UCAN: Unified Convolutional Attention Network for Expansive Receptive Fields in Lightweight Super-Resolution

Supplementary Material

A. More details about Large Kernel Distillation

We provide a detailed architectural breakdown of our proposed block, including formal definitions of its parallel branches, a rigorous derivation of the Effective Receptive Field (ERF) for the large-kernel branch, and the final feature fusion strategy.

Let the input feature map be $\boldsymbol{X} \in \mathbb{R}^{H \times W \times C}$. The block processes $\boldsymbol{X}$ via three parallel branches.

A.1. Parallel Branch Definitions

Hierarchical Large Kernel (HLK) Branch. This branch captures long-range dependencies. We first define its core building blocks:

- A **separable depthwise convolution block**, $\mathcal{S}(\cdot)$, with kernel size k:

$$\mathcal{S}(\boldsymbol{X}, k) = f_{dw}^{k\times 1}\big(f_{dw}^{1\times k}(\boldsymbol{X})\big)$$

- A **dilated separable depthwise convolution block**, $\mathcal{S}_d(\cdot)$, with kernel k and dilation d:

$$\mathcal{S}_d(\boldsymbol{X}, k, d) = f_{dw}^{k\times 1,\, d}\Big(f_{dw}^{1\times k,\, d}(\boldsymbol{X})\Big)$$

The HLK branch has two configurations, both built as a stack. The first stage is always $\mathcal{S}(\boldsymbol{X}, k_{core})$, which forms a dense feature core.

1. **Standard Configuration ($\mathcal{F}_{LK\text{-S}}$):** This is a two-stage stack, used for smaller receptive fields.

$$\mathcal{F}_{LK\text{-S}}(\boldsymbol{X}) = \mathcal{S}_d(\mathcal{S}(\boldsymbol{X}, k_{core}), k_{core}, d) \tag{12}$$

2. **Large Configuration ($\mathcal{F}_{LK\text{-L}}$):** To achieve maximum receptive fields, this configuration extends the standard block into a three-stage stack. The first two stages are identical to the Standard Configuration, after which a third dilated separable depthwise convolution block using k_{extra} is appended.

$$\mathcal{F}_{LK\text{-L}}(\boldsymbol{X}) = \mathcal{S}_d(\mathcal{S}_d(\mathcal{S}(\boldsymbol{X}, k_{core}), k_{core}, d), k_{extra}, d) \tag{13}$$

The final output is $\boldsymbol{X}_{lk}^{\text{out}} = \mathcal{F}_{LK}(\boldsymbol{X})$, where $\mathcal{F}_{LK}$ is either $\mathcal{F}_{LK\text{-S}}$ or $\mathcal{F}_{LK\text{-L}}$.

Channel Branch (CB). The channel branch first computes channel-wise attention using a simple linear projection to weigh the importance of each feature map:

$$\boldsymbol{X}_c = f_{Linear}(\boldsymbol{X}), \tag{14}$$

This $\boldsymbol{X}_c$ is the output of channel branch.

Local Context (LC) Branch. This branch, denoted $\mathcal{F}_{LC}(\boldsymbol{X})$, captures fine-grained local details using a bottleneck (hourglass-style) block. Given a channel reduction factor r:

$$\begin{aligned} \boldsymbol{X}_l^{(1)} &= f_{\text{GELU}}\big(f_{\text{Conv}}^{1\times 1, C\to C/r}(\boldsymbol{X})\big) \\ \boldsymbol{X}_l^{(2)} &= f_{\text{GELU}}\big(f_{\text{Conv}}^{3\times 3}(\boldsymbol{X}_l^{(1)})\big) \\ \mathcal{F}_{LC}(\boldsymbol{X}) \equiv \boldsymbol{X}_l^{(\text{out})} &= f_{\text{Conv}}^{1\times 1, C/r\to C}(\boldsymbol{X}_l^{(2)}) \end{aligned} \tag{15}$$

A.2. Effective Receptive Field (ERF) Derivation

We provide a ERF analysis for the HLK branch, which is defined by the composition of the 1D horizontal convolutions ($1 \times k$). The ERF of a stack of convolutions is $\text{ERF}_{\text{out}} = \text{ERF}_{\text{in}} + (k-1) \times d$, where d is the dilation rate.

Standard Configuration ($\mathcal{F}_{LK\text{-S}}$). This configuration stacks $\mathcal{S}(\cdot, k_{core})$ and $\mathcal{S}_d(\cdot, k_{core}, d)$.

1. The base $\mathcal{S}(\cdot, k_{core})$ block (specifically $f_{dw}^{1\times k_{core}}$) establishes an $\text{ERF}_{\text{in}} = k_{core}$.
2. The second stage, $\mathcal{S}_d(\cdot, k_{core}, d)$, (specifically $f_{dw}^{1\times k_{core}, d}$) adds $(k_{core}-1)d$.

The total ERF is therefore:

$$\text{ERF}_{\text{S}} = k_{core} + (k_{core} - 1)d \tag{16}$$

Large Configuration ($\mathcal{F}_{LK\text{-L}}$). This configuration stacks $\mathcal{S}(\cdot, k_{core})$ and $\mathcal{S}_d(\cdot, k_{extra}, d)$.

1. The base $\mathcal{S}(\cdot, k_{core})$ block establishes an $\text{ERF}_{\text{in}} = k_{core}$.
2. The second stage, $\mathcal{S}_d(\cdot, k_{core}, d)$, adds $(k_{core}-1)d$.
3. The third stage, $\mathcal{S}_d(\cdot, k_{extra}, d)$, adds $(k_{extra}-1)d$.

The total ERF is therefore:

$$\text{ERF}_{\text{L}} = k_{core} + (k_{core} - 1)d + (k_{extra} - 1)d \tag{17}$$

This derivation confirms the formulas used to generate the configurations in Table 4.

A.3. Feature Fusion and Final Output

Finally, the outputs from the three branches are fused. The local and large-kernel spatial features are combined additively, and the result is modulated by the channel attention vector $\boldsymbol{X}_c$.

$$\begin{aligned} \boldsymbol{X}_{\text{spatial}} &= \mathcal{F}_{LC}(\boldsymbol{X}) + \mathcal{F}_{LK}(\boldsymbol{X}) \\ \boldsymbol{X}_{\text{final}} &= \boldsymbol{X}_c \odot \boldsymbol{X}_{\text{spatial}} \end{aligned} \tag{18}$$

where $\odot$ denotes element-wise multiplication, with $\boldsymbol{X}_c$ being broadcast along the spatial dimensions. This allows

Table 4. LKSA module configurations and resulting effective kernel sizes (1D ERF). For rows with $k_{\text{extra}} = -$, we use the $\mathcal{F}_{LK\text{-S}}$ formula (16); otherwise, we use the $\mathcal{F}_{LK\text{-L}}$ formula (17).

k_{core}	Dilation d	k_{extra}	Final ERF
3	1	–	5
5	1	–	9
5	**2**	–	**13**
5	3	11	47
5	3	13	53
5	3	17	65

$\boldsymbol{X}_l^{(\text{out})}$ to contribute fine-grained details, $\boldsymbol{X}_{lk}^{(\text{out})}$ to provide long-range context, and $\boldsymbol{X}_c$ to reweight the fused features based on channel-wise saliency.

B. Limitations of ReLU and ELU $+ 1$ in Linear Attention

Recent linear attention architectures often adopt simple non-negative feature maps such as ReLU or $\phi(x) = \text{ELU}(x) + 1$. However, we identify that both choices are suboptimal for single-image super-resolution, albeit for different reasons.

ReLU feature map. The ReLU activation is defined as $\text{ReLU}(x) = \max(0, x)$. For a query–key pair $q_i, k_j \in \mathbb{R}^d$, the linearized attention kernel becomes:

$$\phi_{\text{ReLU}}(q_i)^\top \phi_{\text{ReLU}}(k_j) = \sum_{c=1}^{d} \max(0, q_{i,c}) \max(0, k_{j,c}). \tag{19}$$

In contrast, standard softmax attention relies on the dot product $q_i^\top k_j = \sum_c q_{i,c} k_{j,c}$, where negative components are fully preserved: if $q_{i,c} < 0$ and $k_{j,c} < 0$, their product is positive and contributes to similarity. By forcing non-negativity, ReLU zeroes out these dimensions, discarding potentially informative negative–negative alignments. In super-resolution, where high-frequency details often rely on subtle, signed correlations across channels, this information loss degrades the approximation of the softmax kernel.

ELU$+1$ feature map. Let $\sigma(\cdot) = \text{ELU}(\cdot)$. An alternative feature map is $\phi_{\text{ELU}+1}(x) = \sigma(x) + 1$. For a single channel c, the interaction expands as:

$$[\sigma(q_{i,c})+1][\sigma(k_{j,c})+1] = \sigma(q_{i,c})\sigma(k_{j,c}) + \sigma(q_{i,c}) + \sigma(k_{j,c}) + 1. \tag{20}$$

Summing over all d channels yields the vector-form kernel:

$$\begin{aligned} \phi(q_i)^\top \phi(k_j) &= \sum_{c=1}^{d} [\sigma(q_{i,c})\sigma(k_{j,c}) + \sigma(q_{i,c}) + \sigma(k_{j,c}) + 1] \\ &= \langle \sigma(q_i), \sigma(k_j) \rangle + \mathbf{1}^\top \sigma(q_i) + \mathbf{1}^\top \sigma(k_j) + d, \end{aligned} \tag{21}$$

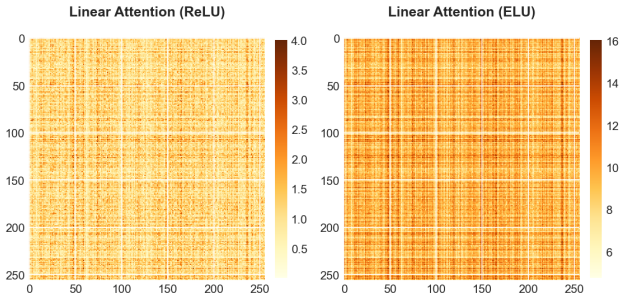


Figure 7. Visualization of attention maps for Linear Attention using ReLU and ELU $+ 1$ (computed with sequence length $N = 256$). The lack of normalization leads to high-magnitude artifacts.

where $\mathbf{1} \in \mathbb{R}^d$ is the all-ones vector. Crucially, the last three terms in Eq. (21) do not encode pairwise similarity; they represent global query/key biases and a constant offset d. In high-dimensional backbones, these bias terms (order $\mathcal{O}(d)$) often dominate the true similarity term $\langle \sigma(q_i), \sigma(k_j) \rangle$. This effectively compresses the attention variation into a narrow range, reducing the contrast between similar and dissimilar tokens and blurring high-frequency reconstruction details.

Figure 7 provides empirical validation of our analysis. We visualize the attention scores computed by the standard softmax operation (scaled by $1/\sqrt{d}$) versus the unnormalized linear kernels. Due to the lack of intrinsic normalization, both ReLU and ELU $+ 1$ produce attention scores with significantly larger magnitudes. This effect is particularly pronounced for ELU$+1$, where scores surge to values around 14. This empirical evidence corroborates the derivation above, confirming that the global bias terms in the ELU $+ 1$ kernel dominate the pairwise similarity signal.

Ranking enhancement. To validate the hypothesis that discarding negative components degrades representational power, we design a symmetric ReLU variant that explicitly preserves negative directionality via concatenation:

$$\phi_{\text{sym}}(x) = [\text{ReLU}(x), \text{ReLU}(-x)]. \tag{22}$$

As illustrated in Fig. 8, this symmetric formulation significantly improves the output ranking consistency compared to the standard ReLU baseline. However, this modification does not address the fundamental limitations of ReLU in linear attention, namely its unbounded linear growth and lack of intrinsic normalization, which can lead to training instability.

Consequently, we adopt the Hedgehog feature map. Unlike the fixed ReLU basis, Hedgehog employs a learnable Multi-Layer Perceptron (MLP) combined with a Softmax activation. This design offers two critical advantages: (1) the Softmax ensures normalized, stable attention weights, and (2) the MLP provides the flexibility to adaptively emphasize informative features across both positive and negative

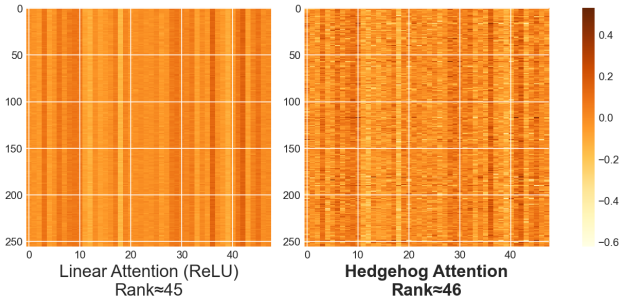


Figure 8. **Ranking consistency analysis.** We compare the output ranking of Linear Attention using standard ReLU, Symmetric ReLU, and the Hedgehog Feature Map (sequence length $N = 256$). While adding negative information (Sym-ReLU) improves consistency, Hedgehog achieves superior performance through learnable stability.

regimes. This analysis confirms that our architectural choice is principled rather than heuristic, positioning our method as a robust, theoretically grounded alternative to recent Mamba-based [8, 10] or standard Softmax [27] attention architectures.

C. More comparison with SOTA methods

Table 6 presents a comprehensive quantitative comparison with state-of-the-art lightweight methods. The results demonstrate that our proposed UCAN achieves a superior efficiency-performance trade-off by significantly reducing computational overhead while maintaining competitive restoration quality. Notably, at scale $\times 2$, UCAN requires only 146.3G FLOPs. This represents a reduction of approximately $\mathbf{47}\%$ compared to PFT-light and a significant decrease relative to SwinIR-light. This efficiency advantage persists at scale $\times 4$ where UCAN operates with merely 38.1G FLOPs, which is nearly half the computational cost of PFT-light. Despite this reduced complexity, UCAN delivers robust performance on challenging benchmarks such as Set14, BSDS100, and Manga109. For instance, on the Set14 dataset at scale $\times 2$, UCAN achieves a leading PSNR of 34.27 dB. Moreover, our scalable variant UCAN-L consistently secures the best or second-best results across complex scenarios. On Manga109 at scale $\times 2$, UCAN-L surpasses the previous best method with a score of 39.66 dB, further validating the effectiveness of our architecture in recovering fine structural details and textures.

Table 5. More ablation study. We train all models on DIV2K for 400K iterations, and test on Set5 and Urban100 ($\times 2$).

Case	Configuration	Set5		Urban100	
		PSNR	SSIM	PSNR	SSIM
Hybrid Block (HB)	depth=3	38.30	0.9617	33.05	0.9363
	depth=2	38.24	0.9615	32.95	0.9353
	depth=1	38.16	0.9612	32.71	0.9336
Large Kernel Distillation (LKD)	LKD depth=4	38.33	0.9618	33.12	0.9369
	LKD depth=3	38.33	0.9618	33.08	0.9366
	LKD depth=2	38.31	0.9618	33.04	0.9363
	LKD depth=1	38.31	0.9617	32.99	0.9358
	Kernel size=5	38.33	0.9618	33.12	0.9369
	Kernel size=53	38.33	0.9618	33.18	0.9375
	Kernel size=65	38.33	0.9618	33.19	0.9375
UCAN	**Base**	**38.34**	**0.9618**	**33.22**	**0.9379**

D. More ablation

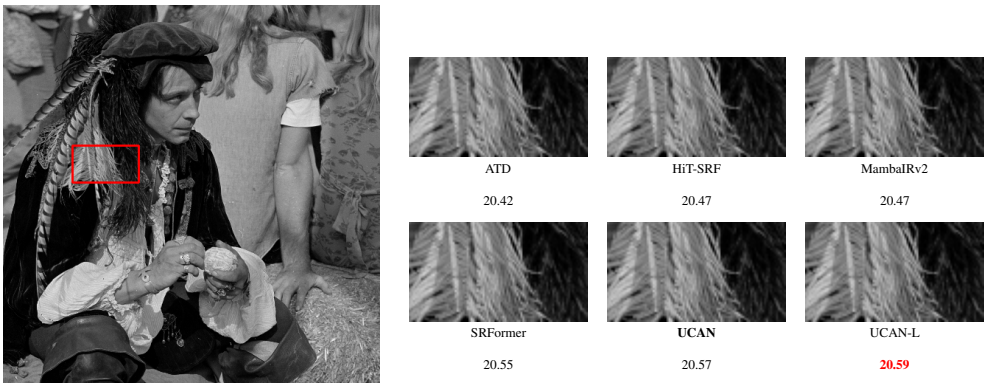
We investigate the contribution of each core component by systematically analyzing the depth of the Hybrid Blocks (HB), the Large Kernel Distillation (LKD) modules, and the choice of kernel size. First, varying the **HB depth** from 3 to 1 reveals a steady performance decline; specifically, on Urban100, PSNR drops from 33.05 dB to 32.71 dB, confirming that multiple HB stages are essential for iteratively integrating local and global contexts. A similar trend is observed when pruning the **LKD depth** from 4 to 1, where performance degrades from 33.12 dB to 32.99 dB, suggesting that deep distillation layers are critical for reconstructing intricate patterns. Finally, we examine the impact of **kernel size** ($k \in \{5, 53, 65\}$) within the LKD. While performance on the simpler Set5 dataset saturates across all sizes (38.33 dB), larger kernels prove advantageous on the complex Urban100 benchmark, where increasing k from 5 to 65 improves PSNR from 33.12 dB to 33.19 dB. Based on these findings, our final UCAN Base model incorporates the optimal configuration, achieving a peak performance of 33.22 dB.

Table 6. Quantitative comparison on ***lightweight image super-resolution*** with state-of-the-art methods. The best and second-best results are shown in **bold** and underlined, respectively.

Method	scale	#param	FLOPs	Set5		Set14		BSDS100		Urban100		Manga109	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SwinIR-light [18]	2×	910K	244.2G	38.14	0.9611	33.86	0.9206	32.31	0.9012	32.76	0.9340	39.12	0.9783
ELAN [42]	2×	621K	245.2G	38.27	0.9616	33.94	0.9207	32.30	0.9012	32.76	0.9340	39.11	0.9782
IPG-Tiny [34]	2×	621K	203.1G	38.17	0.9611	34.24	0.9236	32.35	0.9018	33.04	0.9359	39.31	0.9786
PFT-light [23]	2×	776K	278.3G	38.36	**0.9620**	34.19	0.9232	32.43	0.9030	**33.67**	**0.9411**	39.55	**0.9792**
UCAN (Our)	2×	689K	146.3G	38.34	0.9618	**34.27**	**0.9242**	32.39	0.9025	33.22	0.9379	39.54	0.9790
UCAN-L (Our)	2×	886K	182.4G	**38.37**	0.9619	34.19	0.9224	**32.44**	**0.9031**	33.39	0.9393	**39.66**	0.9789
SwinIR-light [18]	3×	918K	111.2G	34.62	0.9289	30.54	0.8463	29.20	0.8082	28.66	0.8624	33.98	0.9478
ELAN [42]	3×	629K	90.1G	34.61	0.9288	30.55	0.8463	29.21	0.8081	28.69	0.8624	34.00	0.9478
IPG-Tiny [34]	3×	878K	109.0G	34.64	0.9292	30.61	0.8470	29.26	0.8097	28.93	0.8666	34.30	0.9493
PFT-light [23]	3×	783K	123.5G	34.81	0.9305	**30.75**	0.8493	29.33	0.8116	**29.43**	**0.8759**	34.60	0.9510
UCAN (Our)	3×	696K	64.6G	**34.83**	0.9308	30.72	0.8493	29.32	0.8121	29.15	0.8712	34.62	0.9508
UCAN-L (Our)	3×	893K	81.3G	34.81	**0.9311**	**30.75**	**0.8500**	**29.34**	**0.8127**	29.29	0.8738	**34.79**	**0.9516**
SwinIR-light [18]	4×	930K	63.6G	32.44	0.8976	28.77	0.7858	27.69	0.7406	26.47	0.7980	30.92	0.9151
ELAN [42]	4×	640K	54.1G	32.43	0.8975	28.78	0.7858	27.69	0.7406	26.54	0.7982	30.92	0.9150
IPG-Tiny [34]	4×	887K	61.3G	32.51	0.8987	28.85	0.7873	27.73	0.7418	26.78	0.8050	31.22	0.9176
PFT-light [23]	4×	792K	69.6G	32.63	0.9005	28.92	0.7891	27.79	0.7445	**27.20**	**0.8171**	31.51	0.9204
UCAN (Our)	4×	705K	38.1G	32.65	0.9010	28.95	0.7899	27.79	0.7454	26.89	0.8097	31.50	0.9200
UCAN-L (Our)	4×	902K	48.4G	**32.68**	**0.9015**	**28.99**	**0.7917**	**27.80**	**0.7459**	27.06	0.8134	**31.63**	**0.9212**

Set 5 - Baby

Figure 9. Visual comparison between the ground truth and different methods on Set5 - baby.

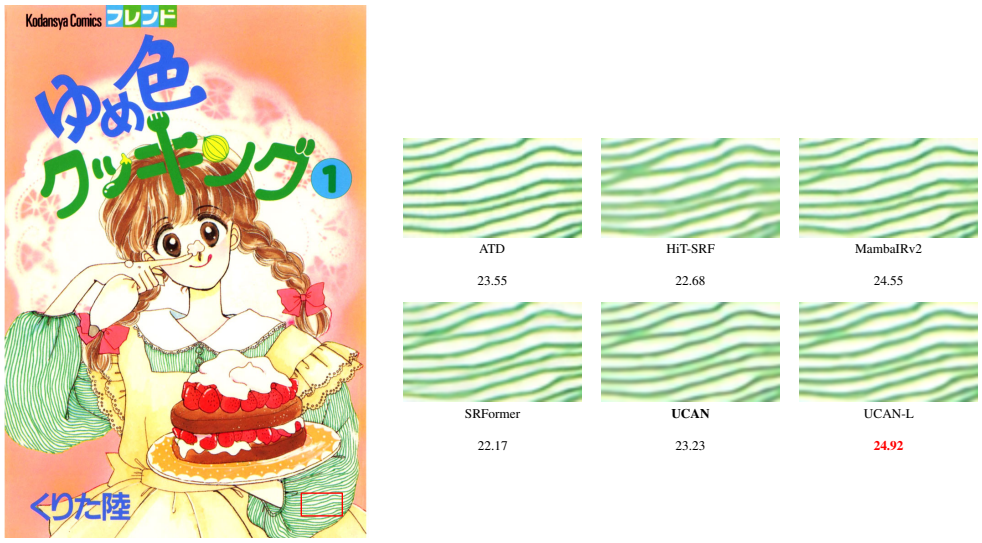


Set 14 - Man

Figure 10. Visual comparison between the ground truth and different methods on Set14 - man.

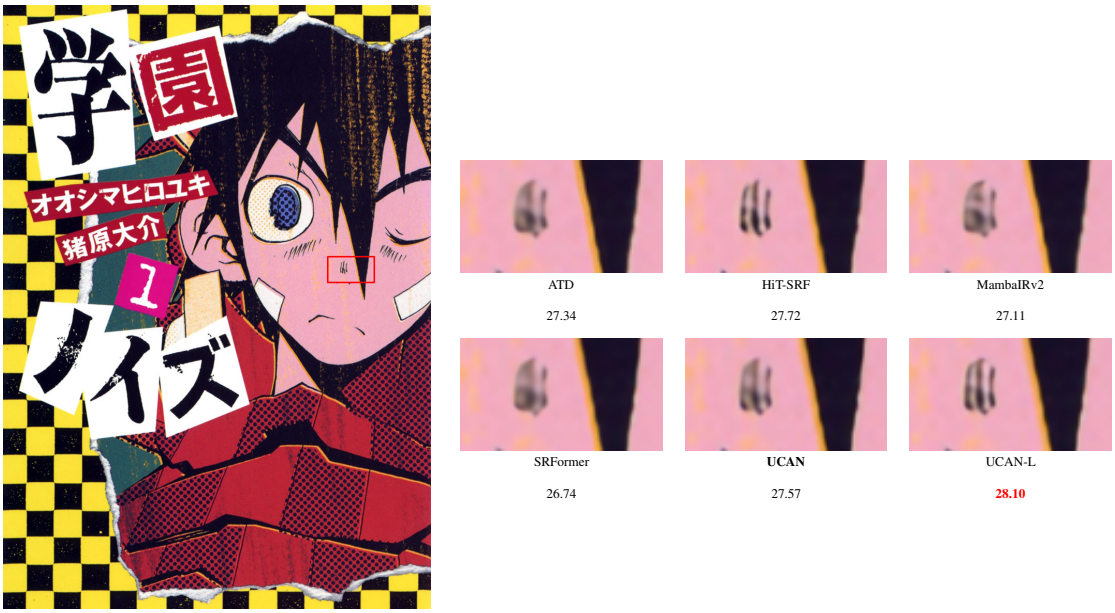
B100 - 300091

Figure 11. Visual comparison between the ground truth and different methods on B100 - 300091.



Manga109 - Yumeko Cooking

Figure 12. Visual comparison between the ground truth and different methods on Manga109 - Yumeko Cooking.

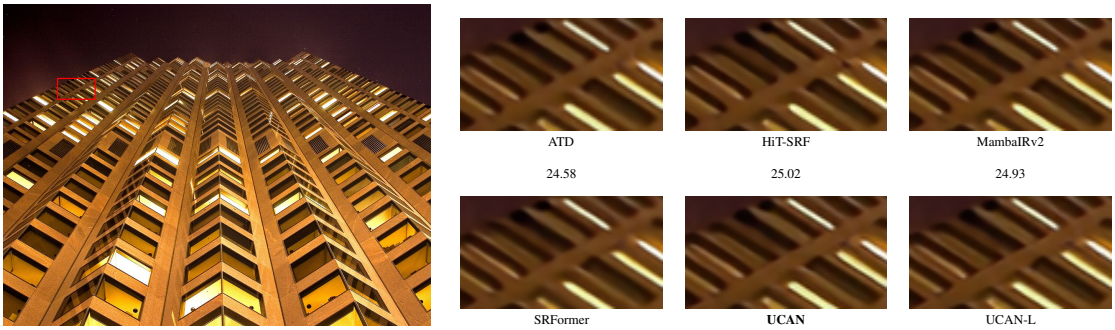


Manga109 - Gakuen Noise

Figure 13. Visual comparison between the ground truth and different methods on Manga109 - Gakuen Noise.

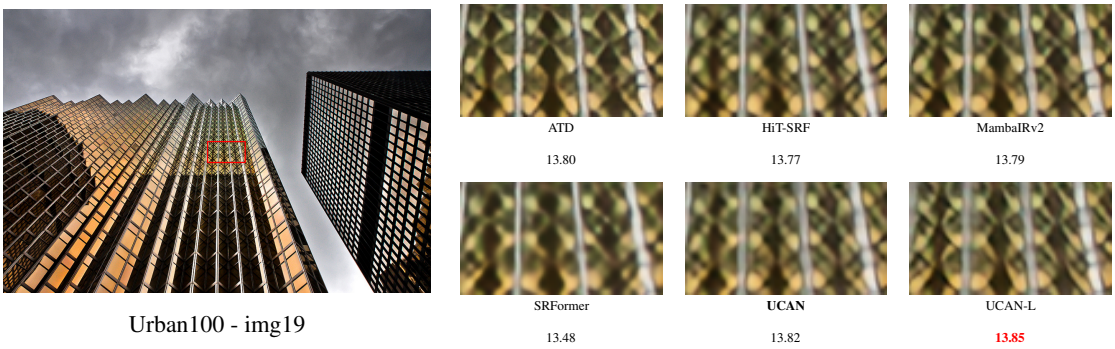
Manga109 - Yasakii Akuma

Figure 14. Visual comparison between the ground truth and different methods on Manga109 - Yasakii Akuma.



Urban100 - img15

Figure 15. Visual comparison between the ground truth and different methods on Urban100 - 015.

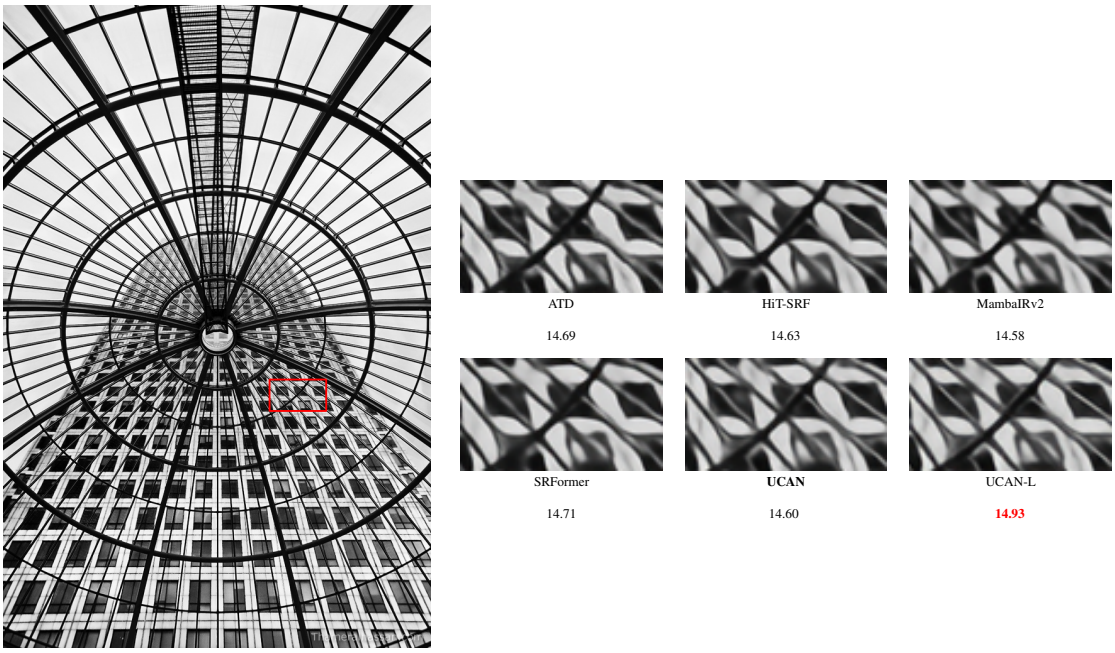


Urban100 - img19

Figure 16. Visual comparison between the ground truth and different methods on Urban100 - img19.

Urban100 - img24

Figure 17. Visual comparison between the ground truth and different methods on Urban100 - img24.



Urban100 - img72

Figure 18. Visual comparison between the ground truth and different methods on Urban100 - img72.