ICCV 2019

Evaluating Robustness of Deep Image Super-Resolution Against Adversarial Attacks

Jun-Ho Choi, Huan Zhang, Jun-Hyuk Kim, Cho-Jui Hsieh, and Jong-Seok Lee

Yongsei University, Korea; UCLA, US

Slides compiled by Mengzhou Li

Introduction

- Deep learning-based single-image Super-Resolution (SR) is one of the popular research areas in recent years.
- Various investigations report the vulnerability of deep networks under intended attacks in the classification tasks.

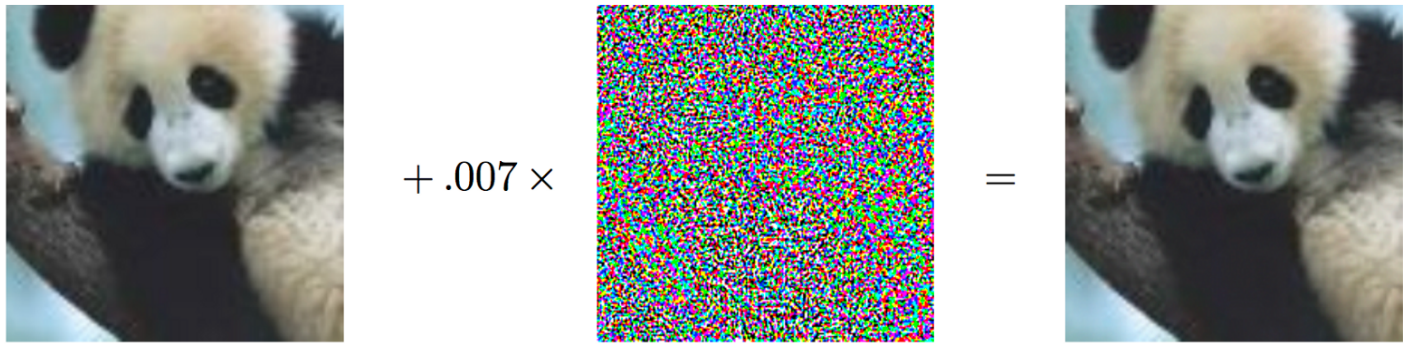
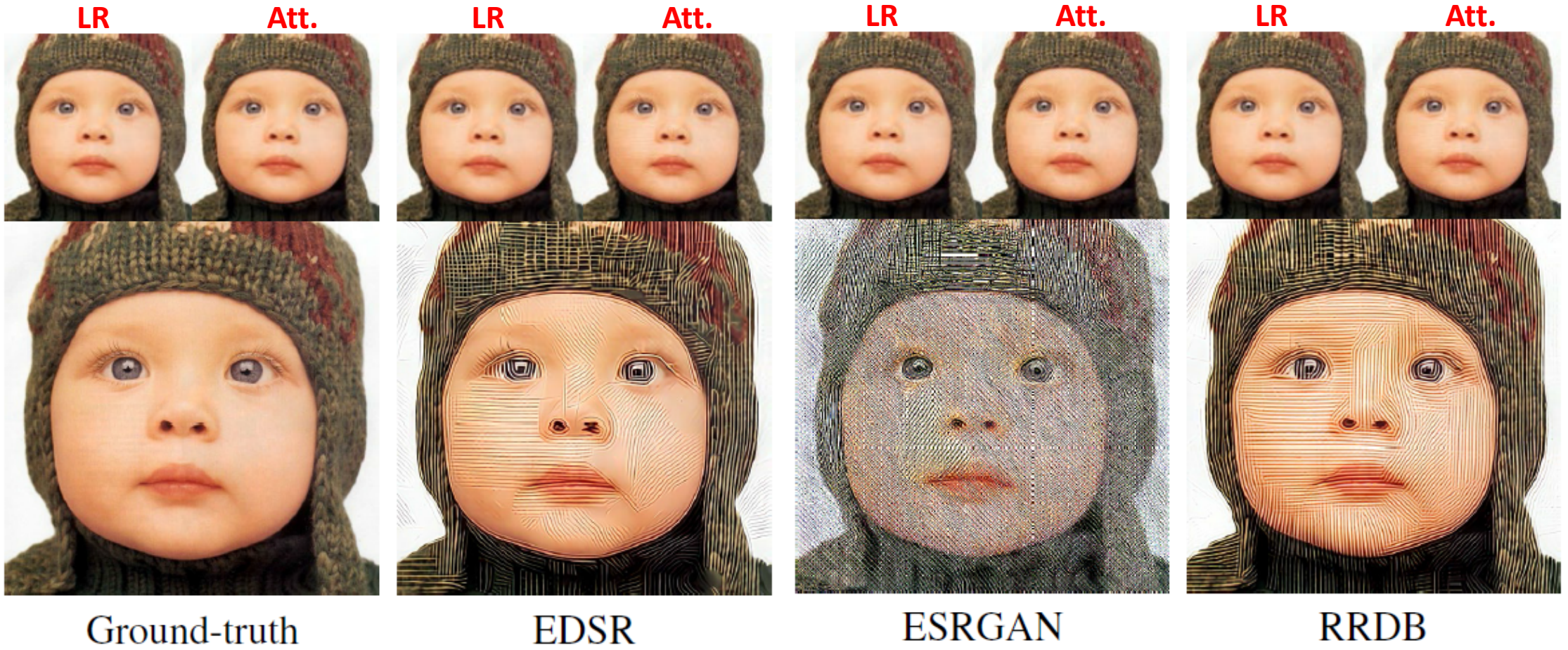


Figure from: Explaining and Harnessing Adversarial Examples, Goodfellow et al, ICLR 2015

Introduction

- Similar issue can be raised for SR applications.



Contributions

- Propose three adversarial attack methods for SR
- Present thorough analysis of the robustness of the SR methods with experimental results
- Investigate the relation of robustness to the model properties and measure the transferability

Attacks on SR

The goal is to inject a small amount of perturbation in the input images so that the quality of outputs is largely deteriorated.

- Basic attacks
- Universal attacks
- Partial attacks

Attacks on SR

The goal is to inject a small amount of perturbation in the input images so that the quality of outputs is largely deteriorated.

- Basic attack (Image specific)

  Based on the Iterative Fast Gradient Sign method (I-FGSM) which is widely used strong attacks for classification models.

  $\boldsymbol{X}_0$: the original low-resolution input; $\boldsymbol{X}$: the attacked version; $\boldsymbol{f}$: trained SR model

  Objective: maximize $\mathcal{L}(\mathbf{X}, \mathbf{X}_0) = ||f(\mathbf{X}) - f(\mathbf{X}_0)||_2$ Maximize deterioration

  S.T. $\|\mathbf{X} - \mathbf{X}_0\|_\infty \leq \alpha$ Constrain perturbation

Attacks on SR

- Basic attack (Image specific perturbation)

  $\boldsymbol{X}_0$: the original low-resolution input; $\boldsymbol{X}$: the attacked version; $\boldsymbol{f}$: trained SR model

  Objective: maximize $\mathcal{L}(\mathbf{X}, \mathbf{X}_0) = ||f(\mathbf{X}) - f(\mathbf{X}_0)||_2$

  S.T. $\|\mathbf{X} - \mathbf{X}_0\|_\infty \leq \alpha$

$$\widetilde{\mathbf{X}}_{n+1} = \mathrm{clip}_{0,1}\Big(\mathbf{X}_n + \frac{\alpha}{T}\,\mathrm{sgn}\big(\nabla\mathcal{L}(\mathbf{X}_n, \mathbf{X}_0)\big)\Big)$$ Similar to BP

$$\mathbf{X}_{n+1} = \mathrm{clip}_{-\alpha,\alpha}(\widetilde{\mathbf{X}}_{n+1} - \mathbf{X}_0) + \mathbf{X}_0$$ Inf norm constraint on perturbation

$$\mathrm{clip}_{a,b}(\mathbf{X}) = \min\big(\max(\mathbf{X}, a), b\big).$$ Clip X to the range (a, b)

Final adversarial example is $\bar{\mathbf{X}} = \mathbf{X}_T$

Attacks on SR

- Universal attack (Image agnostic perturbation for K images)

  $\boldsymbol{X}_0^k$: the k-th image; $\boldsymbol{f}$: trained SR model; Δ:universal perturbation;

  $\boldsymbol{X}^k$: $\mathbf{X}^k = \mathrm{clip}_{0,1}(\mathbf{X}_0^k + \Delta)$

  Objective: maximize $\mathcal{F}(\Delta) = \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}(\mathbf{X}^k, \mathbf{X}_0^k)$ S.T. $\|\Delta\|_\infty \leq \alpha$

  $\Delta_0 = 0$

  $\Delta_{n+1} = \mathrm{clip}_{-\alpha,\alpha}\left(\Delta_n + \frac{\alpha}{T}\,\mathrm{sgn}\left(\nabla\mathcal{F}(\Delta_n)\right)\right)$

  The final universal perturbation is $\Delta = \Delta_T$

Attacks on SR

- Partial attack

  attack a region, measure the amount of deterioration outside the attacked region, hence to examine the perturbation permeates into the adjacent regions during SR.

  M*:** binary mask of the perturbation, 1 for attacked region; ***X: the attacked version;

  $\boldsymbol{X}_0$: the original low-resolution input; ***f***: trained SR model; $\mathbf{M}_H$: the high-resolution counter part of ***M***;

  Objective: maximize $\mathcal{L}_{\mathrm{M}}(\mathbf{X}, \mathbf{X}_0) = \left|\left| \left(f(\mathbf{X}) - f(\mathbf{X}_0)\right) \circ (1 - \mathbf{M}_H) \right|\right|_2$

  S.T. $\|\Delta \circ \mathbf{M}\|_\infty \leq \alpha$ and $\Delta \circ (1 - \mathbf{M}) = 0$

$$\widetilde{\mathbf{X}}_{n+1} = \mathrm{clip}_{0,1}\left(\mathbf{X}_n + \frac{\alpha}{T}\,\mathrm{sgn}\left(\nabla \mathcal{L}_{\mathrm{M}}(\mathbf{X}_n, \mathbf{X}_0)\right) \circ \mathbf{M}\right)$$

$$\mathbf{X}_{n+1} = \mathrm{clip}_{-\alpha,\alpha}(\widetilde{\mathbf{X}}_{n+1} - \mathbf{X}_0) + \mathbf{X}_0$$

The final universal perturbation is $\bar{\mathbf{X}} = \mathbf{X}_T$

Experiments

- Architectures (scale factor of 4, original pretrained models)

Method	# parameters	# layers	GAN-based
EDSR [12]	43.1M	69	-
EDSR-baseline [12]	1.5M	37	-
RCAN [25]	15.6M	815	-
4PP-EUSR [6]	6.3M	95	✓
ESRGAN [19]	16.7M	351	✓
RRDB [19]	16.7M	351	-
CARN [3]	1.1M	34	-
CARN-M [3]	0.3M	43	-

Table 1. Properties of the super-resolution methods.

- Enhanced Deep Super-Resolution (EDSR)
- EDSR-baseline
- Residual Channel Attention Network (RCAN)
- Four-Pass Perceptual Enhanced Upscaling Super-Resolution (4PP-EUSR)
- Enhanced Super-Resolution Generative Adversarial Network (ESRGAN)
- Residual in Residual Dense Block (RRDB)
- Cascading Residual Network (CARN)
- CARN-M

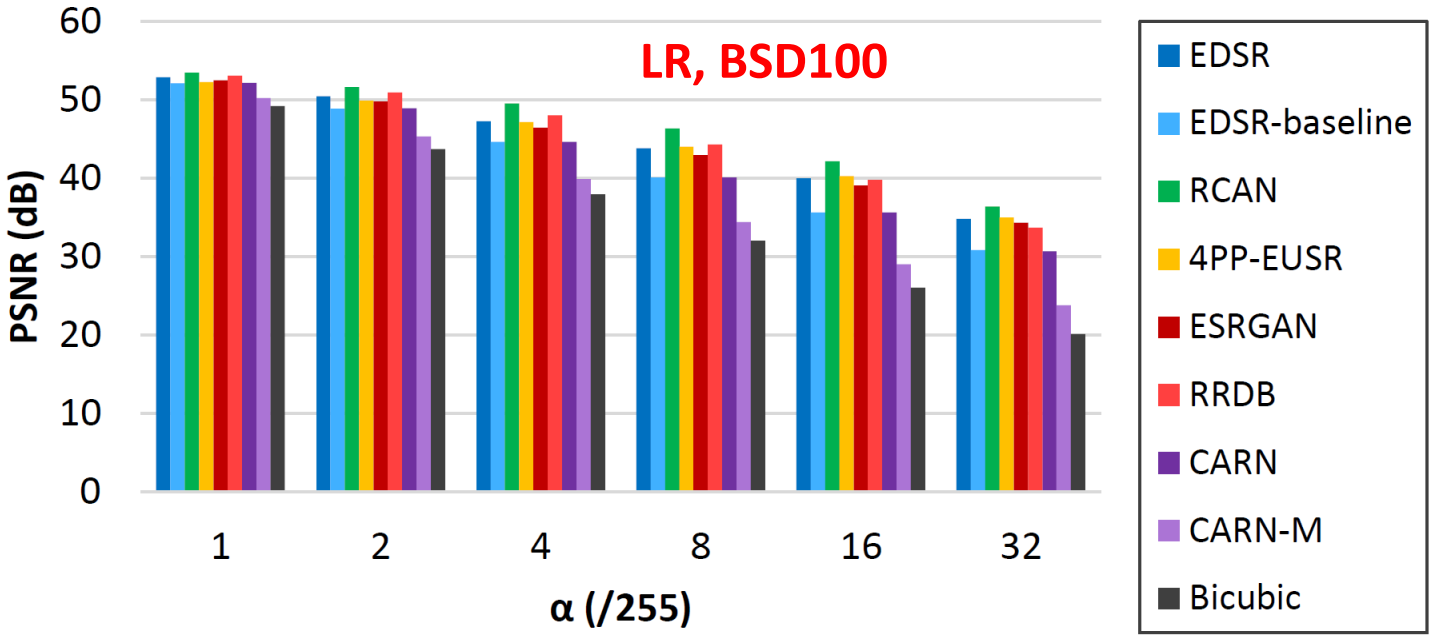
Experiments

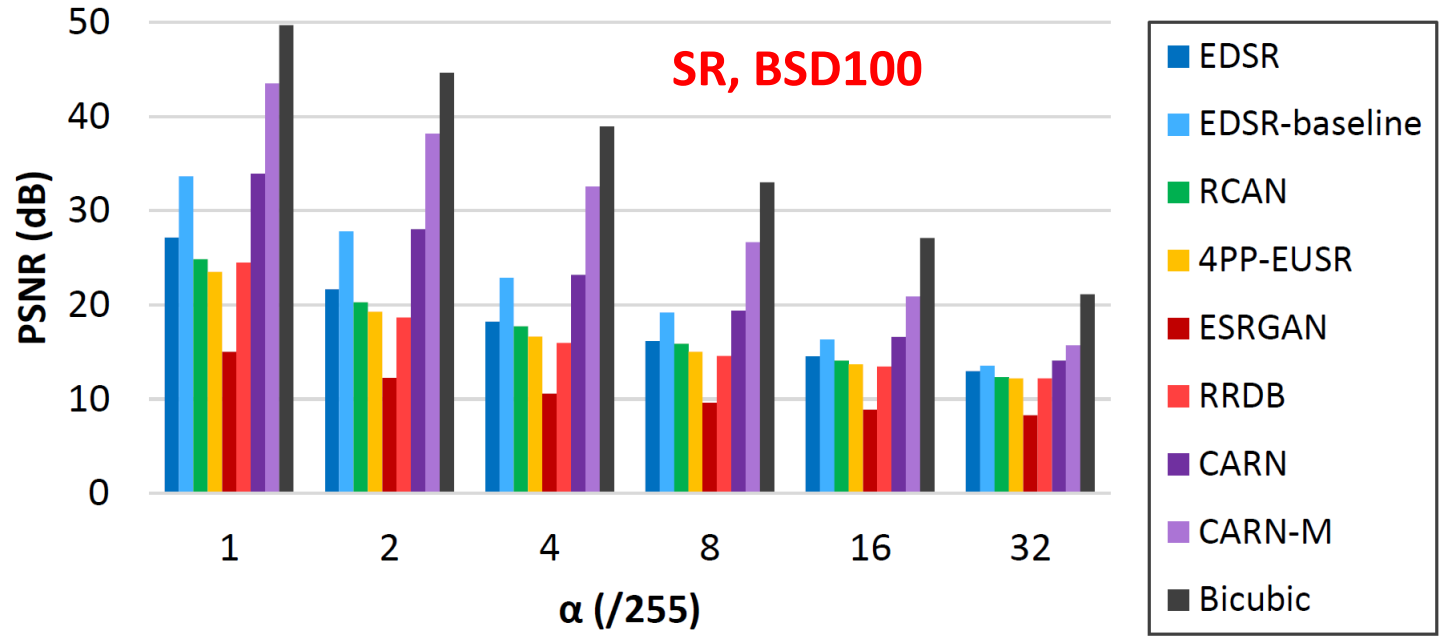
- Datasets
  - Three image datasets: Set5, Set14, and BSD100. Each dataset consists of 5, 14, and 100 images, respectively.
- Performance measurement
  - Measure the robustness in terms of PSNR
    - LR: original image vs attacked image
    - SR: original output vs attacked output
  - Report averaged PSNR values for each dataset.

Basic attack

PSNR>30db => taken as two visually identical images

- As α increases, degradation becomes severe in both the LR and SR images;
- Degradation more significant in SR than LR except for the bicubic interpolation;
- ESRGAN lowest score, Bicubic most robust;
- All the deep learning-based SR methods are highly vulnerable against the adversarial attack.





$\alpha = 8/255$

Att. ~= LR.

Fingerprint-like artifacts

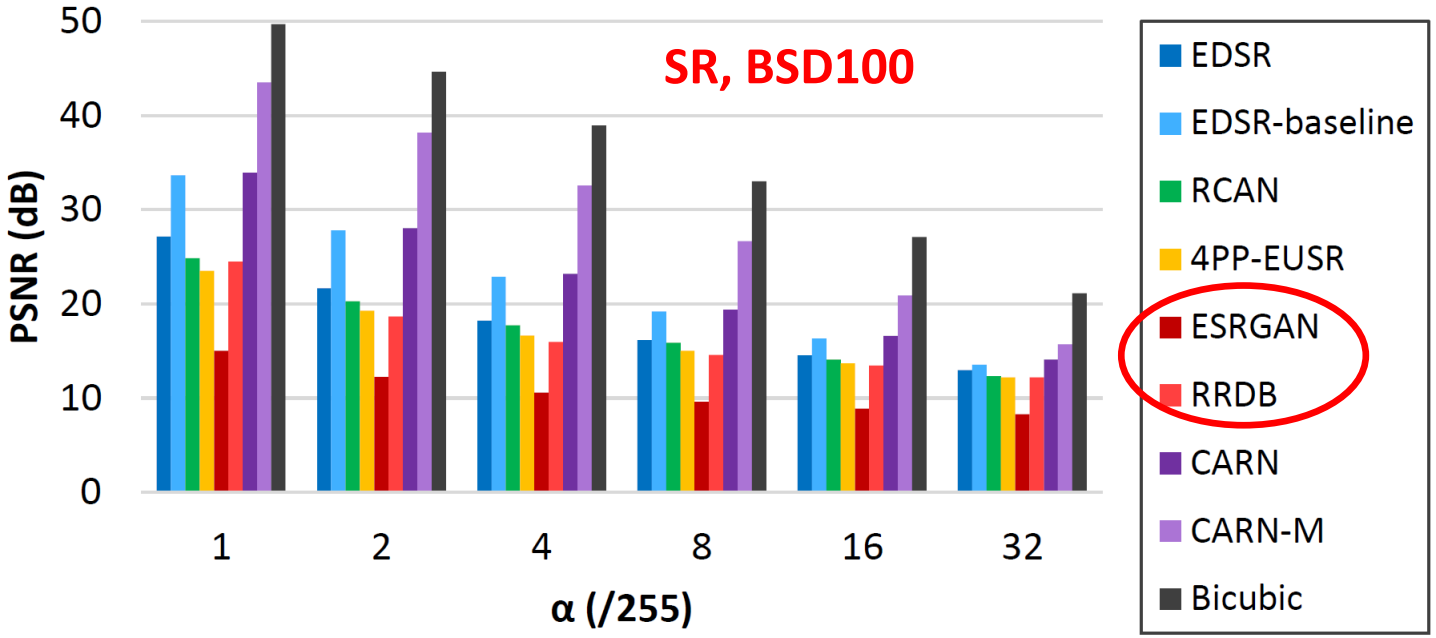
ESRGAN
4PP-EUSR
(the two with GANs)
Worst

Bicubic
RCAN-M
Best

Basic attack

- **Relation to model objectives:**
  - ESRGAN = RRDB + GAN
  - 4PP-EUSR & ESRGAN

  => More vulnerable with GANs



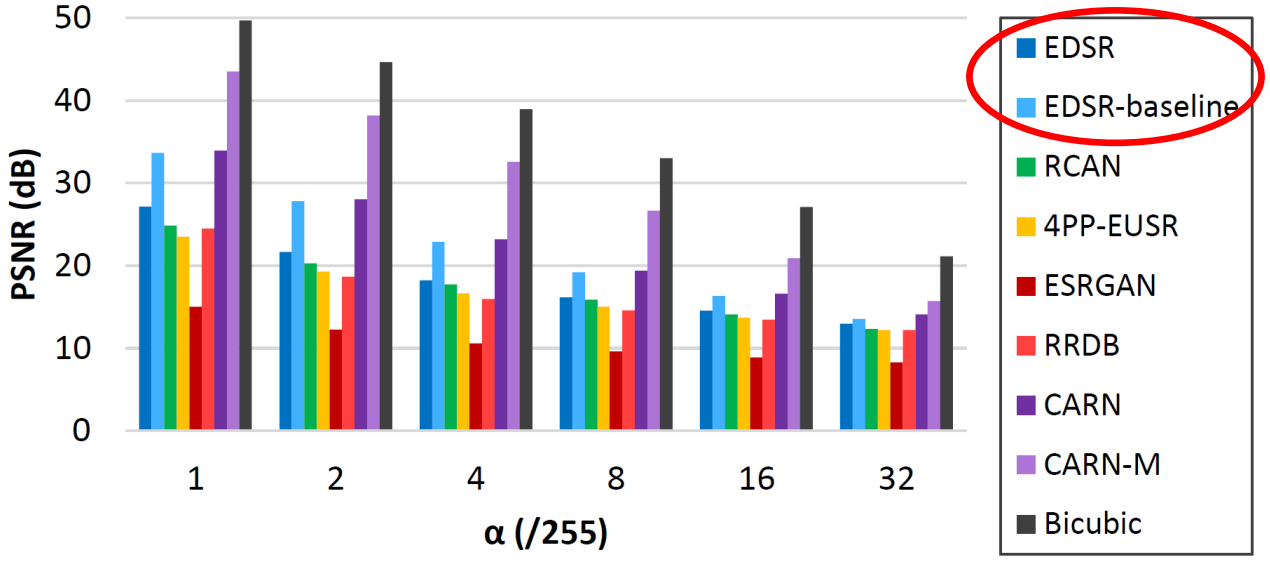
RRDB ESRGAN 4PP-EUSR

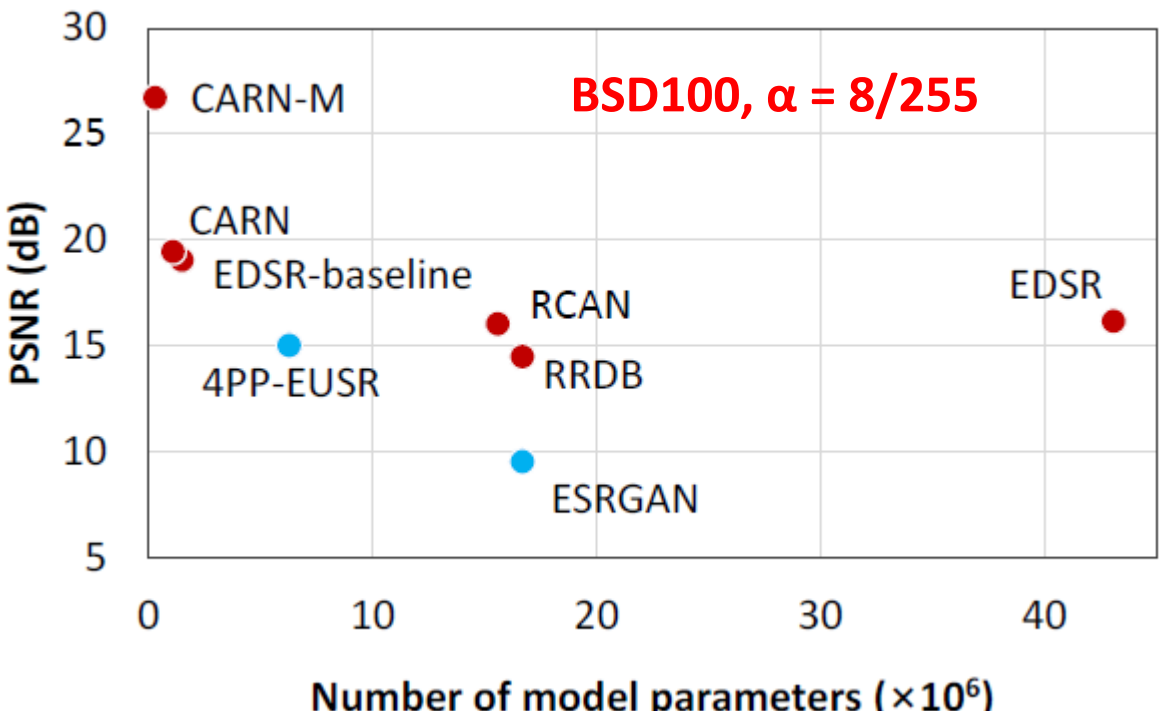
Method	# parameters	# layers	GAN-based
EDSR [12]	43.1M	69	-
EDSR-baseline [12]	1.5M	37	-
RCAN [25]	15.6M	815	-
4PP-EUSR [6]	6.3M	95	✓
ESRGAN [19]	16.7M	351	✓
RRDB [19]	16.7M	351	-
CARN [3]	1.1M	34	-
CARN-M [3]	0.3M	43	-

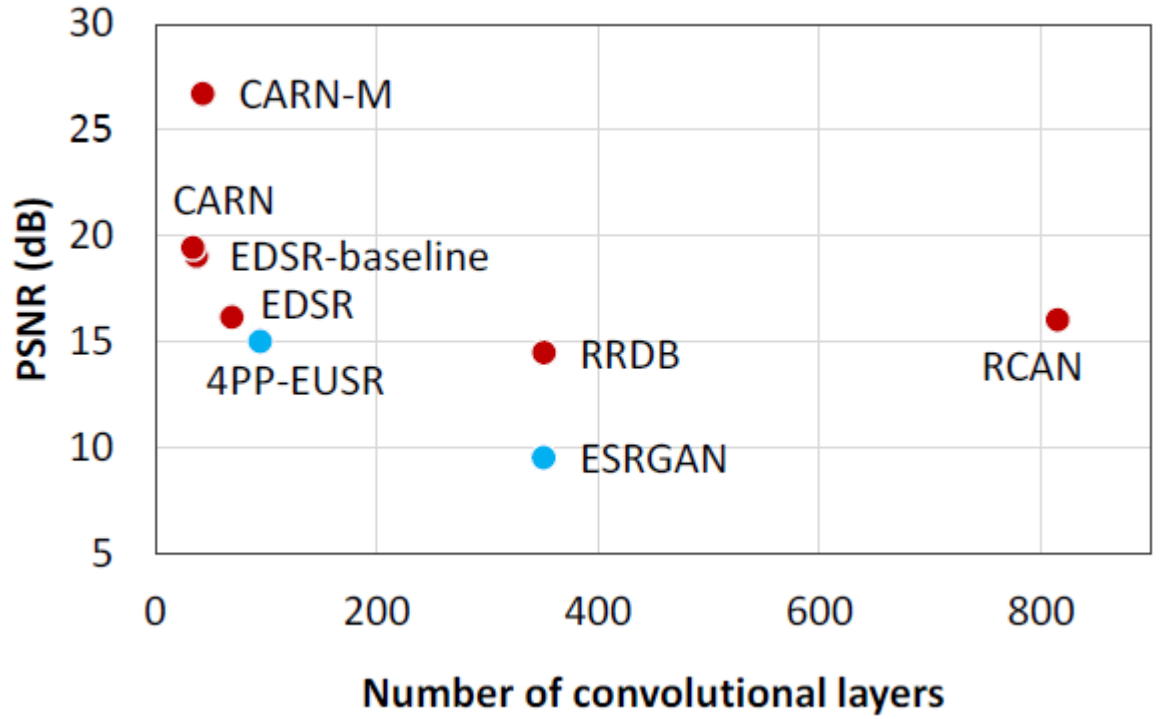
Table 1. Properties of the super-resolution methods.

Basic attack

- **Relation to model sizes:**
  EDSR-baseline = smaller version of EDSR
  PSNR negatively correlates with # parameters and # conv layers
  ⇒ More parameters or conv layers, more vulnerable







Basic attack

PSNR (dB)		Target model: EDSR	EDSR-baseline	RCAN	4PP-EUSR	ESRGAN	RRDB	CARN	CARN-M
Source model	EDSR	16.14	32.88	25.82	23.86	16.26	23.57	32.80	37.74
	EDSR-baseline	24.44	19.19	23.65	21.23	15.15	22.29	26.82	33.62
	RCAN	30.57	35.49	15.89	26.60	18.94	29.74	35.57	40.46
	4PP-EUSR	27.25	32.76	26.83	15.02	16.16	24.71	32.87	37.97
	ESRGAN	28.64	33.11	28.59	24.28	9.57	24.46	33.30	36.56
	RRDB	25.55	33.09	25.31	23.77	15.86	14.59	32.91	38.11
	CARN	24.12	26.05	23.83	21.45	15.24	22.15	19.40	33.51
	CARN-M	27.34	28.20	27.20	23.49	16.27	26.77	28.20	26.66

Figure 4. Comparison of the transferability in terms of PSNR for the BSD100 dataset [14] when $\alpha = 8/255$. Red and blue colors indicate the lowest and highest PSNR values (except the diagonal cells) for each target model, respectively.

- **Transferability:**
  To check whether att. images generated with a source model work on another target model

Everything works on ESRGAN
Nothing works on CARN, CARN-M

Basic attack

PSNR (dB)	Target model: EDSR	EDSR-baseline	RCAN	4PP-EUSR	ESRGAN	RRDB	CARN	CARN-M
Source model: EDSR	16.14	32.88	25.82	23.86	16.26	23.57	32.80	37.74
EDSR-baseline	24.44	19.19	23.65	21.23	15.15	22.29	26.82	33.62
RCAN	30.57	35.49	15.89	26.60	18.94	29.74	35.57	40.46
4PP-EUSR	27.25	32.76	26.83	15.02	16.16	24.71	32.87	37.97
ESRGAN	28.64	33.11	28.59	24.28	9.57	24.46	33.30	36.56
RRDB	25.55	33.09	25.31	23.77	15.86	14.59	32.91	38.11
CARN	24.12	26.05	23.83	21.45	15.24	22.15	19.40	33.51
CARN-M	27.34	28.20	27.20	23.49	16.27	26.77	28.20	26.66

Figure 4. Comparison of the transferability in terms of PSNR for the BSD100 dataset [14] when $\alpha = 8/255$. Red and blue colors indicate the lowest and highest PSNR values (except the diagonal cells) for each target model, respectively.

- **Transferability:**
  To check whether att. images generated with a source model work on another target model

Everything works on ESRGAN
Nothing works on CARN and CARN-M

Generated for EDSR-baseline and CARN work on majority of others
Those for RCAN do not work on others

CARN and EDSR-baseline are highly transferable.

Basic attack

PSNR (dB)	Target model: EDSR	EDSR-baseline	RCAN	4PP-EUSR	ESRGAN	RRDB	CARN	CARN-M
Source model: EDSR	16.14	32.88	25.82	23.86	16.26	23.57	32.80	37.74
EDSR-baseline	24.44	19.19	23.65	21.23	15.15	22.29	26.82	33.62
RCAN	30.57	35.49	15.89	26.60	18.94	29.74	35.57	40.46
4PP-EUSR	27.25	32.76	26.83	15.02	16.16	24.71	32.87	37.97
ESRGAN	28.64	33.11	28.59	24.28	9.57	24.46	33.30	36.56
RRDB	25.55	33.09	25.31	23.77	15.86	14.59	32.91	38.11
CARN	24.12	26.05	23.83	21.45	15.24	22.15	19.40	33.51
CARN-M	27.34	28.20	27.20	23.49	16.27	26.77	28.20	26.66

Figure 4. Comparison of the transferability in terms of PSNR for the BSD100 dataset [14] when $\alpha = 8/255$. Red and blue colors indicate the lowest and highest PSNR values (except the diagonal cells) for each target model, respectively.

- **Transferability:**
  To check whether att. images generated with a source model work on another target model

Everything works on ESRGAN
Nothing works on CARN and CARN-M

Generated for EDSR-baseline and CARN work on majority of others
Those for RCAN do not work on others

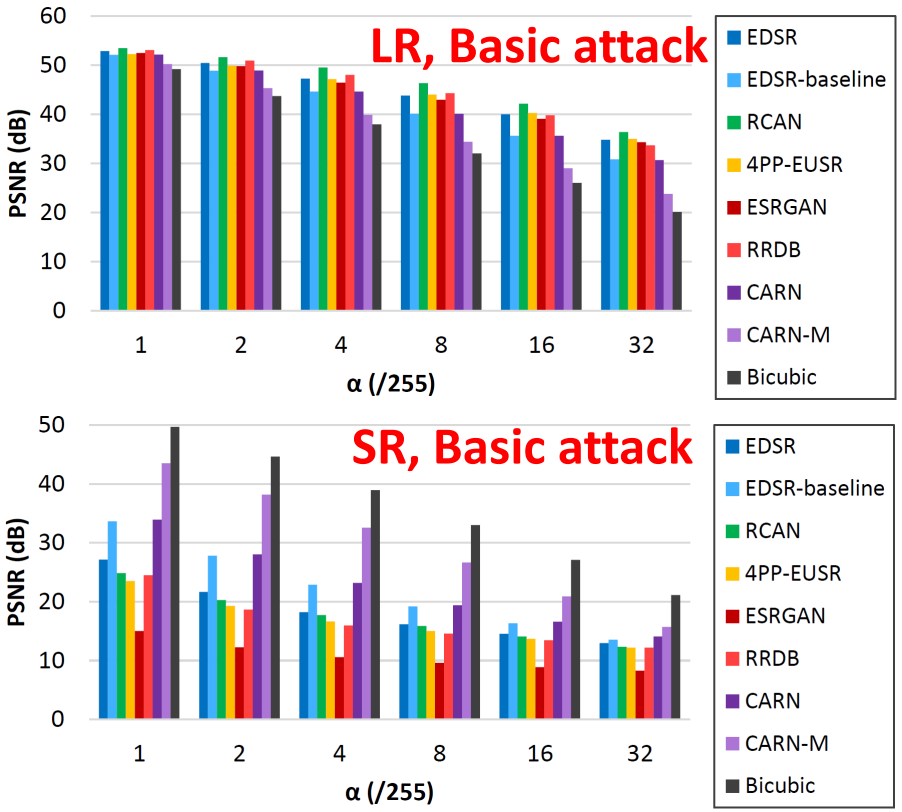
CARN and EDSR-baseline are highly transferable.

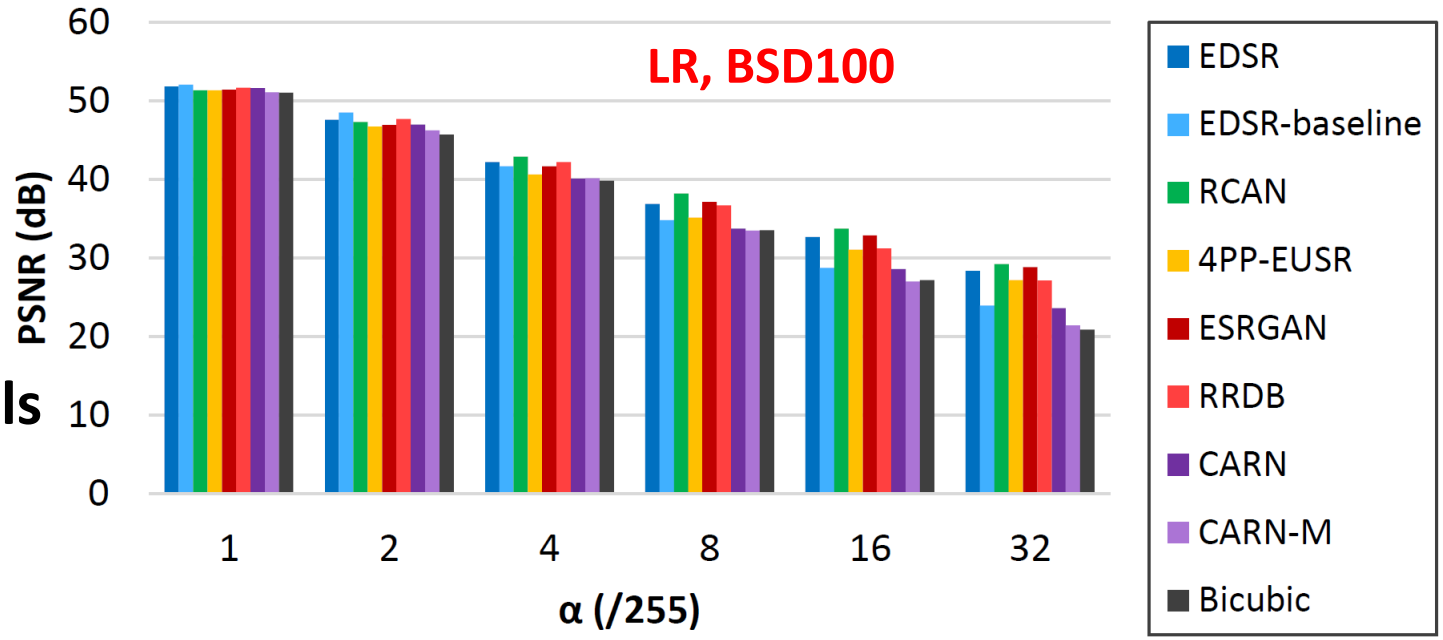
Hence, the level of transferability differs depending on the combination of the source and target models.

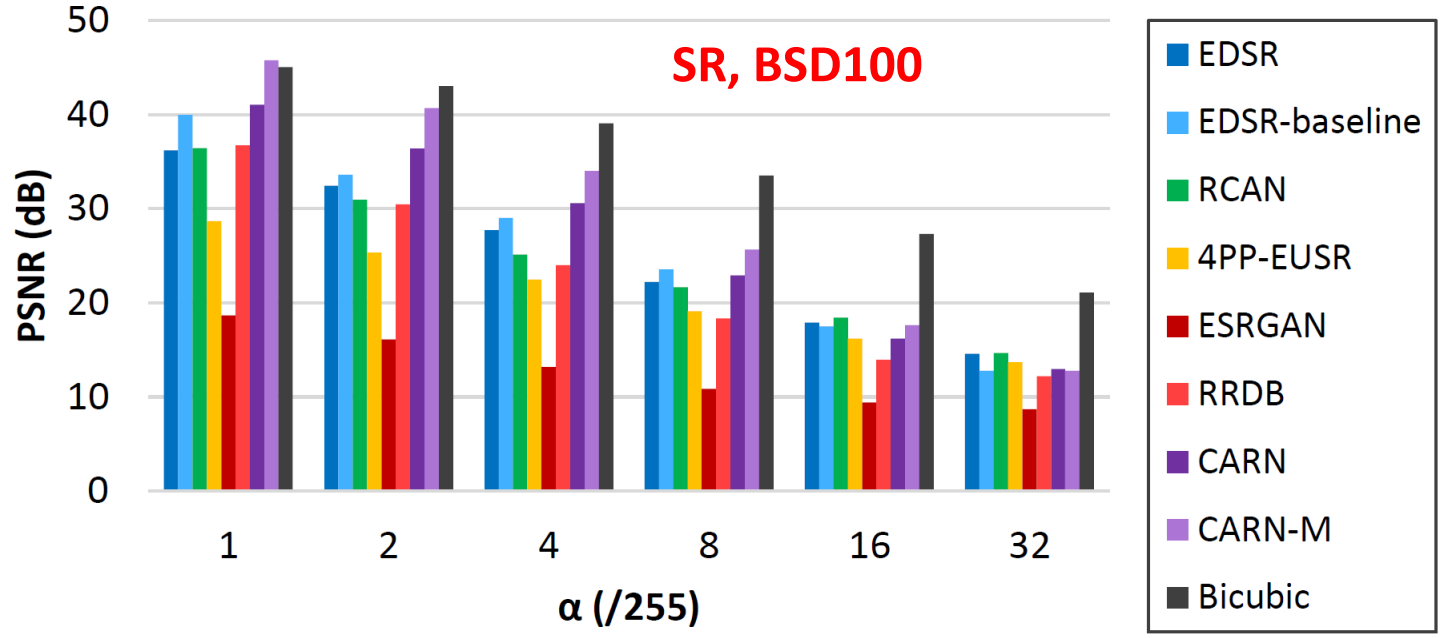
Universal attack

Compared to image-specific attack

- **Same vulnerability rank for models**
- **Requires larger perturbations**
- **Less powerful**







Universal attack

Compared to image-specific attack

- Same vulnerability rank for models
- Requires larger perturbations
- Less powerful

Significant artifacts appears in SR results with the same perturbation while attacked LR images demonstrate hardly noticeable differences from the original images.

Deep learning SR models also vulnerable to the universal attack.

LR, BSD100, α = 8/255

Perturbation

Attacked LR

Attacked SR

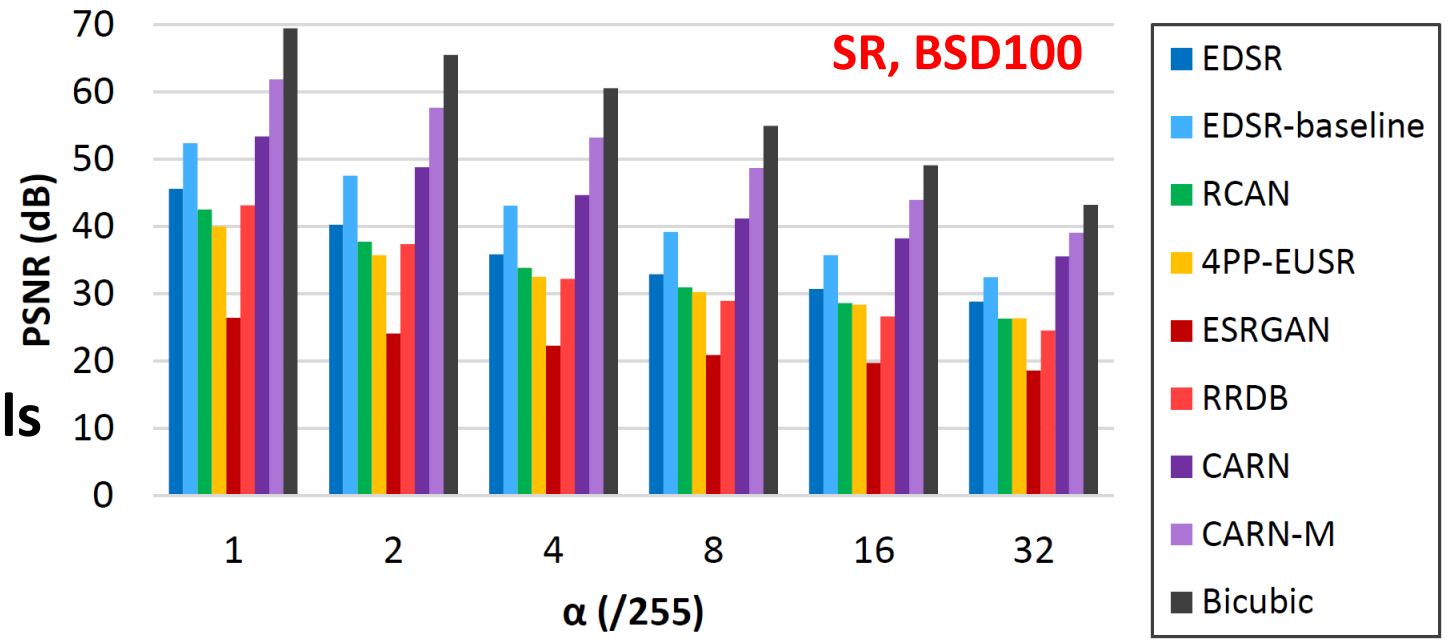
Attacked SR

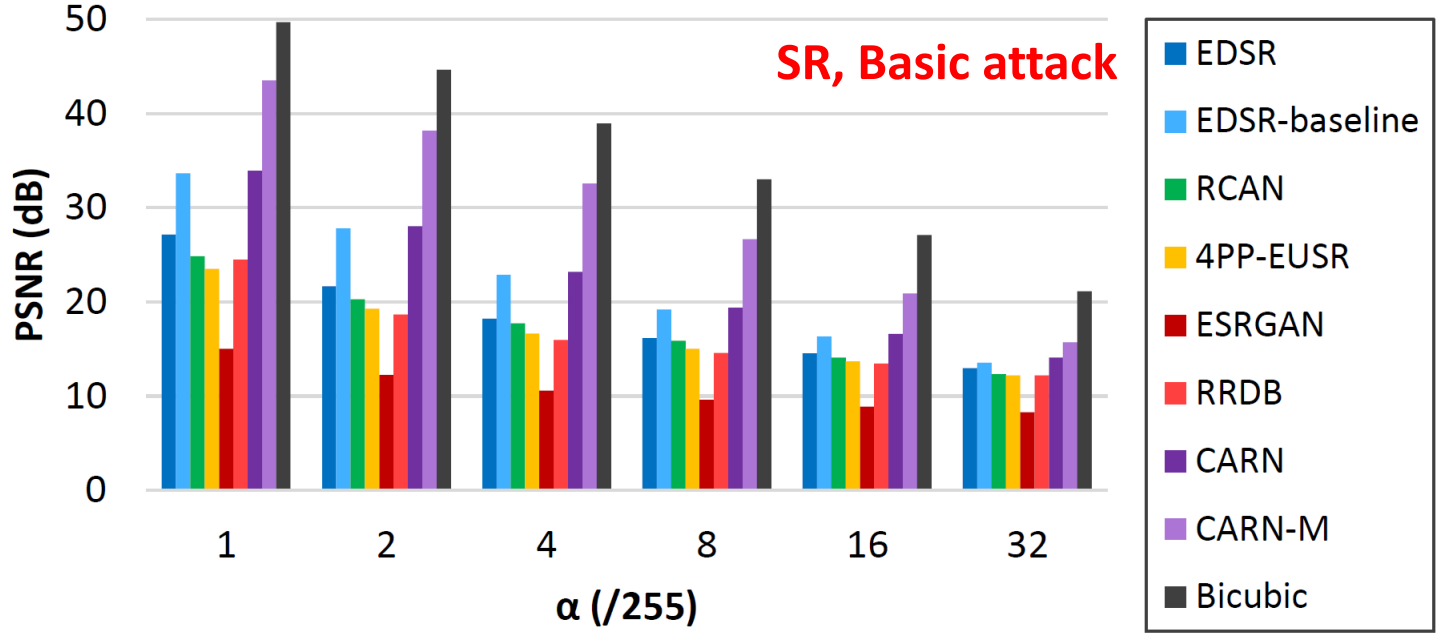
Attacked SR

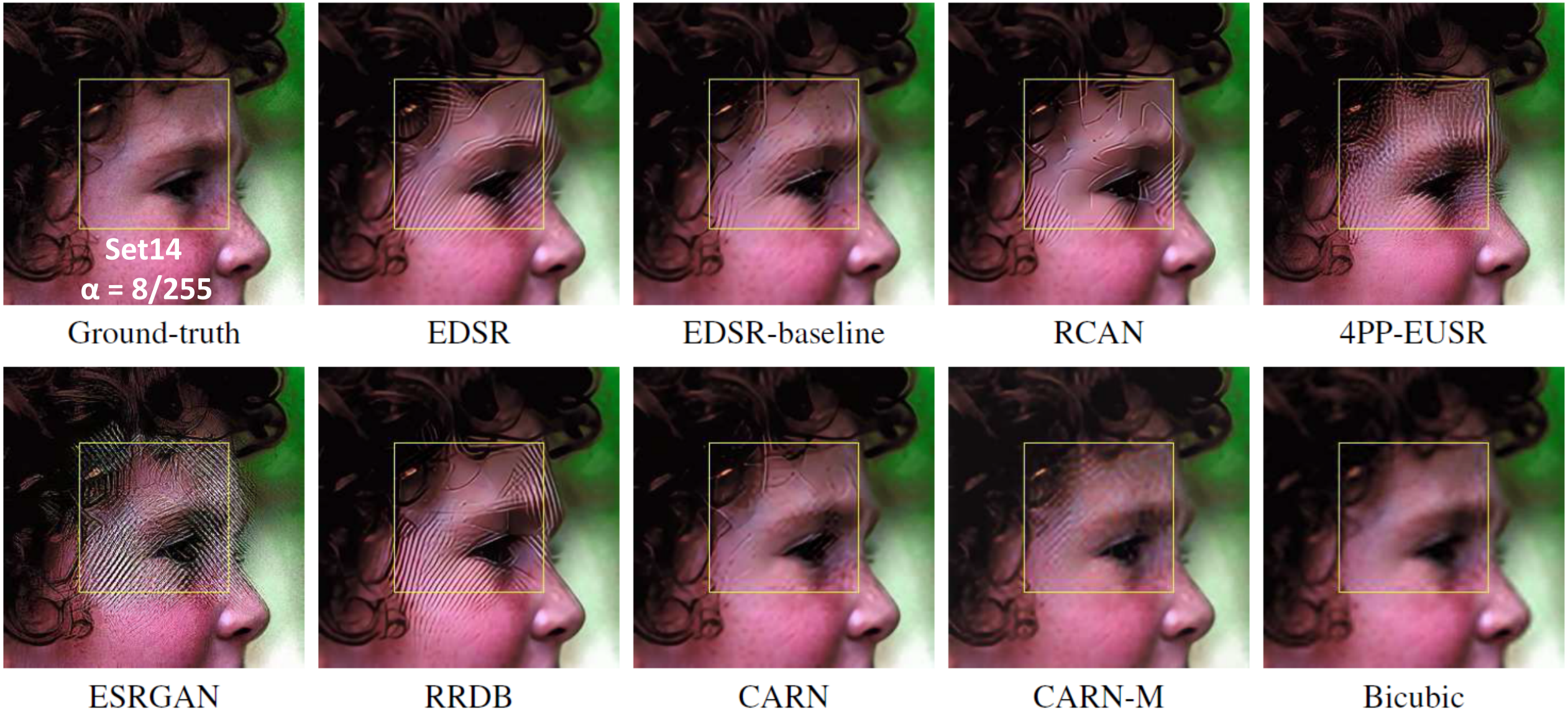
Partial attack

Compared to image-specific attack

- **Same vulnerability rank for models**
- **Higher PSNR in SR due to the special measurement method**







- **Attack leakage due to convolution kernels.**
- **ESRGAN and RRDB worst**

Related to the network structures.

Advanced Topics- Defense

Two simple defense methods against attacks:

- Resizing method by reducing the size of the attacked input image by one pixel of the attacked input image by one pixel and then resizing it back to the original resolution.
  PSNR for EDSR $\alpha = 8/255$, 16.14 => 25.01dB
- Geometric self-ensemble method:
  Geometric transformation augmentation (flip or rotate), then inverse transform the SR results of the augmentations, finally average to obtain the result.
  PSNR for EDSR $\alpha = 8/255$, 16.14 => 23.47dB

Take-homes

- GANs make the network more vulnerable
- Network structures have effects on vulnerability
- Generally more layers and more parameters tend to increase vulnerability
- Simple tricks can defense the attack by adding additional steps to corrupt the attacks.

Thanks for your attention!