amenable to on-chip cache, yielding competitive accuracy with over **5× lower inference latency** on Raspberry 4B than FSRCNN.

02

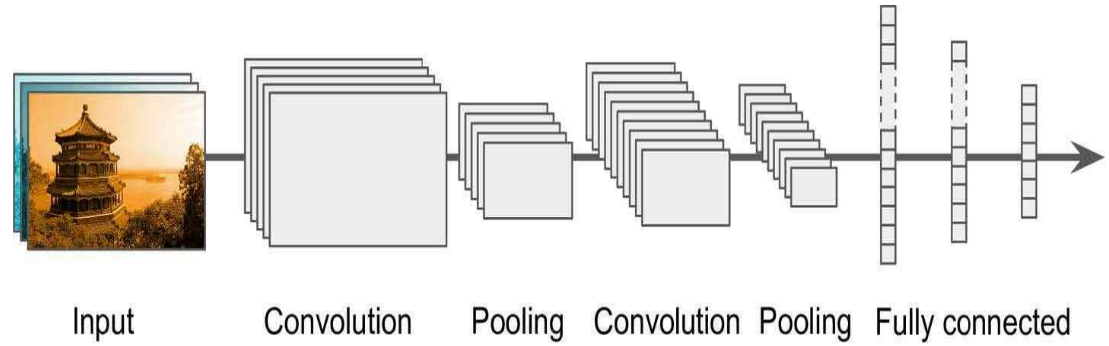
Related Works

Related Works

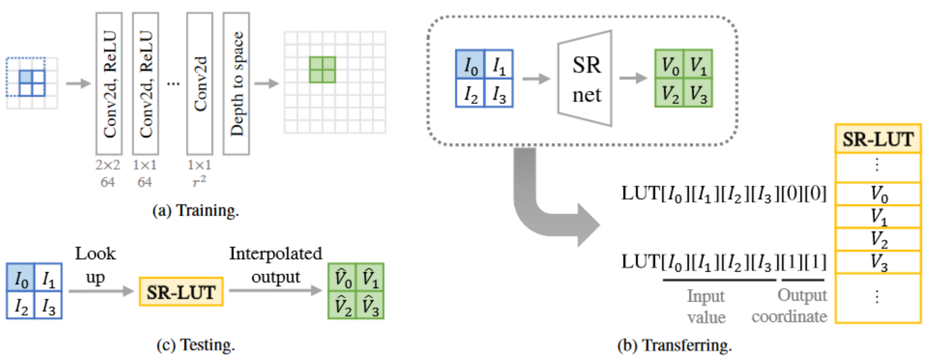
Interpolation

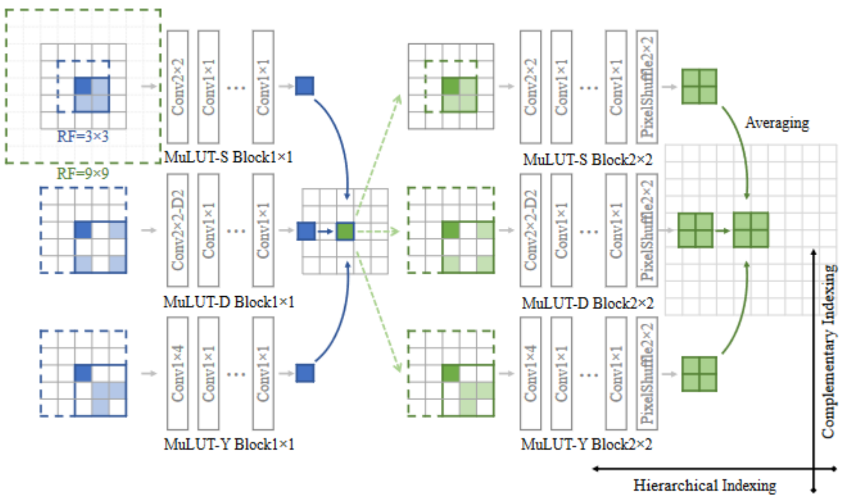


CNN-based Methods



LUT-based Methods



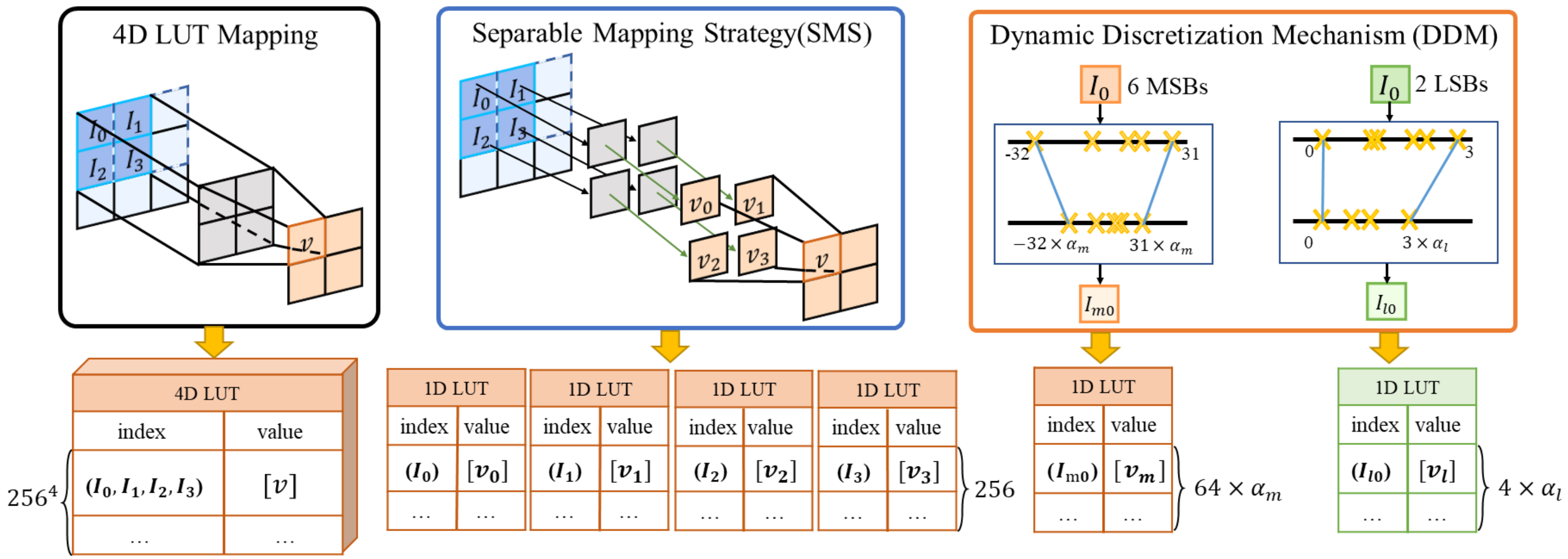


03

TinyLUT Method

TinyLUT Method

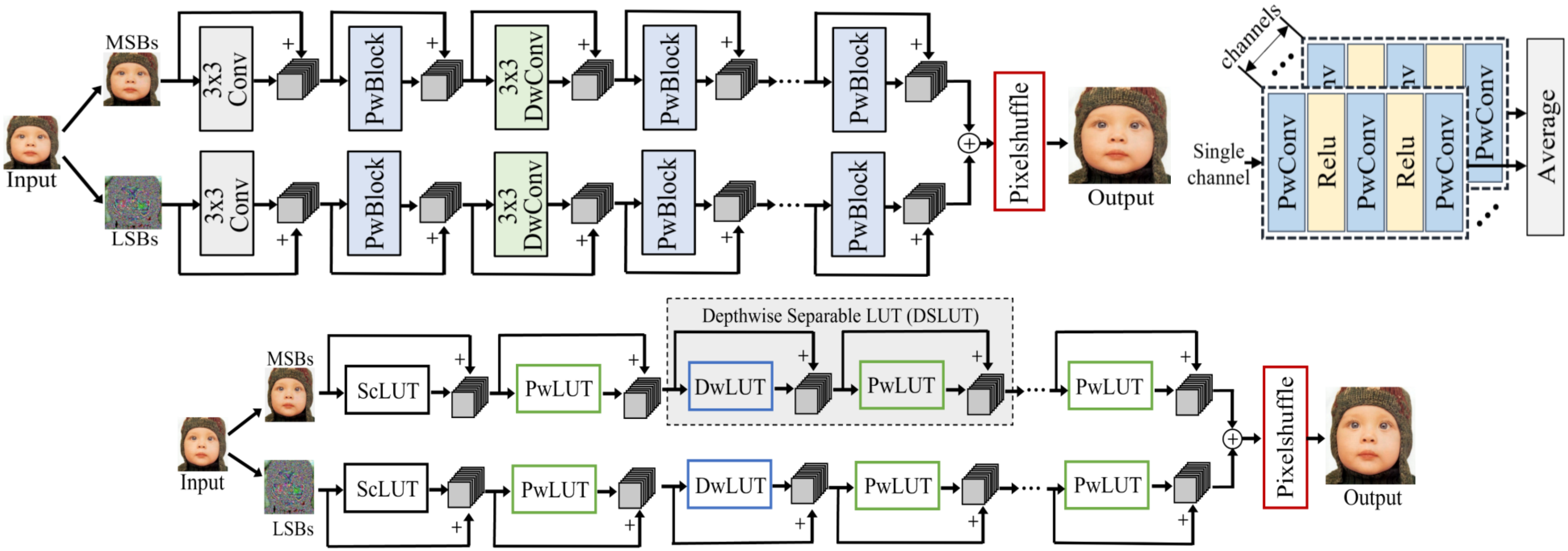
- **Separable Mapping Strategy(SMS) and Dynamic Discretization Mechanism (DDM)**



- We propose **separable mapping strategy(SMS)** to decouple 2 × 2 kernel into four 1 × 1 kernels which reduces the size of the LUT required to 256 × 4 × 8bit= 1KB.
- The **dynamic discretization mechanism (DDM)** adaptively explores the parameterized clipping level of quantization range for each channel based on the gradients with Straight-Through Estimator (STE), which significantly optimize the equilibrium of accuracy and LUT size.

TinyLUT Method

➢ Architecture



- The CNN model is built on standard convolution, **depthwise convolution (DwConv)** and **PwBlock**.

- Based on SMS and DDM, the standard convolution, DwConv and PwBlock are mapped by LUT respectively to build **standard convolution mapped LUT (ScLUT)** , **depthwise LUT (DwLUT)** and **pointwise LUT (PwLUT)**.

04

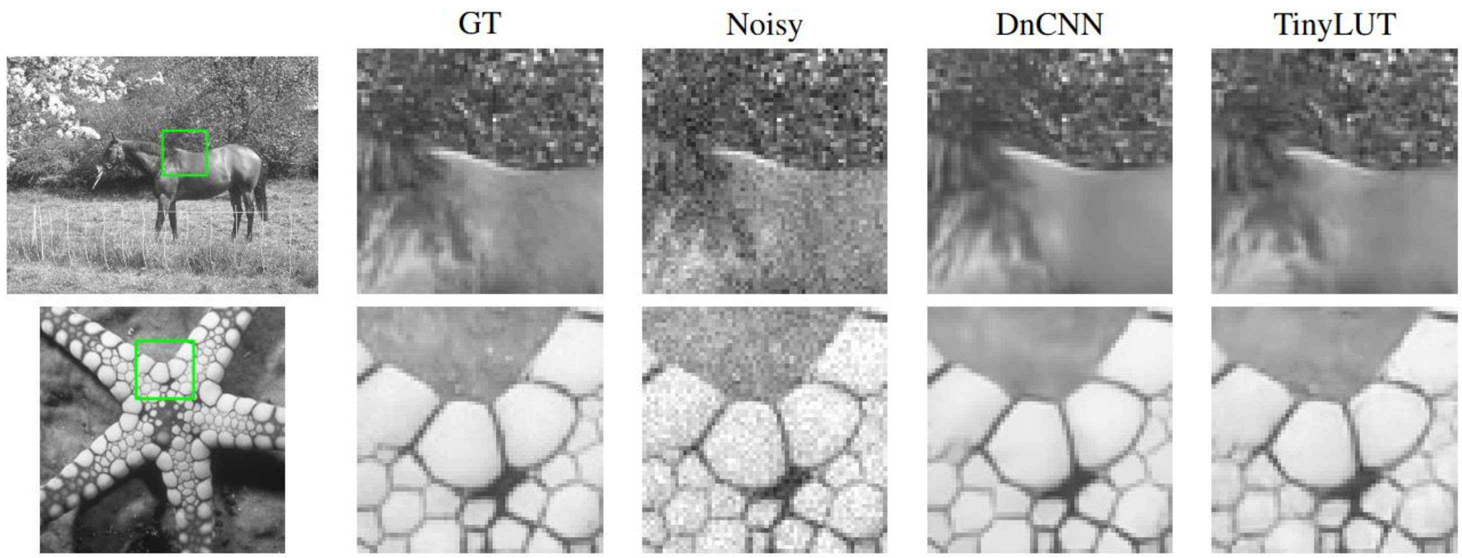
Experiment

➢ Super Resolution

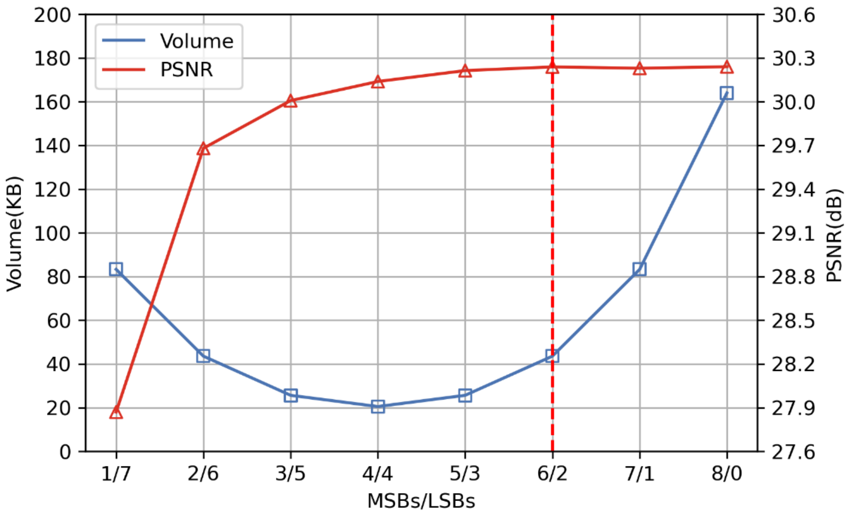
	Method	Storage	Runtime		Set5	Set14	Urban100	BSD100	Manga109	Average
			Xiaomi 11	Raspberry 4B						
LUTs	SRLUT-S [15]	1304KB	137ms	247ms	29.82/0.8478	27.01/0.7355	24.02/0.6990	26.53/0.6953	26.80/0.8380	26.84/0.7631
	SPLUT-L [13]	18432KB	265ms	456ms	30.52/0.8630	27.54/0.7520	24.46/0.7191	26.87/0.7090	27.70/0.8581	27.42/0.7802
	MuLUT-SDY-X2 [14]	4159KB	242ms	403ms	30.60/0.8653	27.60/0.7541	24.46/0.7194	26.86/0.7110	27.90/0.8633	27.48/0.7808
	RCLUT [34]	1549KB	232ms	-	30.72/0.8677	27.67/0.7577	24.57/0.7253	26.95/0.7154	28.05/0.8655	27.59/0.7863
	SPF-LUT+DFC [35]	2066KB	-	-	31.05/0.8755	27.88/**0.7632**	24.81/0.7357	27.08/**0.7190**	28.58/0.8779	27.88/0.7943
	TinyLUT-S	**37KB**	**29ms**	**88ms**	30.22/0.8535	27.33/0.7450	24.19/0.7066	26.71/0.7042	27.21/0.8458	27.13/0.7710
	TinyLUT-F	171KB	146ms	387ms	**31.18/0.8771**	**28.01**/0.7630	**24.92/0.7397**	**27.13**/0.7184	**28.83/0.8798**	**28.01/0.7956**
DNN	SRCNN [12]	228KB*	-	27448ms	30.48/0.8628	27.50/0.7513	24.52/0.7221	26.90/0.7101	27.10/0.8457	27.30/0.7784
	VDSR [24]	2660KB*	-	106972ms	31.35/0.8830	28.02/0.7680	25.18/0.7540	27.29/0.7260	28.50/0.8812	28.07/0.8024
	FSRCNN [8]	48KB*	350ms	2143ms	30.71/0.8656	27.60/0.7543	24.61/0.7263	26.96/0.7129	27.90/0.8610	27.56/0.7840
	CARN-M [26]	1648KB*	3300ms	17609ms	31.82/0.8898	28.29/0.7747	25.62/0.7694	27.42/0.7350	29.85/0.8993	28.60/0.8136

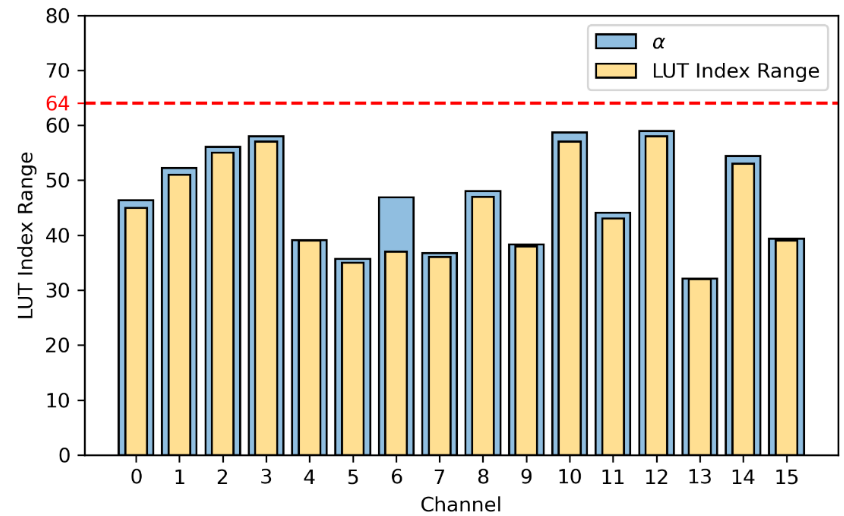
➢ Denoise

Method	Storage	Runtime		Set12			BSD68			Average
		Xiaomi 11	Raspberry 4B	15	25	50	15	25	50	
SRLUT [15]	82KB	**7ms**	**21ms**	30.42	27.19	22.62	29.78	26.85	22.39	26.54
MuLUT-SDY-X2 [14]	289KB	26ms	44ms	31.50	28.94	25.46	30.63	28.18	24.97	28.28
MuLUT-SDYEHO-X2 [48]	978KB	51ms	89ms	31.77	29.18	25.47	30.89	28.34	24.96	28.44
TinyLUT-S	**22KB**	20ms	27ms	31.10	28.26	24.29	30.24	27.48	23.83	27.53
TinyLUT-F	187KB	146ms	254ms	**32.22**	**29.69**	**26.27**	**31.20**	**28.65**	**25.53**	**28.93**
DnCNN [7]	2220KB*	635ms	6859ms	32.86	30.44	27.18	31.73	29.23	26.23	29.61



➢ Ablation Studies





Model	Method	Set5	Set14	Urban100	BSD100	Manga109	LUT Size
TinyLUT-S	Original	30.35	27.42	26.77	24.31	27.41	9.7×10^{24}PB
	Uniformly Sampled [15]	29.82	27.01	26.53	24.02	26.80	1.274MB
	SMS	30.24	27.33	26.72	24.19	27.23	**164KB**

- The accuracy of SMS is competitive with corresponding mapped neural network with 8bit data in SISR, while yielding over **7× storage reduction**.
- The images above illustrate the reasons for DDM selecting **6 MSBs** and **2 LSBs**. They also demonstrate the feasibility of **reducing storage** overhead through activation quantization in DDM.

05

Conclusion

Conclusion

In this paper, we analyze previous successful LUT-based deep learning approaches and summarize the key problem of storage explosion, which limits the further popularization of LUT in image restoration on edge devices. To address the storage explosion, we propose the separable mapping strategy (SMS) and dynamic discretization mechanism(DDM) to decompose the kernel and activation, respectively.

In particular, we design the TinyLUT framework based on SMS and DDM. By seamlessly integrating these innovations, TinyLUT-F sets a new record for SISR by achieving over 31dB PSNR on the Set5 dataset at just 171KB LUT storage. Overall, extensive experiments across seven benchmark datasets and two classic image restoration tasks demonstrate the effectiveness and efficiency of TinyLUT on resource-constrained devices.

西安电子科技大学
XIDIAN UNIVERSITY

Thanks