Interpreting Super-Resolution Networks

Chao Dong

XPixel-Group

Cooperators: Jinjin Gu, Xintao Wang
Yihao Liu, Liangbin Xie, Anran Liu, etc.

Interpreting Super-Resolution Networks

Interpretability in Low-level Vision

- Pixel: What pixels contribute most to restoration?
- Feature: Where can we find semantics in SR-net?
- Filters: Whether learned filters are discriminative?

Interpreting Super-Resolution Networks

Interpretability
in Low-level Vision

Pixel: What pixels contribute most to restoration?

Feature: Where can we find semantics in SR-net?

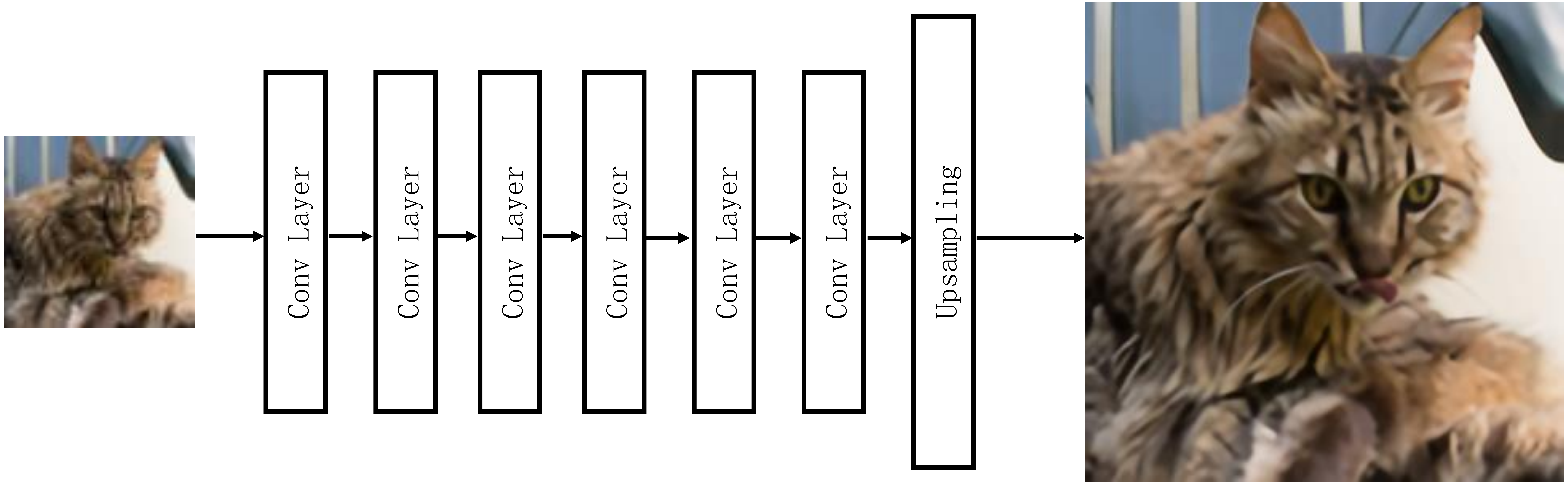
Filters: Whether learned filters are discriminative?

Interpreting Super-Resolution Networks with **L**ocal **A**ttribution **M**aps

Jinjin Gu, Chao Dong

The University of Sydney,
Shenzhen Institutes of Advanced Technology, CAS

Super-Resolution Networks



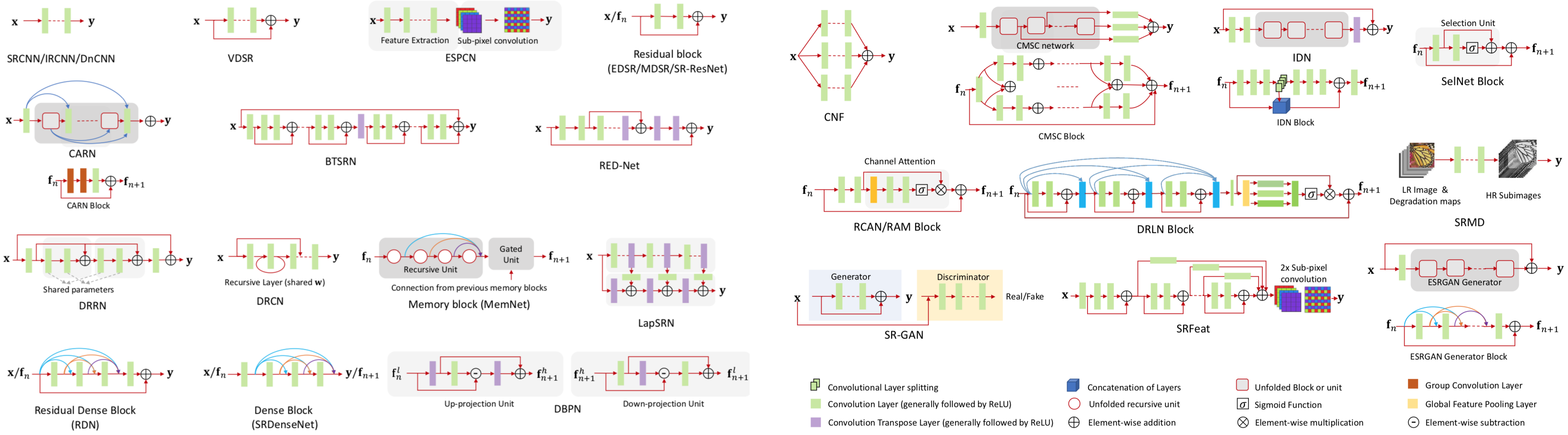
SR networks build up of convolutional layers and upsampling blocks, with parameter θ.

Similar structures can be found in denoising, deblurring, deraining, etc.

Super-Resolution Networks

Many SR network architectures have been proposed.

What makes their different performance?



[Anwar, S., Khan, S., & Barnes, N. (2019). A Deep Journey into Super-resolution: A survey. arXiv preprint arXiv:1904.07523.]

SR networks are still mysterious

Have you met these scenarios?

- Do you need multi-scale architecture or a larger receptive field?
- Does non-local attention module work as you want?
- Why different SR networks perform differently?

We lack scientific understanding
and also research tools

Information usage in SR networks

In the past, we only have one metric to study SR networks:

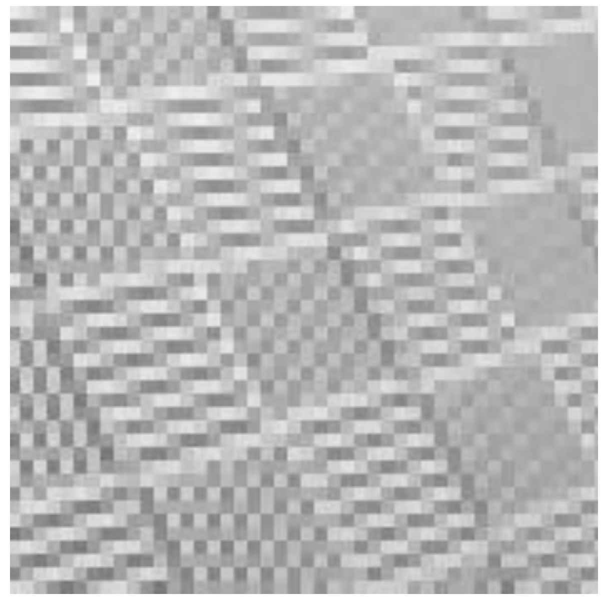
The Performance

Add module A,
seems good

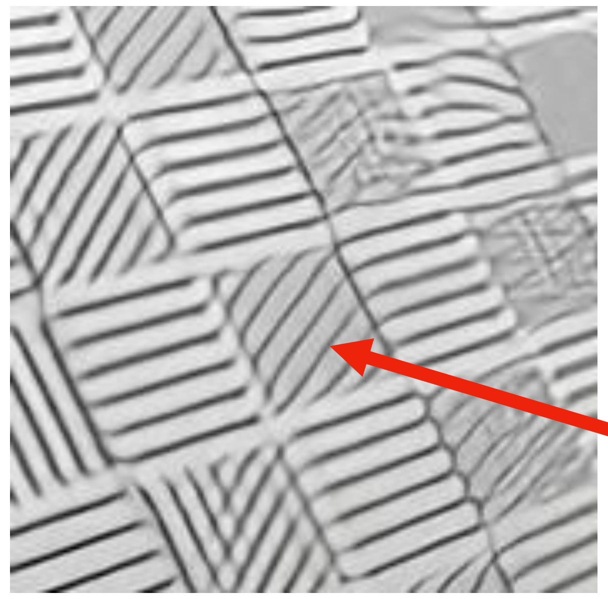
Add module B,
seems good

Combine A and B,
not good

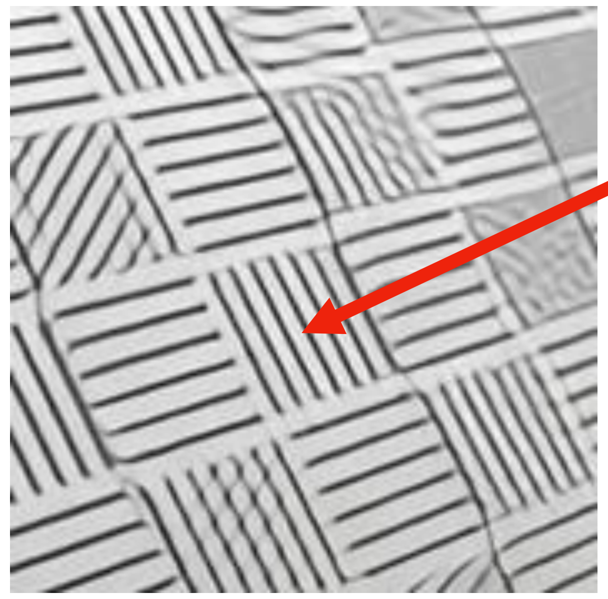
Attribution Analysis



Input image



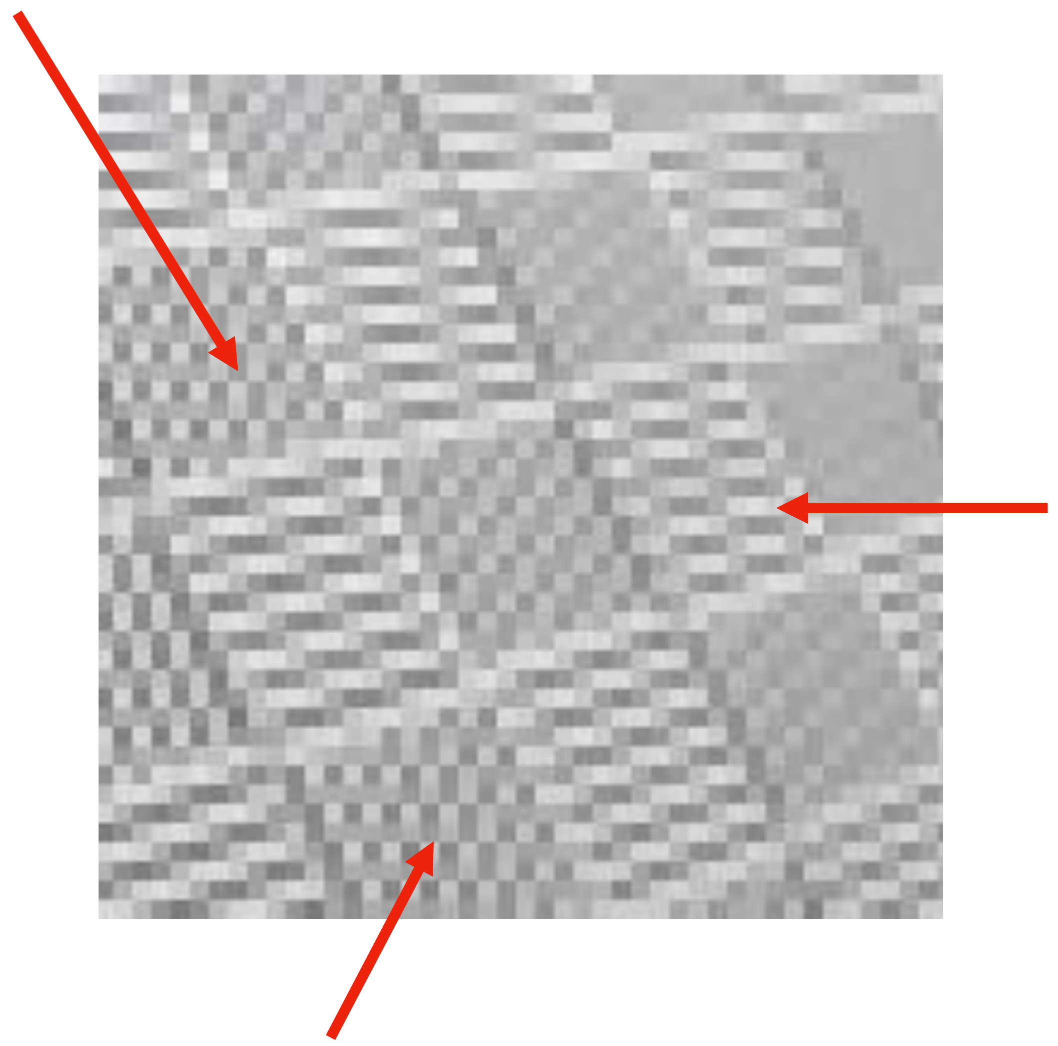
EDSR



RNAN

Why RNAN gives correct results in the center?

Attribution Analysis



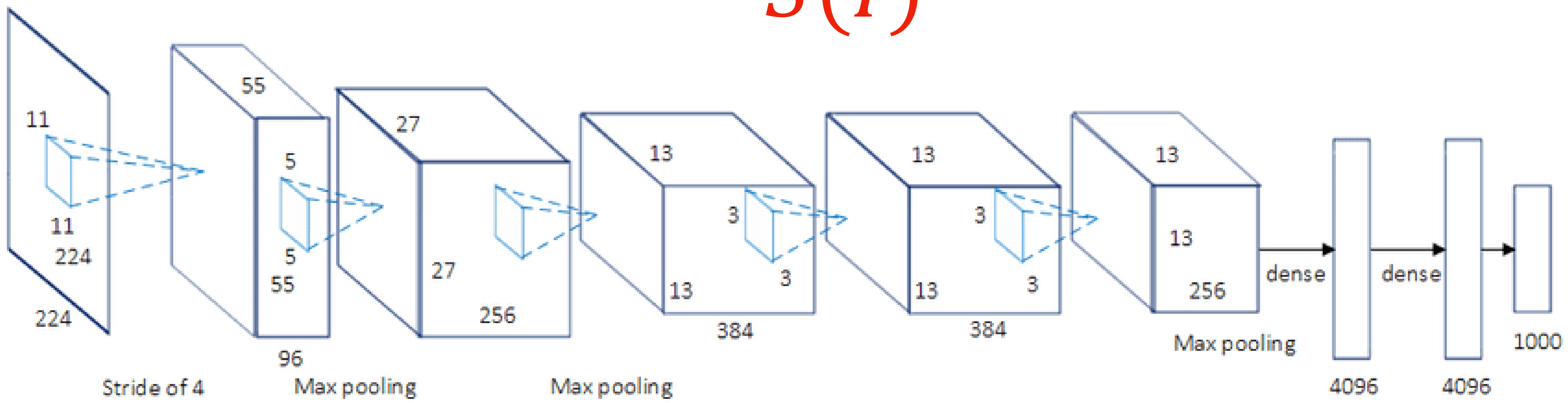
What does RNAN notice from the input?

Does EDSR notice this information?

Attribution Analysis for High-level Networks

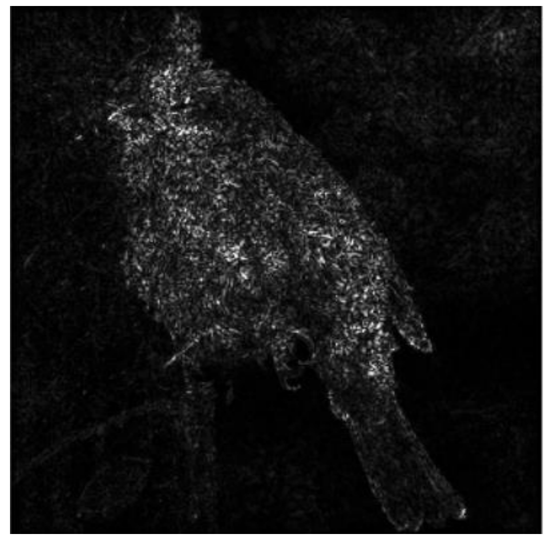
$S(I)$

I



house finch

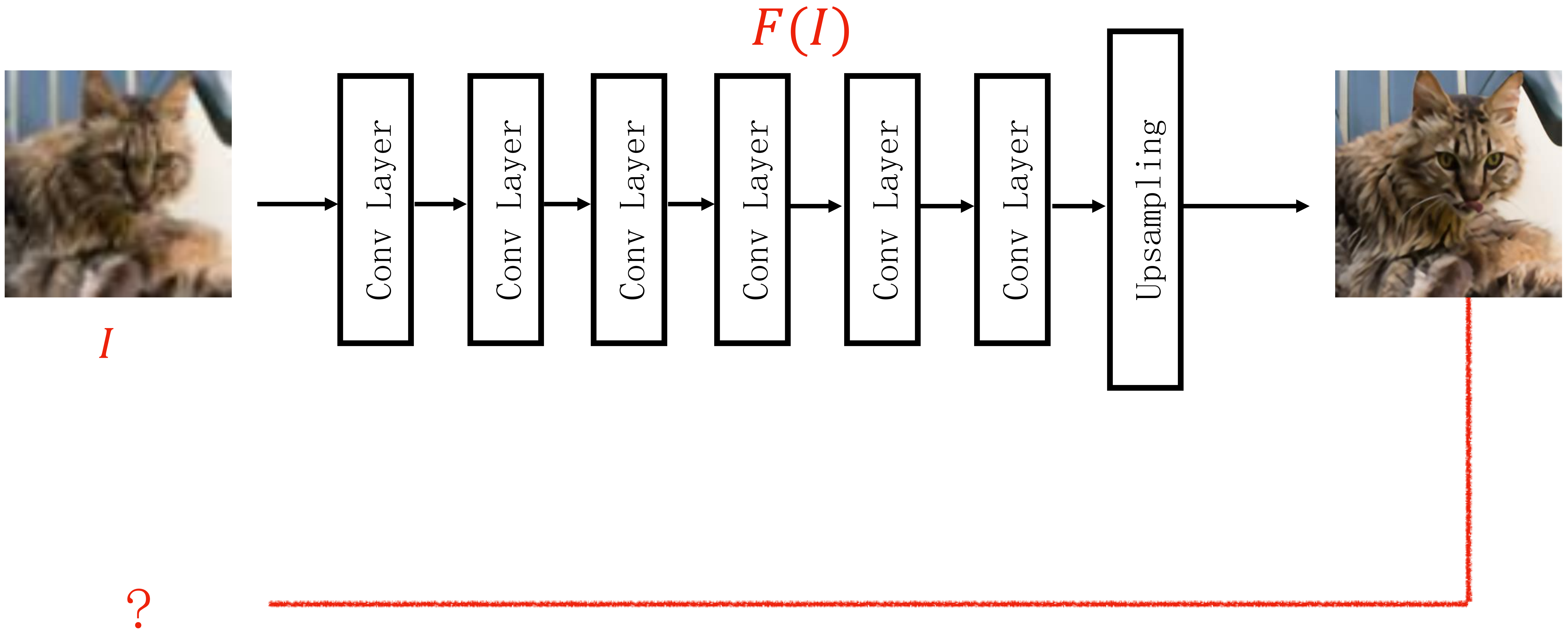
Backprop methods: gradient



The visualized attribution map

$$\mathrm{Grad}_S(I) = \frac{\partial S(I)}{\partial I}$$

Attribution Analysis for Low-level Networks



How to calculate gradients for low-level networks?

Auxiliary Principles

We introduce auxiliary principles for interpreting low-level networks:

- Interpreting local not global

SR networks can not be interpreted globally

Auxiliary Principles

We introduce auxiliary principles for interpreting low-level networks:

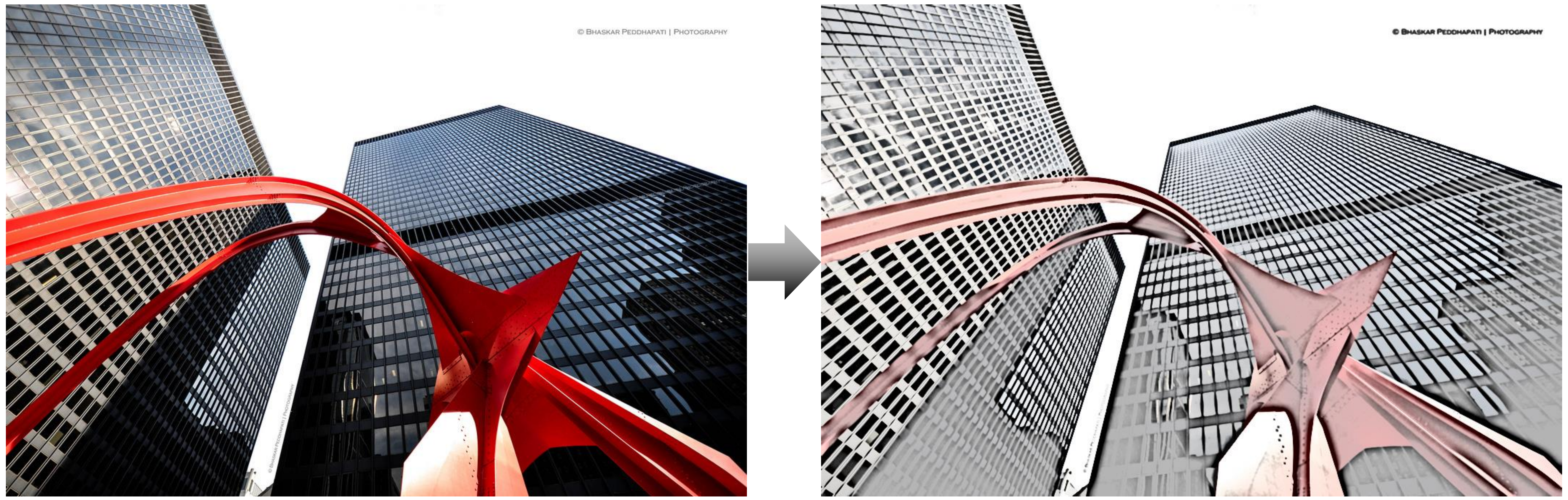
- Interpreting local not global
- Interpreting hard not simple

Interpreting simple cases can provide limited help

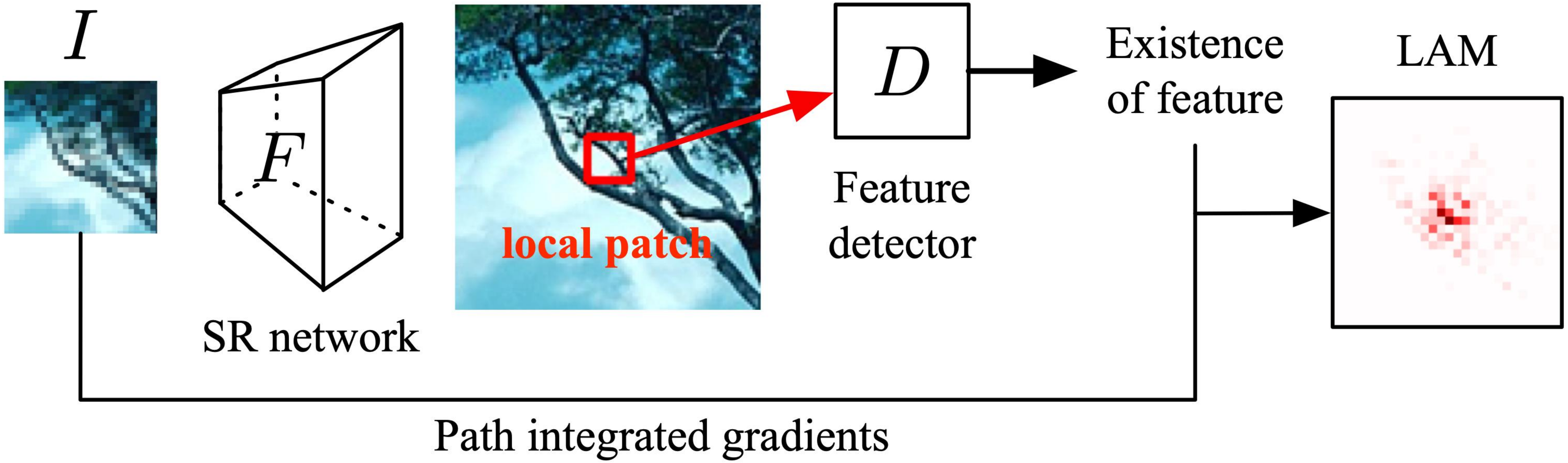
Auxiliary Principles

We introduce auxiliary principles for interpreting low-level networks:

- Interpreting local not global
- Interpreting hard not simple
- Interpreting features not pixels



Local Attribution Maps (LAM)



Local Attribution Maps (LAM)

We employ Path Integral Gradient

$$\mathrm{LAM}_{F,D}(\gamma)_i := \int_0^1 \frac{\partial D(F(\gamma(\alpha)))}{\partial \gamma(\alpha)_i} \times \frac{\partial \gamma(\alpha)_i}{\partial \alpha} d\alpha.$$

SR Network F

Feature Detector D

Path function $\gamma(\alpha), \alpha \in \mathbb{R}$

Baseline Input $\gamma(0) = I'$

Input $\gamma(1) = I$

Local Attribution Maps (LAM)

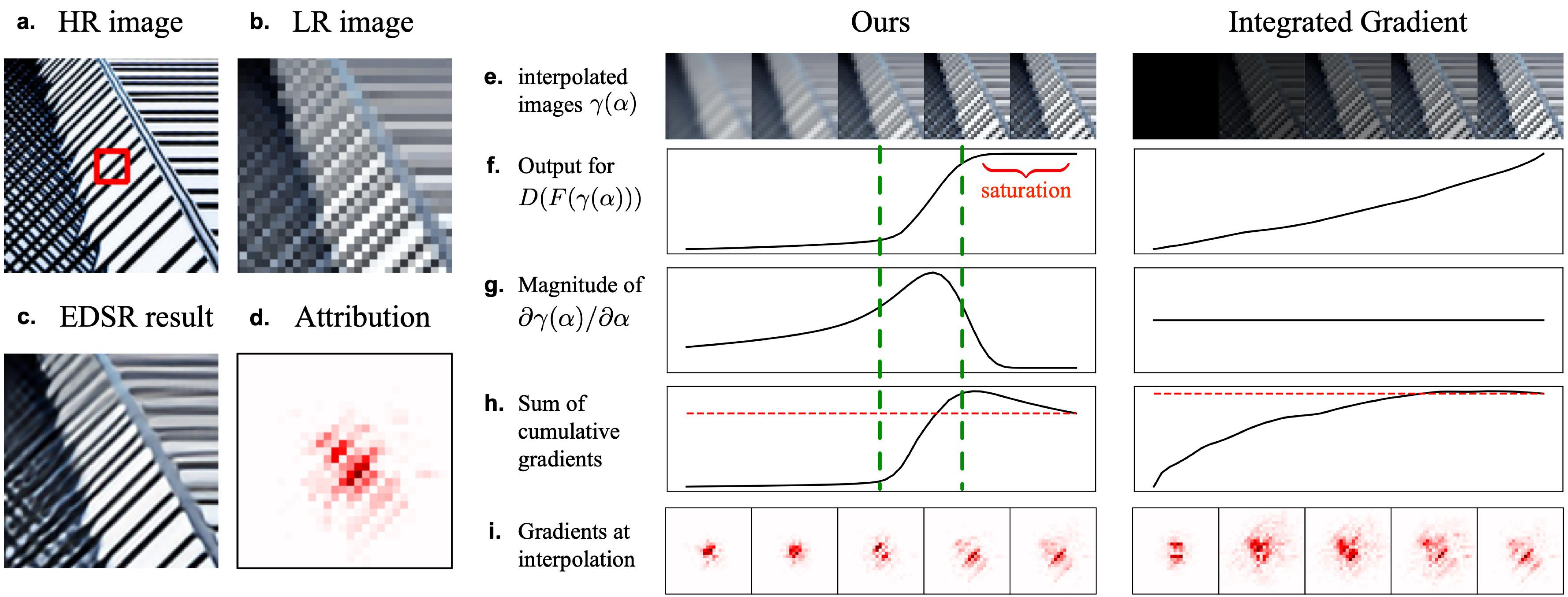
We design the Baseline Input and Path function especially for SR networks.

Blurred image as baseline input: $I' = \omega(\sigma) \otimes I$

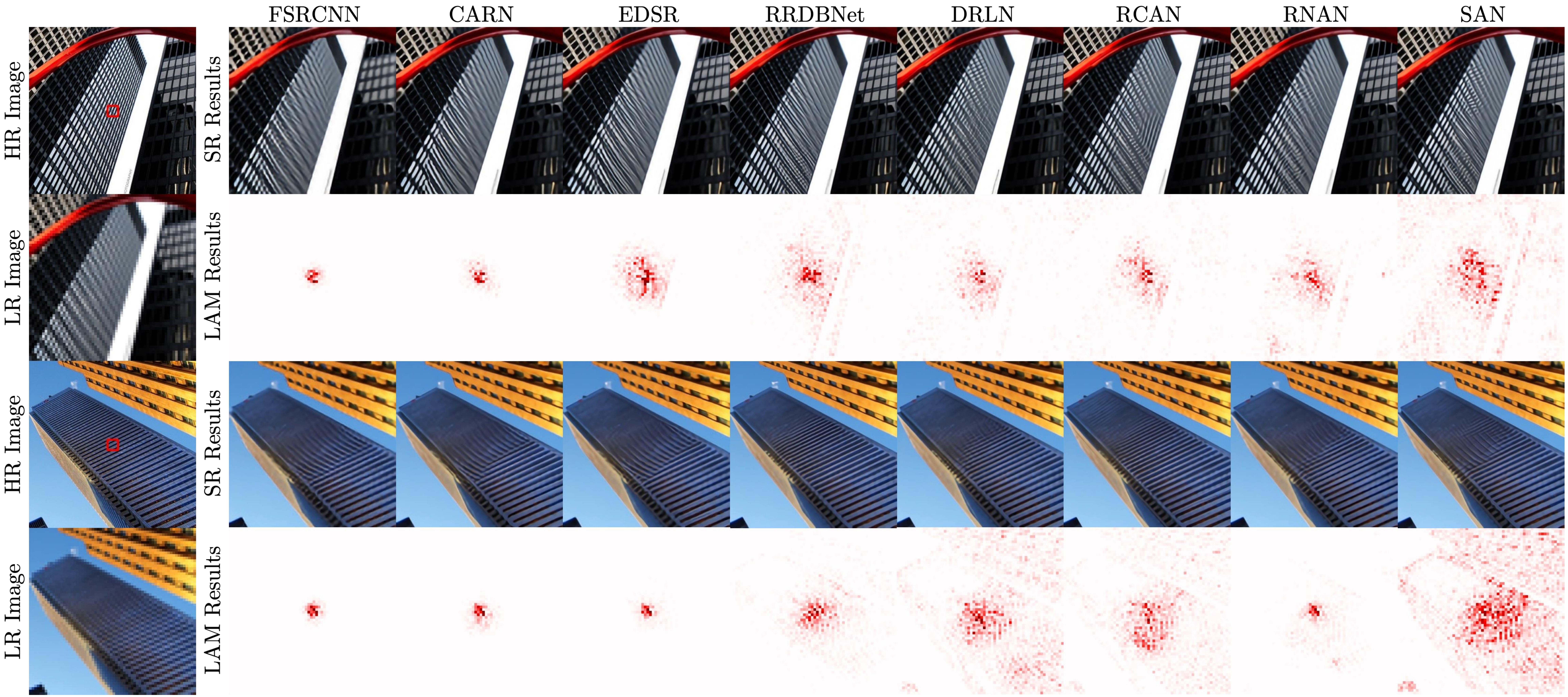
Progressive blurring path function: $\gamma_{\mathrm{pb}}(\alpha) = \omega(\sigma - \alpha\sigma) \otimes I$

Local Attribution Maps (LAM)

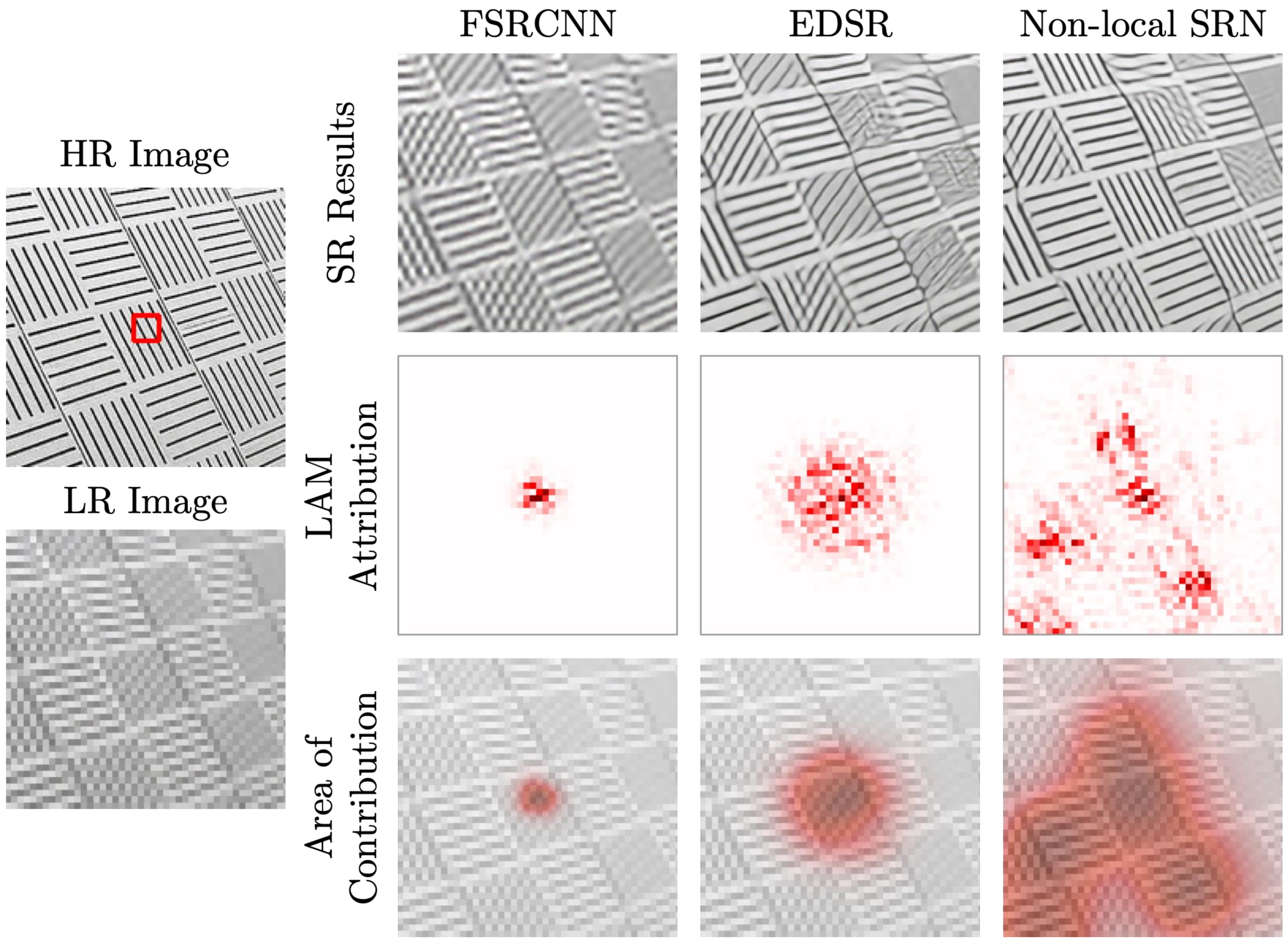
Why using path integral gradient: Gradient Saturation



Local Attribution Maps Results



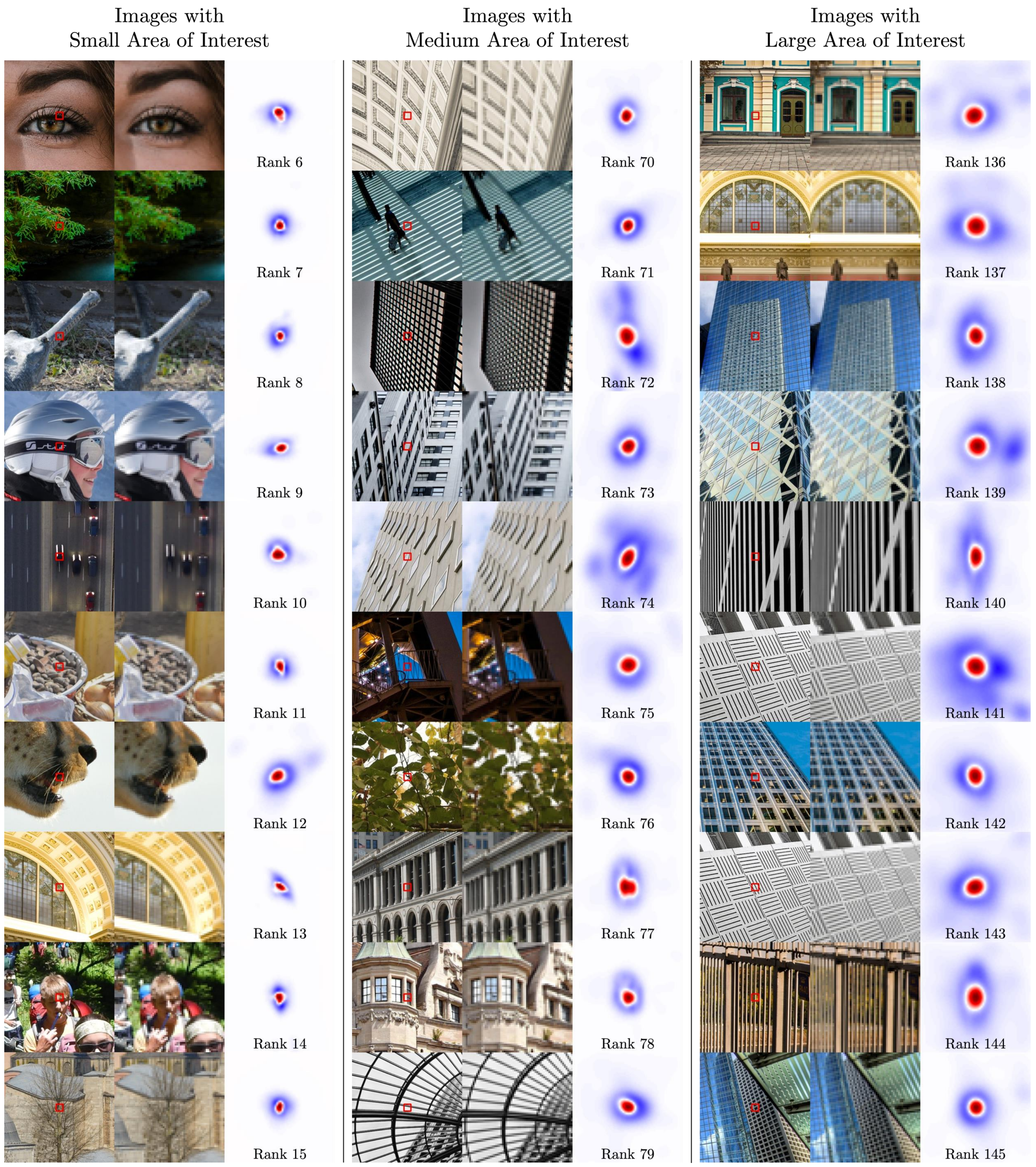
Local Attribution Maps Results



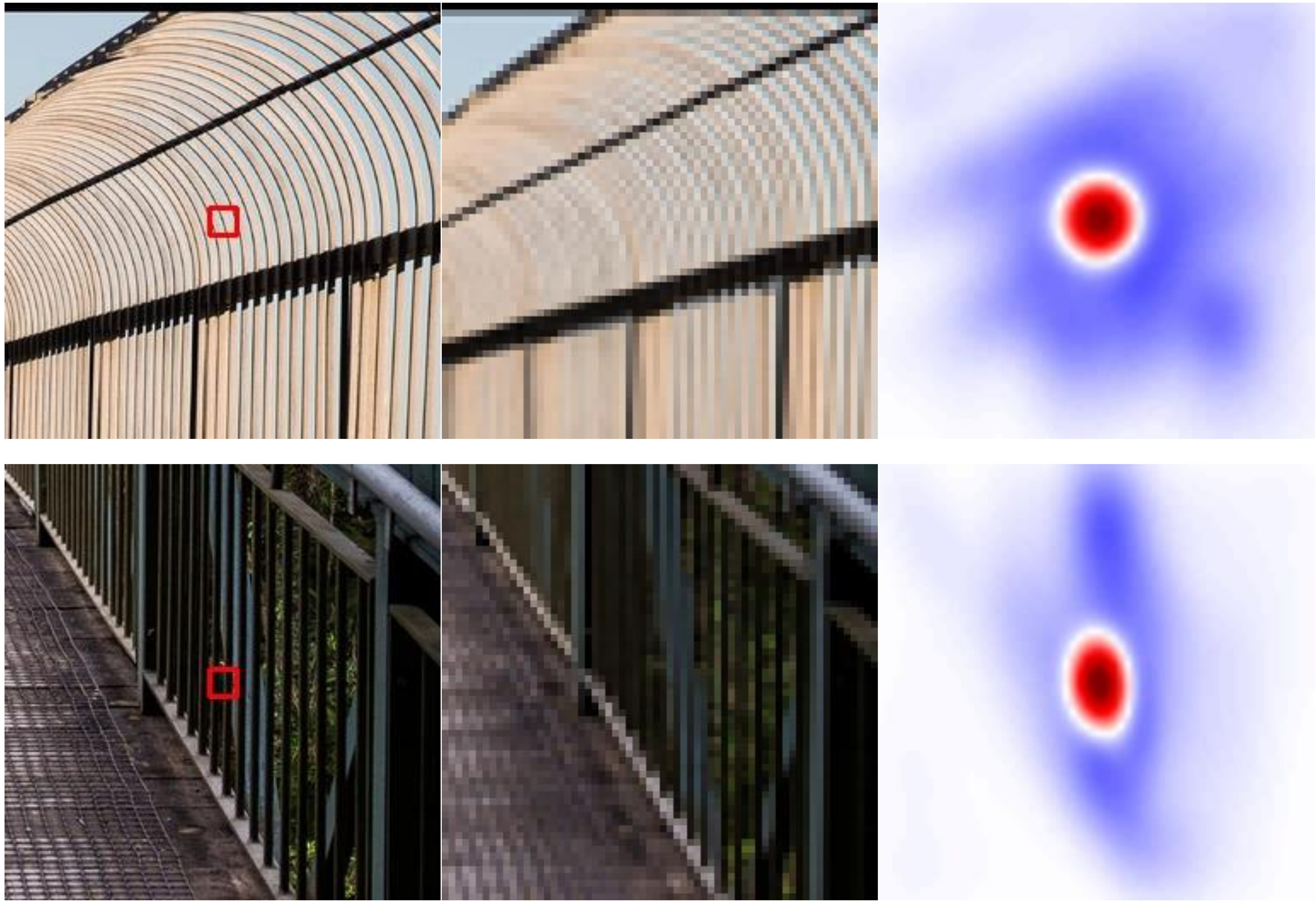
Informative Areas

The similarities and differences of LAM results for different SR networks

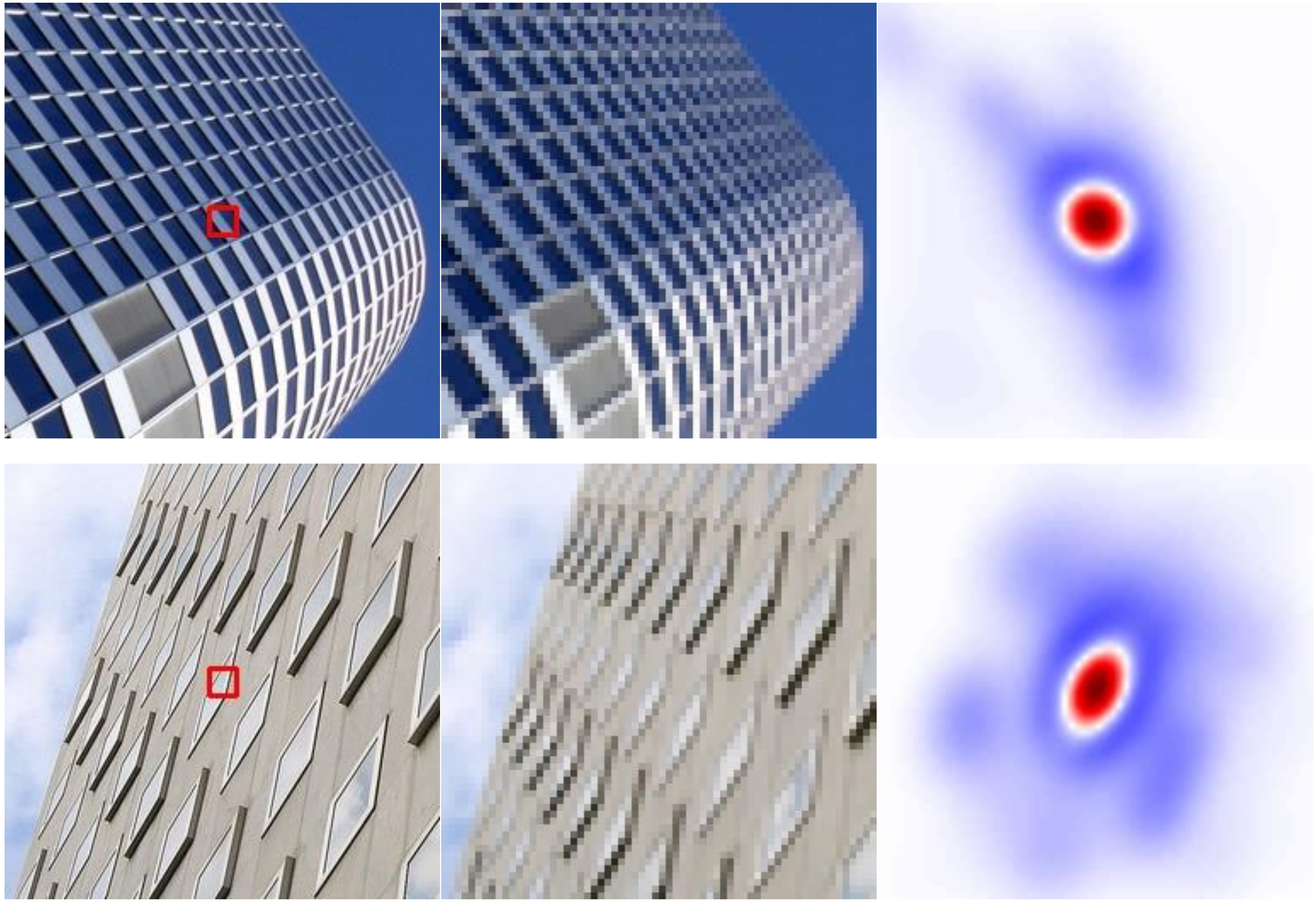
- Red areas can be used for the most preliminary level of SR
- Blue areas show the potential informative areas



Informative Areas



Informative Areas



Exploration with LAM

We use Gini Index to indicate the range of involved pixels:

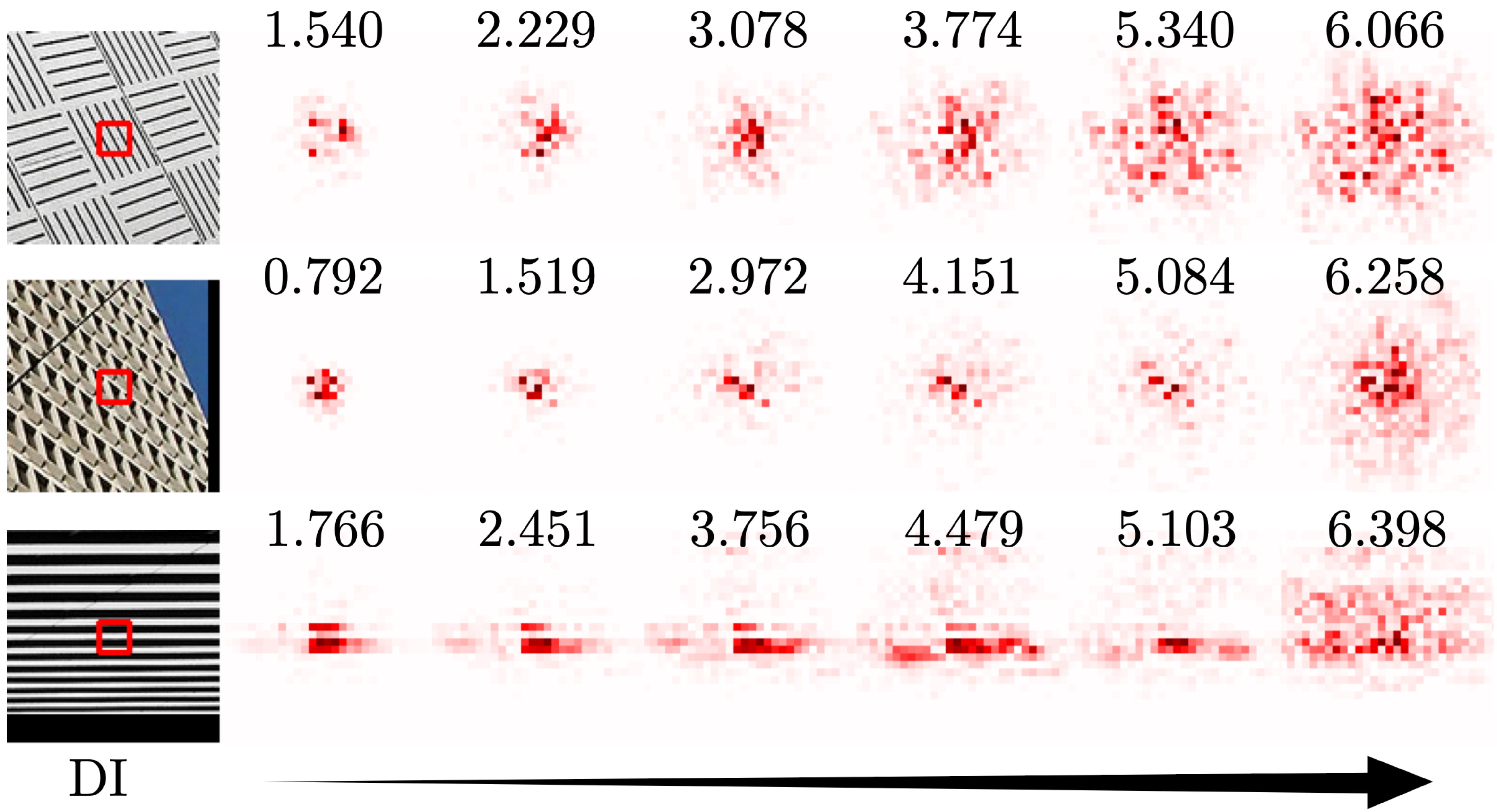
$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|g_i - g_j|}{2n^2\overline{g}}$$

We propose Diffusion Index for quantitative analysis:

$$\mathrm{DI} = (1 - G) \times 100$$

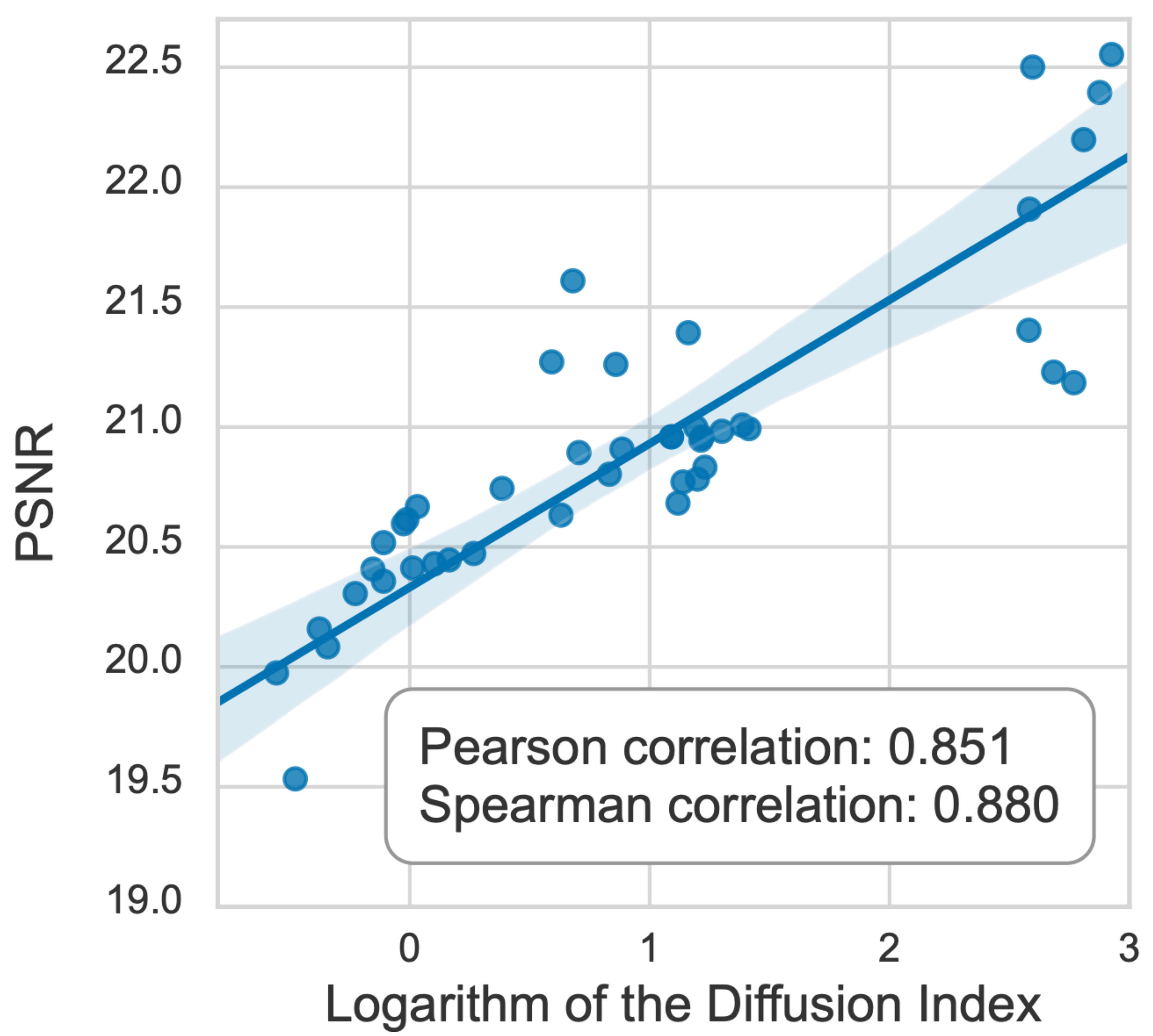
Exploration with LAM

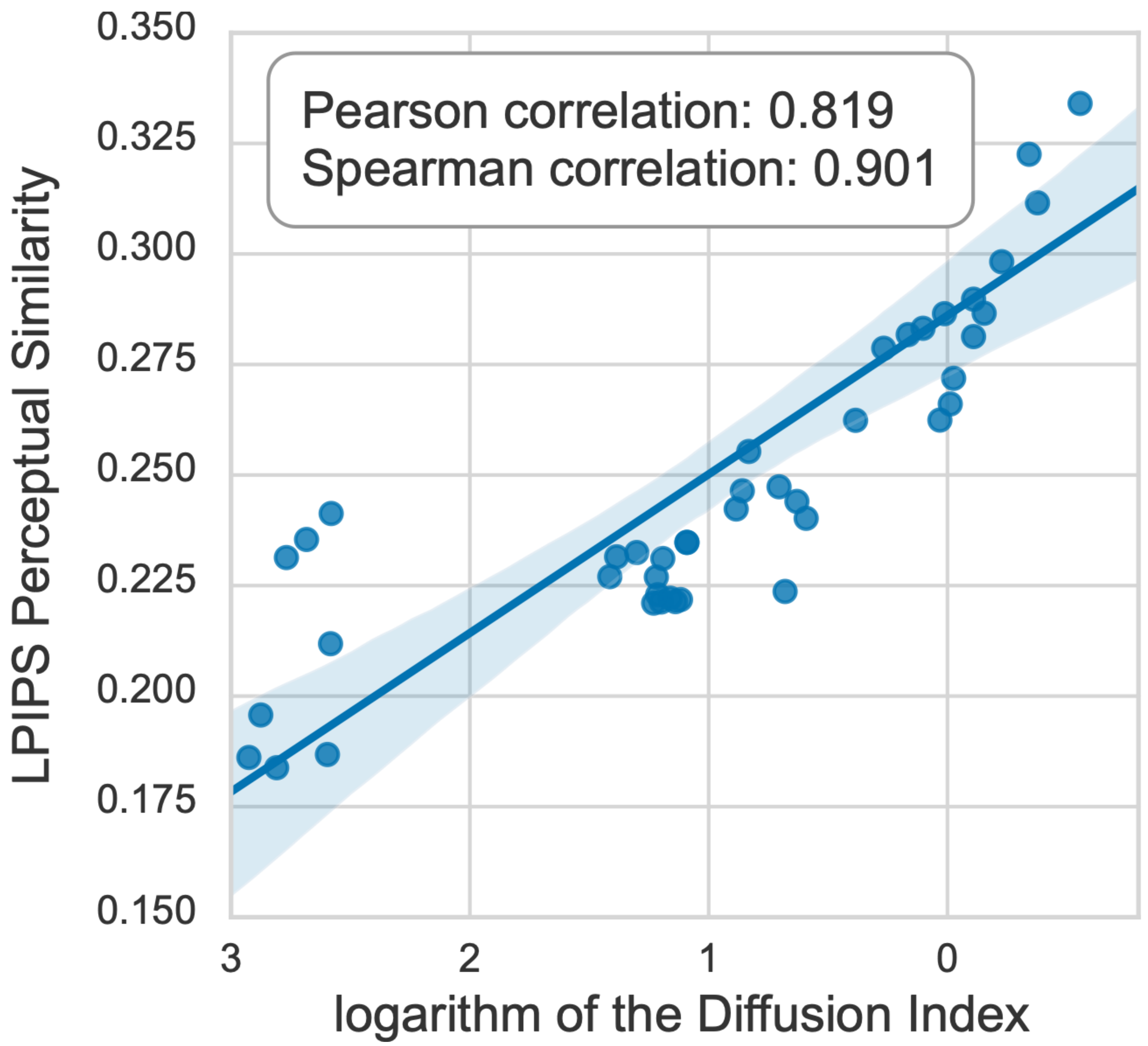
Diffusion Index for Quantitative Analysis:



Exploration with LAM

Diffusion Index vs. Network Performances.





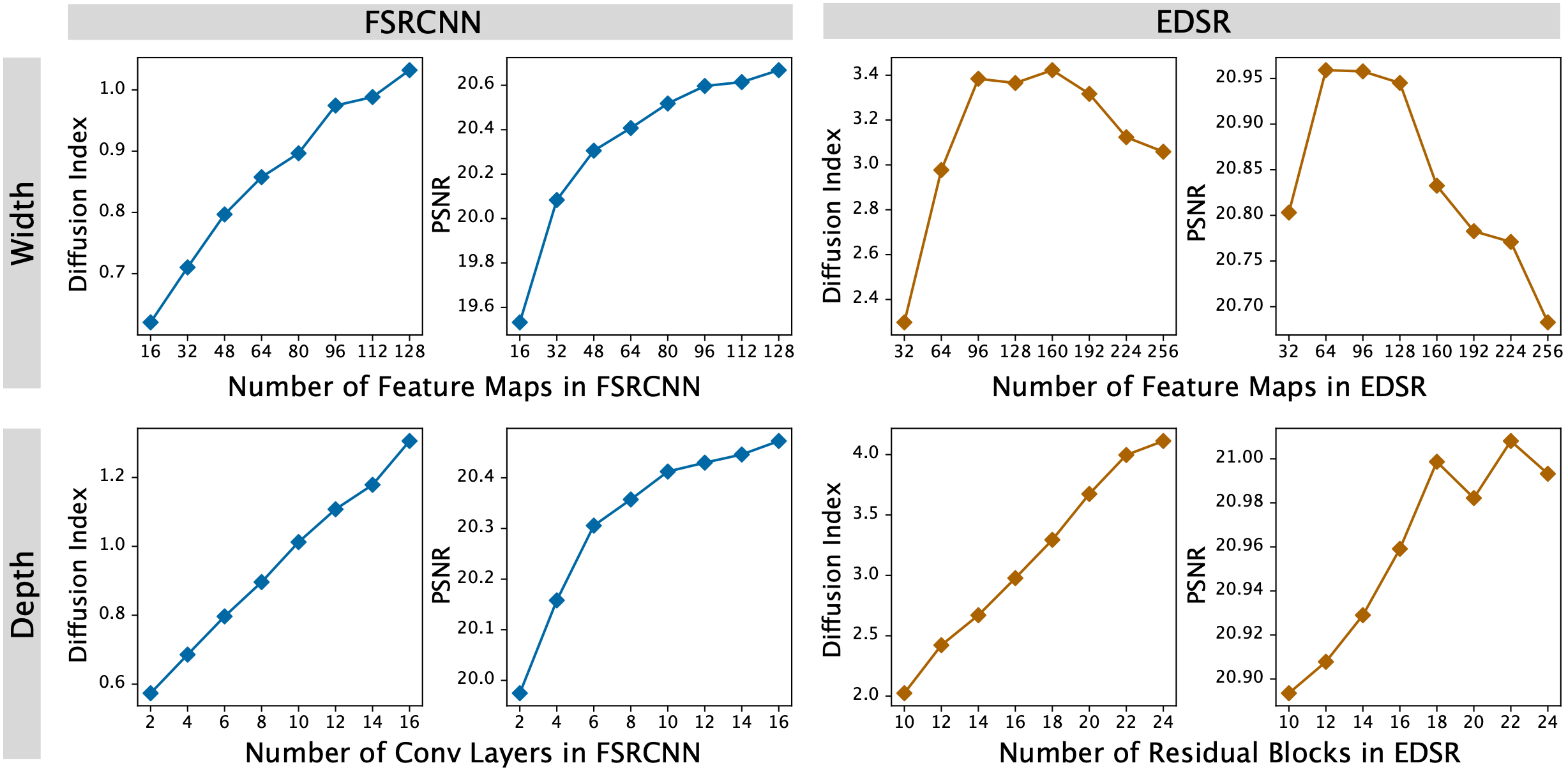
Exploration with LAM

Diffusion Index vs. Receptive Field.

Model	Recpt. Field	PSNR	DI	Remark
FSRCNN	17×17	20.30	0.797	Fully convolution network.
CARN	45×45	21.27	1.807	Residual network.
EDSR	75×75	20.96	2.977	Residual network.
MSRN	107×107	21.39	3.194	Residual network.
RRDBNet	703×703	20.96	13.417	Residual network.
IMDN	global	21.23	14.643	Global pooling.
RFDN	global	21.40	13.208	Global pooling.
RCAN	global	22.20	16.596	Global pooling.
RNAN	global	21.91	13.243	Non-local attention.
SAN	global	22.55	18.642	Non-local attention.

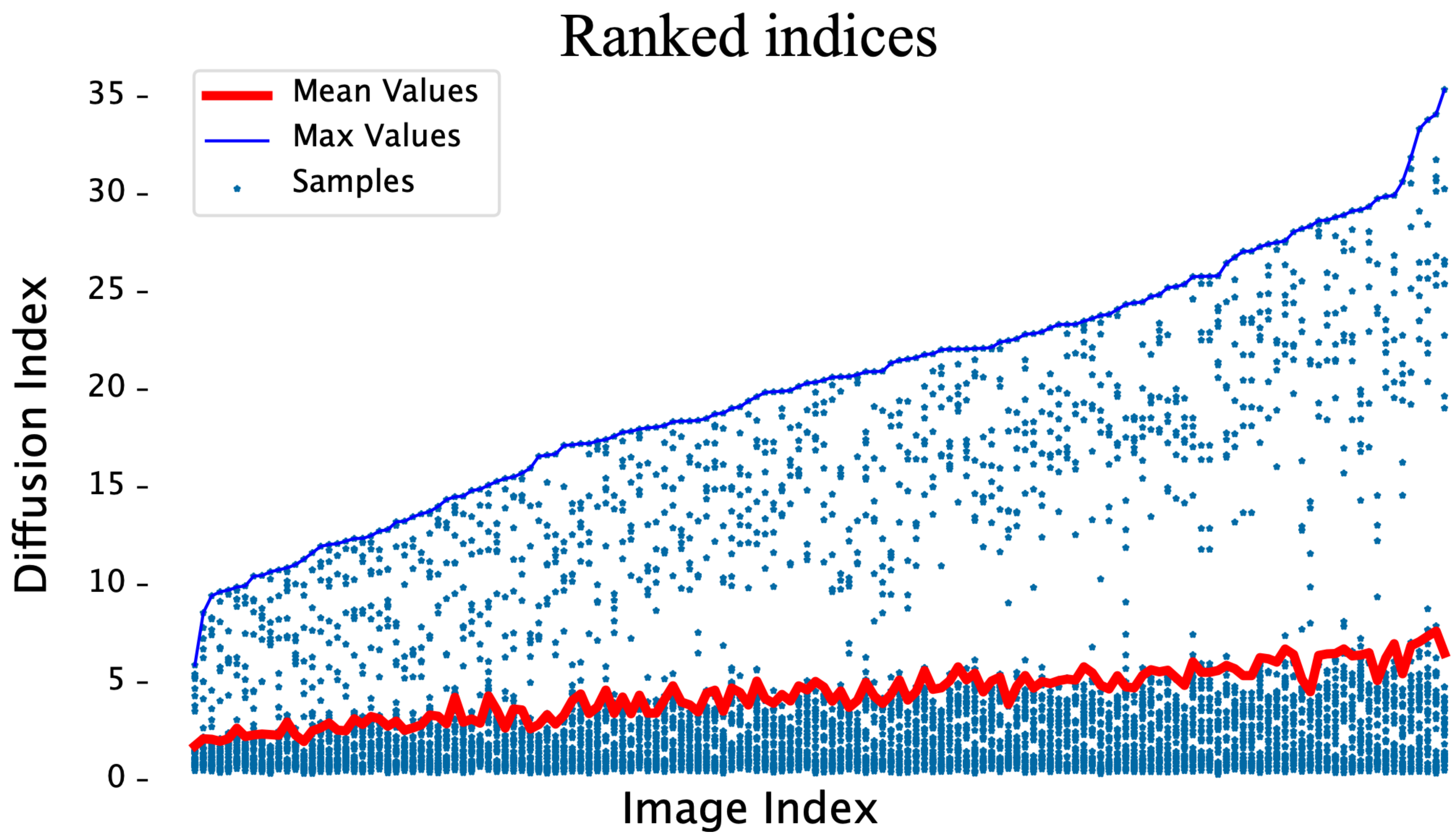
Exploration with LAM

Diffusion Index vs. Network Scale.

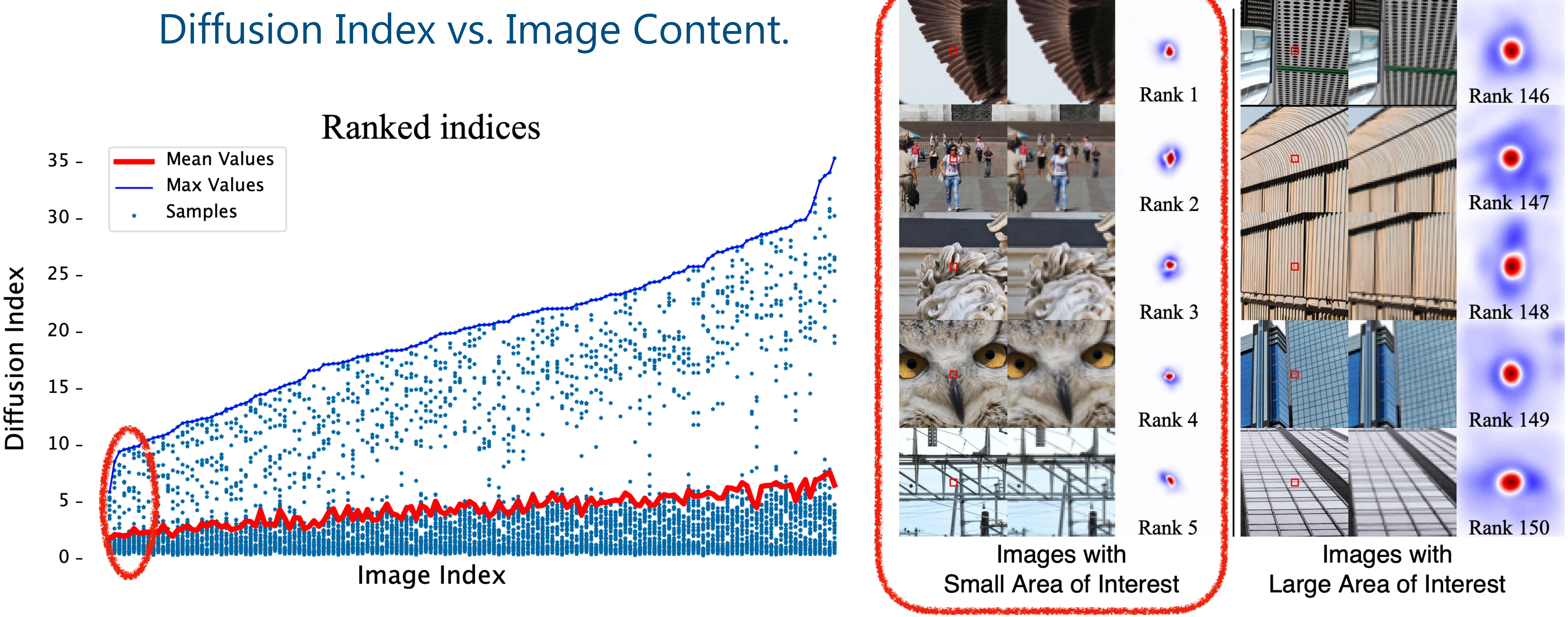


Exploration with LAM

Diffusion Index vs. Image Content.

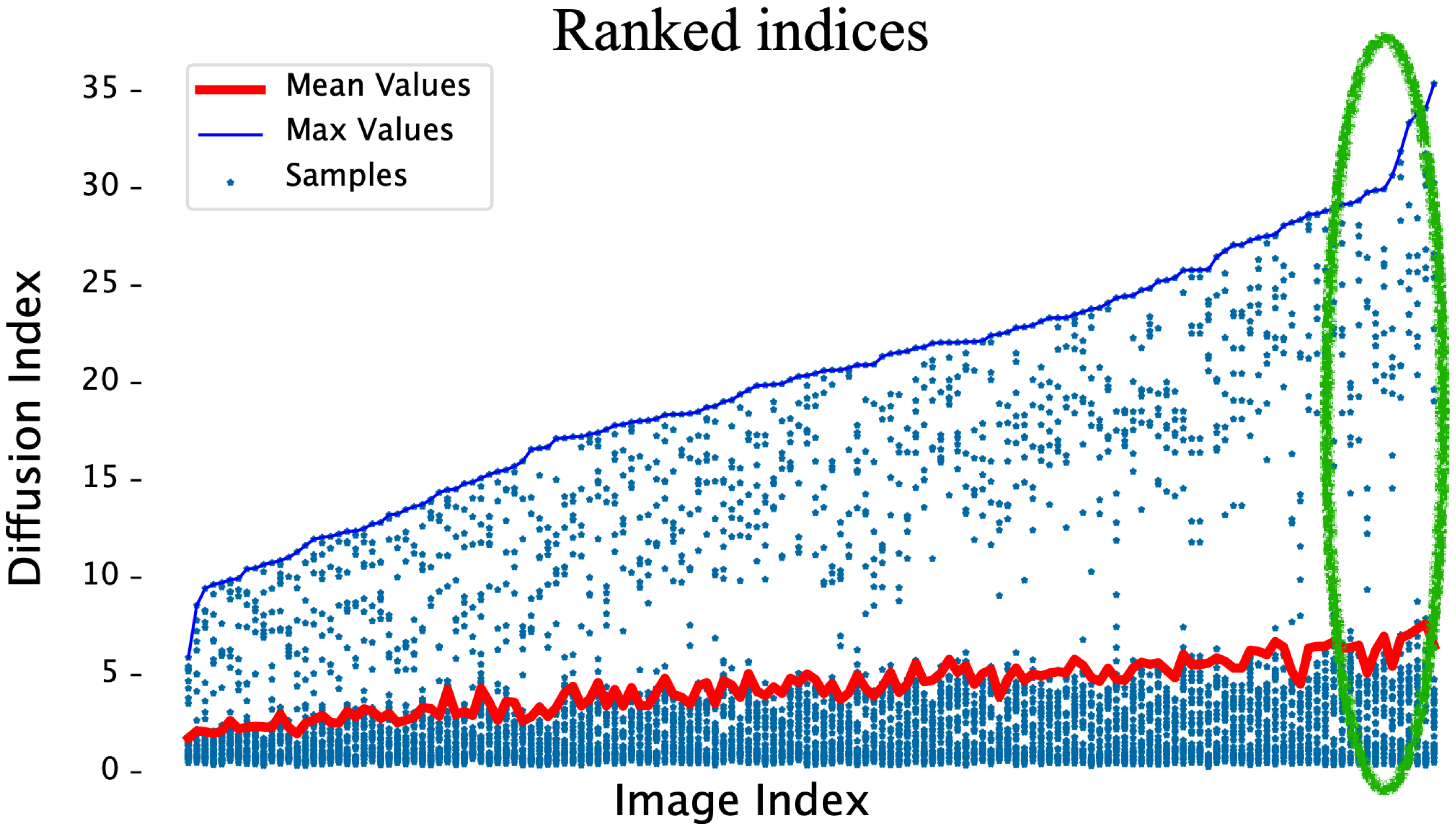


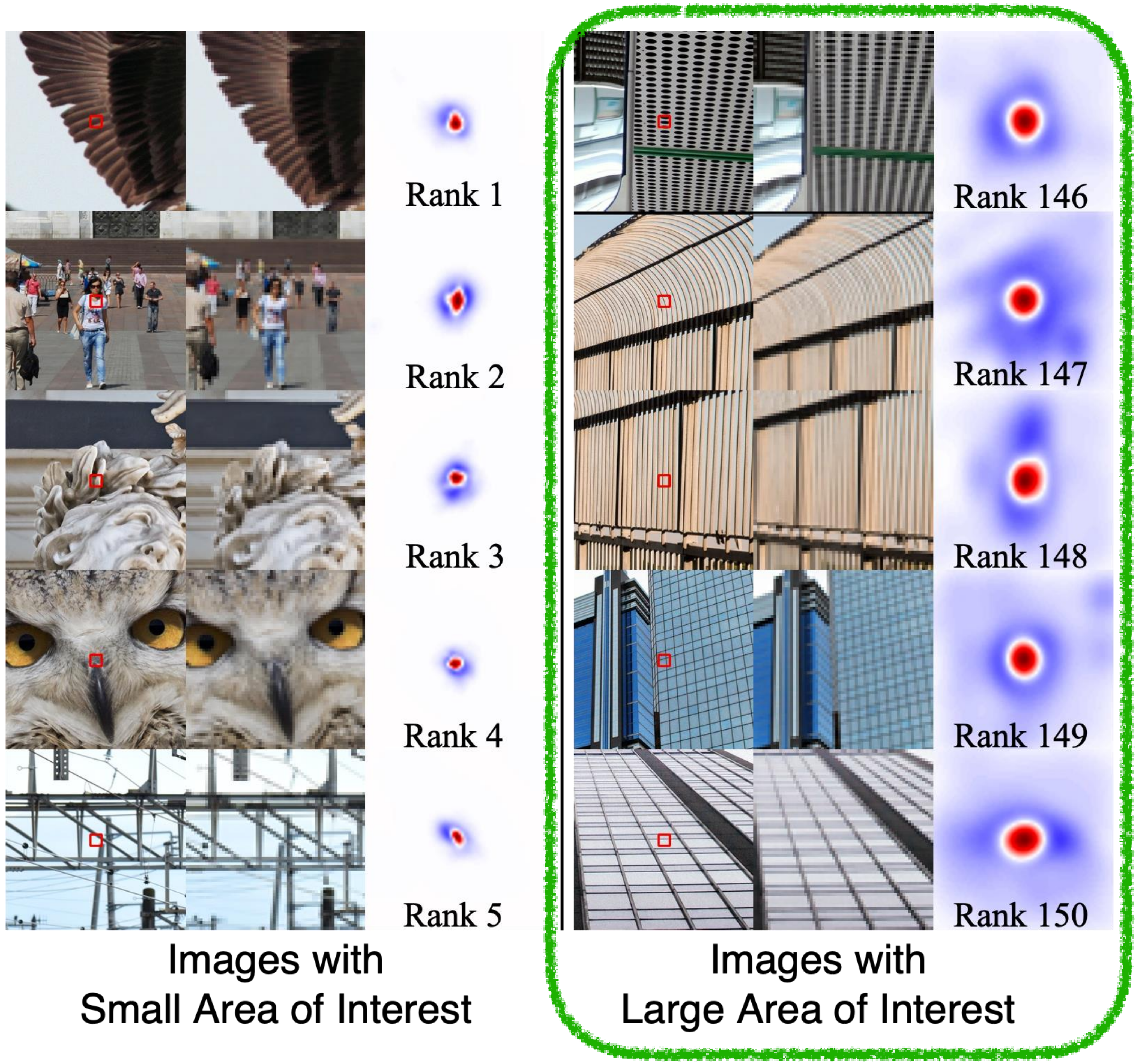
Exploration with LAM



Exploration with LAM

Diffusion Index vs. Image Content.





Interpreting Super-Resolution Networks

Interpretability
in Low-level Vision

Pixel: What pixels contribute most to restoration?

Feature: Where can we find semantics in SR-net?

Filters: Whether learned filters are discriminative?

Discovering "Semantics" in Super-Resolution Networks

Yihao Liu, Anran Liu, Jinjin Gu, Zhipeng Zhang, Wenhao Wu, Yu Qiao, Chao Dong

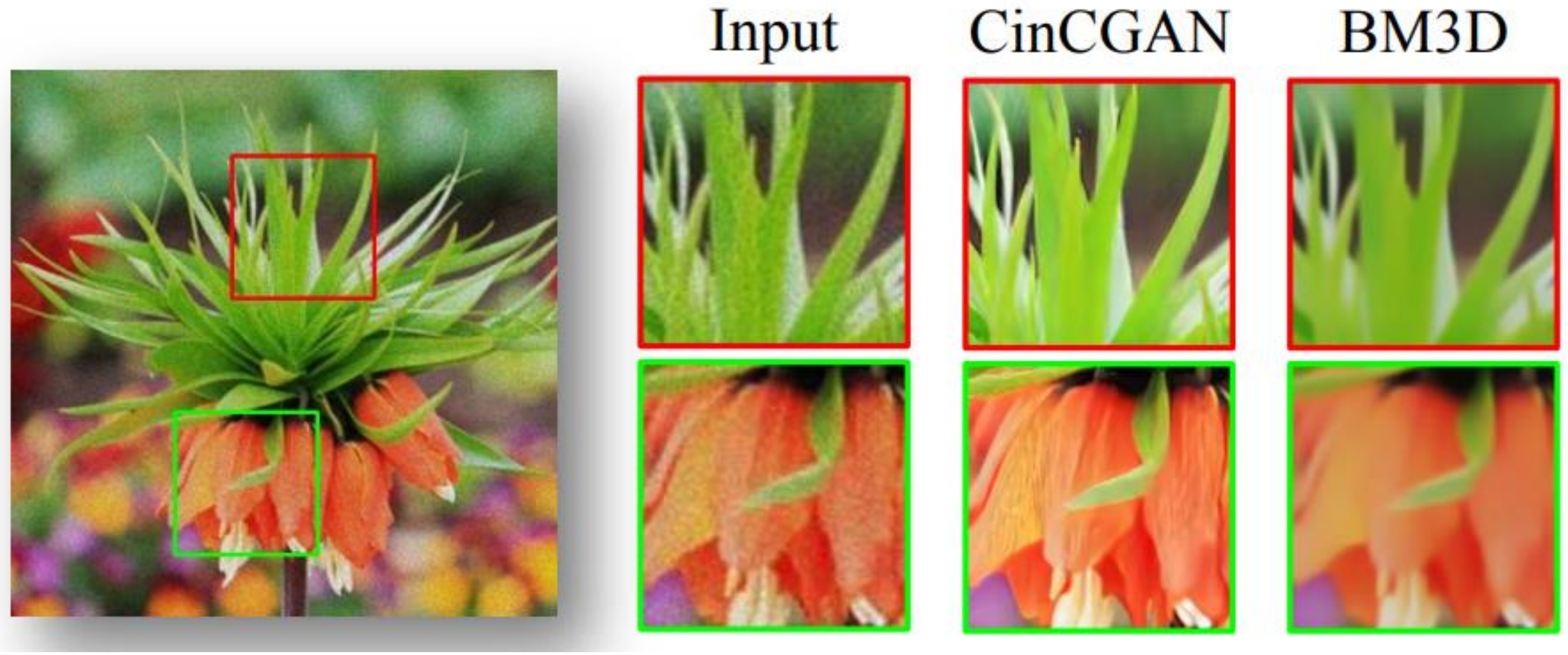
Shenzhen Institute of Advanced Technology, CAS
The University of Hongkong
University of Sydney
Shanghai AI Lab
Institute of Automation, CAS
Baidu Inc.

Discovering “Semantic”

No Semantic → ?? Semantic → Clear Semantic

No Semantic	?? Semantic	Clear Semantic
Traditional method e.g.,Interpolation	Low-level Vision e.g.,Super-resolution	High-level Vision e.g.,Classification

Observation

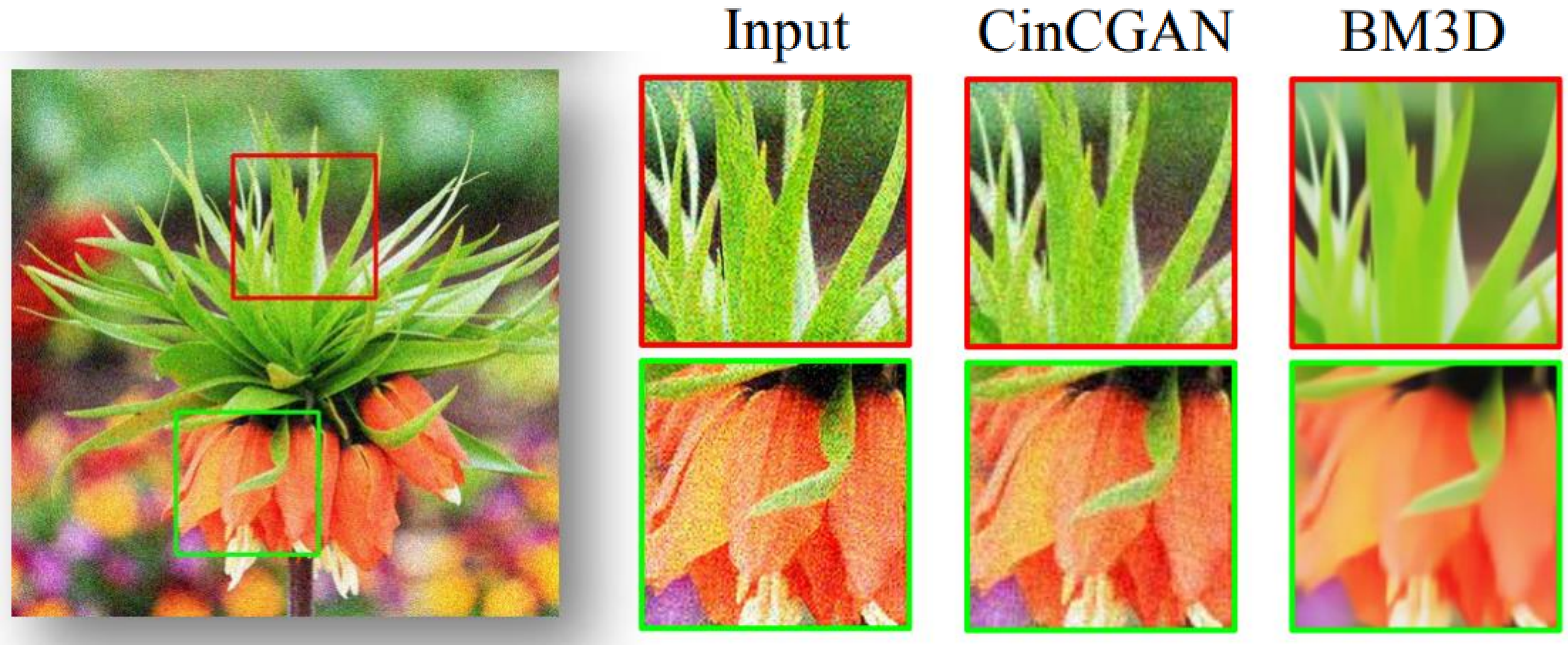


CinCGAN

>>

BM3D

Observation



CinCGAN
<<
BM3D

Observation

CinCGAN

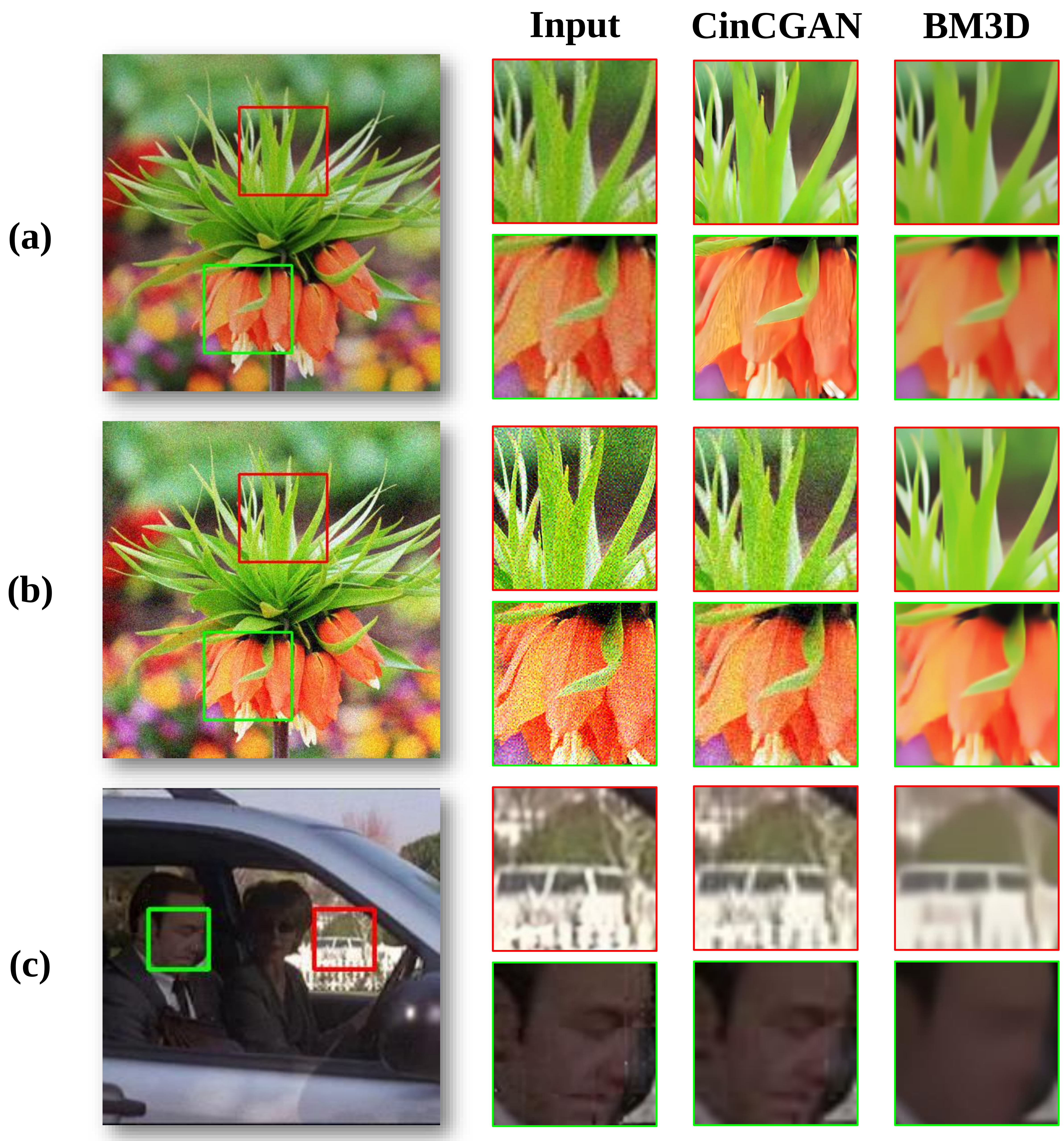
<<

BM3D

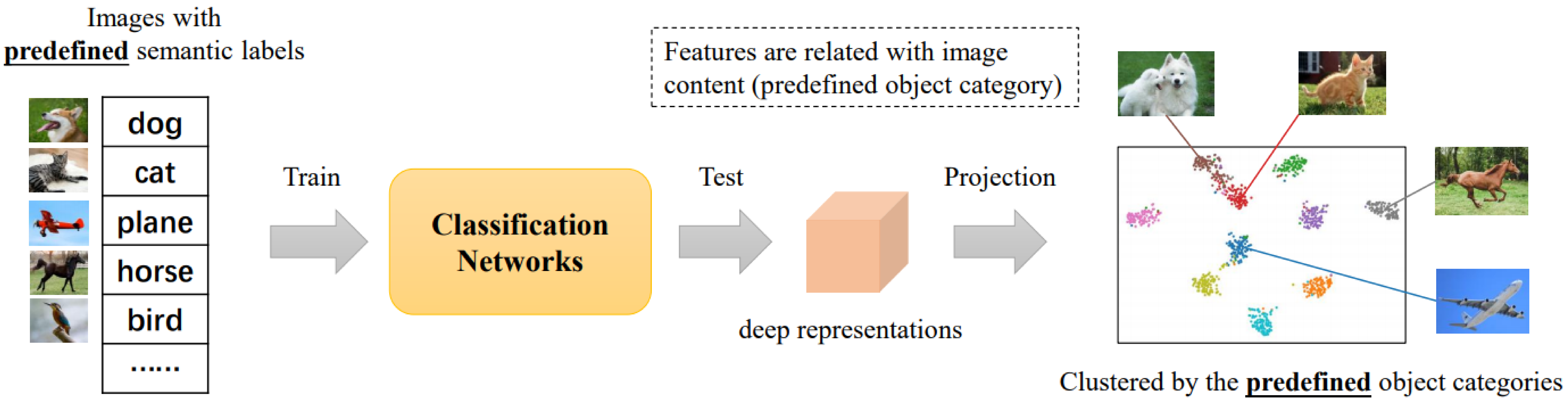
Observation

✓ CinCGAN can figure out the specific degradation types within its training data

✓ The distribution mismatch will make the network "**turn off**" its ability



Analogy to classification



PCA (dimensional reduction)
+ t-SNE (visualization)

Analogy to classification

- Clustering based on pre-defined object categories

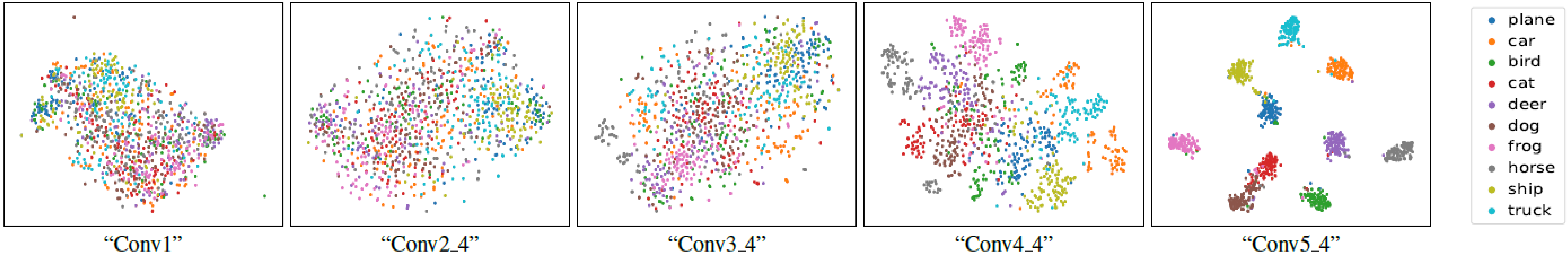
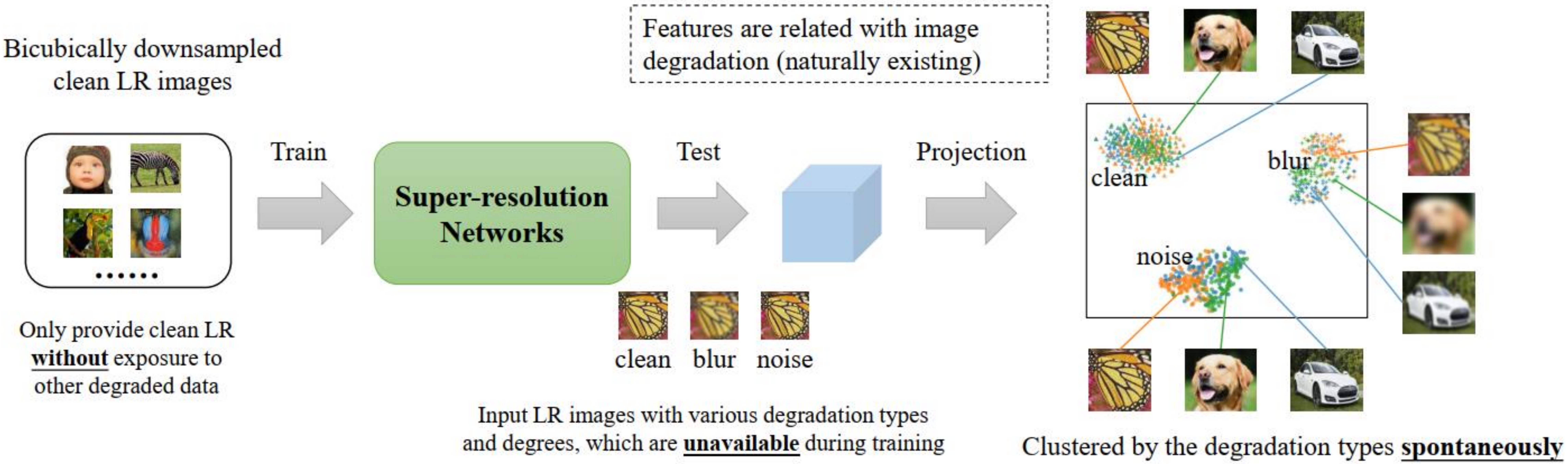


Figure 4. Projected feature representations extracted from different layers of ResNet18 using t-SNE. With the network deepens, the representations become more discriminative to object categories, which clearly shows the semantics of the representations in classification.

Deeper features contain clear semantics

Degradation-related semantics in SR-net

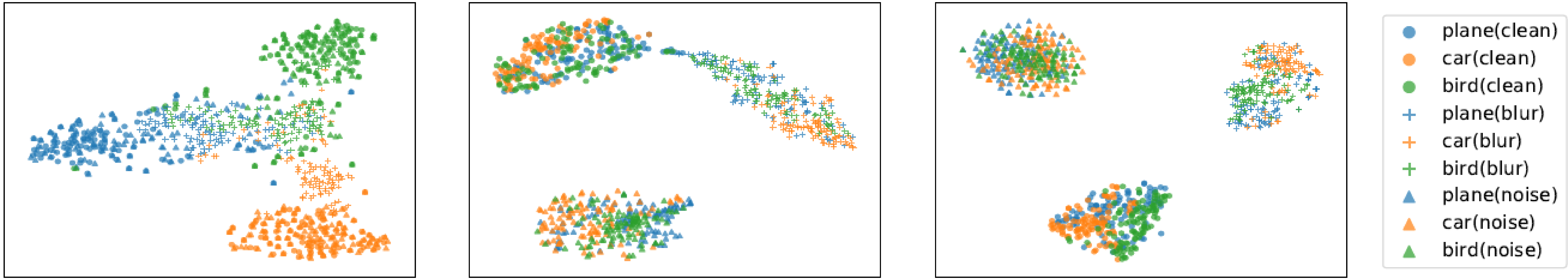


Features are clustered by degradations

They are trained on a single degradation type!

Classification VS. Super-resolution

Semantics in SR networks are in terms of degradation types regardless of the image contents.



(a) ResNet18 (classification) (b) SRResNet-wGR (c) SRGAN-wGR

Figure 5. Feature representation differences between classification and super-resolution networks. The same object category is represented by the same color, and the same image degradation type is depicted by the same marker shape. For the classification network, the feature representations are clustered by the same color, while the representations of SR network are clustered by the same marker shape, suggesting there is a significant difference in feature representations between classification and super-resolution networks. Better viewed on the screen.

Influential Factors

- SR networks with global residual shows discriminability to different degradation types with high-level features.

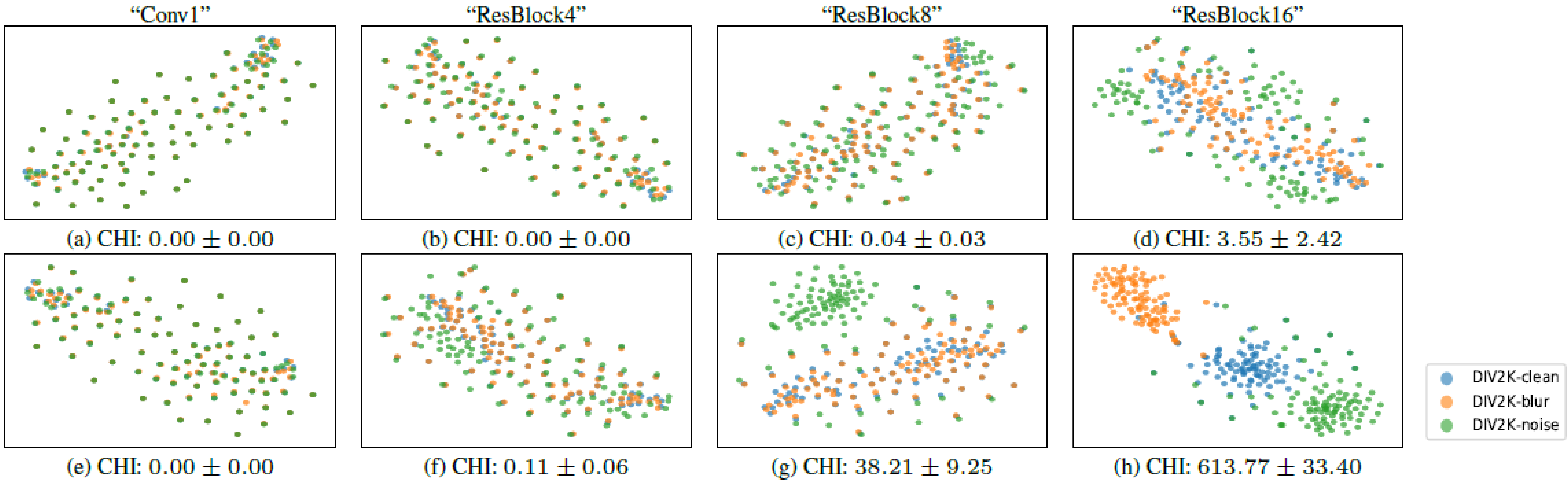


Figure 7. Projected feature representations extracted from different layers of SRResNet-woGR (1st row) and SRResNet-wGR (2nd row) using t-SNE. With image global residual (GR), the representations of MSE-based SR networks show discriminability to degradation types.

Influential Factors

- SR networks trained with discriminator (GAN) shows more obvious discriminability to different degradation types.

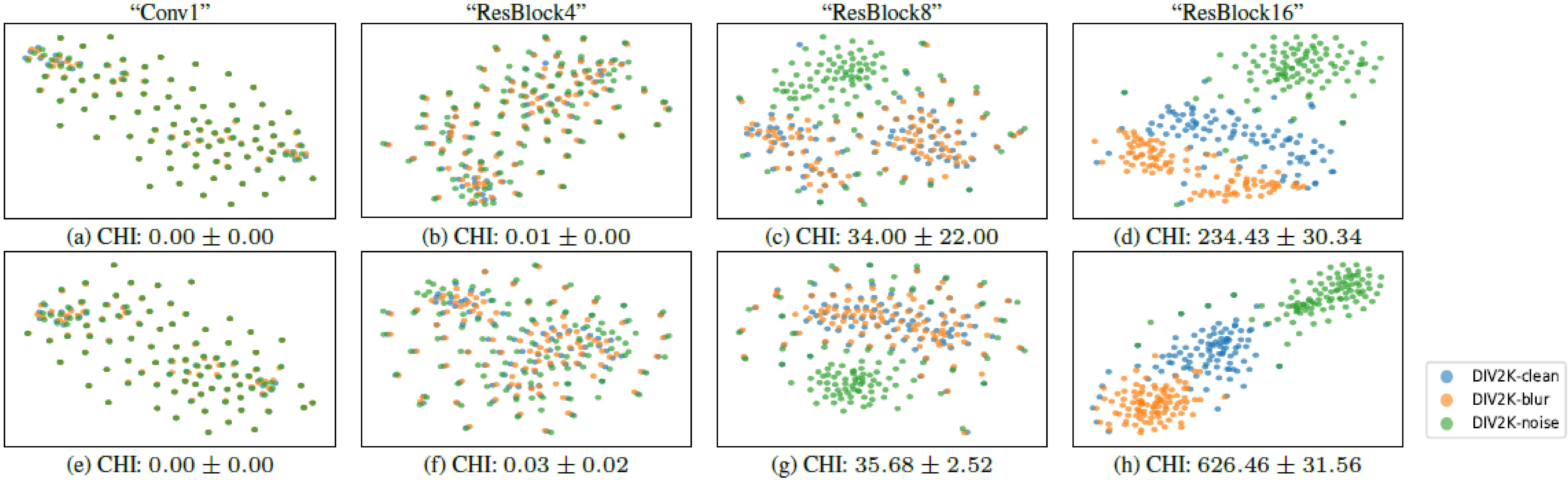


Figure 8. Projected feature representations extracted from different layers of SRGAN-woGR (1st row) and SRGAN-wGR (2nd row) using t-SNE. Even without GR, GAN-based SR networks can still obtain deep degradation representations.

Influential Factors

- Different degradation types/degrees differ a lot in discriminative ablity.

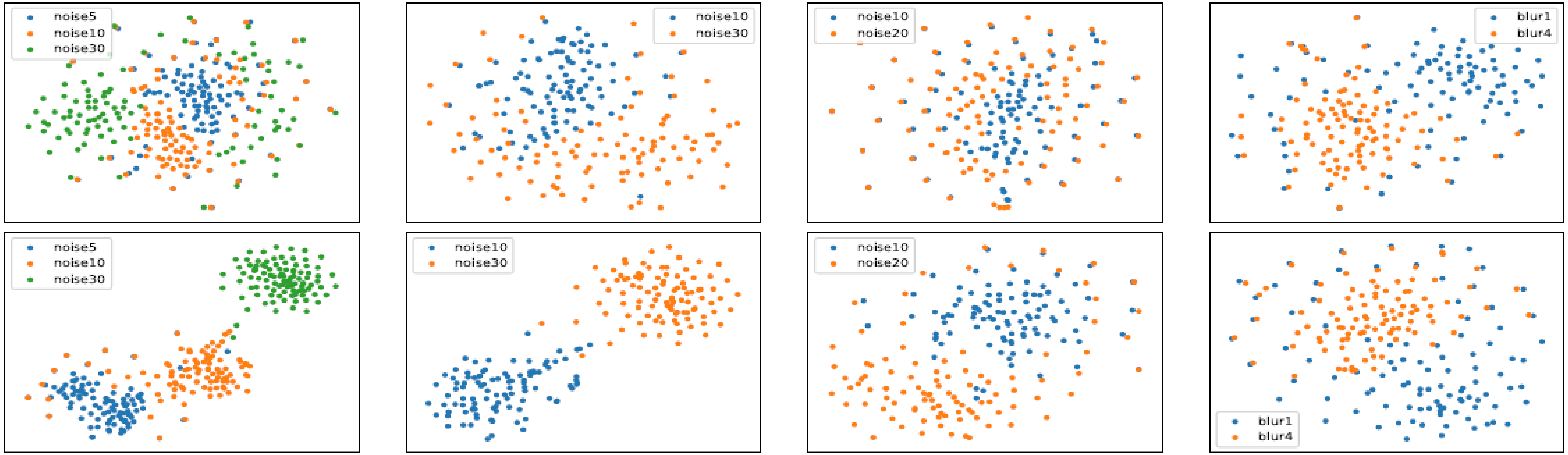


Figure 9. Even for the same type of degradation, different degradation degrees will also cause differences in features. The greater the difference between degradation degrees, the stronger the discriminability. **First row:** SRResNet-wGR. **Second row:** SRGAN-wGR.

Inspirations

- Interpreting the Generalization of SR Networks
- Developing Degradation-adaptive Algorithms
- Disentanglement of Image Content/Degradation

Interpreting Super-Resolution Networks

Interpretability
in Low-level Vision

Pixel: What pixels contribute most to restoration?

Feature: Where can we find semantics in SR-net?

Filters: Whether learned filters are discriminative?

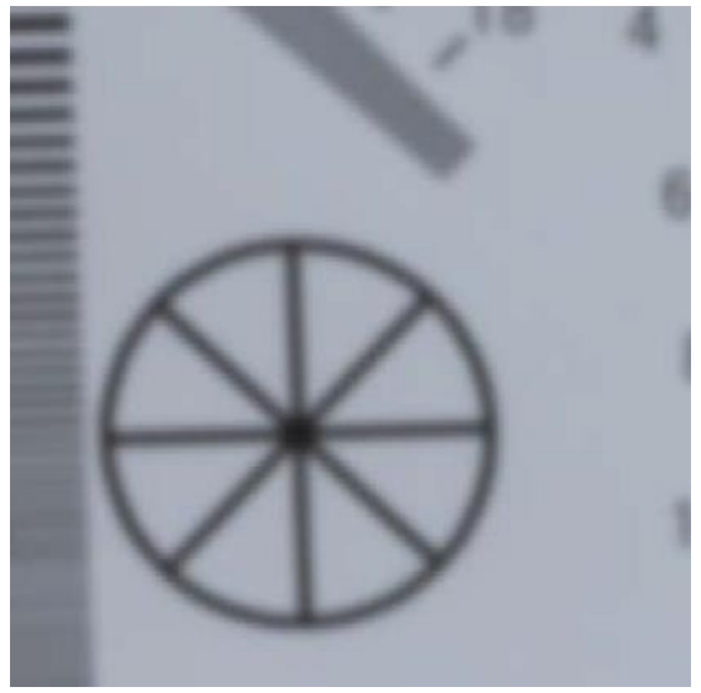
Finding Discriminative Filters for Specific Degradations in Blind Super-Resolution

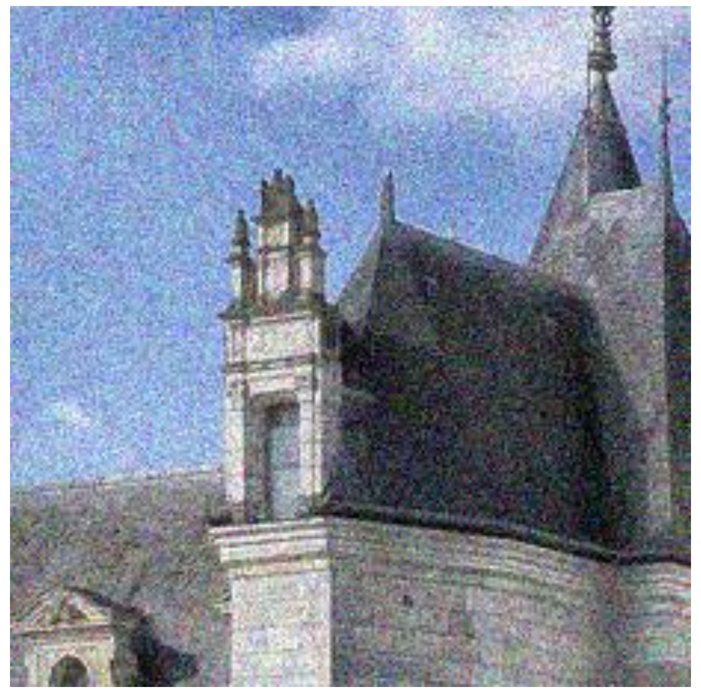
Liangbin Xie, Xintao Wang, Chao Dong
Zhongang Qi, Ying Shan

Applied Research Center (ARC), Tencent PCG
Shenzhen Institute of Advanced Technology, CAS

Background – Blind SR

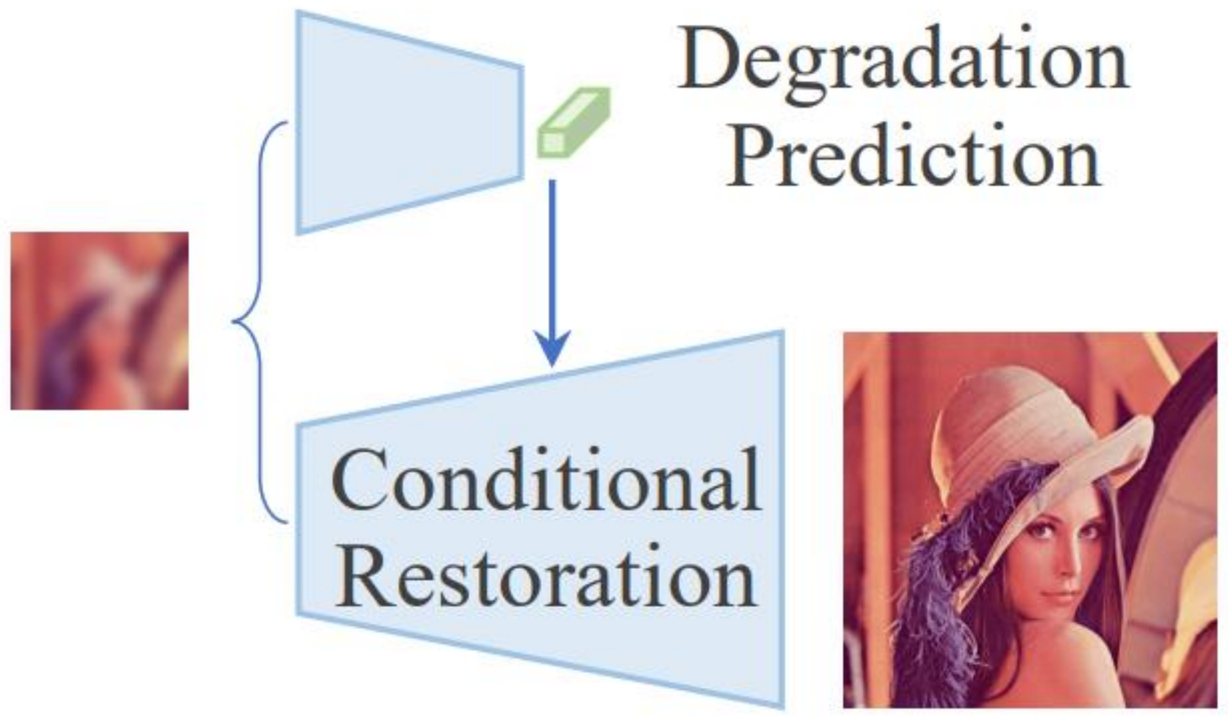
Reconstruct a high-resolution image from its low-resolution counterpart which contains unknown and complex degradations


blur


noise


JPEG

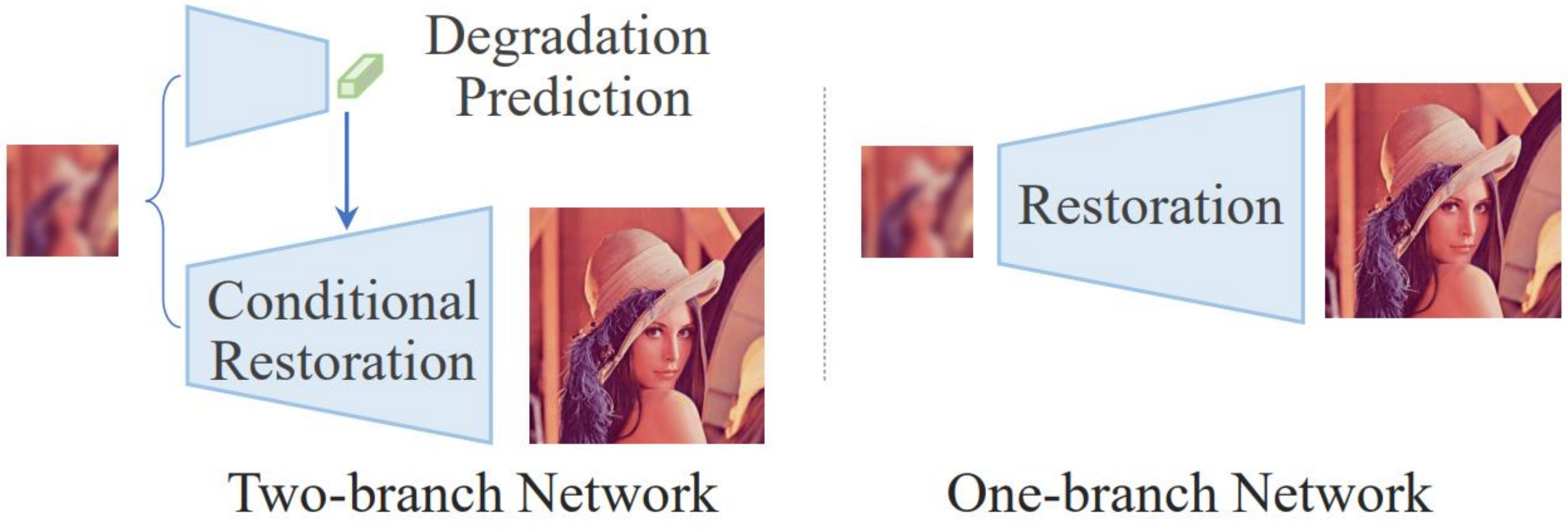
Typically, consists of two branches
- one for degradation prediction
- the other for conditional restorations

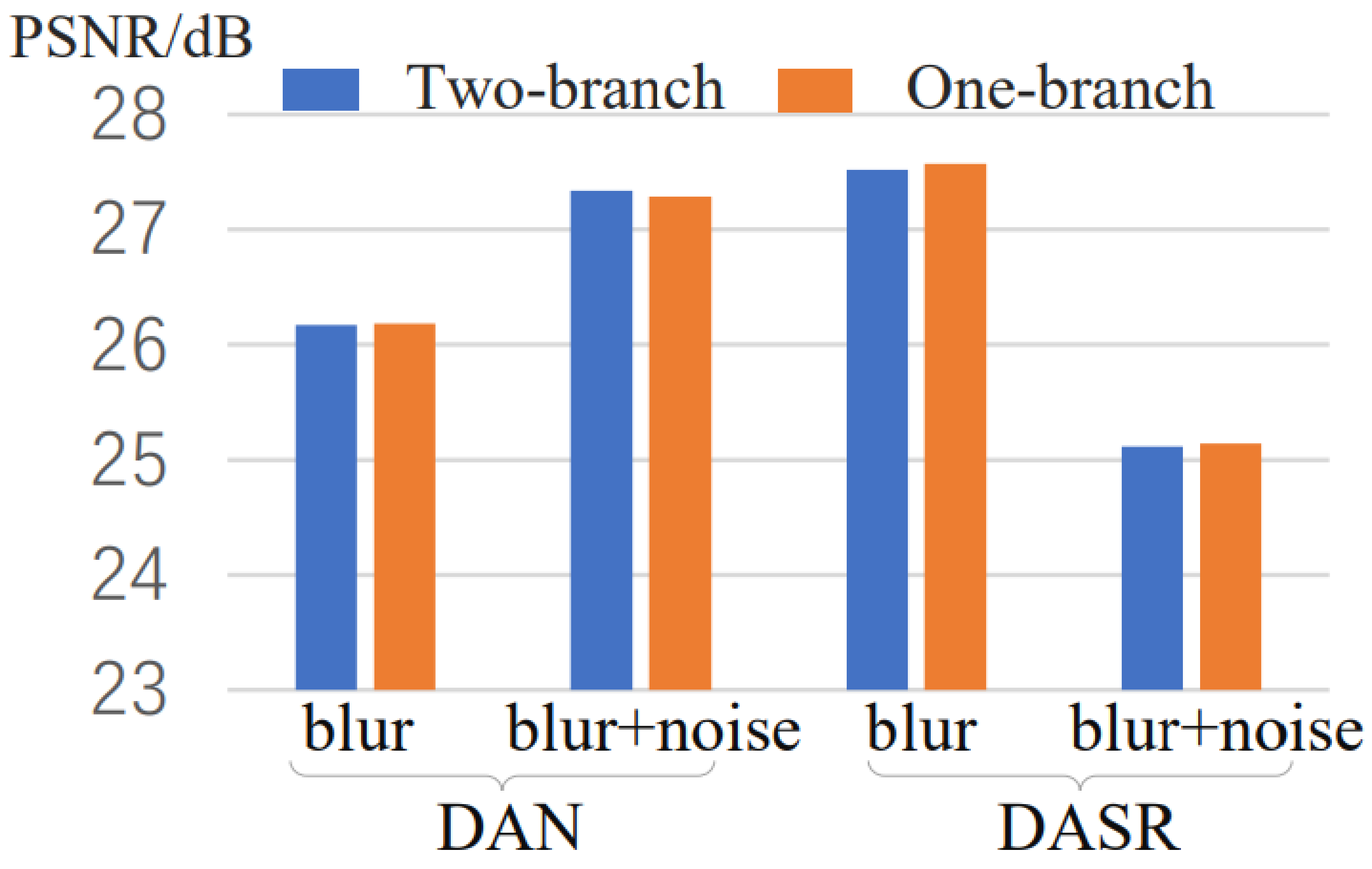


Two-branch Network

Motivation

We conduct preliminary experiments on several state-of-the-art methods: DAN and DASR.





PSNR (dB)	DAN [3]		DASR [5]	
	blur	blur+noise	blur	blur+noise
Official two-branch	26.168±0.009	27.341±0.072	27.518±0.034	25.116±0.012
SRResNet one-branch	26.182±0.011	27.288±0.027	27.573±0.010	25.143±0.013

A unified one-branch network could achieve comparable performance !

Motivation

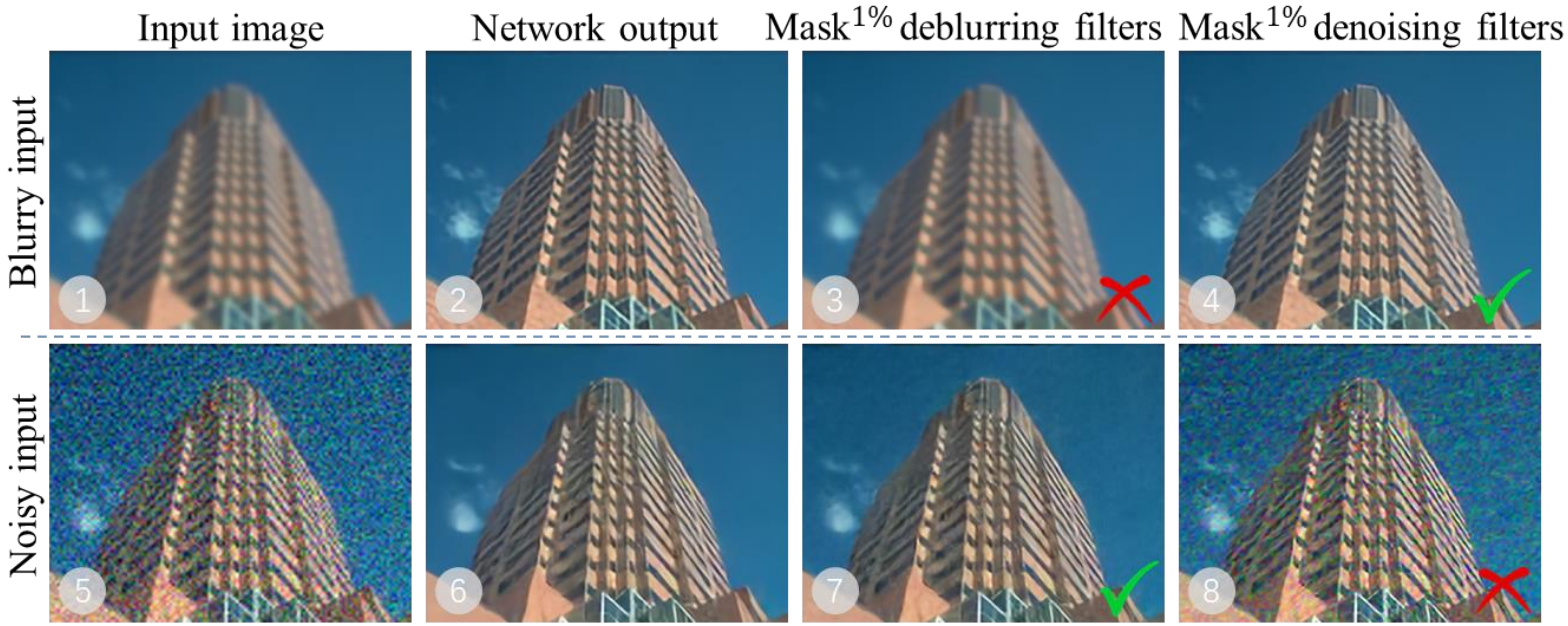
One-branch network - more like a 'black-box'
Two-branch network - delicate designs with higher interpretability

Two key questions:

- ✓ Could one-branch networks automatically learn to distinguish degradations as in two-branch methods?
- ✓ Are there any small sub-network (a set of filters) existing inside the unified network for a specific degradation?

Basic Finding

In one-branch blind SR networks, we are able to find a very small number of (at least to 1%) discriminative filters for each specific degradation (e.g., blur, noise).



Methods – Background

In classification task, Integrated Gradient (IG) is used to attributes the most important input components (e.g., pixels in input images) that affect the network predictions.

$$\mathtt{IG}_i(x) = (x_i - \bar{x}_i) \times \int_{\alpha=0}^{1} \frac{\partial F(\boxed{\bar{x} + \alpha \times (x - \bar{x})})}{\partial x_i} d\alpha,$$

Recall that: Gradient – The fastest changing direction

So IG finds the input pixels that will change the network output largely, i.e., the most import pixels that can interpret the network prediction.

Filter Attribution Integrated Gradients (FAIG)

	Classification	Blind SR
Purpose	Find input pixels that explain **network prediction**	Fine core filters that explain **degradation removal**
Attribute to	Input **pixels**	network **parameters** (filters)
Integral path	Input space	Parameter space
Method	Integrated Gradient	**Filter Attribution Integrated Gradients (FAIG)**

We propose Filter Attribution Integrated Gradients (FAIG) to attribute network functional alterations to filter change.

Methods – FAIG

1. The *baseline network* $F(\bar{\theta})$ is a pure SR network that cannot remove any degradations.

2. The *target network* $F(\theta)$ is a re-trained network that can deal with complex degradations.

3. Given the same input, the changes of the network output can be attributed to the changes of network parameters (*i.e.*, filters).

Methods – FAIG

We quantify the network function of degradation removal by

$$\mathcal{L}(\theta, x) = ||F(\theta, x) - x^{gt}||_2^2$$

We consider a continuous path between the baseline model and the target model

$$\gamma(\alpha) = \bar{\theta} + \alpha \times (\theta - \bar{\theta})$$

We can get the gradient of each dimension of network parameters with FAIG

$$\mathtt{FAIG}_i(\theta, x) = \int_{\alpha=0}^{1} \frac{\partial \mathcal{L}(\gamma(\alpha), x)}{\partial \gamma(\alpha)_i} \times \frac{\gamma(\alpha)_i}{\partial \alpha} d\alpha$$

Methods – FAIG

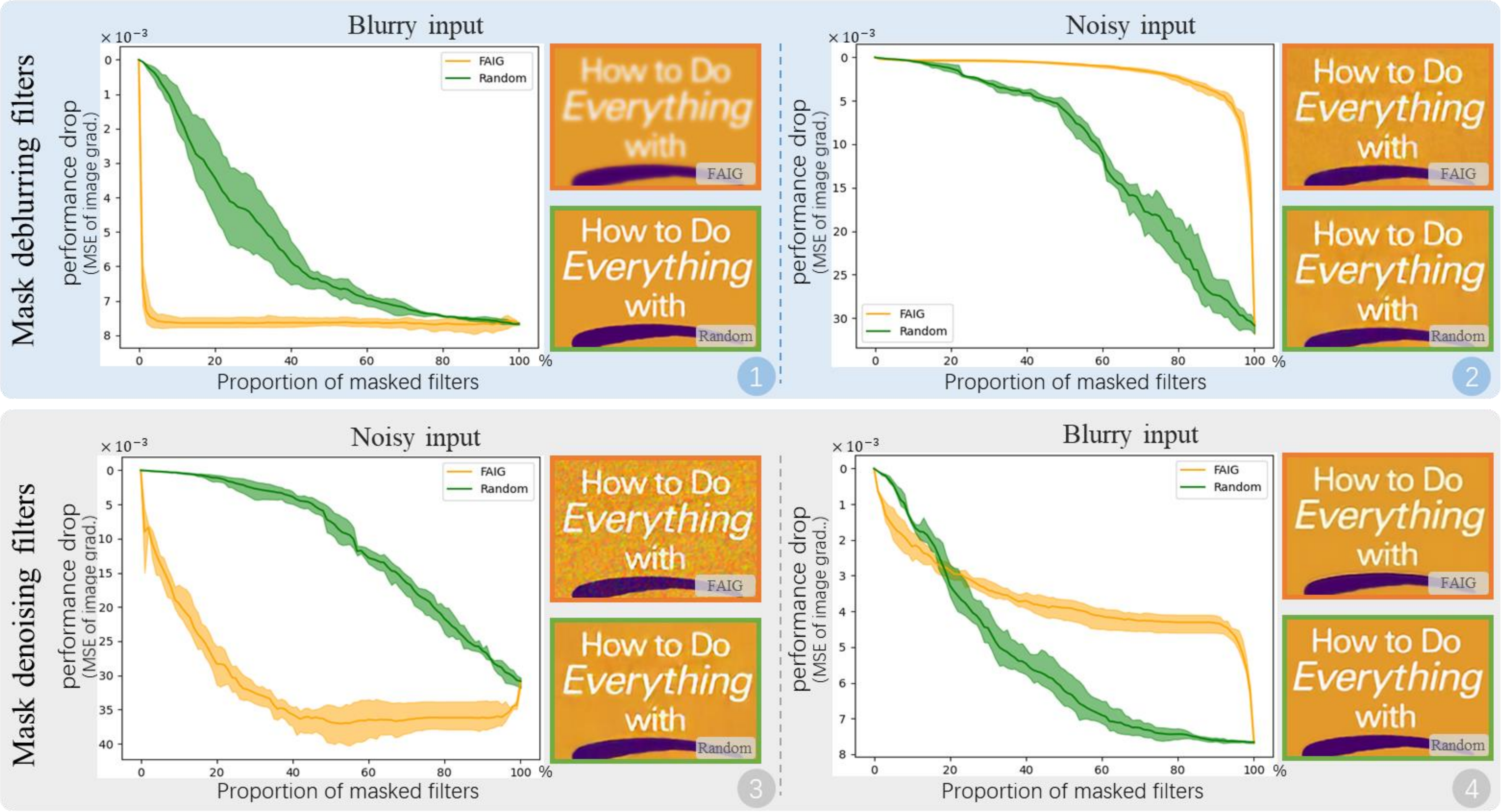
1. We calculate the gradient difference between a specific degradation of interest D and other degradations ~D.

2. We average all the gradient difference in a whole dataset to eliminate the impact of image contents.

$$\mathtt{FAIG}_i^{\mathcal{D}}(\theta) = \frac{1}{\|\mathcal{X}|}(\underbrace{\sum_{x \in \mathcal{X}} |\mathtt{FAIG}_i(\theta, x^{\mathcal{D}})|}_{\text{attribution for degradtion}\mathcal{D}} - \underbrace{\sum_{x \in \mathcal{X}} |\mathtt{FAIG}_i(\theta, x^{\sim\mathcal{D}})|}_{\text{attribution for other degradtions}})$$

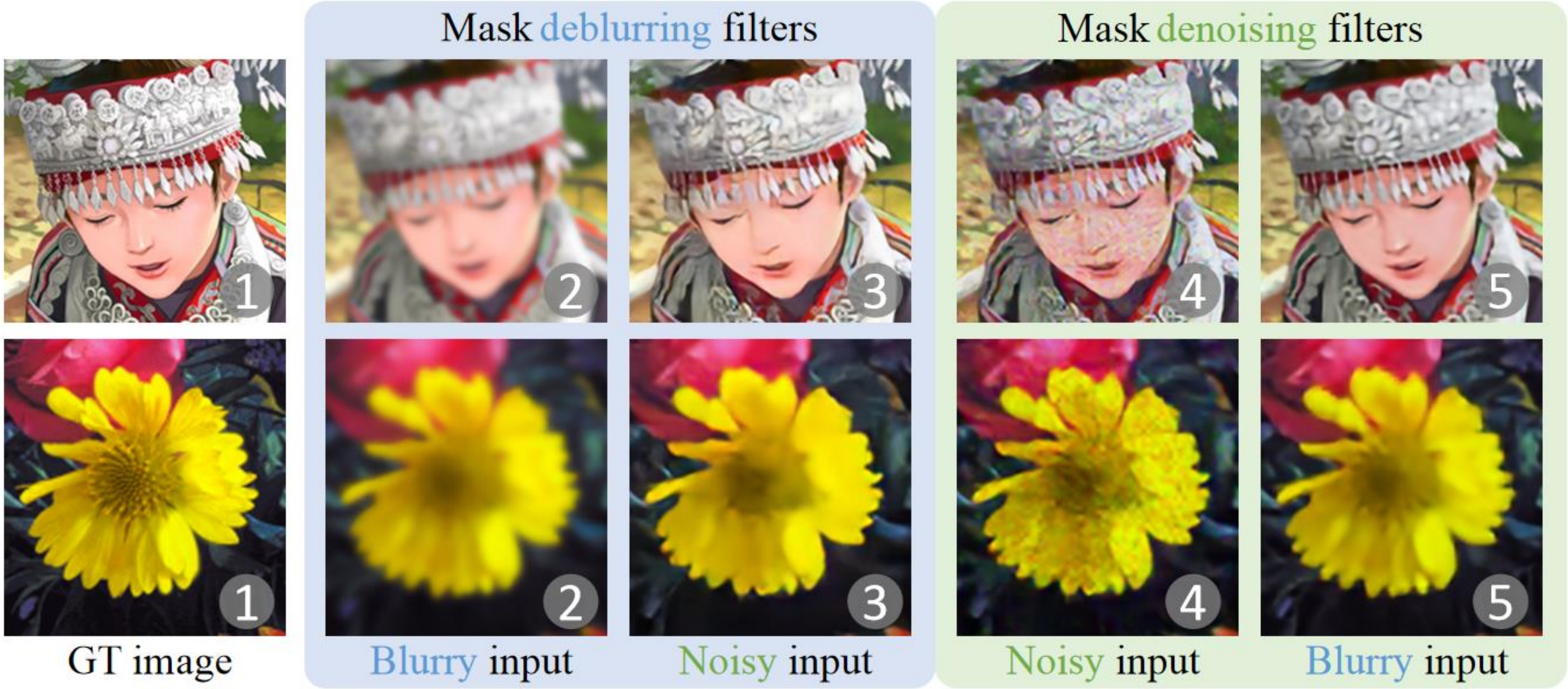
Masking Discovered Filters

We measure the importance of discovered filters by replacing them with the filters in the baseline model (at the same locations).



Masking Discovered Filters

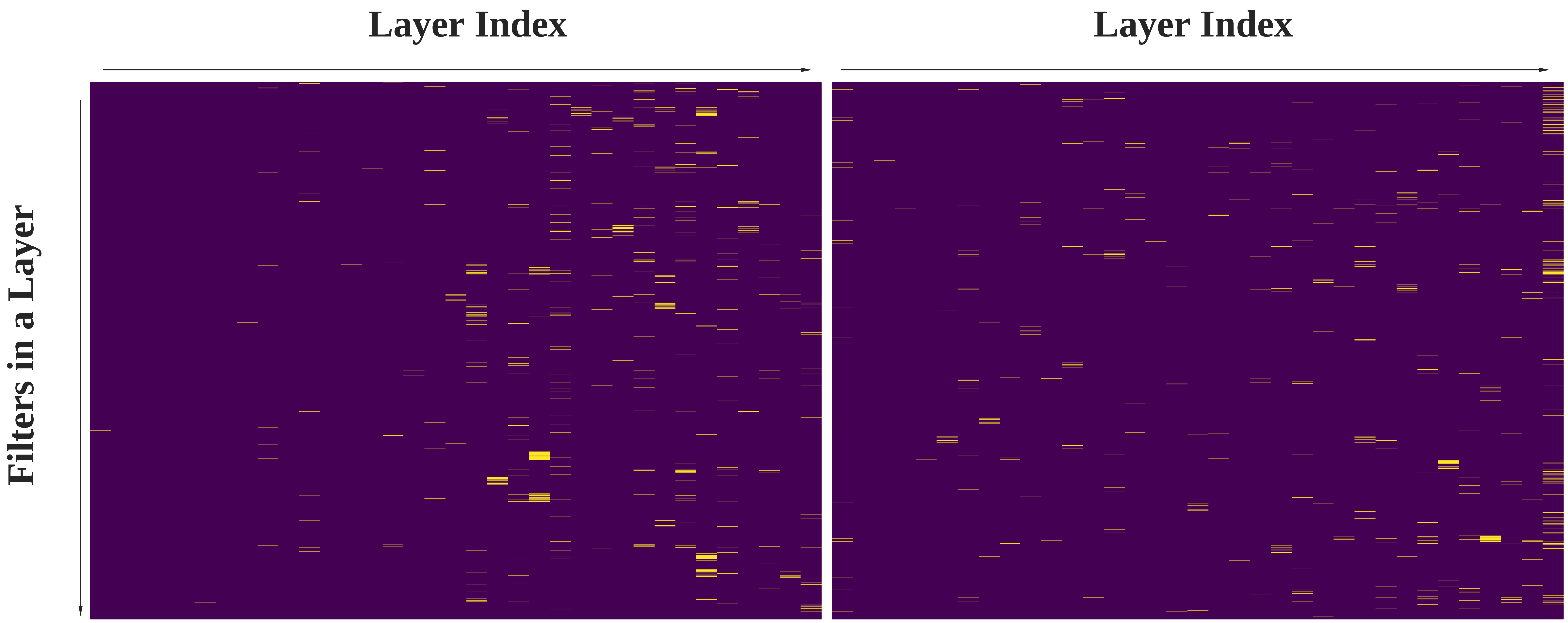
More qualitative results – Mask 1% filters



Masking Discovered Filters

More qualitative results – Mask 5% filters

Distribution of Discovered Filters



Discovered Filters for Deblurring

Discovered Filters for Denoising

The deblurring filters are more located in the back part while denoising filters locate more uniformly.

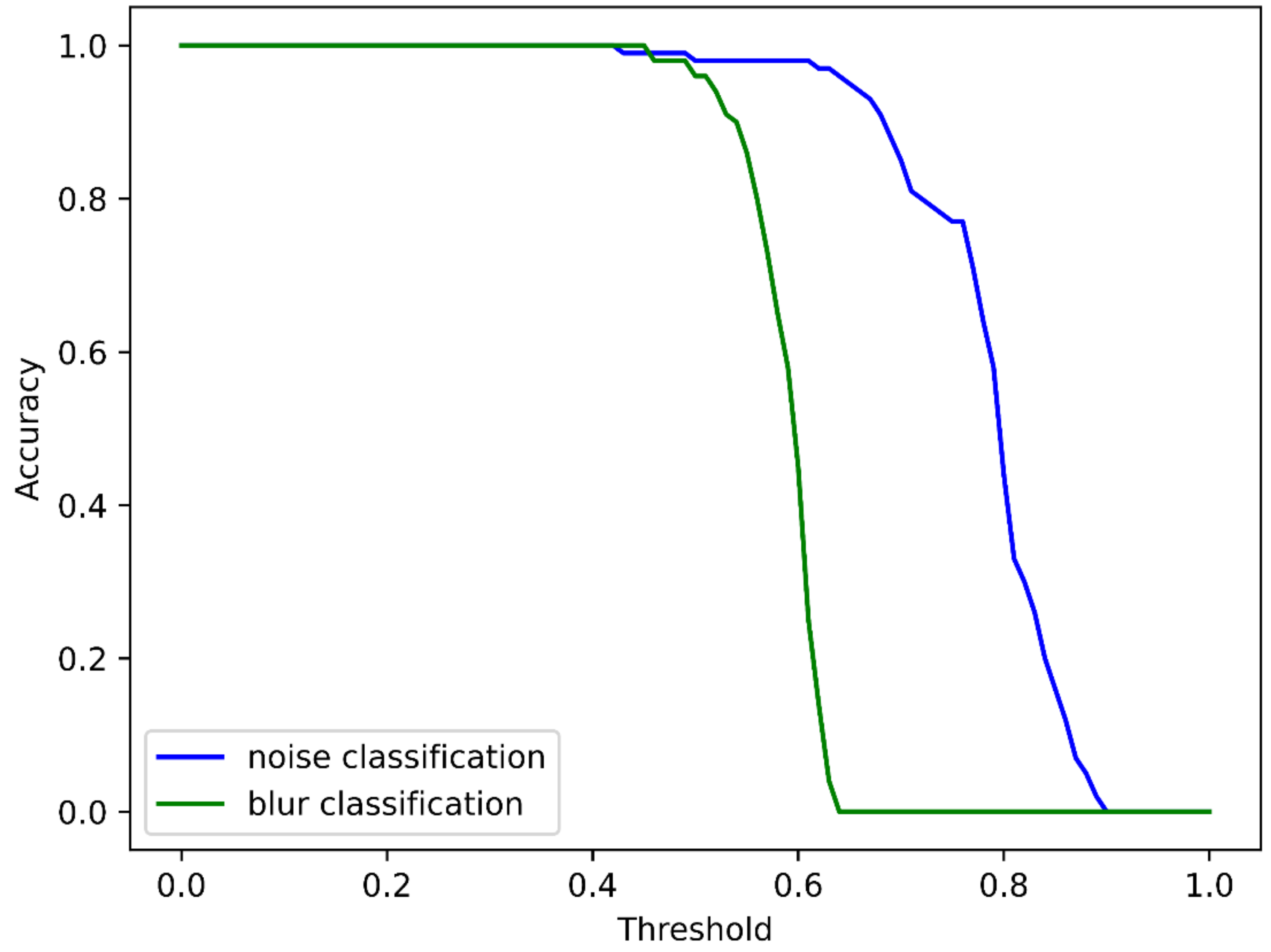
Application — Degradation Classification

Predict the degradation of input images without training in the supervision of degradation labels.

we calculate the overlap score (OS) to measure the intersection of the two sets of filters:

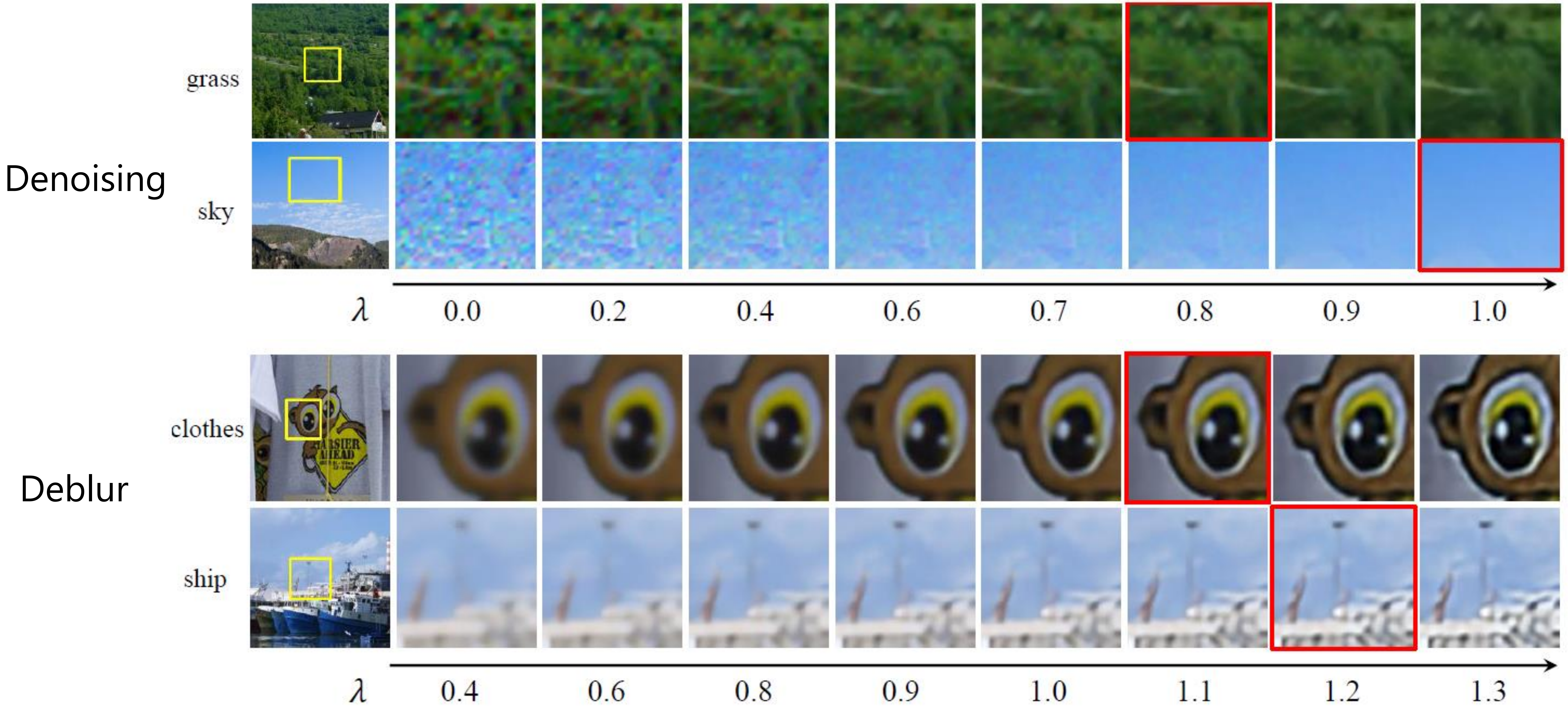
$$\mathrm{OS}(x, \mathcal{D}) = \frac{|\{\texttt{filter}^{\mathcal{D}}\} \cap \{\texttt{filter}^{x}\}|}{|\{\texttt{filter}^{x}\}|}$$

By setting the thresholds: T^noise and T^blur to 0.6 and 0.5, the prediction accuracy can reach 98% and 96%.



Application — Controllable restoration

Interpolate the corresponding parameters (at the same location)

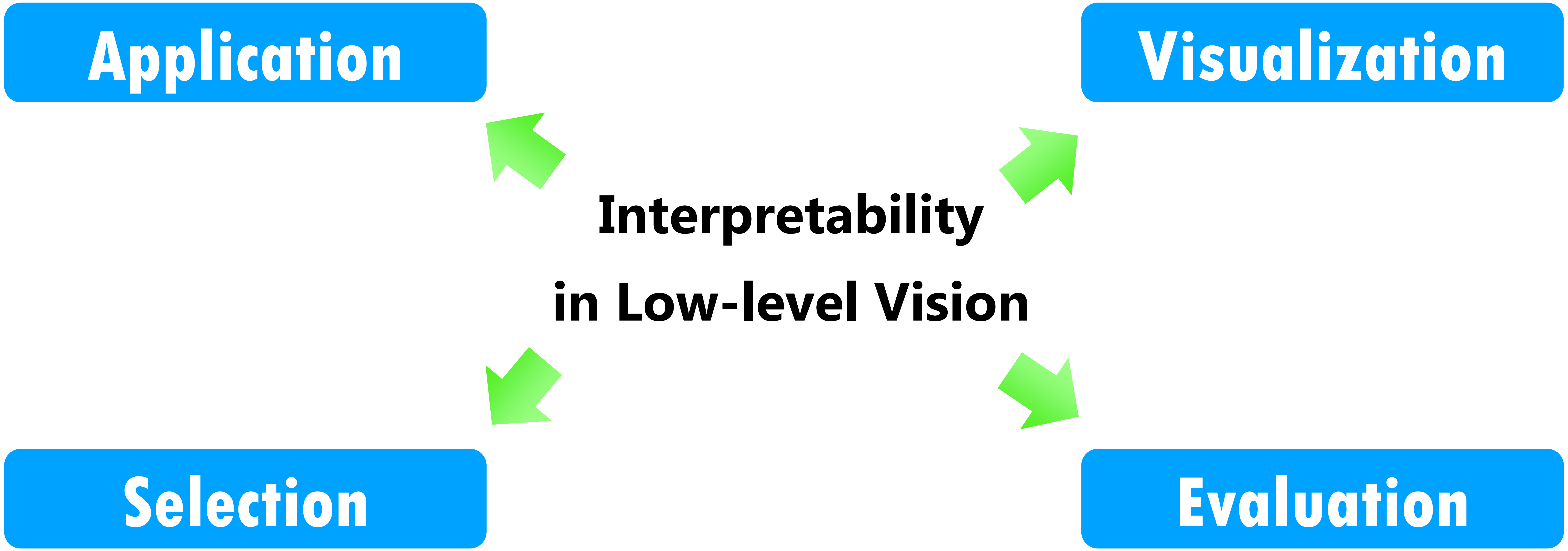


Interpreting Super-Resolution Networks

Interpretability in Low-level Vision

- **Pixel:** What pixels contribute most to restoration?
- **Feature:** Where can we find semantics in SR-net?
- **Filters:** Whether learned filters are discriminative?

Future Work


Application
Visualization
Interpretability
in Low-level Vision
Selection
Evaluation

Thanks

Interpretable
Low-Level Vision

X-Pixel Group

BasicSR