Structure-Resonant Discriminator *for* Image Super-Resolution

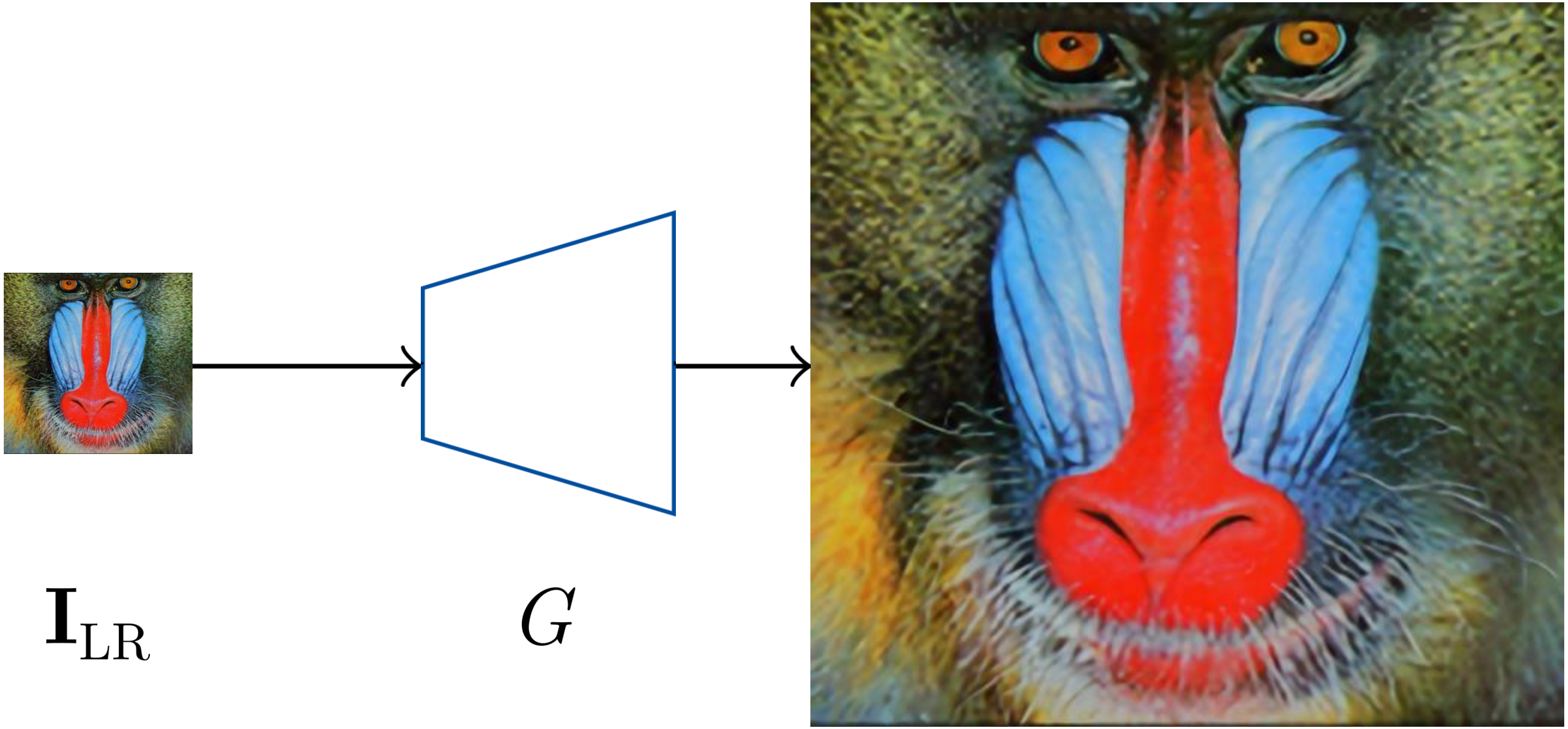
Jaerin Lee and **Kyoung Mu Lee**

Computer Vision Lab

Dept. of ECE, ASRI, Seoul National University

https://cv.snu.ac.kr

Single Image Super-Resolution

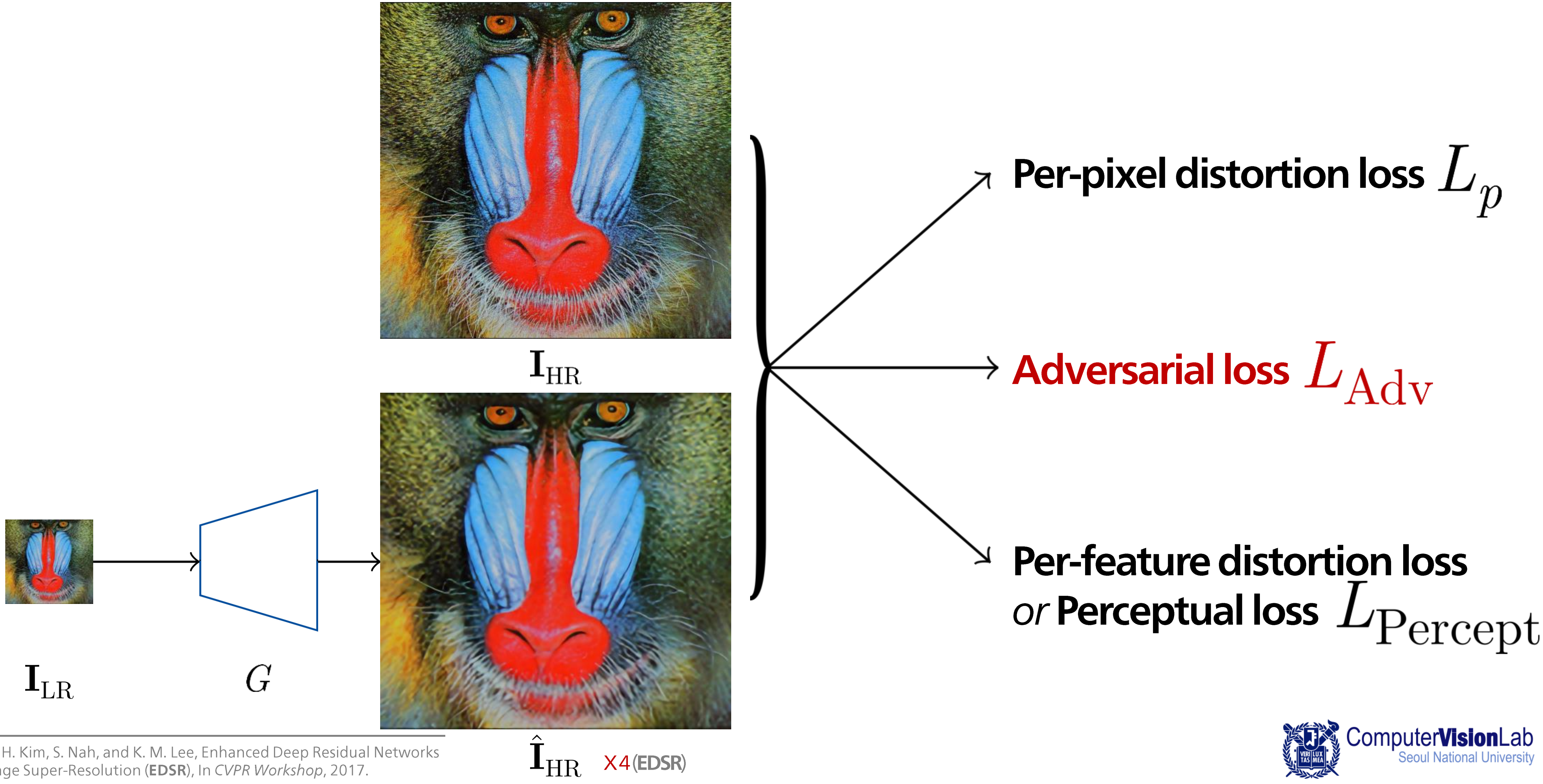


$\mathbf{I}_{\mathrm{LR}}$ $\quad G \quad$ $\hat{\mathbf{I}}_{\mathrm{HR}}$

X 4 (EDSR)

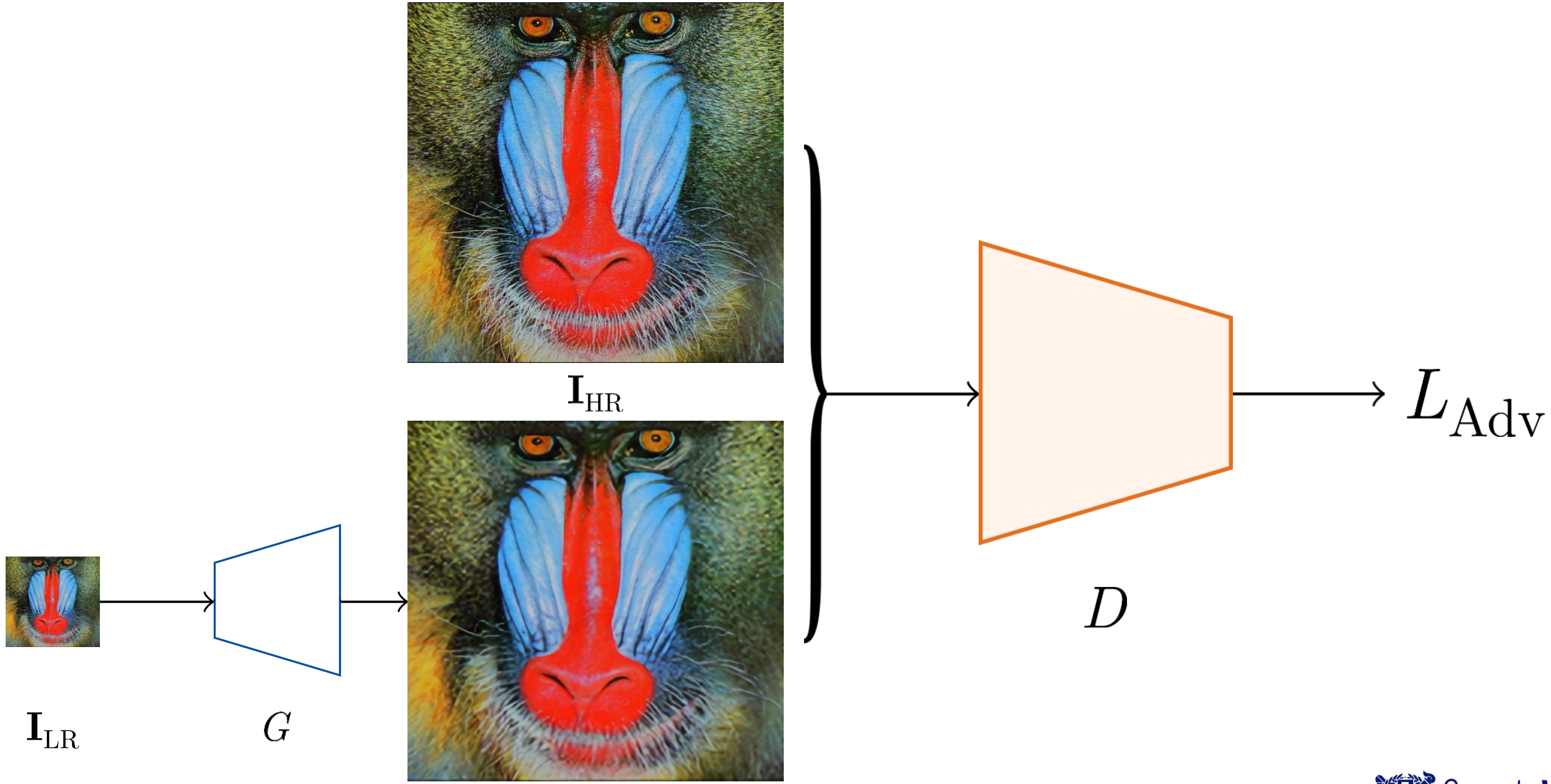
B. Lee, S. Son, H. Kim, S. Nah, and K. M. Lee, Enhanced Deep Residual Networks for Single Image Super-Resolution (**EDSR**), In *CVPR Workshop*, 2017.

Enhancing Perceptual Quality in SISR



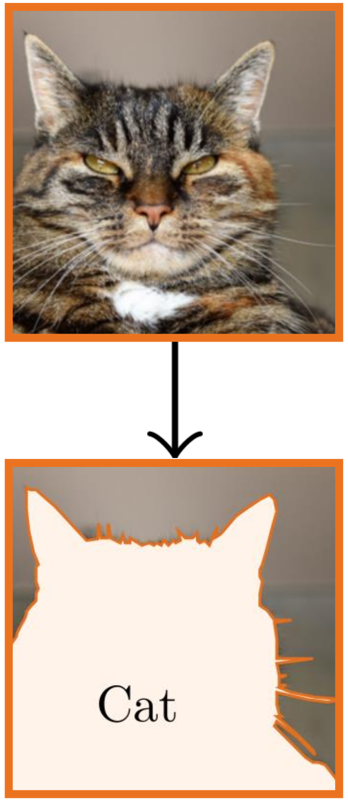
B. Lee, S. Son, H. Kim, S. Nah, and K. M. Lee, Enhanced Deep Residual Networks for Single Image Super-Resolution (**EDSR**), In *CVPR Workshop*, 2017.

Discriminator is a Data Model

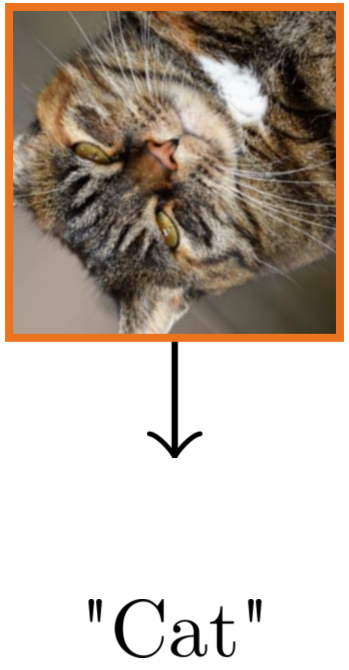


B. Lee, S. Son, H. Kim, S. Nah, and K. M. Lee, Enhanced Deep Residual Networks for Single Image Super-Resolution (**EDSR**), In *CVPR Workshop*, 2017.

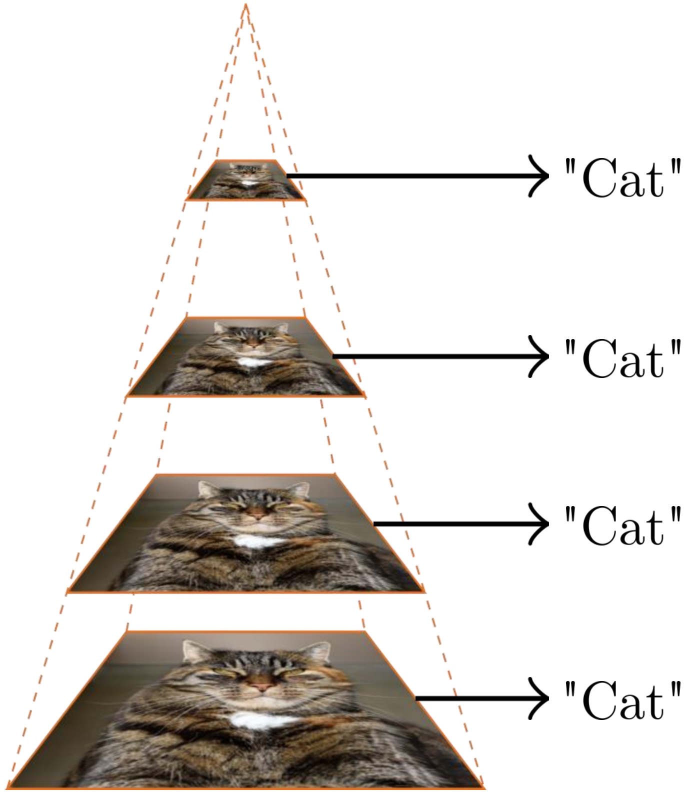
Structural Properties of Natural Images



1. Translation Equivariance



2. Rotation Invariance



3. Hierarchy of Scale

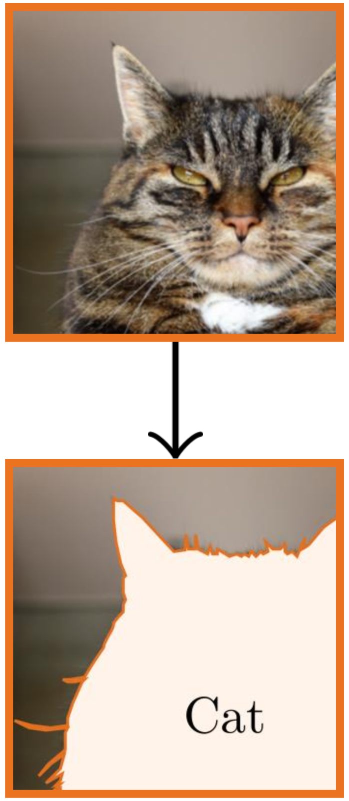
1. Translation Equivariance

- In a natural image, **information is concentrated locally.**
- A **translation equivariant** map:

$$\mathbf{F}' := f(\mathbf{F}), \quad f(\mathtt{translate}(\mathbf{F}, t)) = \mathtt{translate}(\mathbf{F}', t/s).$$

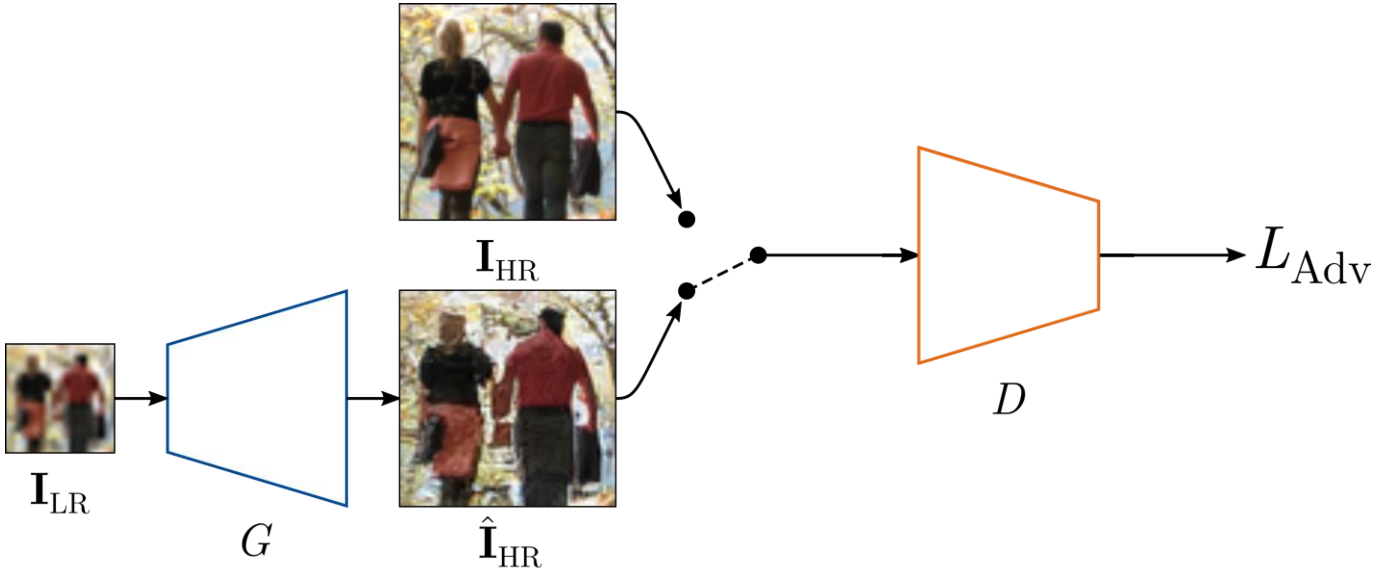
$$\mathbf{F} \in \mathbb{R}^{b \times c \times h \times w}, \quad \mathbf{F}' \in \mathbb{R}^{b \times c \times h/s \times w/s}.$$

  - Translated input produces translated output.
  - Ensures consistent local feature extraction.
  - Maintains positional information.



1. Translation Equivariance

- Layers that are **translation equivariant:**
  - Convolutions
  - Point-wise activations
  - Batch Normalizations
  - Residual blocks
- Layers that are **not:**
  - Pooling layers



2. Rotation Invariance

- In a natural image, objects appear in **arbitrary orientations.**
- Build a **rotation invariant** map as a composition of:
  1) A discrete rotation **equivariant** convolution

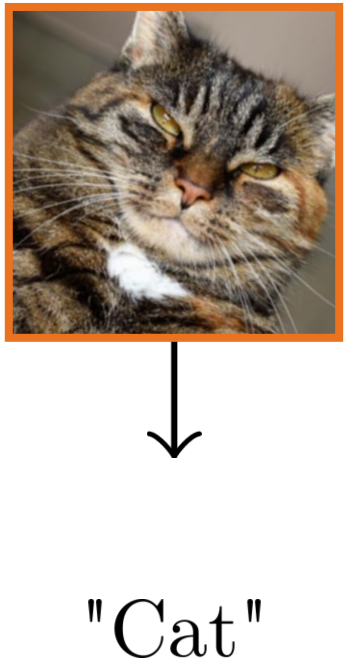
$$[\mathbf{F}']_\theta = [\mathbf{F} *_{\mathbb{Z}^2 \to p4} \mathbf{k}]_\theta := \mathbf{F} * R^\theta \mathbf{k}\,.$$

$$\mathbf{F} \in \mathbb{R}^{b \times c \times h \times w}\,, \quad \mathbf{F}' \in \mathbb{R}^{b \times c \times |H| \times h \times w}\,.$$

  2) A rotation group-wise **pooling** layer

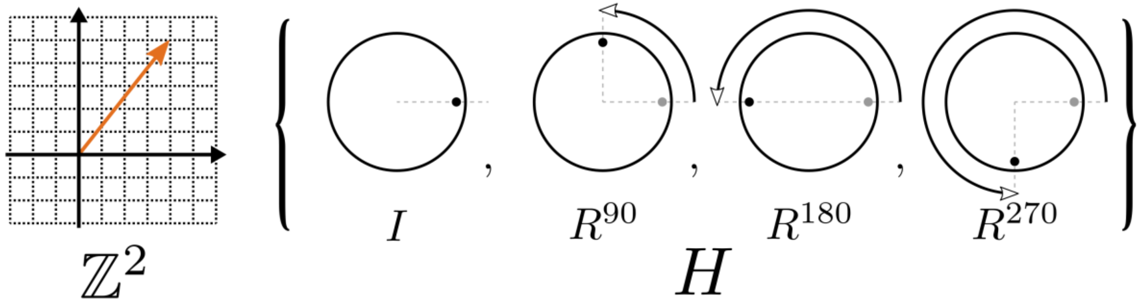
$$\texttt{gmaxpool}(\mathbf{F}')[b, c, u, v] := \max_{g \in \mathcal{H}} \mathbf{F}'[b, c, g, u, v]\,.$$

$$\texttt{gmaxpool}(\mathbf{F}') \in \mathbb{R}^{b \times c \times h \times w}\,.$$



2. Rotation Invariance

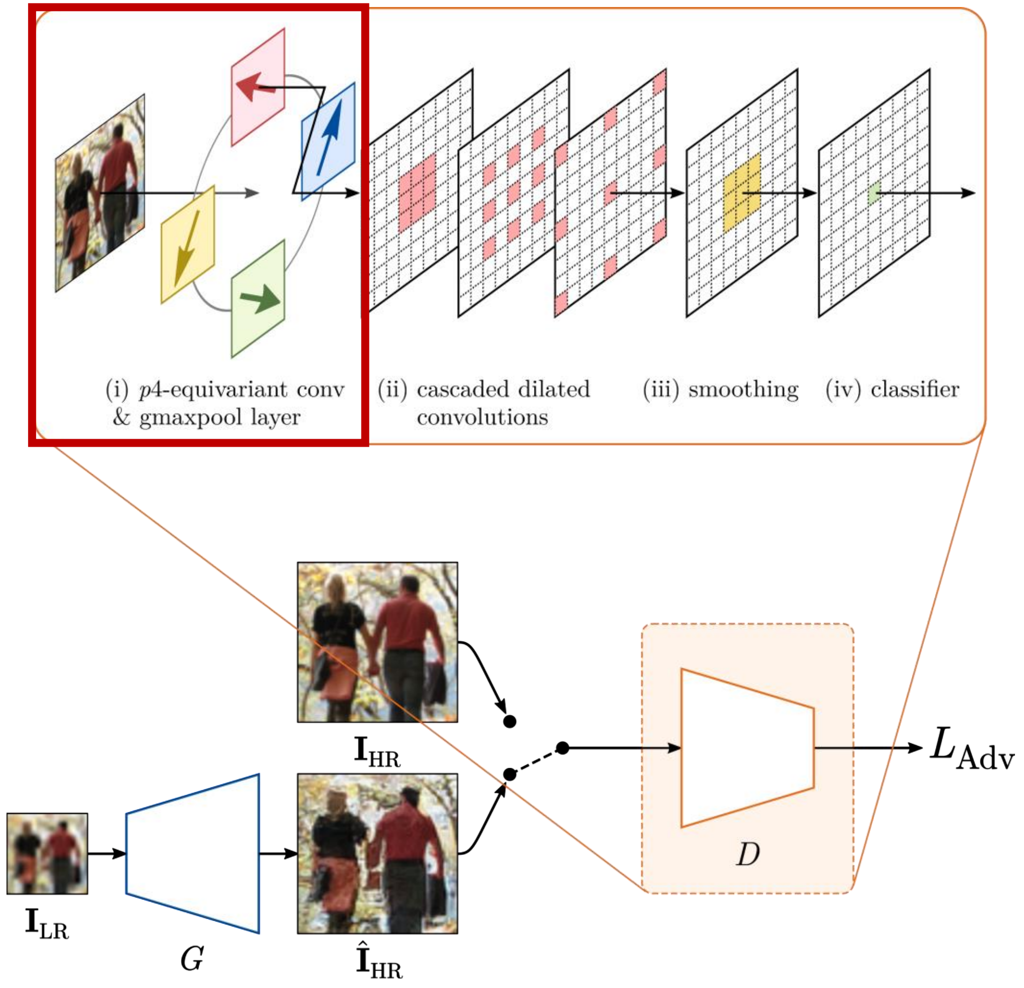
- Discrete roto-translation group ***p4*** is a natural choice for digital images.



$$H := \{I, R^{90}, R^{180}, R^{270}\}, \; p4 := \mathbb{Z}^2 \rtimes H$$

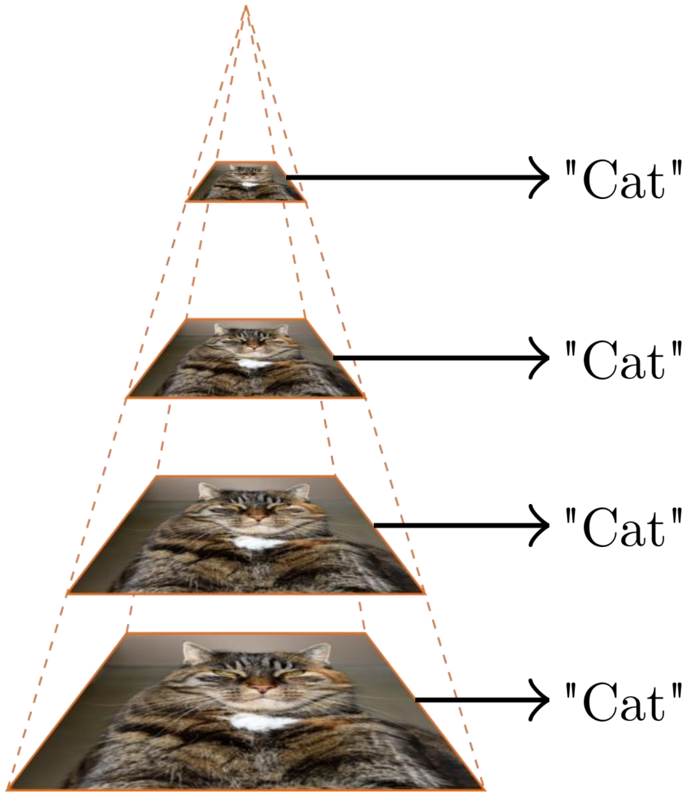
$$[\mathbf{F}']_\theta = [\mathbf{F} *_{\mathbb{Z}^2 \to p4} \mathbf{k}]_\theta := \mathbf{F} * R^\theta \mathbf{k}\,.$$

$$\mathbf{F} \in \mathbb{R}^{b \times c \times h \times w}, \quad \mathbf{F}' \in \mathbb{R}^{b \times c \times |H| \times h \times w}.$$



3. Hierarchy of Scale

- In a natural image, the same object can appear in **various scales.**
- The **scale** of an object
  ⟺ The **receptive field size** of a filter
- ***Two*** design features of the discriminator:
  1) **Multi-branch architecture**
     each branch attends to objects of different scale.
  2) **Cascaded dilated convolutions**
     for parameter-efficient scaling of receptive fields.



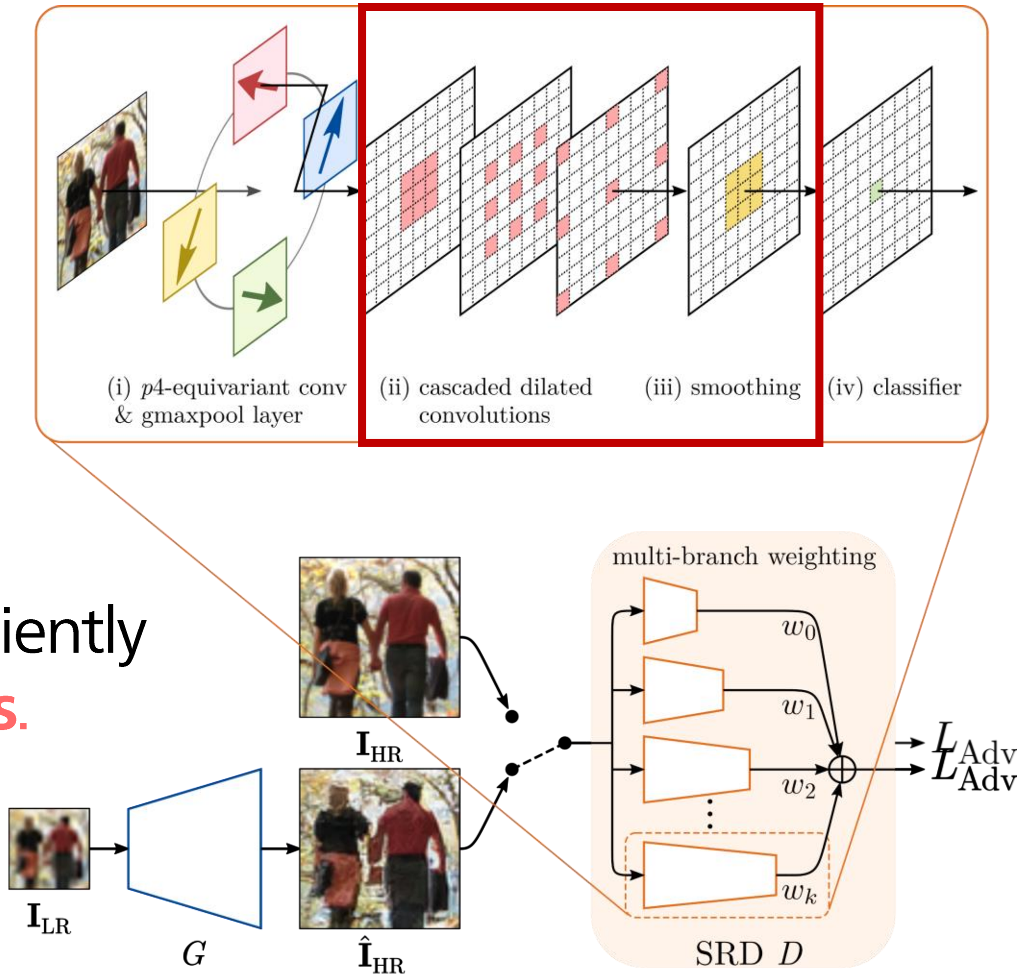
3. Hierarchy of Scale

1) Multi-branch architecture

- Multiple **receptive field sizes** for different **scales**.
- Outputs from different branches are summarized with a **weighted sum**.

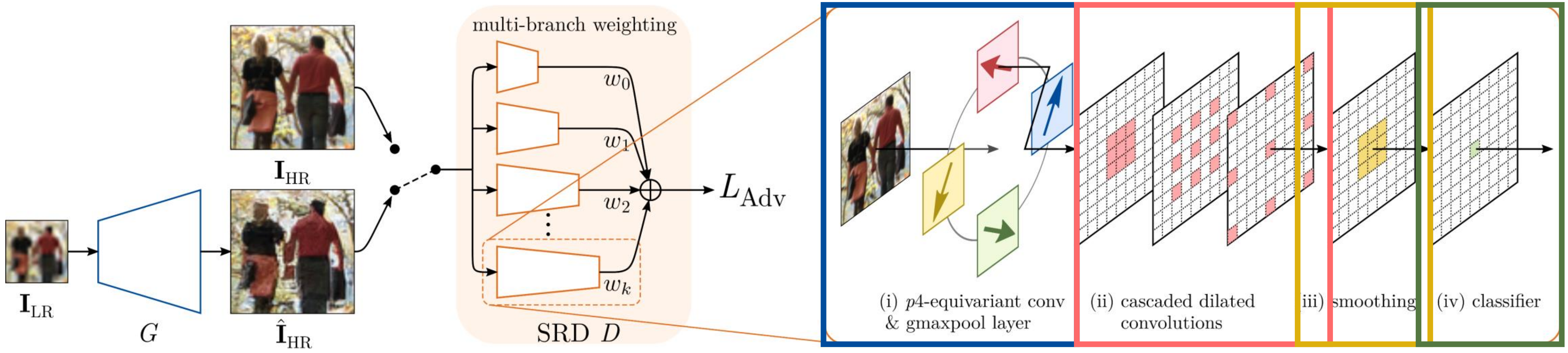
2) Cascaded dilated convolution

- Large receptive field can be achieved efficiently by **exponentially increasing the dilations.**
- The **smoothing convolution** mitigates potential artifact due to sparse filters.



Structure-Resonant Discriminator (SRD)

- Multi-branch structure with:
  1) **Rotation invariant layer:** ***p4***-equivariant conv + group-wise maxpool
  2) **Cascaded dilated convolutions** with exponentially increasing dilations
  3) A single 3X3 **smoothing convolution**
  4) A 1X1 convolution **classifier**



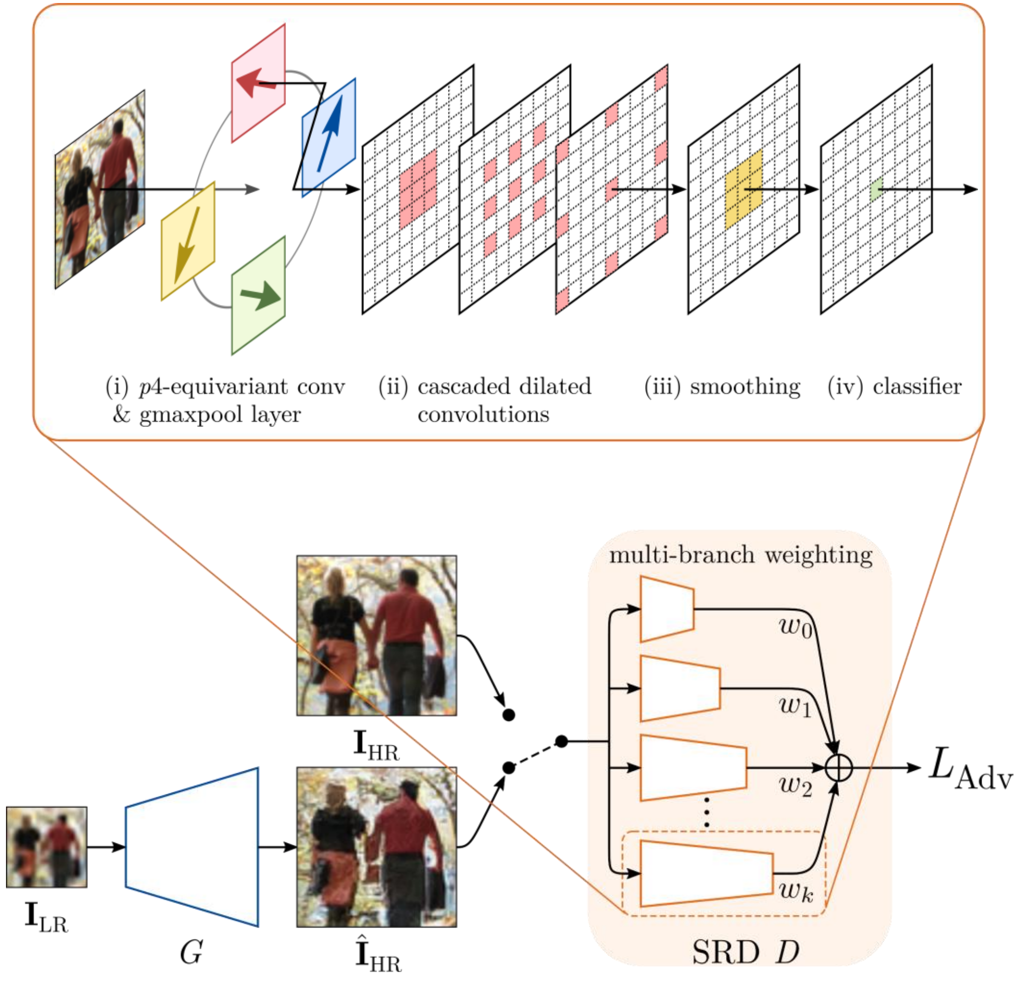
Implementation

1. Structure of a single branch of **SRD**

layer	gconv	conv_1	conv_2	dconv_1	$\cdots$	dconv_l	conv_3	cls
kernel size	3	3	3	3	$\cdots$	3	3	1
stride	1	2	2	1	$\cdots$	1	1	1
dilation	1	1	1	2	$\cdots$	2^l	1	1
receptive field size	3	+2	+4	+16	$\cdots$	$+2^{l+3}$	+8	+0

2. Architecture of **SRD**

branch ID i	depth l	receptive field size	branch weight w_i (rel.)
1	1*	25×25	1
2	2	65×65	1.25^2
3	3	129×129	1.5^2
4	4	257×257	1.75^2
5	5	513×513	2^2



ComputerVisionLab
Seoul National University

Training Procedure

- Loss function: **LSGAN** loss.

$$
\begin{aligned}
L_1(\mathbf{I}_{\mathrm{HR}}, \hat{\mathbf{I}}_{\mathrm{HR}}) &= \mathbb{E}[\|\mathbf{I}_{\mathrm{HR}} - \hat{\mathbf{I}}_{\mathrm{HR}}\|_1]\,, \\
L_{\mathrm{per}}(\mathbf{I}_{\mathrm{HR}}, \hat{\mathbf{I}}_{\mathrm{HR}}) &= \mathbb{E}[\|\mathrm{VGG}(\mathbf{I}_{\mathrm{HR}}) - \mathrm{VGG}(\hat{\mathbf{I}}_{\mathrm{HR}})\|_1]\,, \\
L^D_{\mathrm{adv}} &= 1/2\ \mathbb{E}\|D(\mathbf{I}_{\mathrm{HR}}) - \mathbf{1}\|_2^2 + 1/2\ \mathbb{E}\|D(\mathbf{I}_{\mathrm{SR}}) - \mathbf{0}\|_2^2\,, \\
L^G_{\mathrm{adv}} &= 1/2\ \mathbb{E}\|D(\mathbf{I}_{\mathrm{SR}}) - \mathbf{1}\|_2^2\,, \\
L_{\mathrm{tot}} &= \lambda_1 L_1 + \lambda_{\mathrm{per}} L_{\mathrm{per}} + \lambda_{\mathrm{adv}} L_{\mathrm{adv}}\,.
\end{aligned}
$$

- Dataset: Mixture of **DIV2K** & **Flickr2K**.
- SR Network: **RRDB** (**ESRGAN** baseline)
- Optimizer: **Adam** with default hyperparameter.

X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang and S. P. Smolley, "On the Effectiveness of Least Squares Generative Adversarial Networks" (**LSGAN**), In *TPAMI*, *41*(**12**): 2947-2960, 2019.
E. Agustsson and R. Timofte, "NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study" (**DIV2K Dataset**), In *CVPR Workshop*, 2017.
R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang, "NTIRE 2017 Challenge on Single Image Super-Resolution:, Methods and Results" (**Flickr2K Dataset**), In *CVPR Workshop*, 2017.
X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C. C. Loy, Y. Qiao and X. Tang, "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks", In *ECCV Workshop*, 2018.
D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization", In *ICLR*, 2015.

RESULTS

Quantitative Results

Method	Set5			Set14			BSD100			Urban100		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Bicubic	28.42	0.8104	0.340	26.00	0.7027	0.438	25.96	0.6675	0.524	23.14	0.6577	0.473
EDSR	32.46	0.8968	0.174	28.80	0.7876	0.284	27.71	0.7420	0.372	26.64	0.8033	0.232
RRDB	32.73	0.9011	0.173	28.99	0.7917	0.277	27.85	0.7455	0.366	27.03	0.8153	0.200
EDSR + GAN	29.39	0.8420	0.086	26.47	0.7183	0.146	25.73	0.6662	0.191	23.88	0.7169	0.155
SRGAN (Reproduced)	29.90	0.8486	0.081	26.56	0.7089	0.145	25.49	0.6524	0.177	24.38	0.7305	0.143
ESRGAN	30.44	0.8498	0.075	26.28	0.6981	0.134	25.28	0.6495	0.162	24.34	0.7327	0.123
SROBB	28.93	0.817	0.087	25.43	0.678	0.162	-	-	-	-	-	-
RankSRGAN-NIQE	29.77	0.8363	0.073	26.48	0.7023	0.138	25.49	0.6484	0.177	24.53	0.7279	0.142
RankSRGAN-Ma	28.85	0.8204	0.078	25.79	0.6852	0.145	25.03	0.6390	0.183	24.12	0.7182	0.143
RankSRGAN-PI	29.65	0.8342	0.073	26.46	0.7021	0.137	25.44	0.6484	0.175	24.47	0.7289	0.138
EDSR + RaGAN + **SRD (Ours)**	29.91	0.8473	0.080	26.80	0.7187	0.134	25.78	0.6626	0.184	24.38	0.7326	0.149
RRDB + RaGAN + **SRD (Ours)**	30.46	**0.8523**	0.074	26.73	0.7129	**0.126**	25.80	0.6663	0.164	24.72	0.7441	0.124
RRDB + LSGAN + **SRD (Ours)**	**30.55**	0.8506	**0.063**	**26.93**	**0.7227**	**0.126**	**25.87**	**0.6693**	**0.157**	**25.07**	**0.7542**	**0.117**

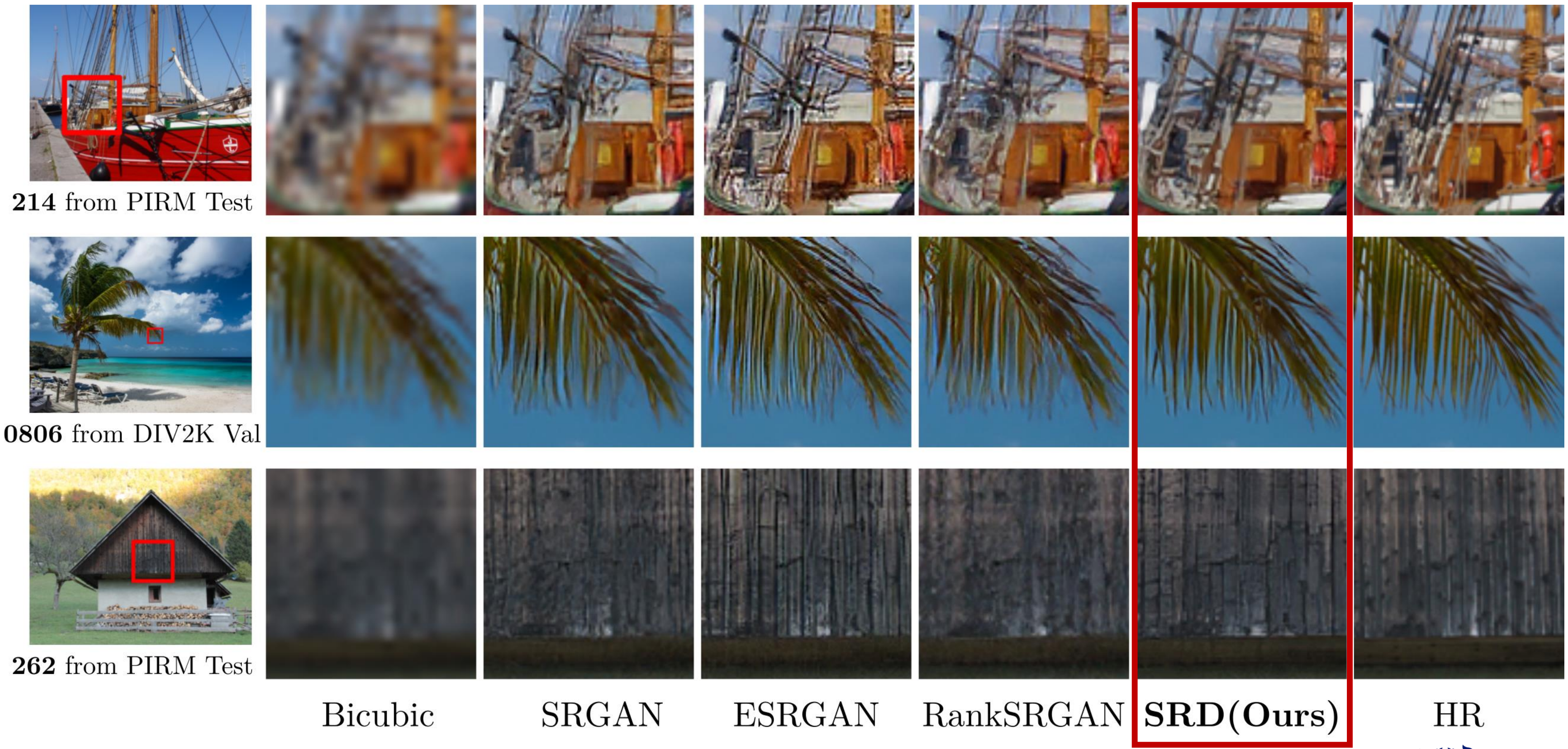
ComputerVisionLab
Seoul National University

Quantitative Results

Method	DIV2K-Val			PIRM-Val			PIRM-Test			OST300		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Bicubic	26.66	0.8521	0.409	26.50	0.6980	0.465	26.45	0.6892	0.481	25.74	0.6647	0.512
EDSR	29.25	0.9017	0.271	28.29	0.7716	0.309	28.23	0.7632	0.325	27.00	0.7289	0.373
RRDB	29.44	0.9043	0.256	28.71	0.7849	0.292	28.61	0.7756	0.310	27.30	0.7411	0.362
EDSR + GAN	26.62	0.8513	0.133	25.88	0.6848	0.141	25.78	0.6712	0.149	25.25	0.6589	0.190
SRGAN (Reproduced)	26.60	0.8481	0.126	25.61	0.6757	0.144	25.47	0.6599	0.153	24.90	0.6461	0.180
ESRGAN	26.61	0.8479	0.115	25.18	0.6599	0.144	25.04	0.6452	0.152	24.63	0.6422	0.169
RankSRGAN-NIQE	26.53	0.8421	0.128	25.76	0.6739	0.139	25.62	0.6584	0.145	24.97	0.6415	0.184
RankSRGAN-Ma	25.60	0.8261	0.145	25.22	0.6610	0.143	25.11	0.6468	0.150	24.55	0.6261	0.192
RankSRGAN-PI	26.48	0.8431	0.122	25.64	0.6724	0.136	25.48	0.6564	0.143	24.91	0.6410	0.180
EDSR + RaGAN + **SRD (Ours)**	26.80	0.8518	0.127	25.92	0.6830	0.141	25.82	0.6688	0.148	25.19	0.6538	0.185
RRDB + RaGAN + **SRD (Ours)**	26.82	0.8471	0.118	25.88	0.6837	0.130	25.79	0.6710	0.137	25.29	0.6611	0.185
RRDB + LSGAN + **SRD (Ours)**	**27.05**	**0.8529**	**0.107**	**26.09**	**0.6935**	**0.123**	**26.00**	**0.6815**	**0.129**	**25.31**	**0.6624**	**0.166**

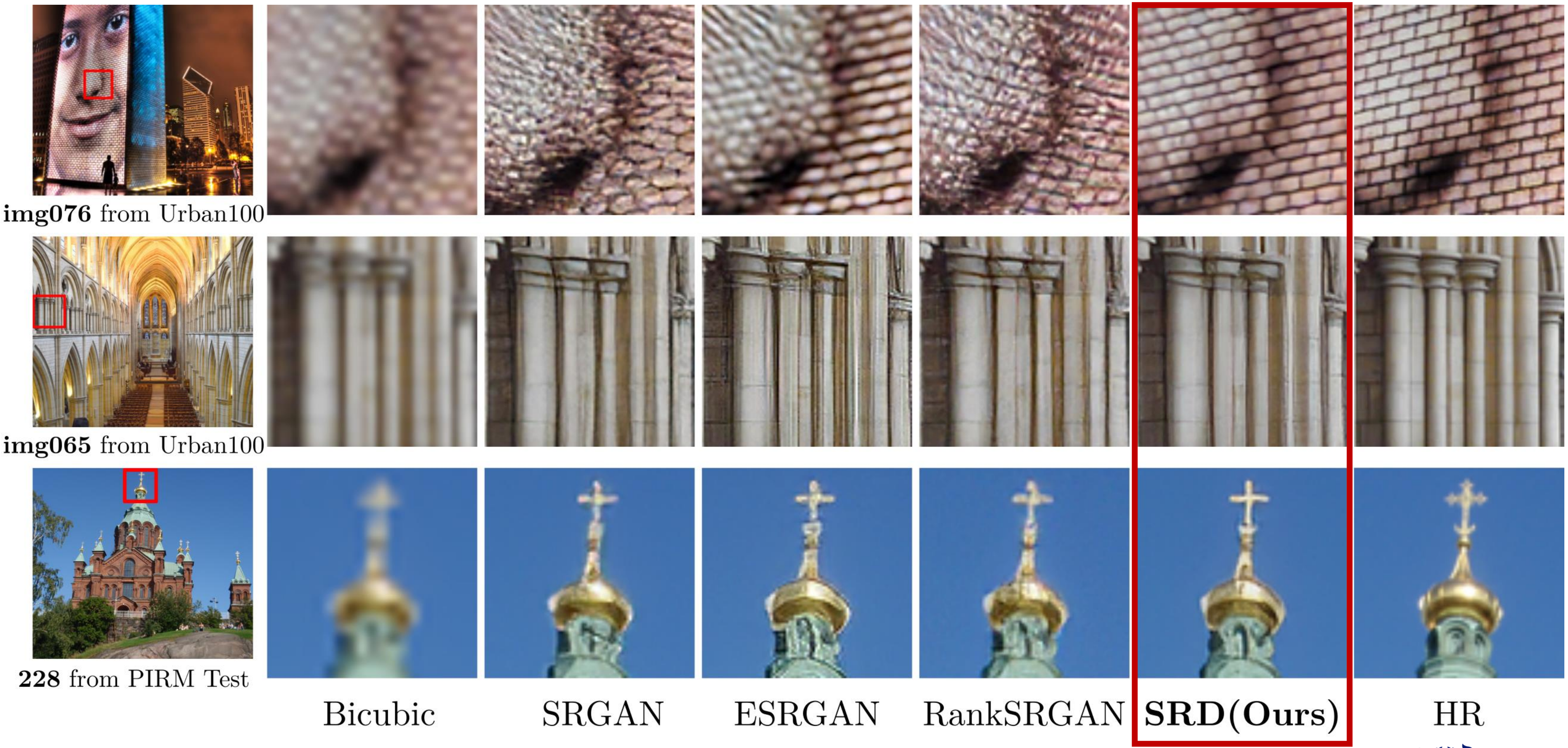
ComputerVisionLab
Seoul National University

Qualitative Results

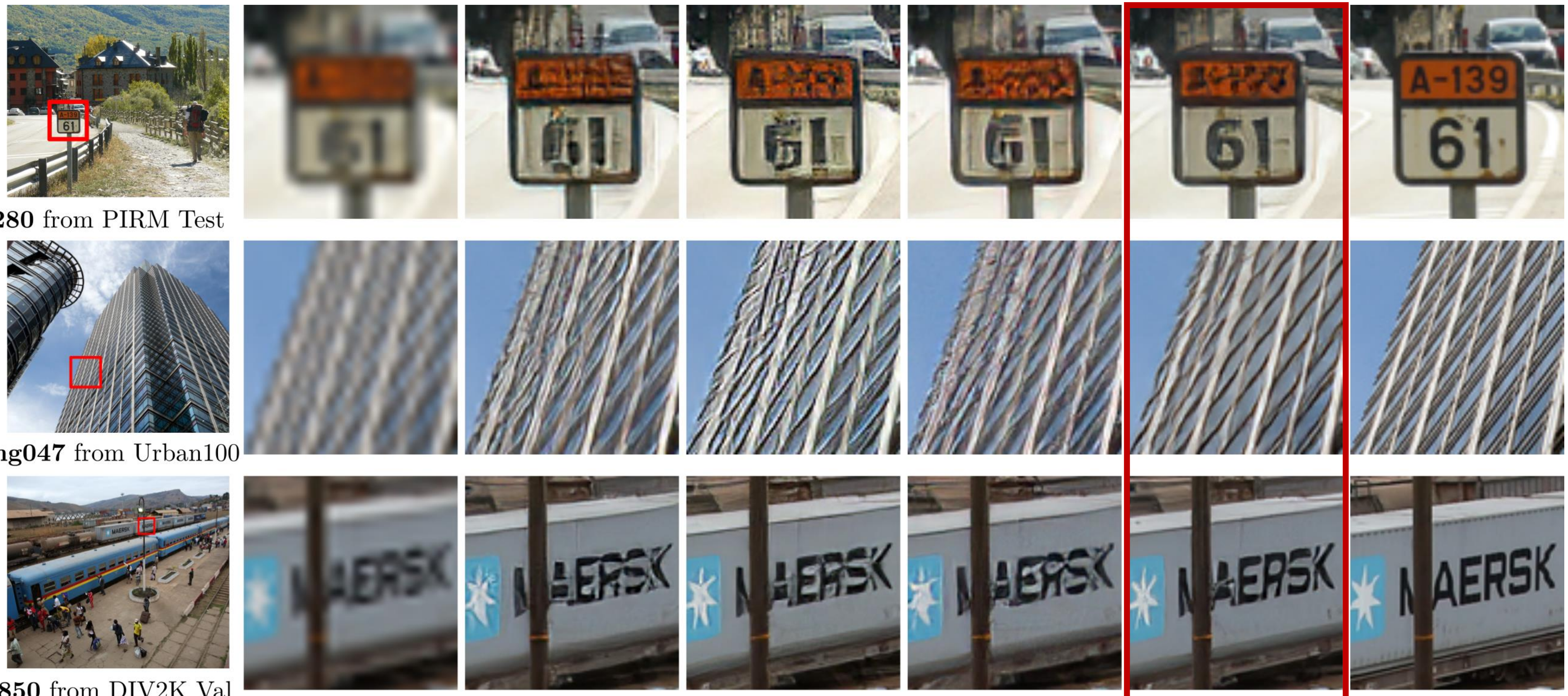


ComputerVisionLab
Seoul National University

Qualitative Results



Qualitative Results



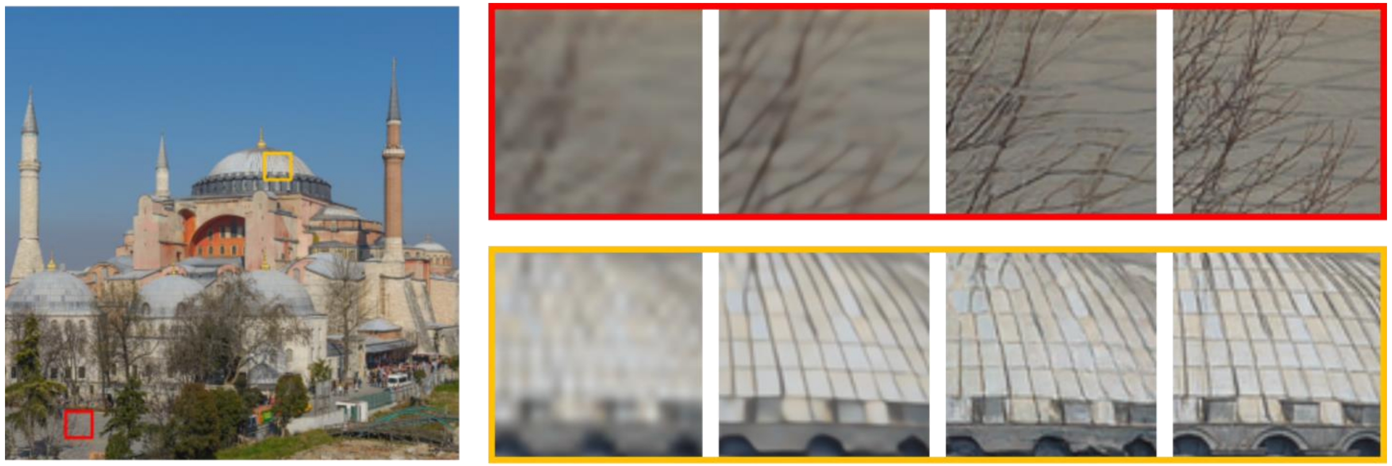
280 from PIRM Test

img047 from Urban100

0850 from DIV2K Val

Bicubic | SRGAN | ESRGAN | RankSRGAN | **SRD(Ours)** | HR

Qualitative Results



0890 from DIV2K Val | Bicubic | RRDB | **SRD (Ours)** | HR

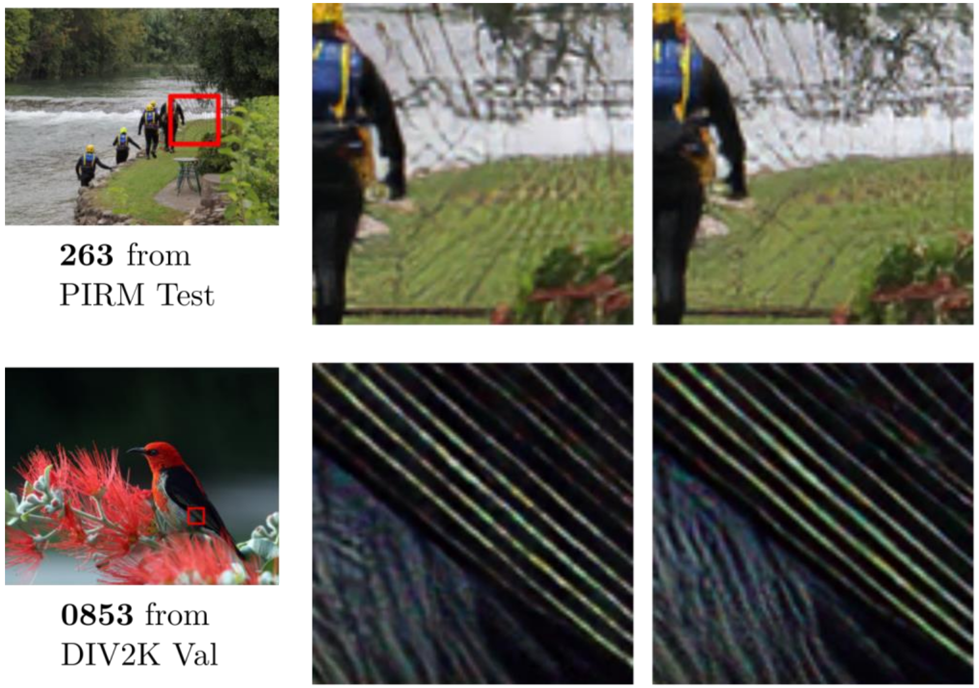
256 from PIRM Test | Bicubic | RRDB | **SRD (Ours)** | HR

Ablation Studies

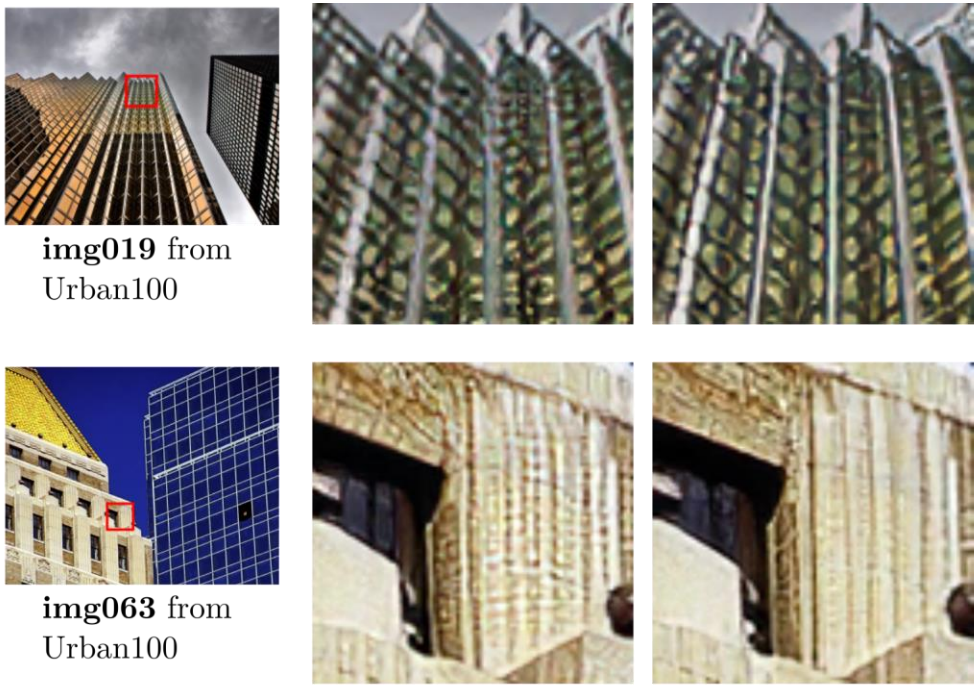
1. **Cascaded dilated convolution** vs. **explicit input scaling** for multi-scale architecture.

2. ***p4*-invariant layer** vs. **standard convolution** for first layer convolution of the discriminator.

Ablation Studies



(a) Different multi-scale architecture.



(b) Effect of the first layer conv of **D**.

Conclusion

- ***Three*** structural features of natural images:
    1. **Translation equivariance**
    2. **Rotation invariance**
    3. **Hierarchy of scale**
- **Discriminator is also a data model**.
- Design them to be compatible to the **structure of natural images**.

THANK YOU

Computer Vision Lab

https://cv.snu.ac.kr