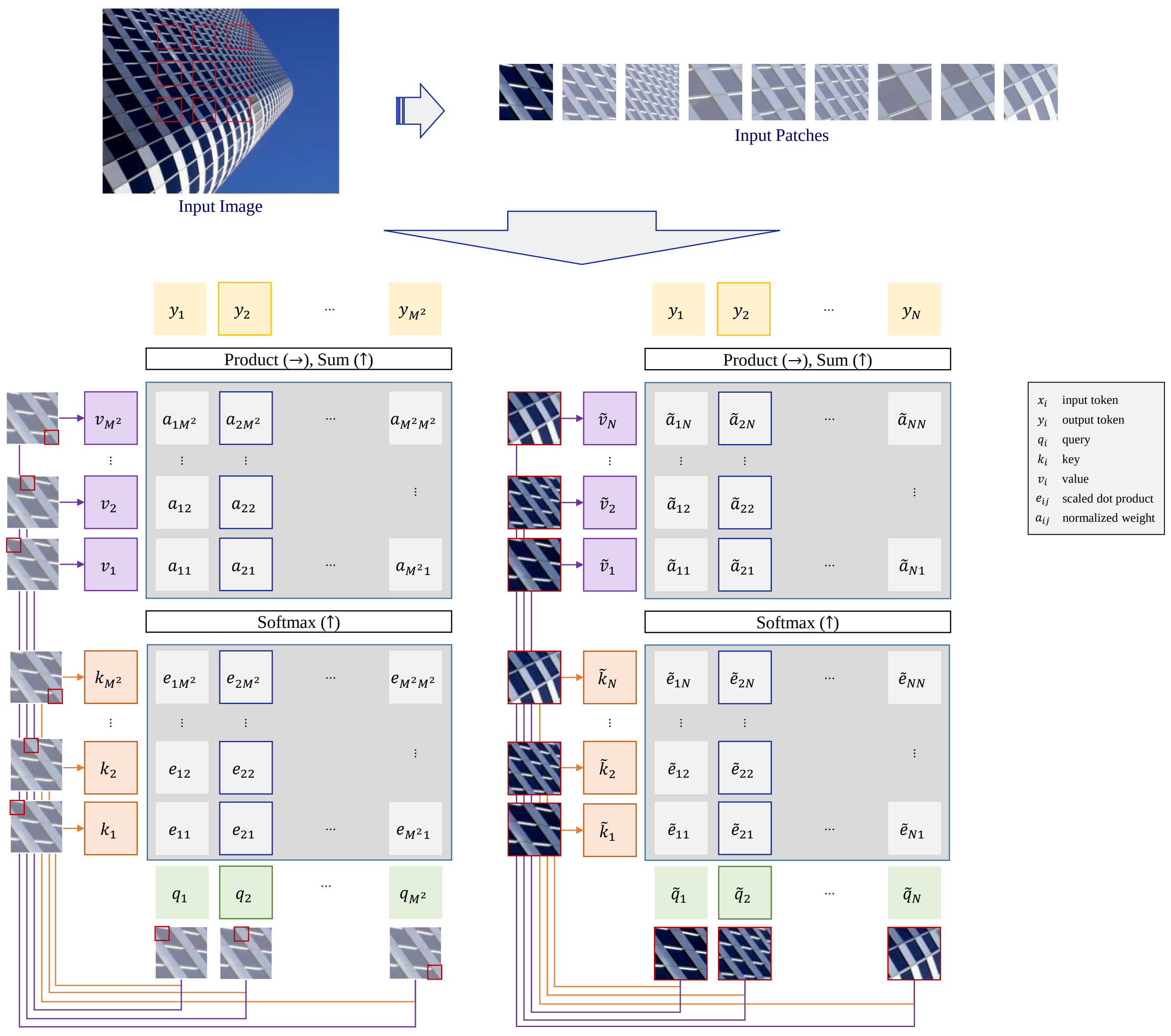
iiTransformer: A Unified Approach to Exploiting Local and Non-Local Information for Image Restoration

Soo Min Kang, Youngchan Song, Hanul Shin, and Tammy Lee
Samsung Research
https://github.com/SamsungLabs/iiTransformer

Introduction

1. Pixels surrounding the degraded pixel often provide useful information, and
2. images in general contain repetitive information

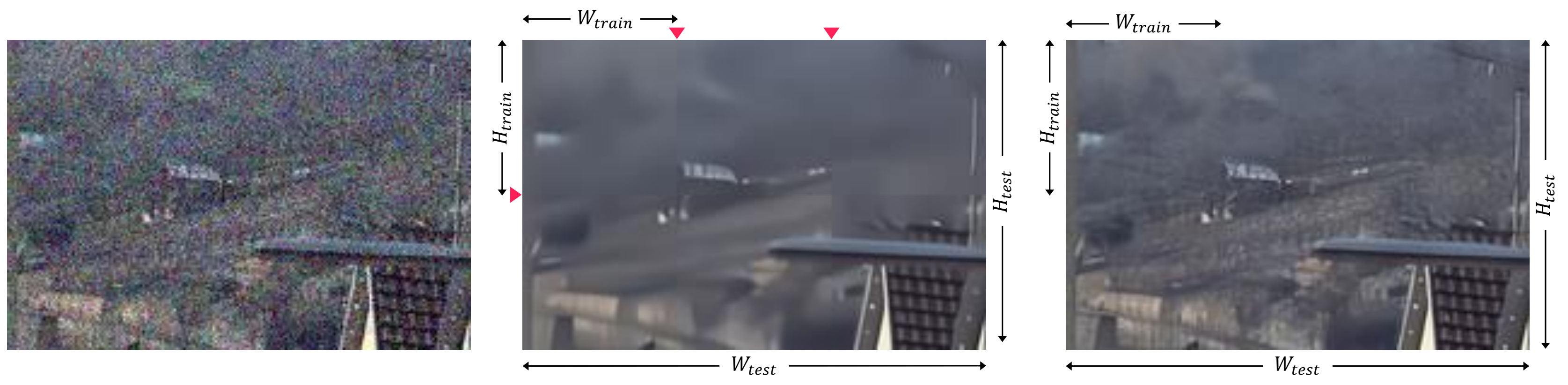
- Local and non-local relationships can be captured by considering long-range dependencies at the pixel- or patch-level using the self-attention module of Transformers



Intra Self-Attention (aMSA)

Inter Self-Attention (eMSA)

- Existing inter SA module-based ViT require resolution of training and inference images to match, resulting in either
  (i) increased computational complexity due to overlapping sliding windows, or
  (ii) patch boundary artifacts



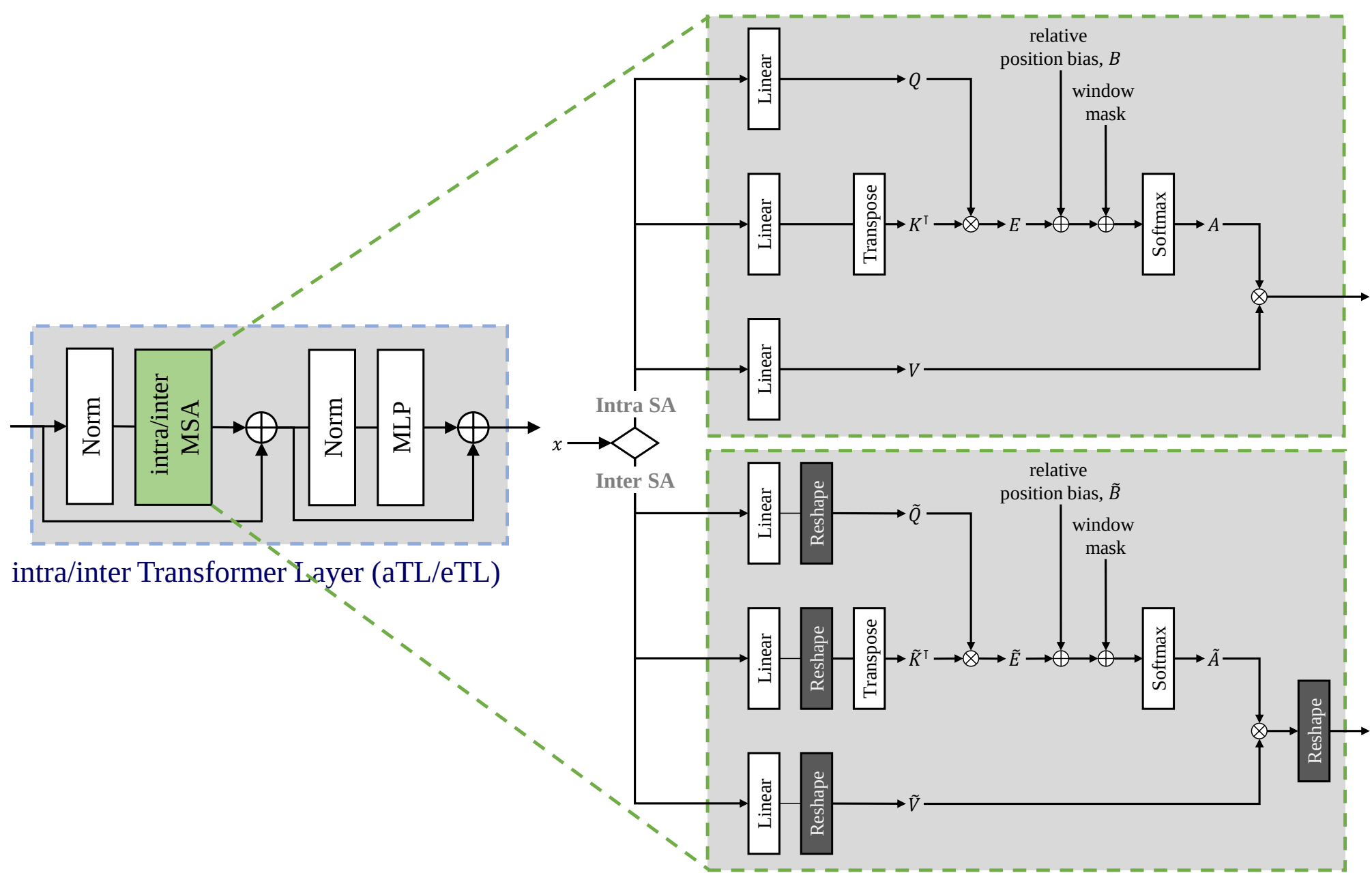
(a) Noisy Input Image (LQ)

(b) Presence of Boundary Artifacts

(c) Absence of Boundary Artifacts

Methodology

1. Exploiting local and non-local information for image restoration
   - intra SA (aMSA): treat pixels as tokens to compute local pixelwise correlations
   - inter SA (eMSA): treat patches as tokens to compute non-local patchwise correlations
   - shape of projected tokens used to compute the attention matrix differ between aMSA and eMSA


Norm
intra/inter MSA
Norm
MLP
Intra SA
Inter SA
Linear
Reshape
Transpose
Softmax
relative position bias, B
window mask
relative position bias, $\tilde{B}$
Q $K^\top$ E A V y
$\tilde{Q}$ $\tilde{K}^\top$ $\tilde{E}$ $\tilde{A}$ $\tilde{V}$ x

intra/inter Transformer Layer (aTL/eTL)

intra/inter Multi-head Self-Attention (aMSA/eMSA)

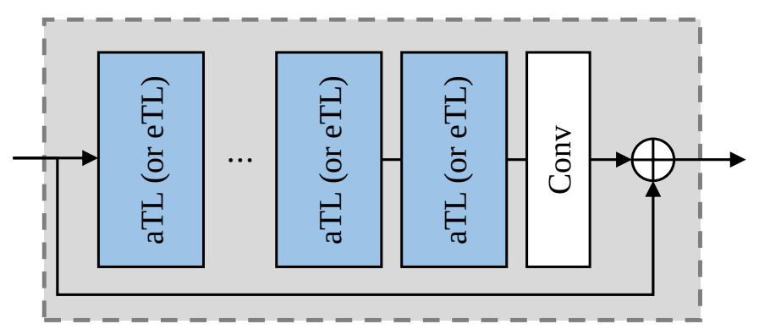
$$Q = XW^Q, K = XW^K, V = XW^V, \qquad Q, K, V \in \mathbb{R}^{N\times M^2\times C}$$

$$aSA(X) = \mathtt{softmax}\left(\frac{QK^P}{\sqrt{C}} + B\right)V, \qquad aSA(X) \in \mathbb{R}^{N\times M^2\times C}$$

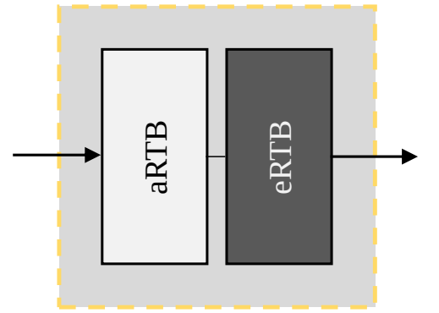
$$\tilde{Q} = \mathtt{reshape}(Q), \tilde{K} = \mathtt{reshape}(K), \tilde{V} = \mathtt{reshape}(V), \qquad \tilde{Q}, \tilde{K}, \tilde{V} \in \mathbb{R}^{N\times M^2 C}$$

$$eSA(\tilde{X}) = \mathtt{softmax}\left(\frac{\tilde{Q}\tilde{K}^P}{\sqrt{C}} + \tilde{B}\right)\tilde{V}, \qquad eSA(\tilde{X}) \in \mathbb{R}^{N\times M^2 C}$$

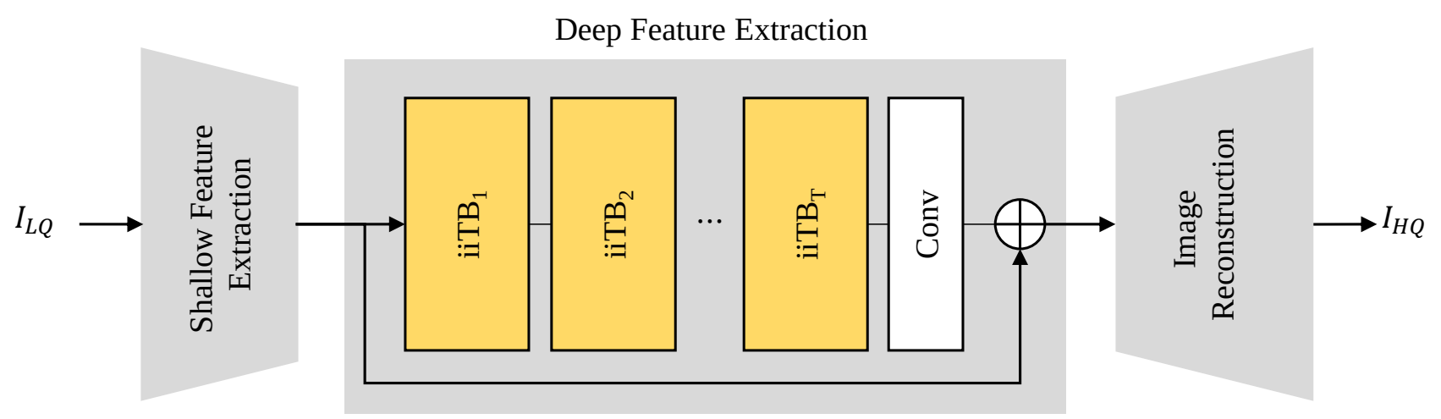
$$eSA(X) = \mathtt{reshape}\left(eSA(\tilde{X})\right), \qquad eSA(X) \in \mathbb{R}^{N\times M^2\times C}$$



intra Residual Transformer Block (aRTB)
(or inter Residual Transformer Block (eRTB))



iiTransformer Block (iiTB)

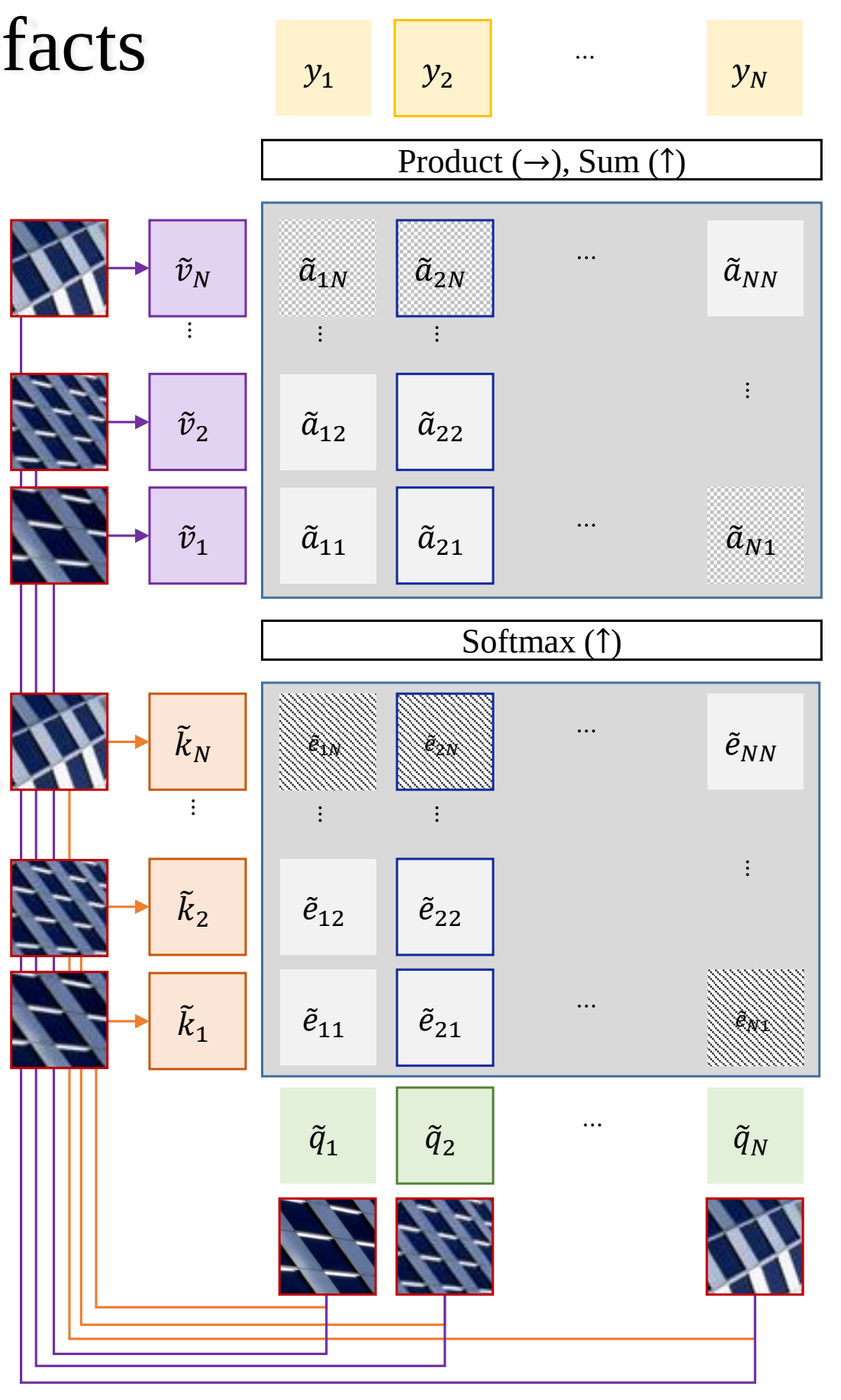


iiTransformer

2. Support arbitrary resolutions without boundary artifacts by masking patch-to-patch distance that exceeds furthest patch-to-patch distance used during training:

$$\tilde{B}_{test}\left[d\left(P_i, P_j\right)\right] = \begin{cases} \tilde{B}_{train}\left[d\left(P_i, P_j\right)\right] & \text{if } d\left(P_i, P_j\right) \le d_{train}^{max} \\ -\infty & \text{otherwise} \end{cases}$$

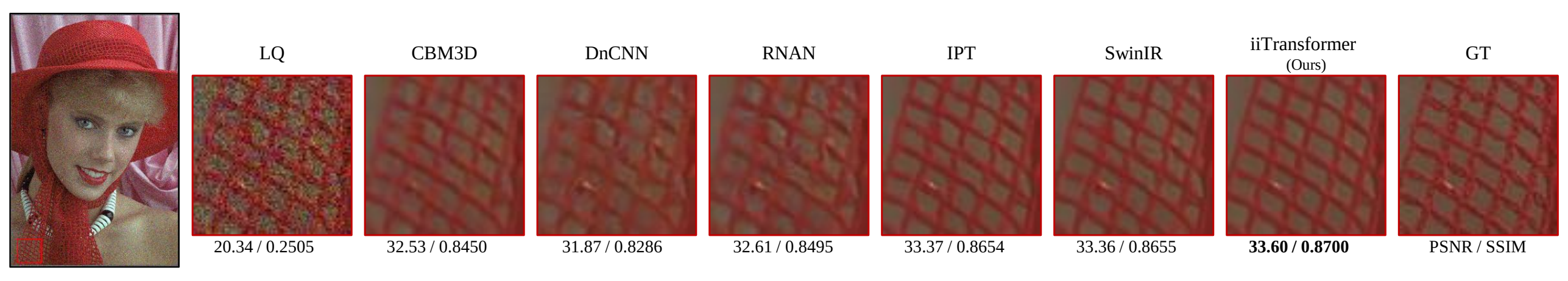
$d\left(P_i, P_j\right)$ is the distance between patches P_i and P_j,
d_{train}^{max} is the distance between furthest patches in a training image, and
$\tilde{B}[k]$ is an element in $\tilde{B}$ indexed at k



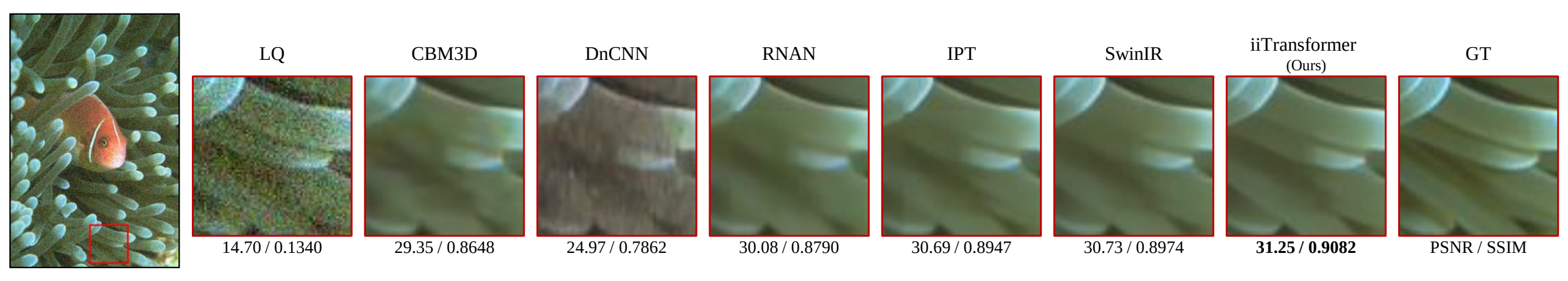
Masked eMSA

Results and Conclusion

AWGN Reduction

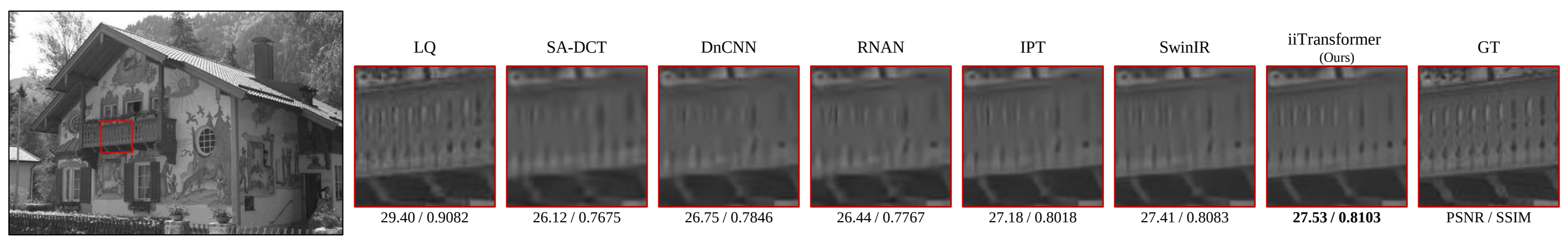


(a) AWGN Image Denoising for $\sigma = 25$ on 'kodim04' from Kodak24

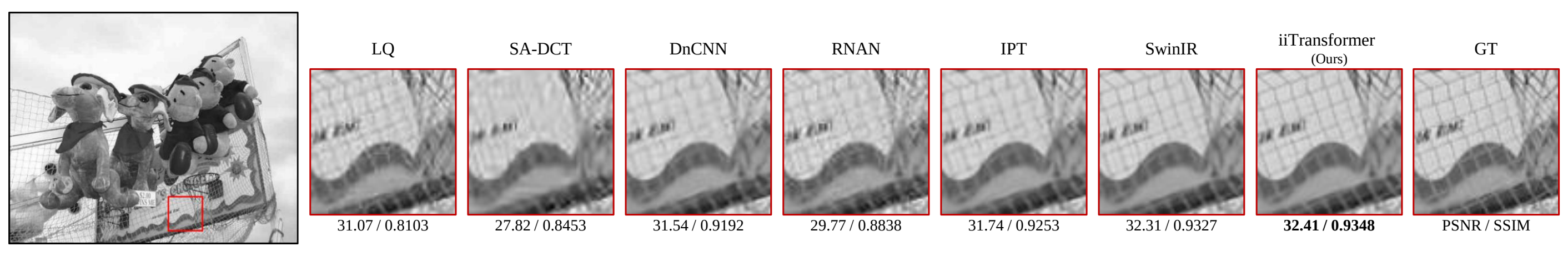


(b) AWGN Image Denoising for $\sigma = 50$ on '210088' from BSDS68

JPEG CAR

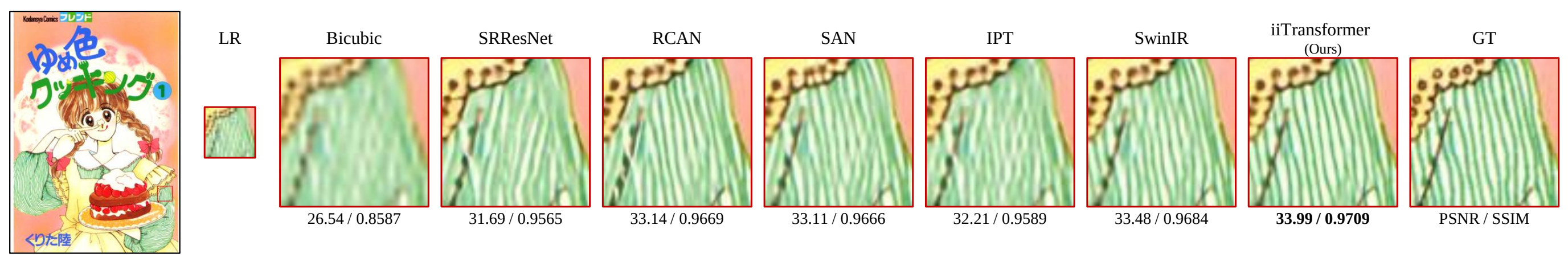


(a) JPEG CAR for q = 10 on 'paintedhouse' from LIVE1

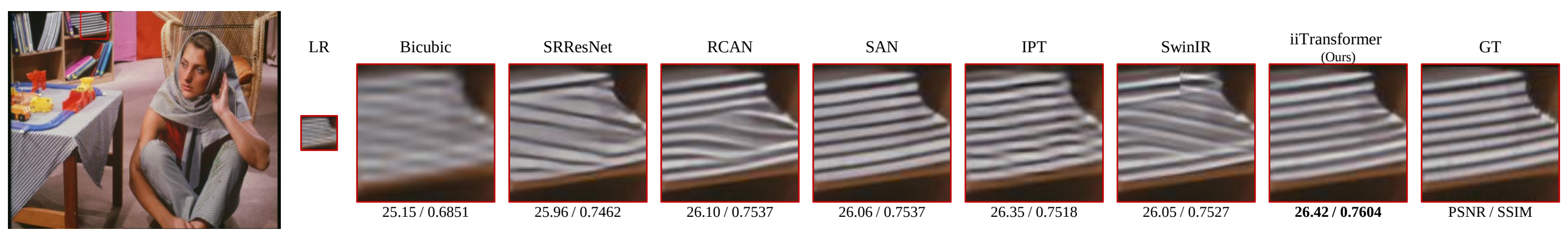


(b) JPEG CAR for q = 20 on 'carnivaldolls' from LIVE1

SISR

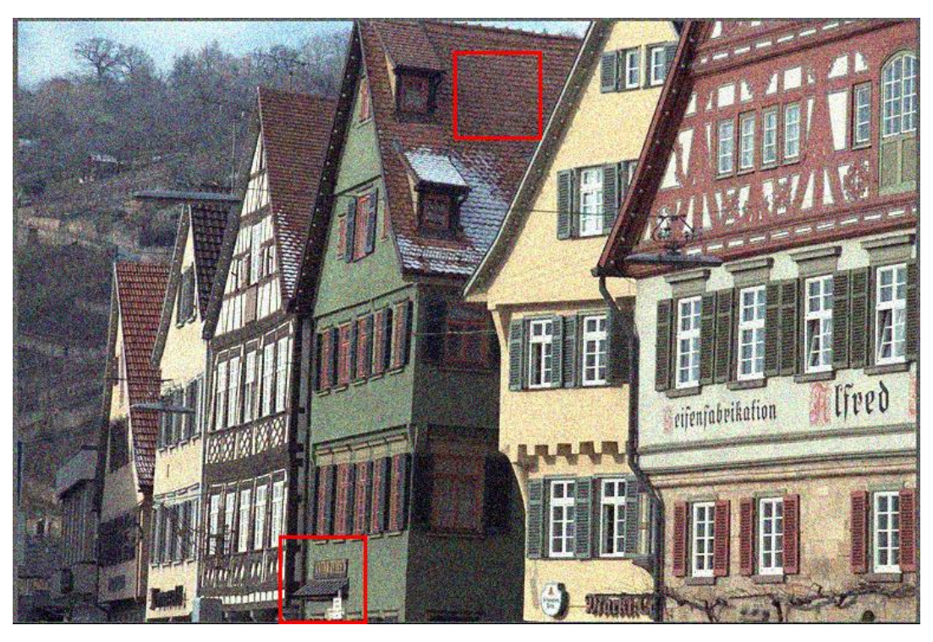


(a) SISR for s = 3 on 'YumeiroCooking' from Manga109

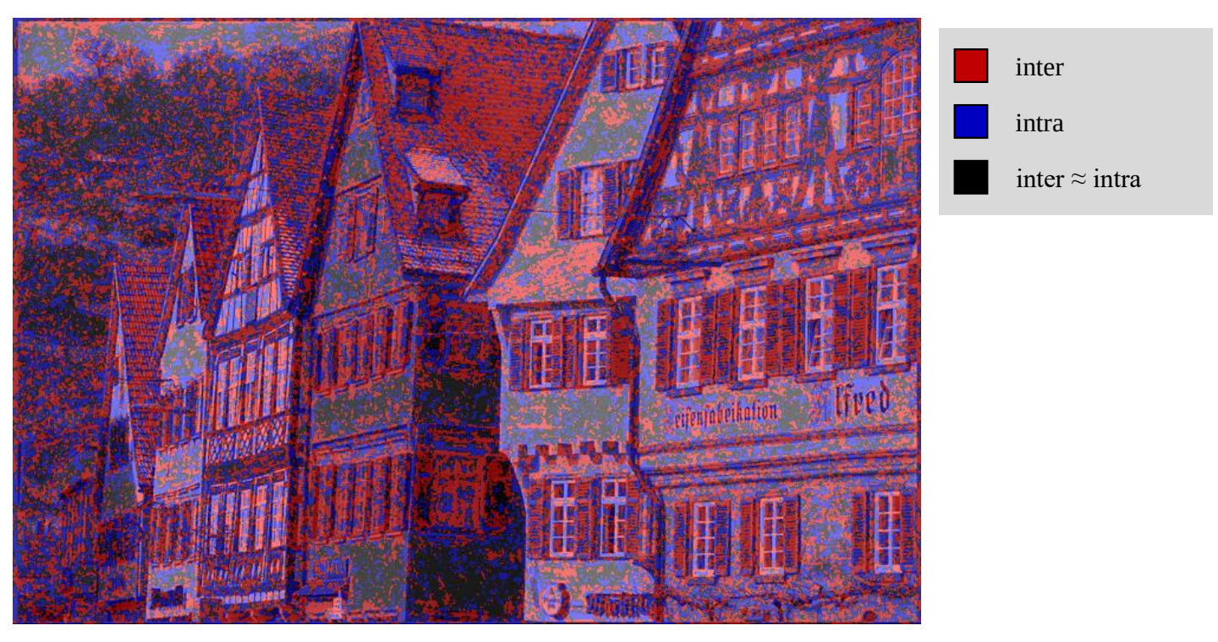


(b) SISR for s = 4 on 'barbara' from Set14

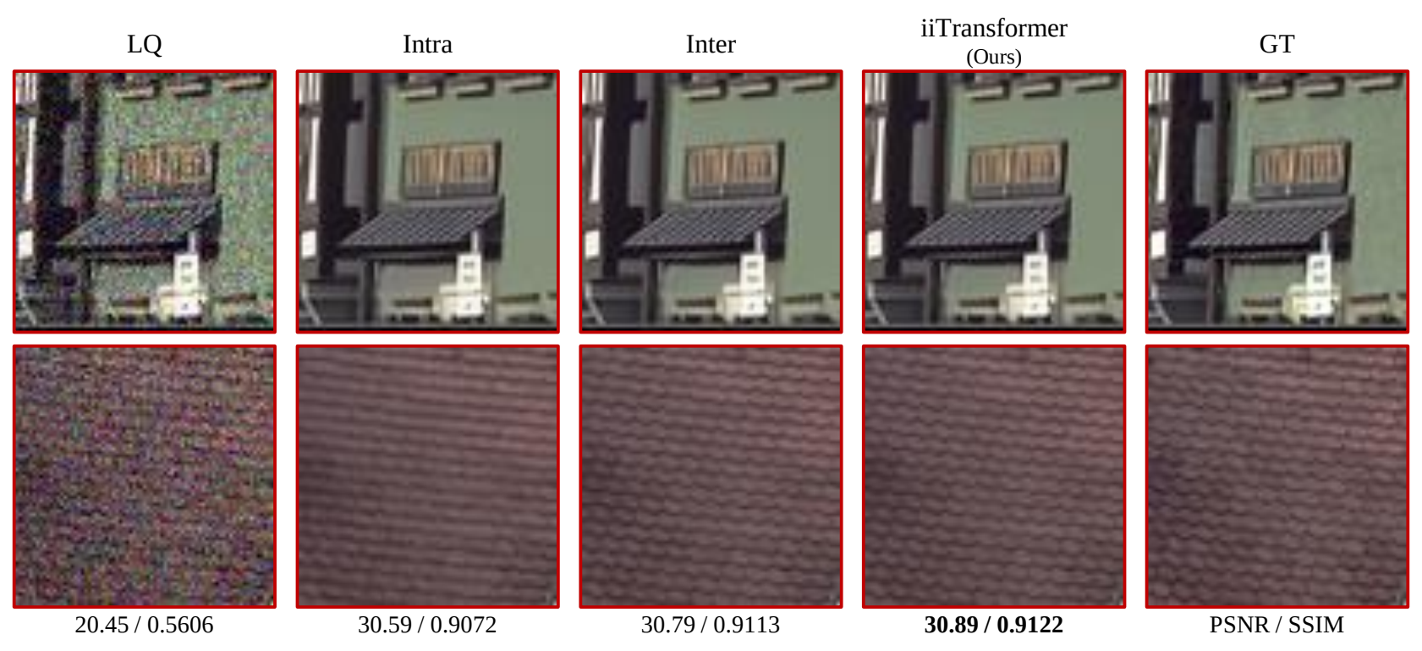
Ablation Study



LQ of 'kodim08' from Kodak24



Preferred Attention Mechanism



Regions Cropped from Each Output

Conclusion

- iiTransformer is a framework that combines local and non-local attention mechanisms to extract features at various sub-region levels of the image
- *Local* context is captured using the *intra* self-attention module and the *internal data repetition* is exploited using the *inter* self-attention module
- The patchwise relative position bias is *masked* to provide a boundary artifact-free solution for images of various resolutions
- State-of-the-art performance is achieved on various restoration tasks

Samsung Research