

Computer Vision; Image Transformation; Super-Resolution, Denoising, and Colorization



[YouTube Playlist](#)

Maziar Raissi

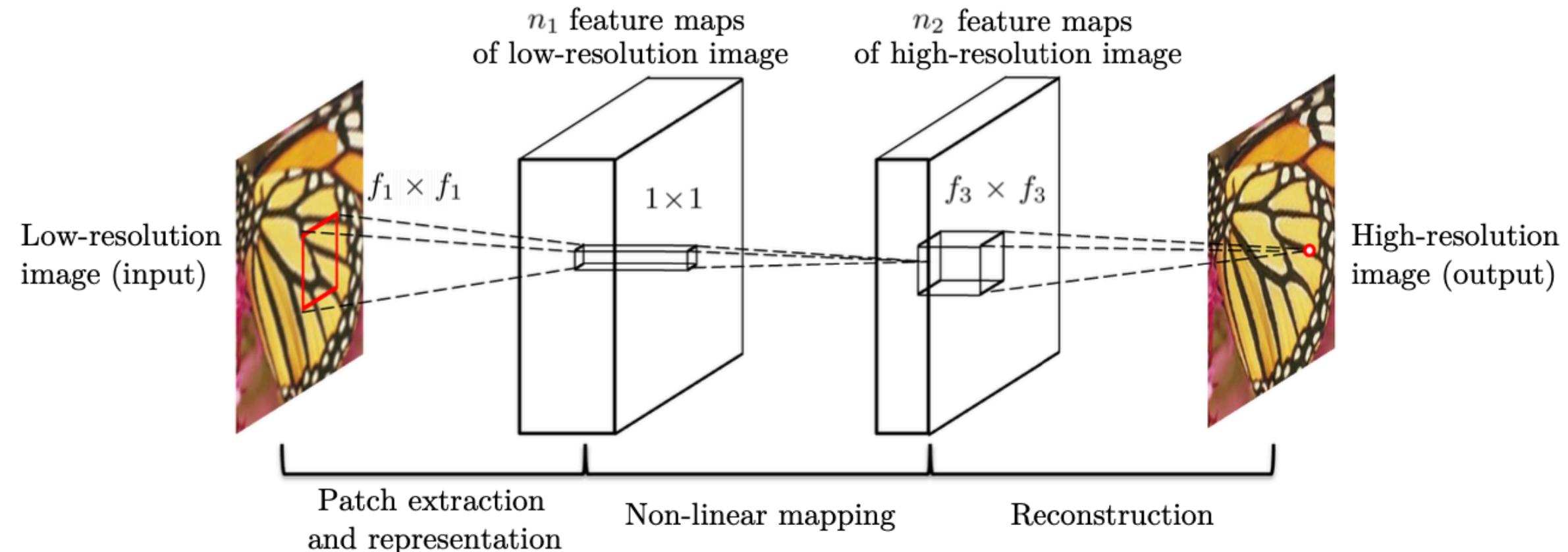
Assistant Professor

Department of Applied Mathematics

University of Colorado Boulder

maziar.raissi@colorado.edu

Learning a Deep Convolutional Network for Image Super-Resolution


[YouTube Video](#)


low resolution image \triangleright bicubic interpolation
 $Y \rightarrow$ interpolated image
 $F(Y) \rightarrow$ an image as similar as possible to X
 $X \rightarrow$ high resolution image

$$F_1(Y) = \max(0, W_1 * Y + B_1)$$

$$W_1 \in \mathbb{R}^{c \times f_1 \times f_1 \times n_1} \quad B_1 \in \mathbb{R}^{n_1}$$

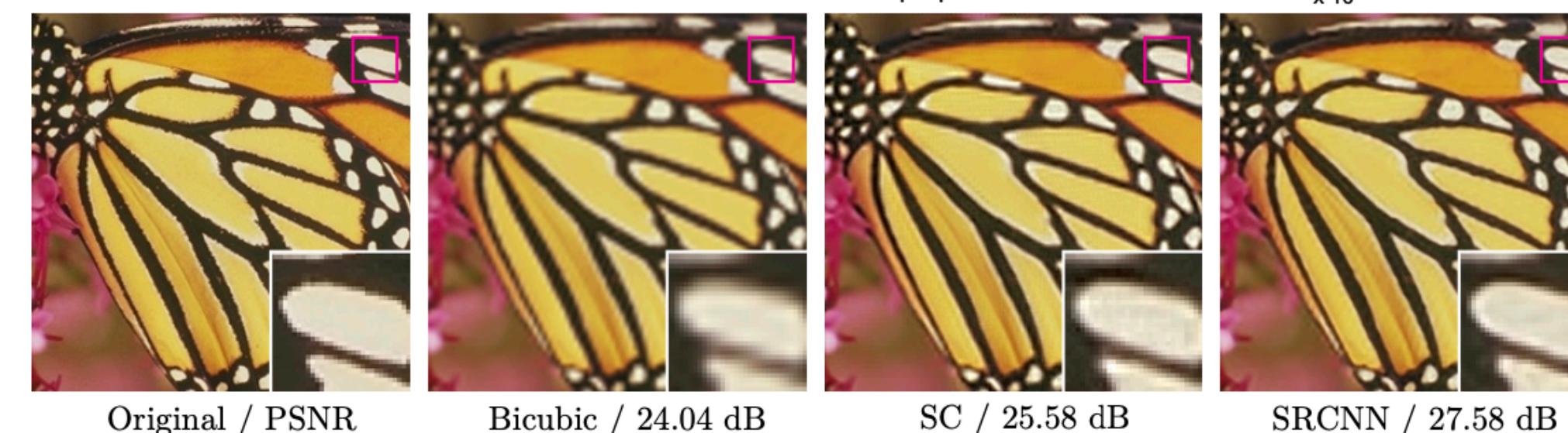
$$F_2(Y) = \max(0, W_2 * F_1(Y) + B_2)$$

$$W_2 \in \mathbb{R}^{n_1 \times 1 \times 1 \times n_2} \quad B_2 \in \mathbb{R}^{n_2}$$

$$F_3(Y) = W_3 * F_2(Y) + B_3$$

$$W_3 \in \mathbb{R}^{n_2 \times f_3 \times f_3 \times c} \quad B_3 \in \mathbb{R}^c$$

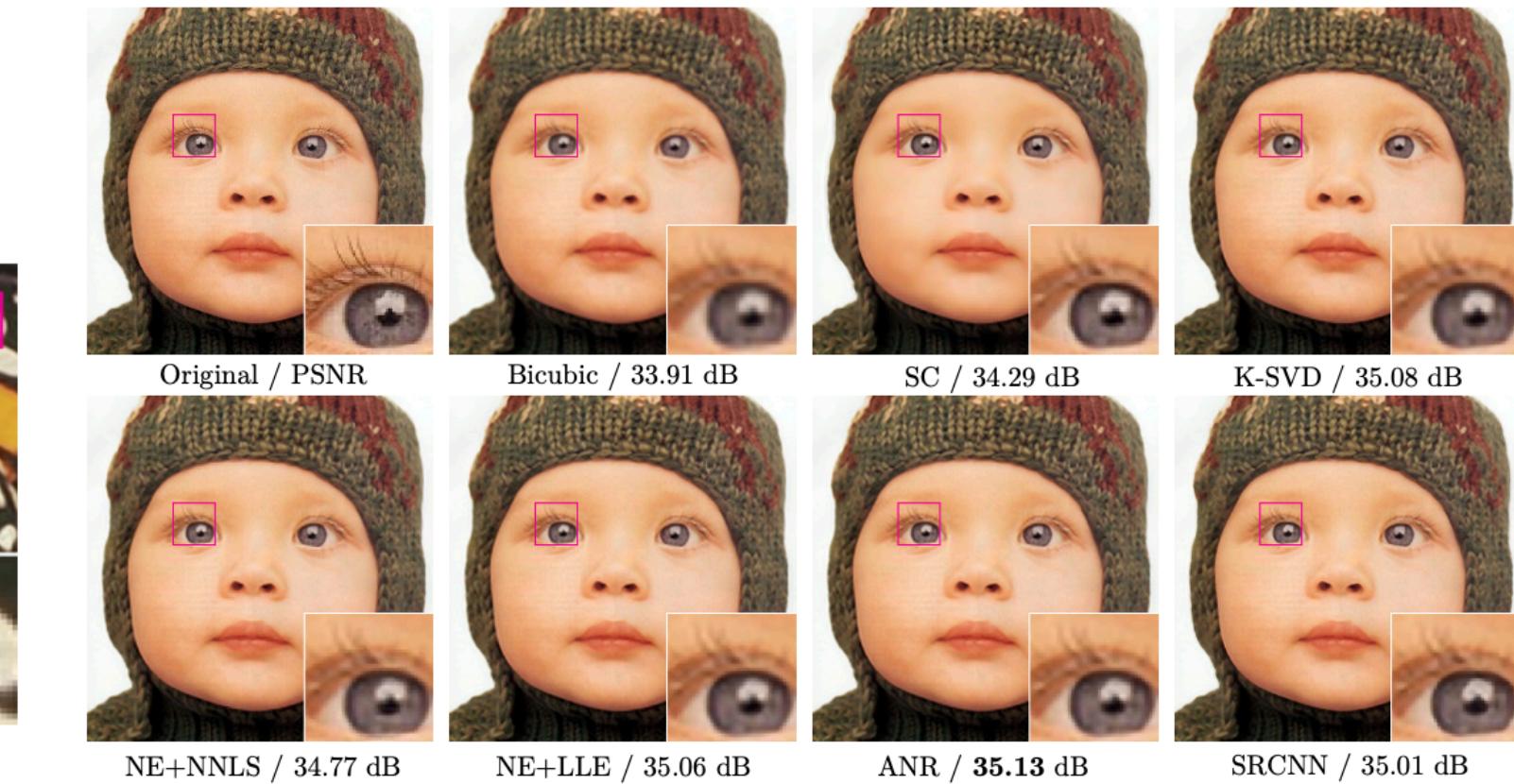
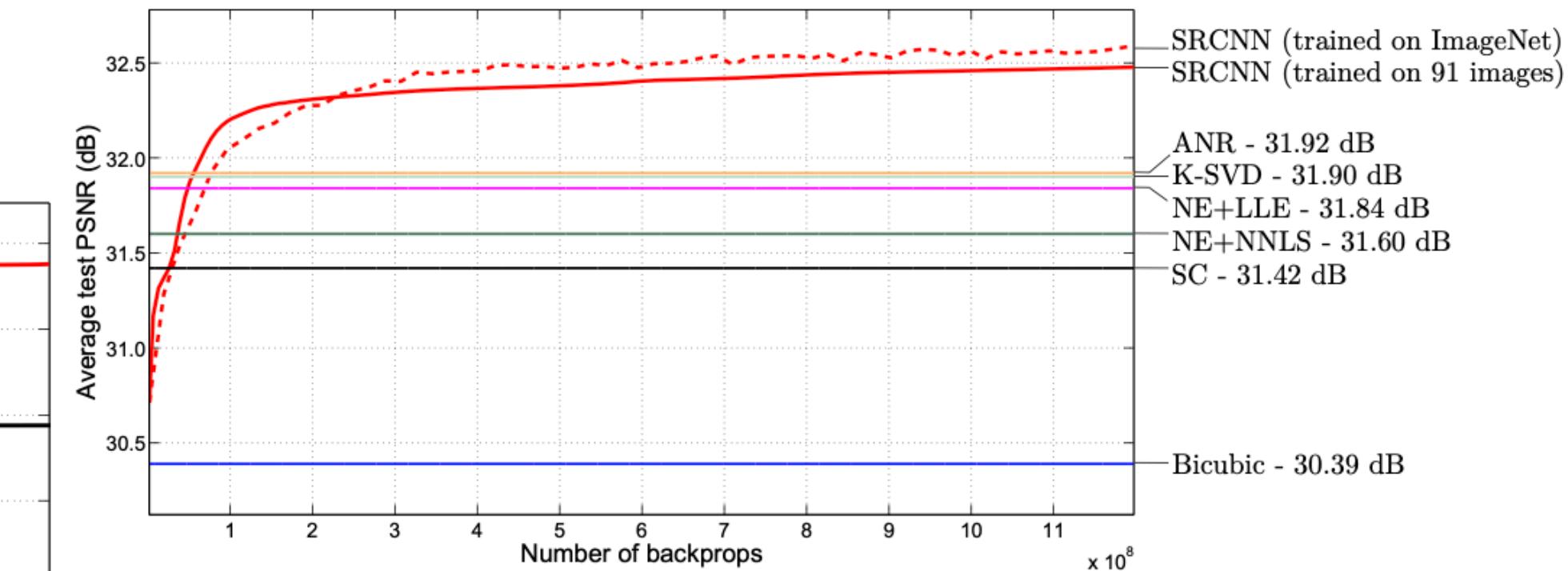
$$L(\Theta) = \frac{1}{n} \sum_{i=1}^n \|F(Y_i; \Theta) - X_i\|^2$$



Super-Resolution Convolutional Neural Network (SRCNN)
Sparse Coding (SC) based method
Peak Signal-to-Noise Ratio (PSNR)

Set5 [2] images	$n_1 = 128, n_2 = 64$		$n_1 = 64, n_2 = 32$		$n_1 = 32, n_2 = 16$	
	PSNR	Time	PSNR	Time	PSNR	Time
	32.60	0.60	32.52	0.18	32.26	0.05

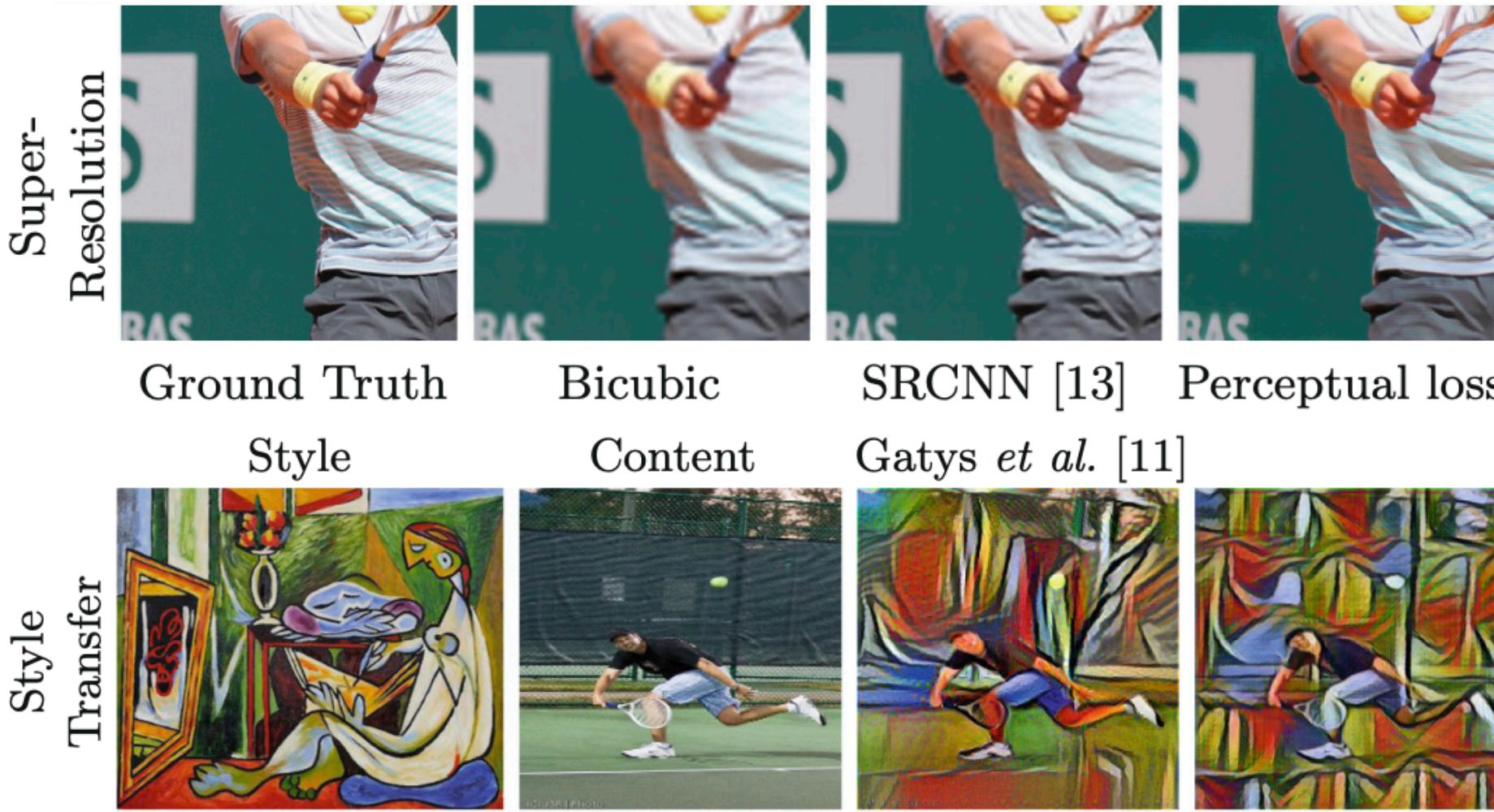
$$\begin{aligned} PSNR &= 10 \cdot \log_{10} \left(\frac{MAX_I^2}{MSE} \right) \\ &= 20 \cdot \log_{10} \left(\frac{MAX_I}{\sqrt{MSE}} \right) \\ &= 20 \cdot \log_{10}(MAX_I) - 10 \cdot \log_{10}(MSE) \\ MAX_I &= 255 \end{aligned}$$



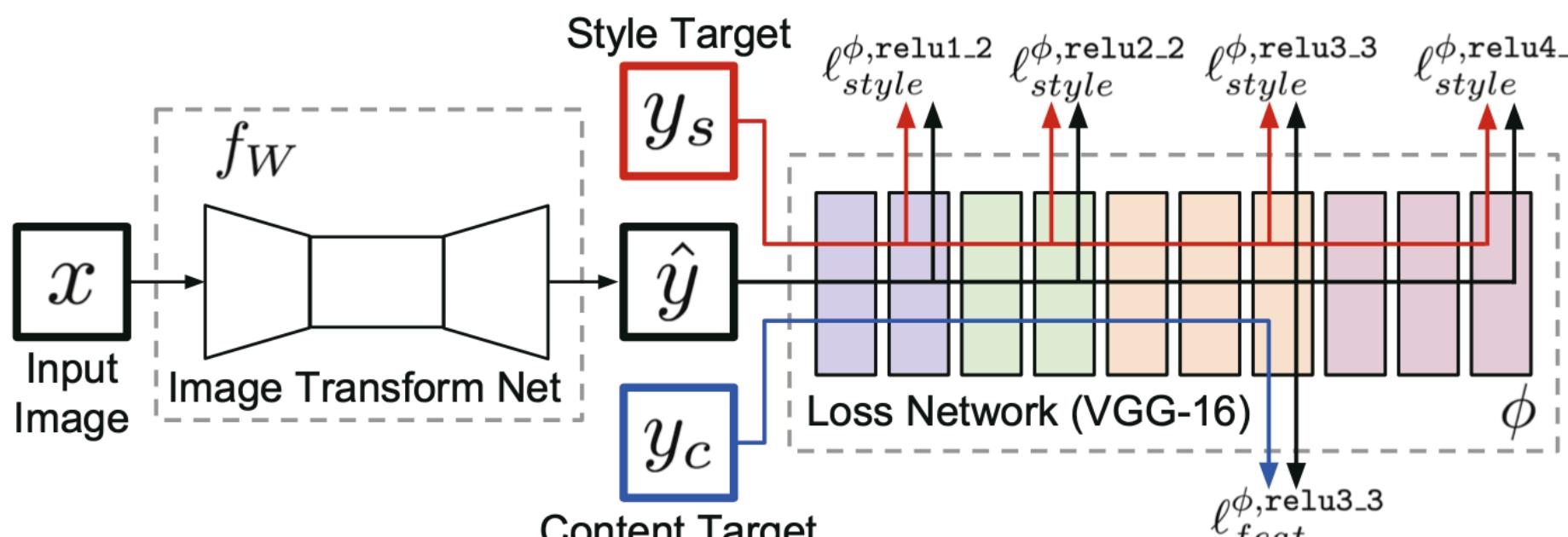
Dong, Chao, et al. "Learning a deep convolutional network for image super-resolution." *European conference on computer vision*. Springer, Cham, 2014.

Dong, Chao, et al. "Image super-resolution using deep convolutional networks." *IEEE transactions on pattern analysis and machine intelligence* 38.2 (2015): 295-307.

Perceptual Losses for Real-Time Style Transfer and Super-Resolution


[YouTube Video](#)


"For example, consider two identical images offset from each other by one pixel; despite their perceptual similarity they would be very different as measured by per-pixel losses."



$$W^* = \arg \min_W \mathbb{E}_{x,y} \left[\sum_i \lambda_i \ell_i(\underbrace{f_W(x)}_{\hat{y}}, y) \right]$$

Style transfer: $y_c = x$ & train one network per style target y_s

Super-resolution: $y_c = \text{high-resolution image}$ & style reconstruction loss is not used

Feature Reconstruction Loss

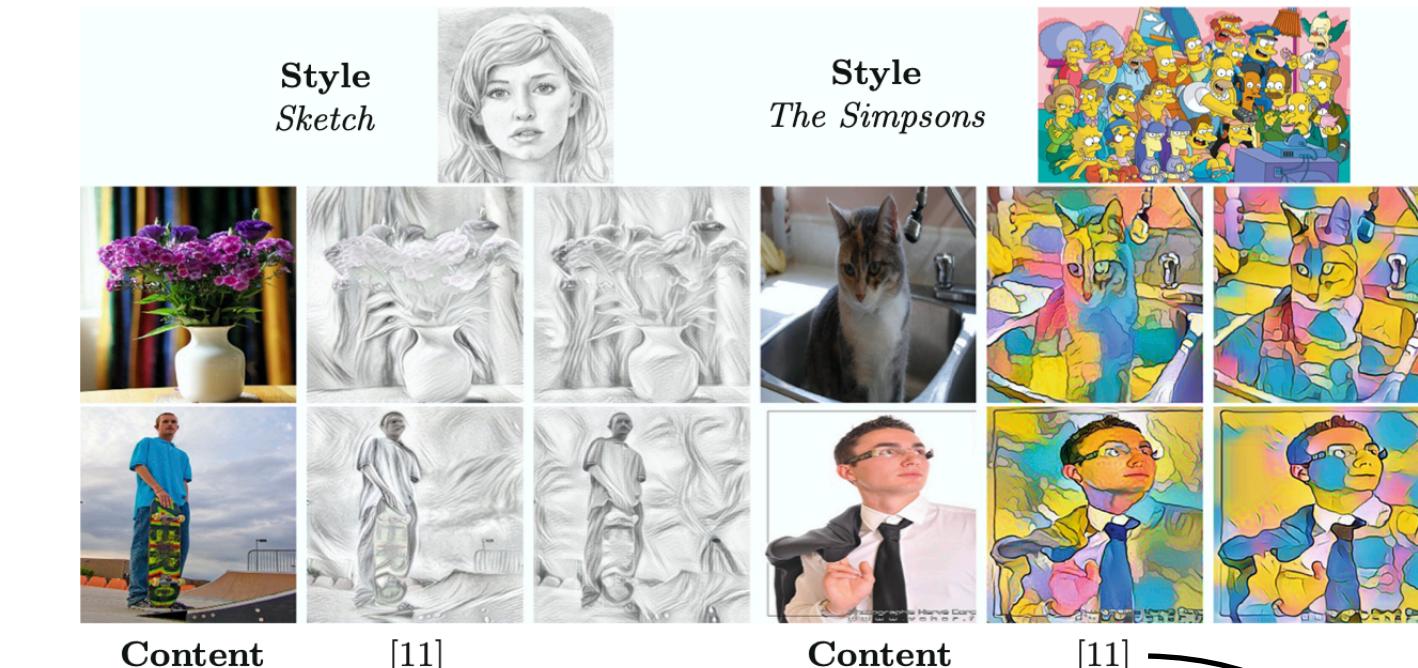
$\phi_j(x) \in \mathbb{R}^{C_j \times H_j \times W_j} \rightarrow \text{activations of the } j\text{-th layer of network } \phi$

$$\ell_{feat}^{\phi,j}(\hat{y}, y) = \frac{1}{C_j H_j W_j} \|\phi_j(\hat{y}) - \phi_j(y)\|_2^2$$

Style Reconstruction Loss

$$G_j^\phi(x)_{c,c'} = \frac{1}{C_j H_j W_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} \phi_j(x)_{h,w,c} \phi_j(x)_{h,w,c'}$$

$G_j^\phi(x) \in \mathbb{R}^{C_j \times C_j} \rightarrow \text{Gram Matrix}$



Gram matrix captures information about which features tend to activate together.

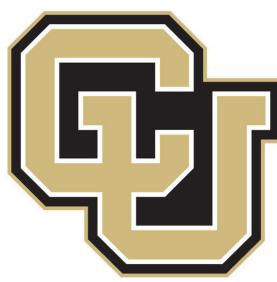
$$\ell_{style}^{\phi,j}(\hat{y}, y) = \|G_j^\phi(\hat{y}) - G_j^\phi(y)\|_F^2$$

$$\hat{y} = \arg \min_y \lambda_c \ell_{feat}^{\phi,j}(y, y_c) + \lambda_s \ell_{style}^{\phi,j}(y, y_s) + \lambda_{TV} \ell_{TV}(y)$$



Speedup over [11] at 100, 300, 500 optimization iterations

Image Size	Speedup		
	100	300	500
256 × 256	212x	636x	1060x
512 × 512	205x	615x	1026x
1024 × 1024	208x	625x	1042x



Boulder

Image Style Transfer Using Convolutional Neural Networks

Content Representation

$p \rightarrow$ photograph

$x \rightarrow$ generated image (randomly initialized)

$P^l \rightarrow$ feature representation of p at layer l

$F^l \rightarrow$ feature representation of x at layer l

$$\mathcal{L}_{\text{content}}(\vec{p}, \vec{x}, l) = \frac{1}{2} \sum_{i,j} (F_{ij}^l - P_{ij}^l)^2$$

$F_{i,j}^l \rightarrow$ activation of i -th filter at position j in layer l

$$F^l \in \mathbb{R}^{N_l \times M_l}$$

$N_l \rightarrow$ number of feature maps

$M_l \rightarrow$ height times width of the feature map

Style Representation

$$G^l \in \mathbb{R}^{N_l \times N_l} \rightarrow \text{Gram matrix}$$

$$G_{ij}^l = \sum_k F_{ik}^l F_{jk}^l \rightarrow \text{inner product between the vectorized feature maps } i \text{ and } j \text{ in layer } l$$

captures texture but not global arrangement

$a \rightarrow$ artwork image $A^l \rightarrow$ style representation of a

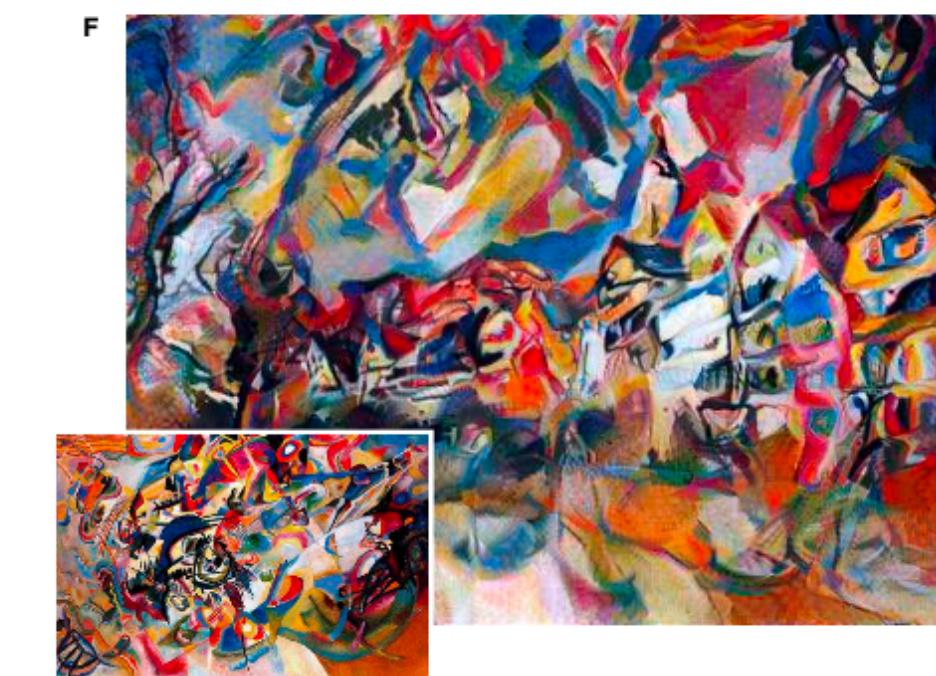
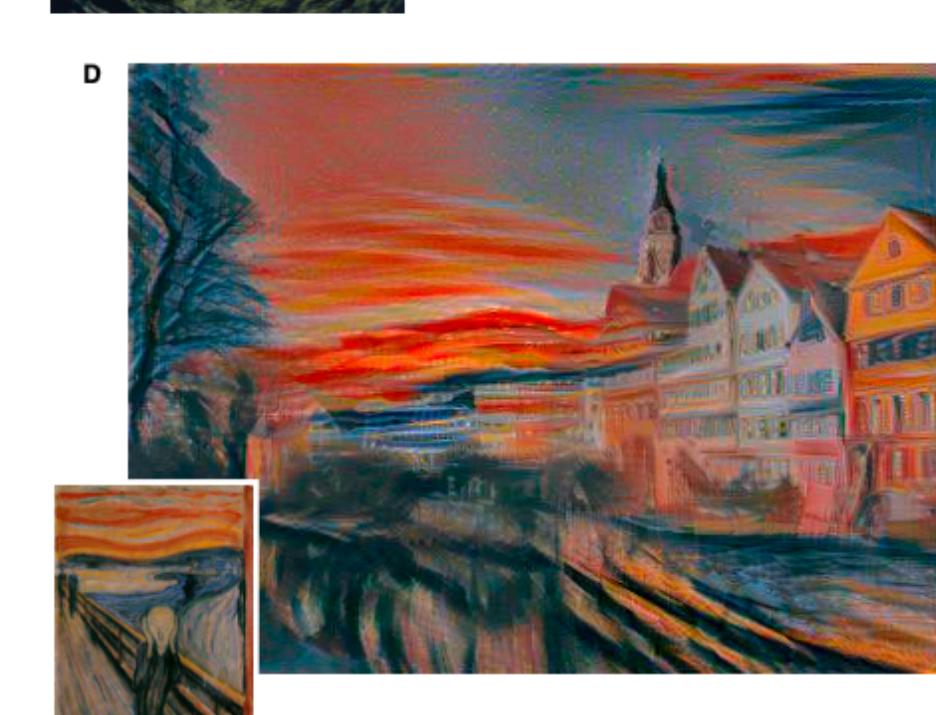
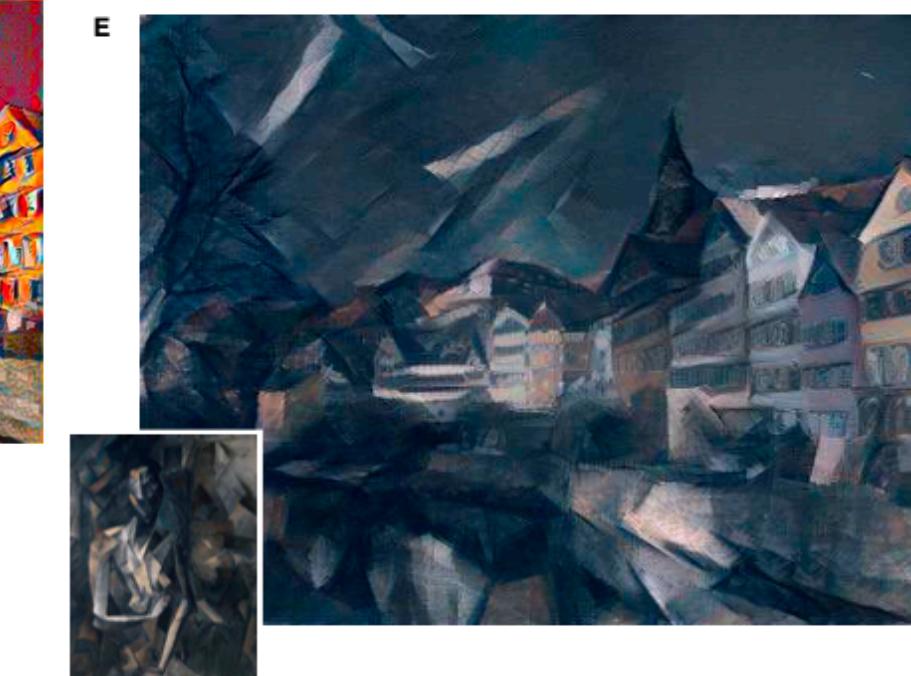
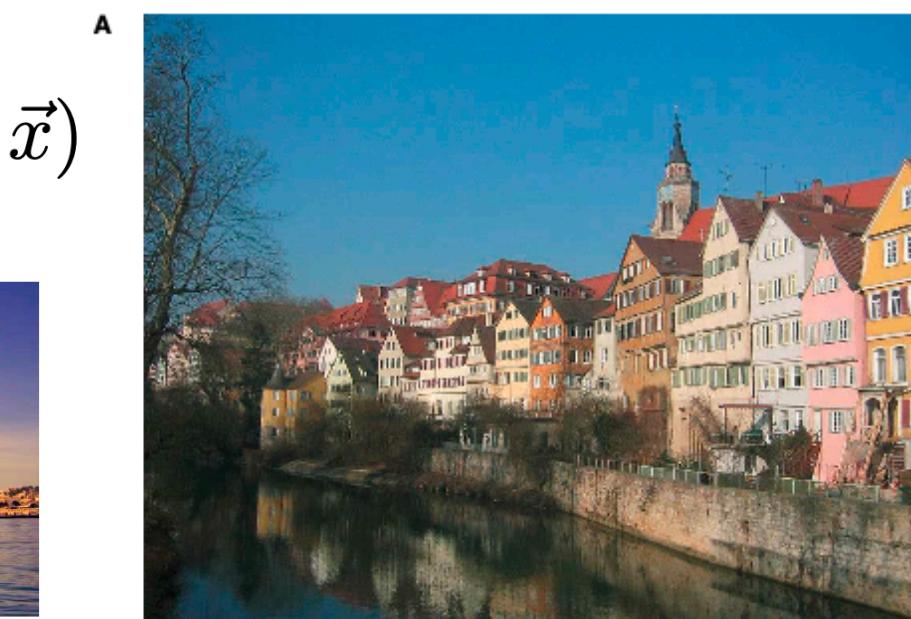
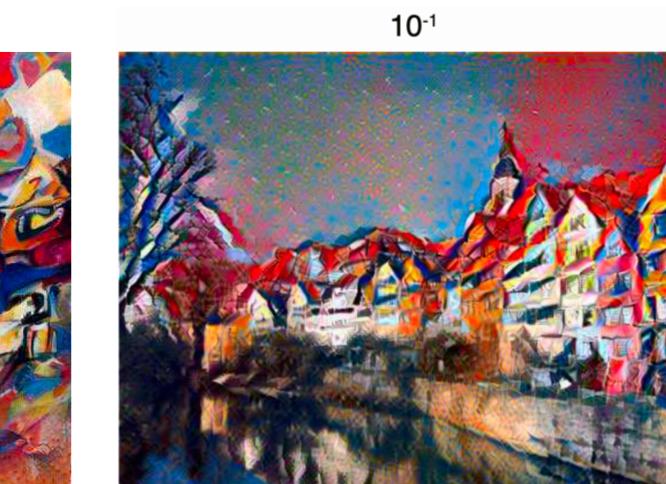
$G^l \rightarrow$ style representation of x

$$\mathcal{L}_{\text{style}}(\vec{a}, \vec{x}) = \sum_{l=0}^L w_l E_l \quad E_l = \frac{1}{4N_l^2 M_l^2} \sum_{i,j} (G_{ij}^l - A_{ij}^l)^2 \quad \boxed{\text{The ratio } \alpha/\beta}$$

Style Transfer

$$\mathcal{L}_{\text{total}}(\vec{p}, \vec{a}, \vec{x}) = \alpha \mathcal{L}_{\text{content}}(\vec{p}, \vec{x}) + \beta \mathcal{L}_{\text{style}}(\vec{a}, \vec{x})$$

$$\frac{\partial \mathcal{L}_{\text{total}}}{\partial \vec{x}}$$



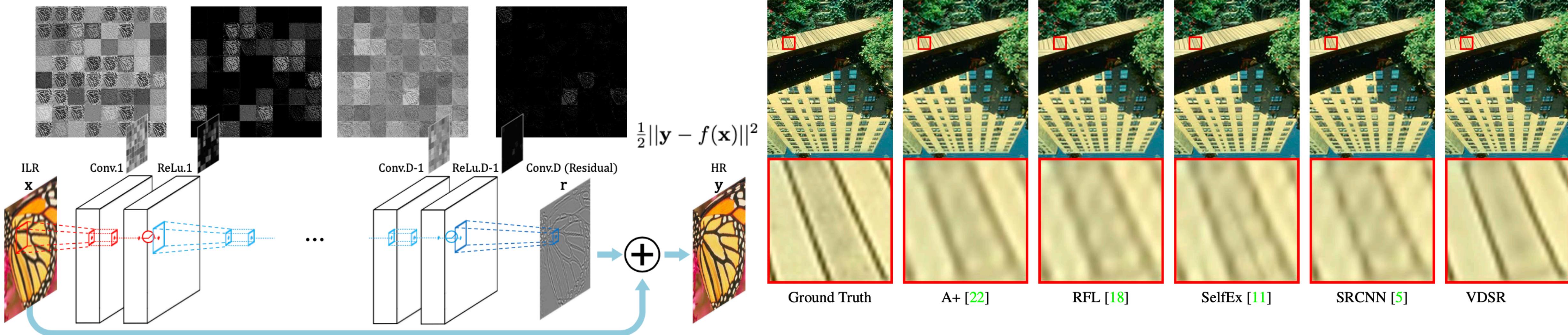
Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "Image style transfer using convolutional neural networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "A neural algorithm of artistic style." *arXiv preprint arXiv:1508.06576* (2015).

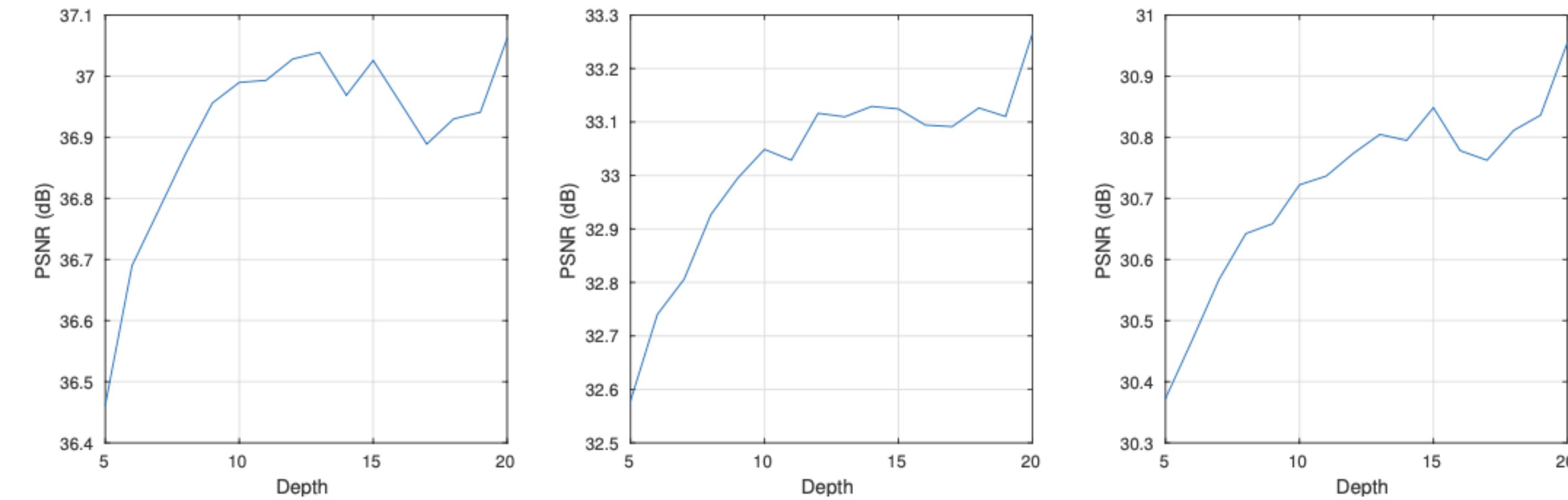
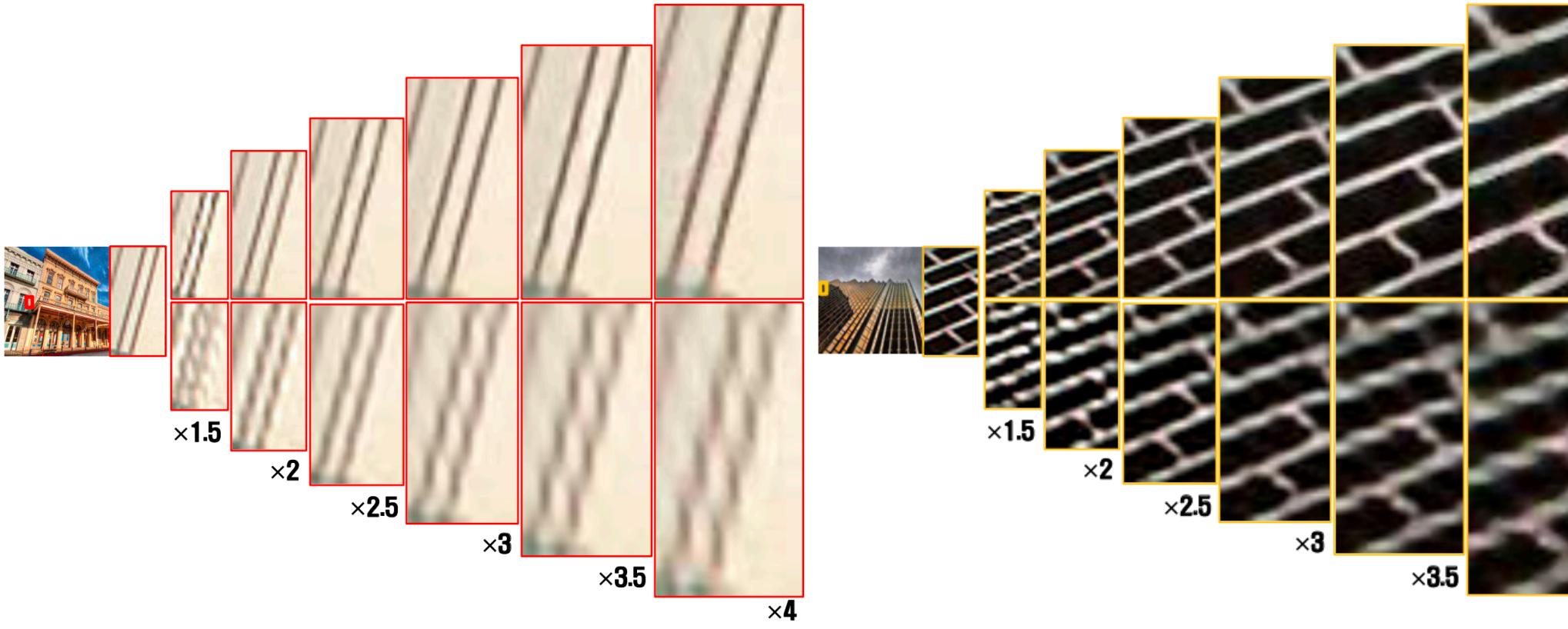
Accurate Image Super-Resolution Using Very Deep Convolutional Networks


[YouTube Video](#)

Single Image Super-Resolution (SISR)



- Residual-Learning
- High Learning Rates
- Adjustable Gradient Clipping
- Multi-scale





Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network

I^{HR} → high-resolution image

I^{LR} → low-resolution image

$I^{\text{LR}} \leftarrow I^{\text{HR}}$
 ▷ Gaussian filter
 ▷ Downsampling

I^{SR} → super-resolved image

r → upscaling ratio

$I^{\text{LR}} \in \mathbb{R}^{H \times W \times C}$

$I^{\text{HR}} \in \mathbb{R}^{rH \times rW \times C}$

avoid upscaling I^{LR}

(i.e., bicubic interpolation)

before feeding into the network

$$f^1(I^{\text{LR}}; W_1, b_1) = \tanh(W_1 * I^{\text{LR}} + b_1)$$

$$f^l(I^{\text{LR}}; W_{1:l}, b_{1:l}) = \tanh(W_l * f^{l-1}(I^{\text{LR}}) + b_l), l = 2, \dots, L-1$$

$$W_l \in \mathbb{R}^{n_{l-1} \times n_l \times k_l \times k_l}, b_l \in \mathbb{R}^{n_l}$$

n_l → number of features at layer l

$n_0 = C, k_l$ → filter size at layer l

Efficient sub-pixel convolution layer

$$\mathbf{I}^{\text{SR}} = f^L(\mathbf{I}^{\text{LR}}) = \mathcal{PS}(W_L * f^{L-1}(\mathbf{I}^{\text{LR}}) + b_L)$$

\mathcal{PS} → periodic shuffling operator

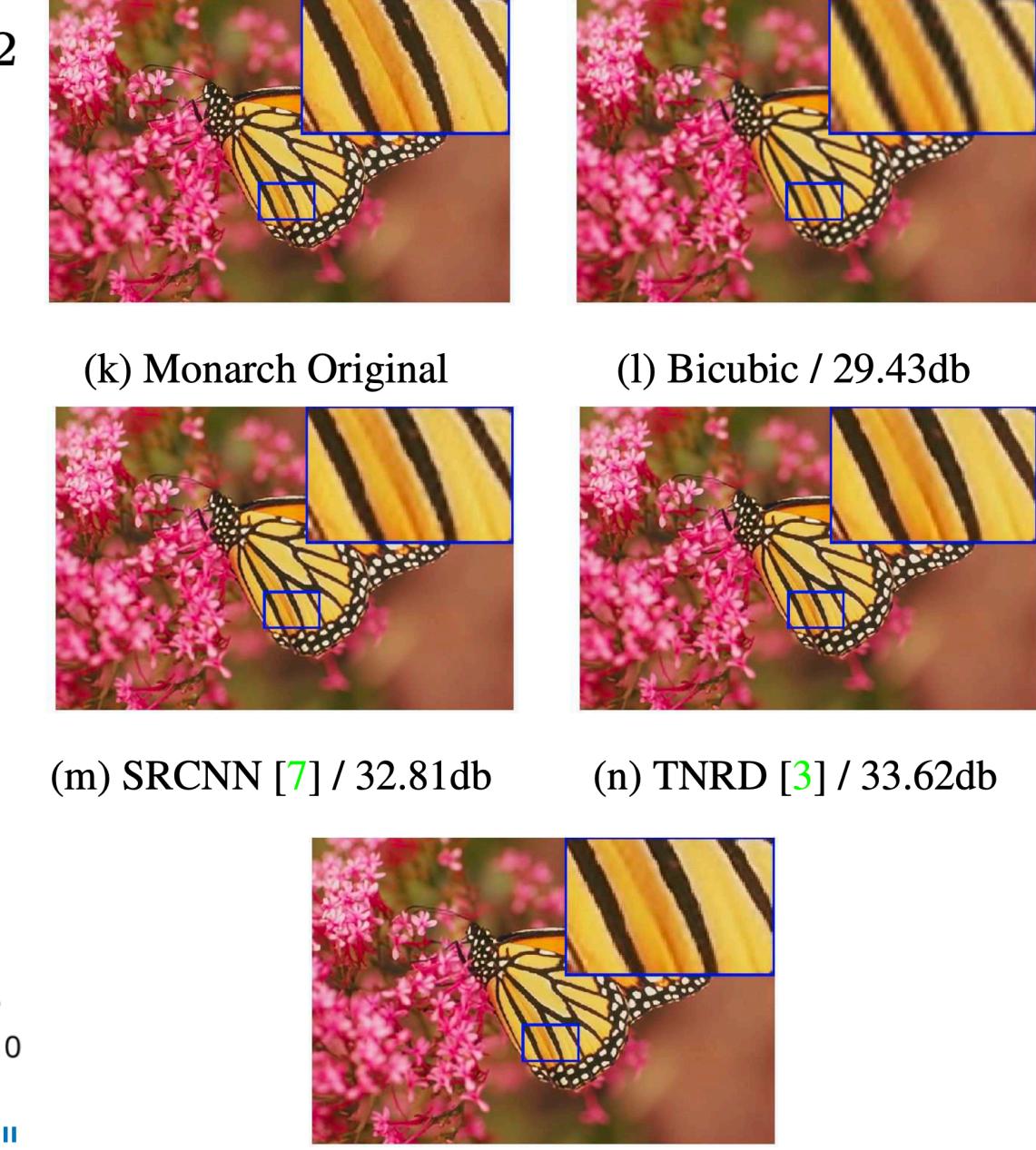
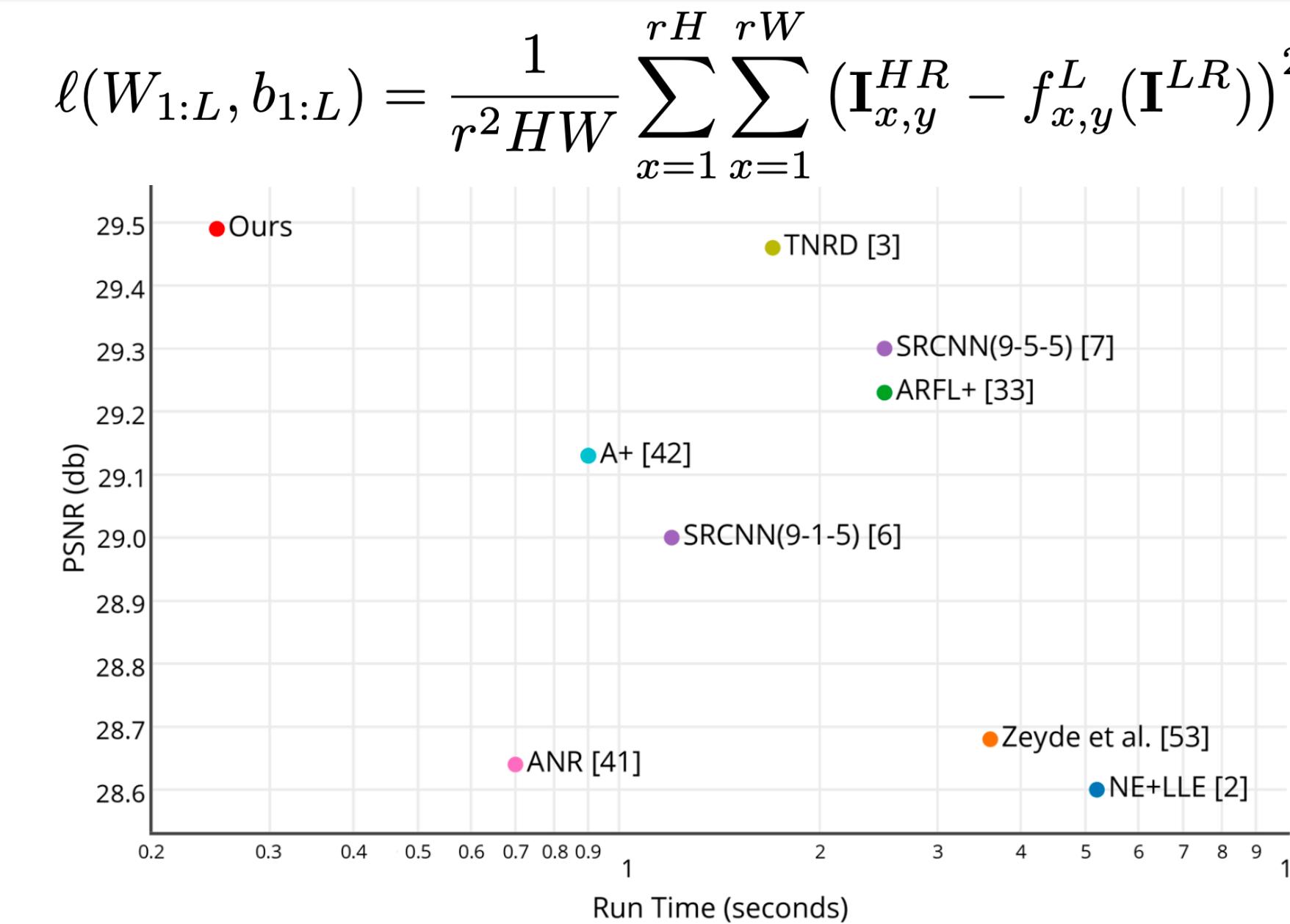
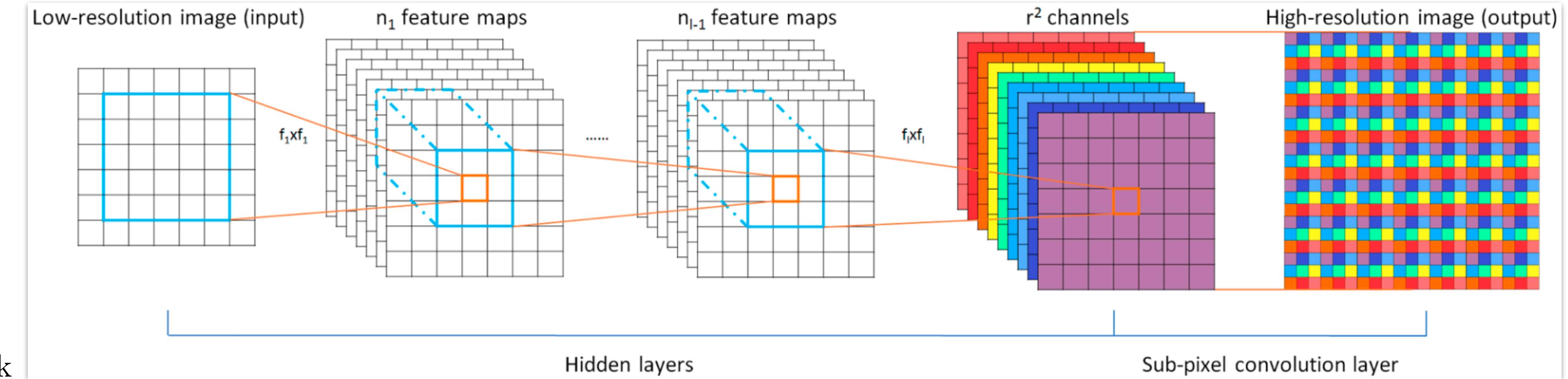
$$H \times W \times C \cdot r^2 \mapsto rH \times rW \times C$$

$$\mathcal{PS}(T)_{x,y,c} = T_{\lfloor x/r \rfloor, \lfloor y/r \rfloor, c \cdot r \cdot \text{mod}(y,r) + c \cdot \text{mod}(x,r)}$$

equivalent to a fractionally-strided convolution

convolution with stride $1/r$ in the LR space

with a filter W_s of size $k_s = rk_L$, but more efficient!



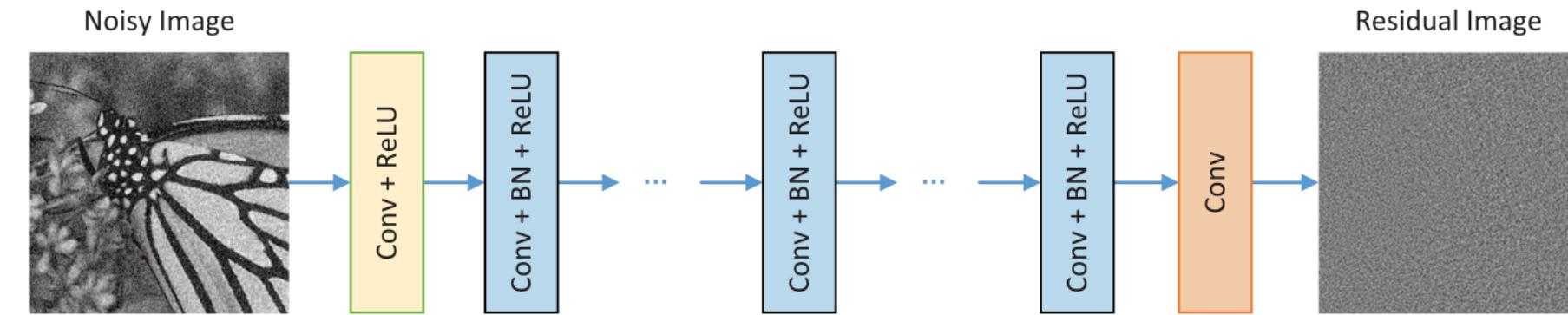
Shi, Wenzhe, et al. "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network."

Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising


[YouTube Video](#)

Denoising Convolutional Neural Networks (DnCNNs)



$x \rightarrow$ clean image

$y \rightarrow$ noisy observation

$y = x + v \rightarrow$ image degradation model

AWGN (Additive White Gaussian Noise)
with standard deviation σ

Receptive Field of DnCNN (3×3 conv)

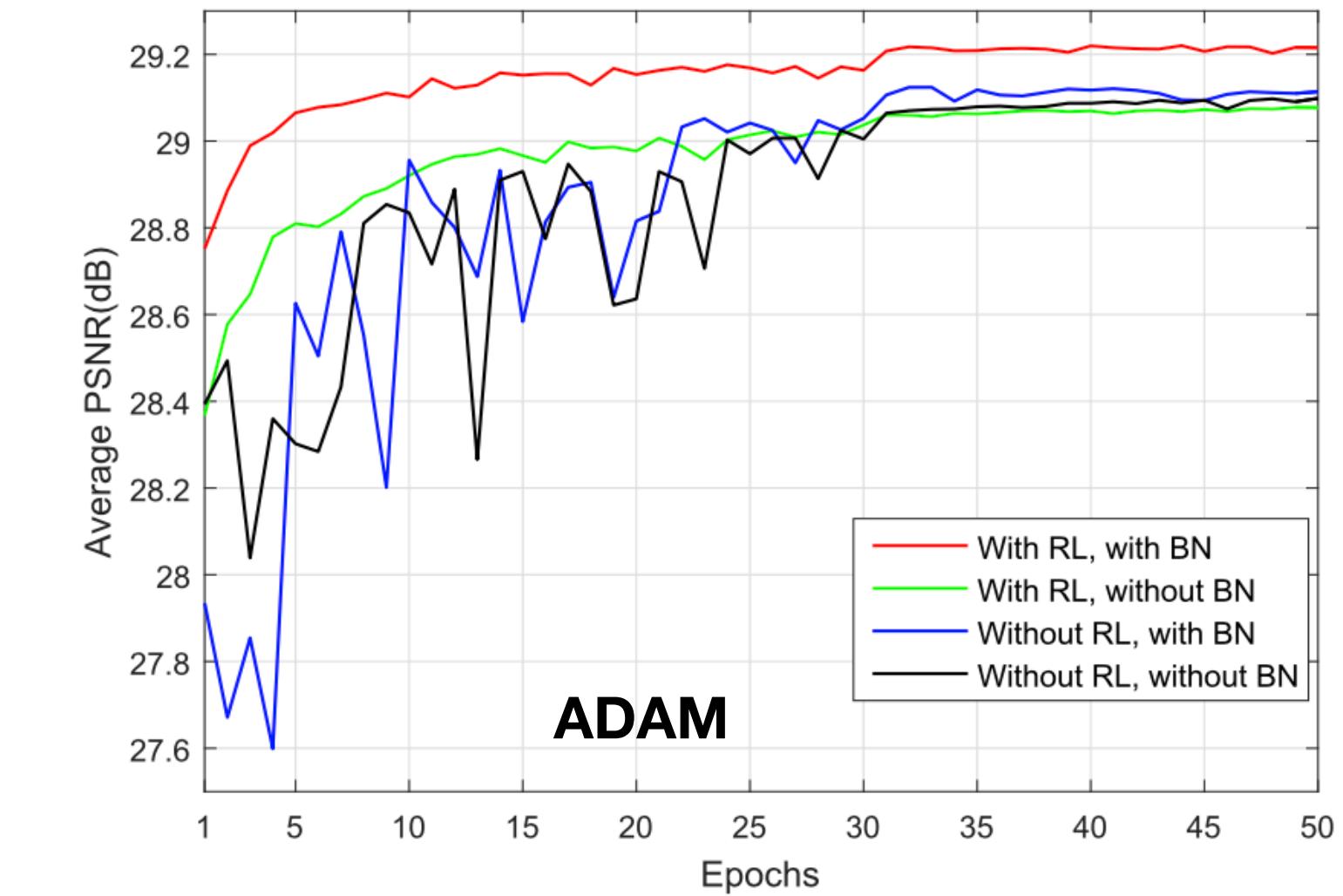
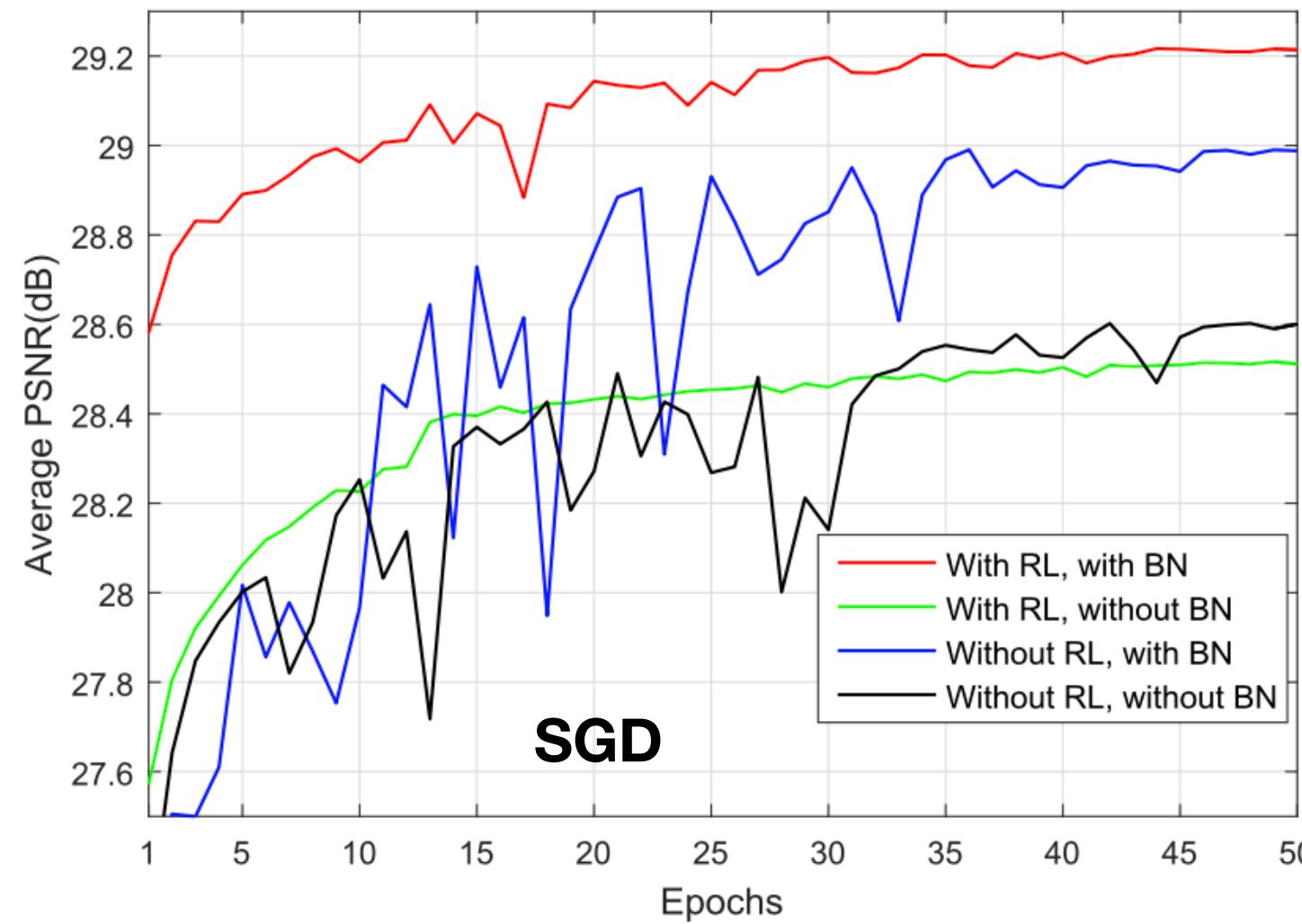
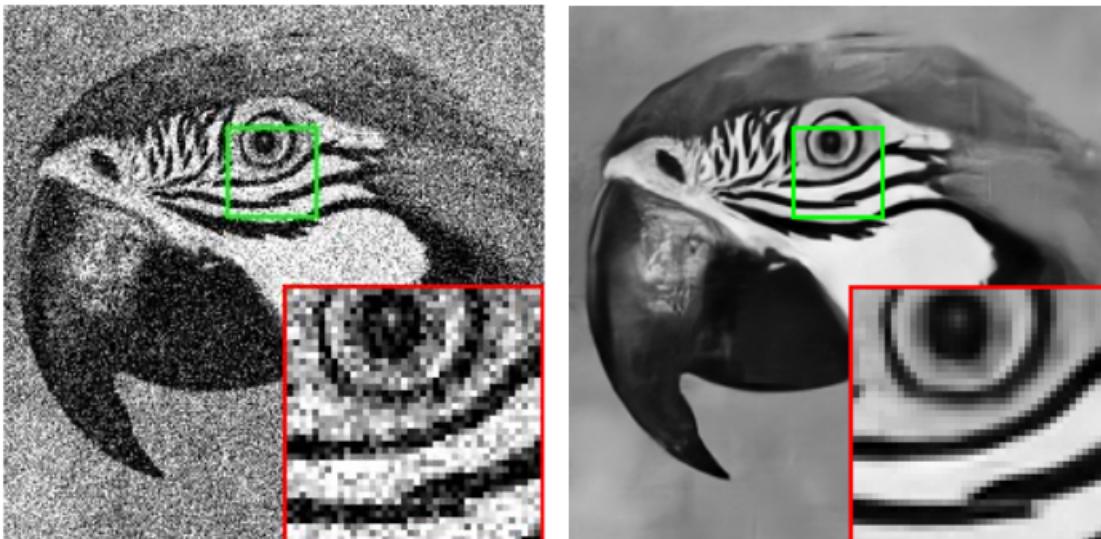
with depth d is $(2d + 1) \times (2d + 1)$

$\mathcal{F}(y) = x$

$\mathcal{R}(y) \approx v \rightarrow$ residual mapping

$\Rightarrow x = y - \mathcal{R}(y)$

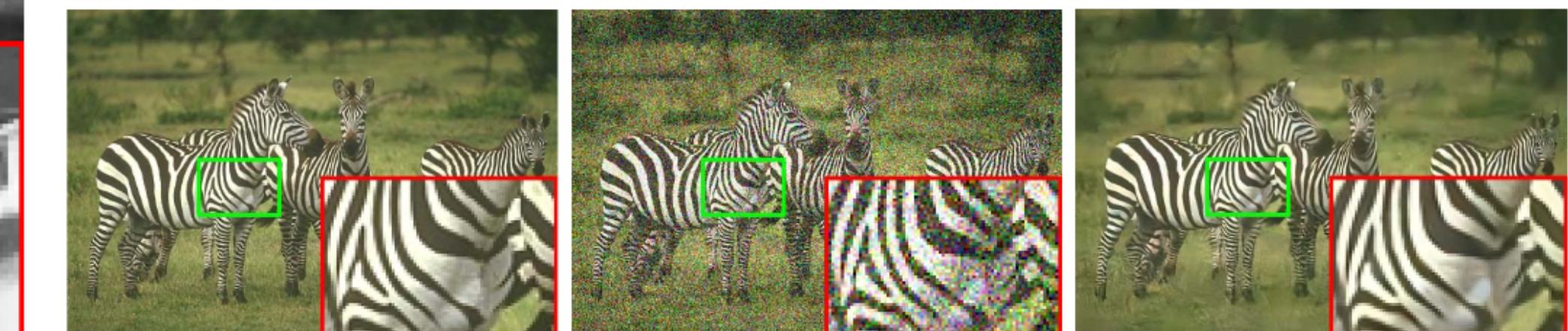
$$\ell(\theta) = \frac{1}{2N} \sum_{i=1}^N \|R(y_i; \theta) - (y_i - x_i)\|_F^2$$



JPEG image de-blocking

Single Image Super-resolution

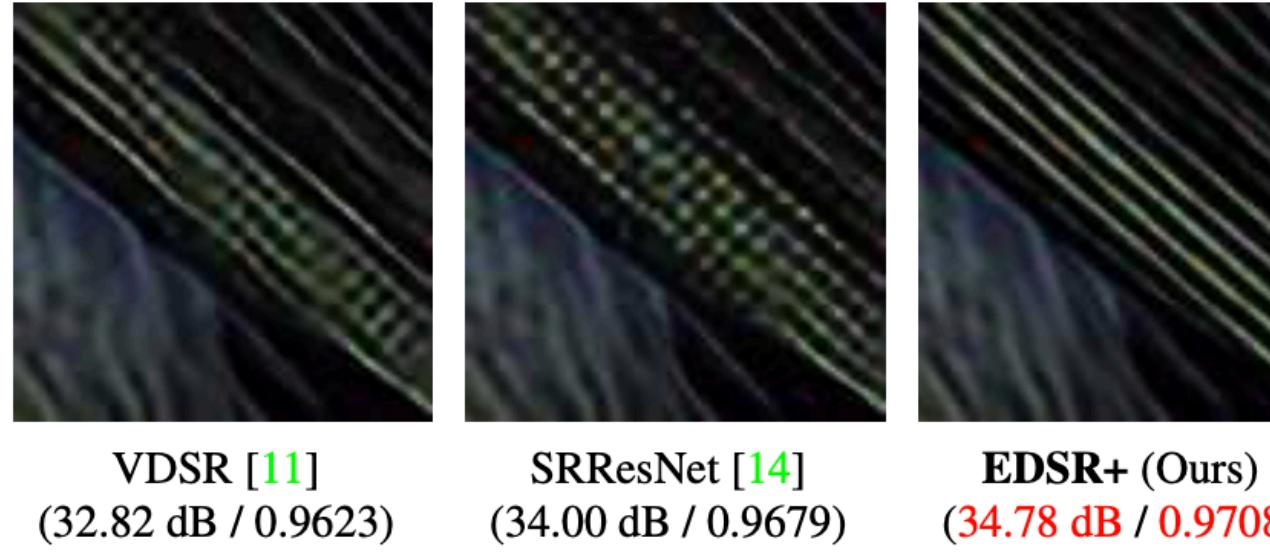
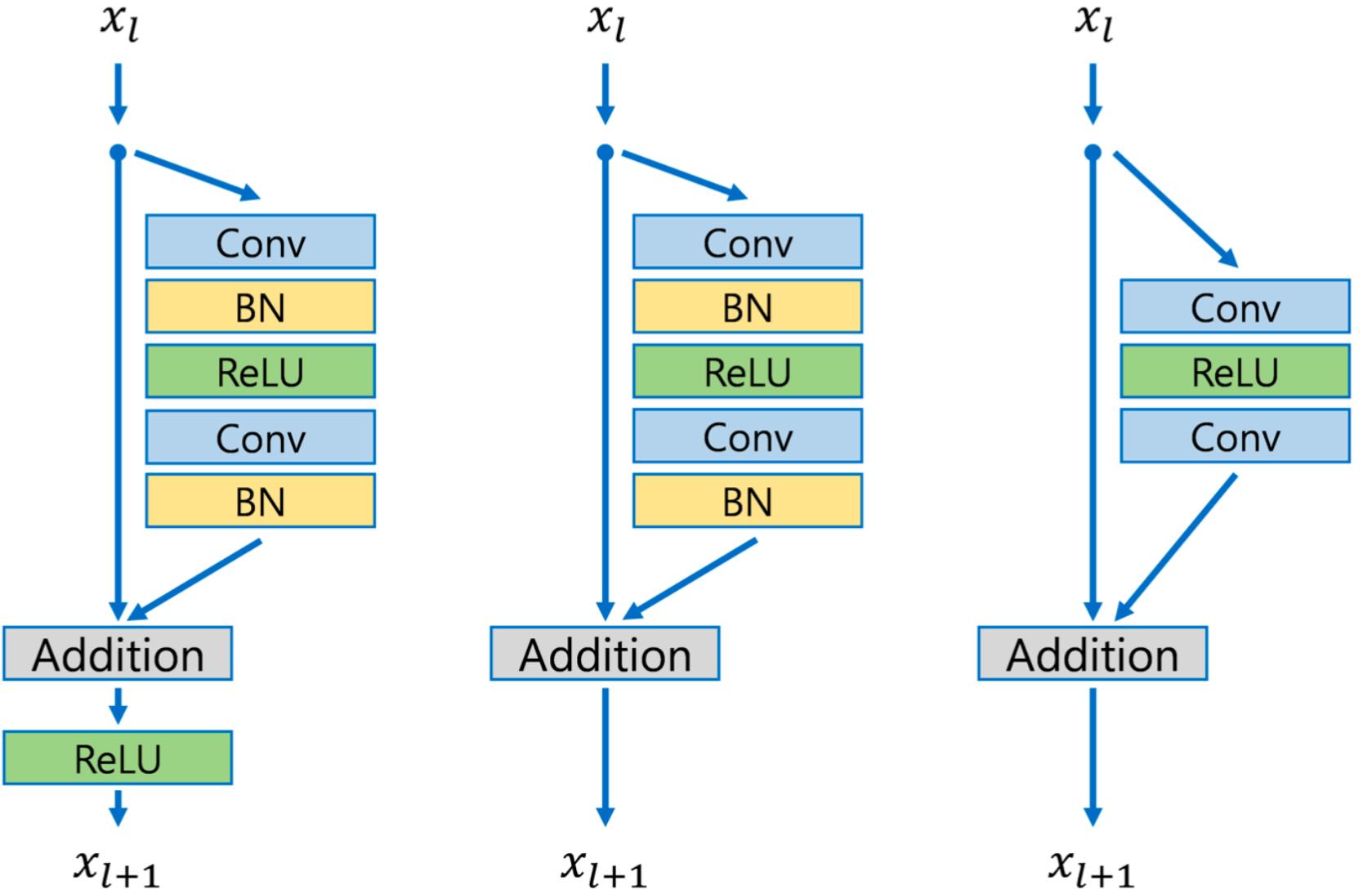
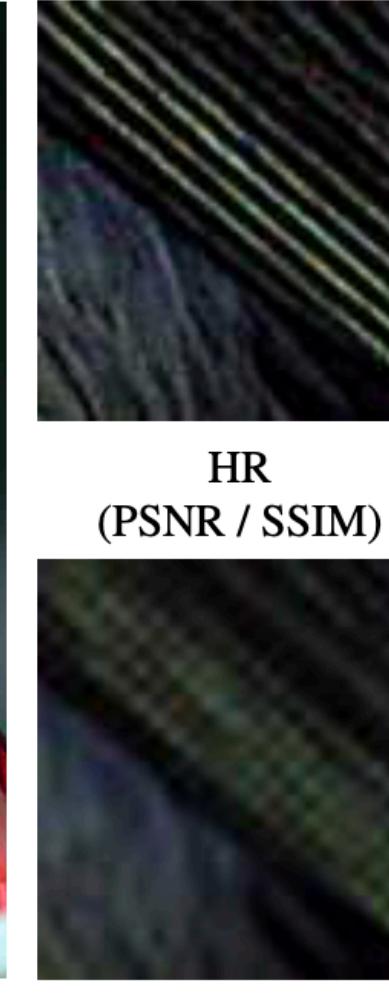
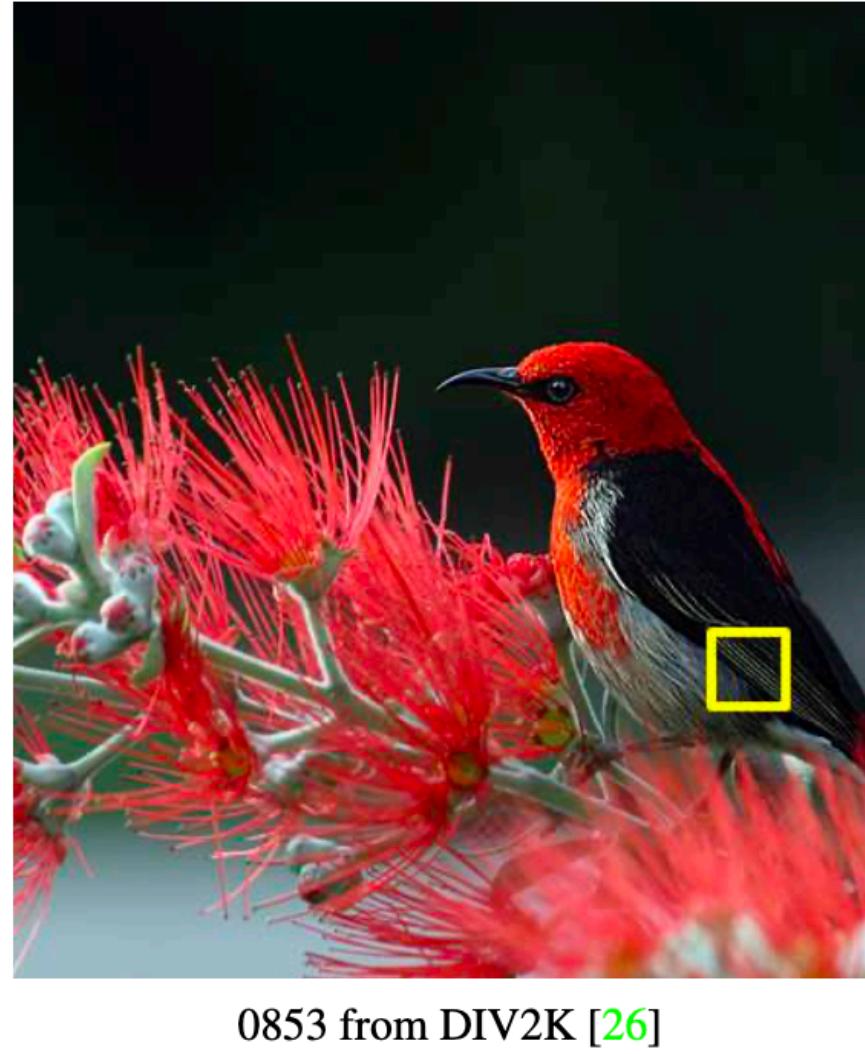
Gaussian Denoising





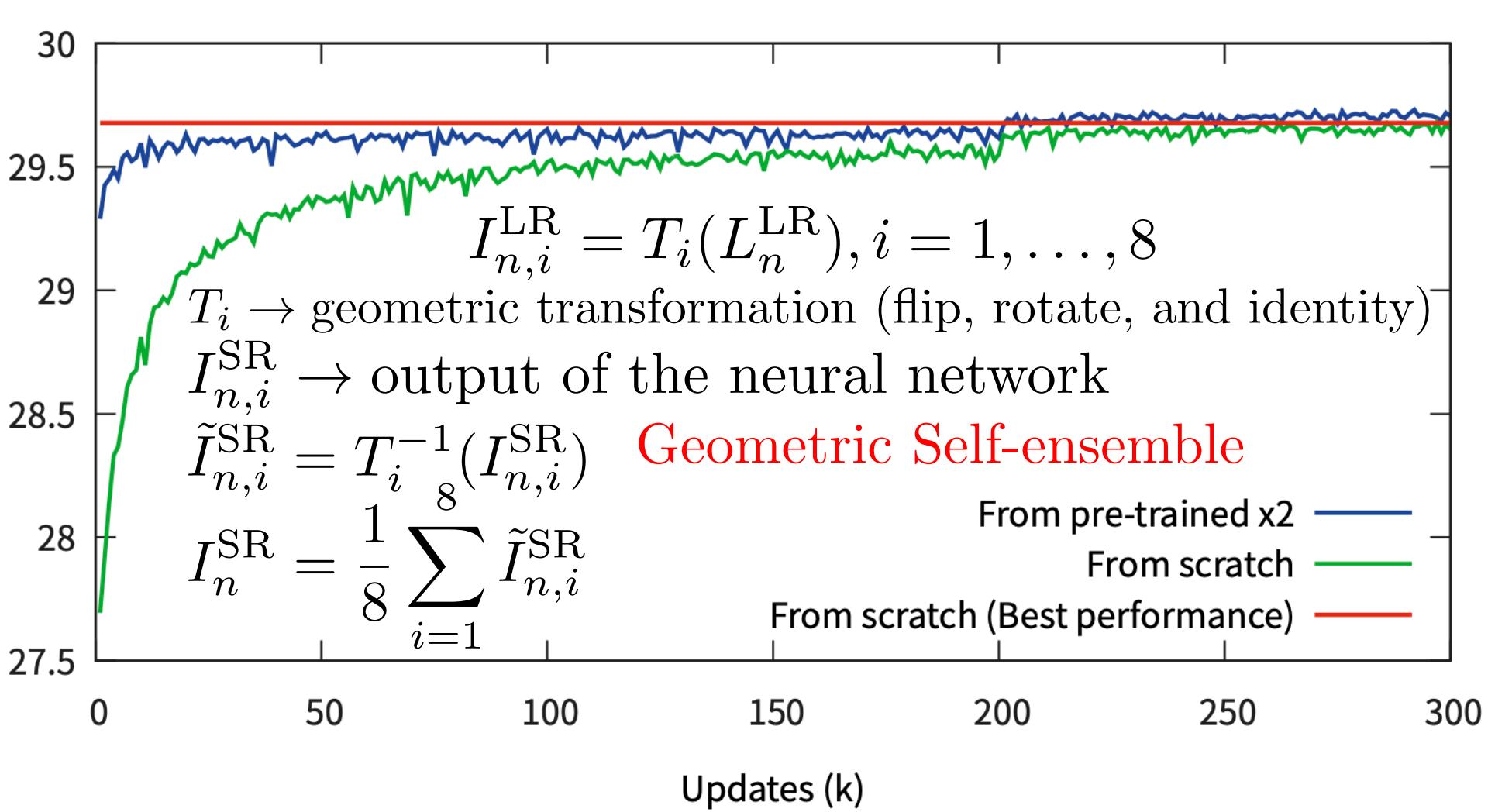
Boulder

Enhanced Deep Residual Networks for Single Image Super-Resolution

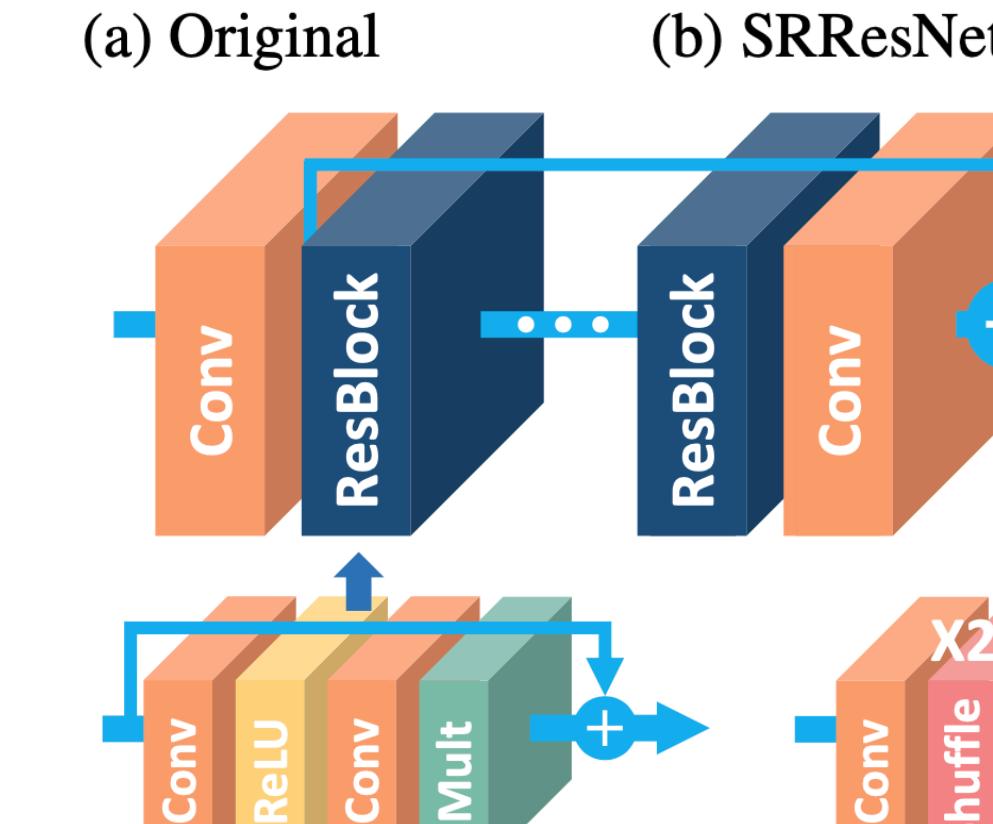


Options	SRResNet [14] (reproduced)	Baseline (Single / Multi)	EDSR	MDSR
# Residual blocks	16	16	32	80
# Filters	64	64	256	64
# Parameters	1.5M	1.5M / 3.2M	43M	8.0M
Residual scaling	-	-	0.1	-
Use BN	Yes	No	No	No
Loss function	L2	L1	L1	L1

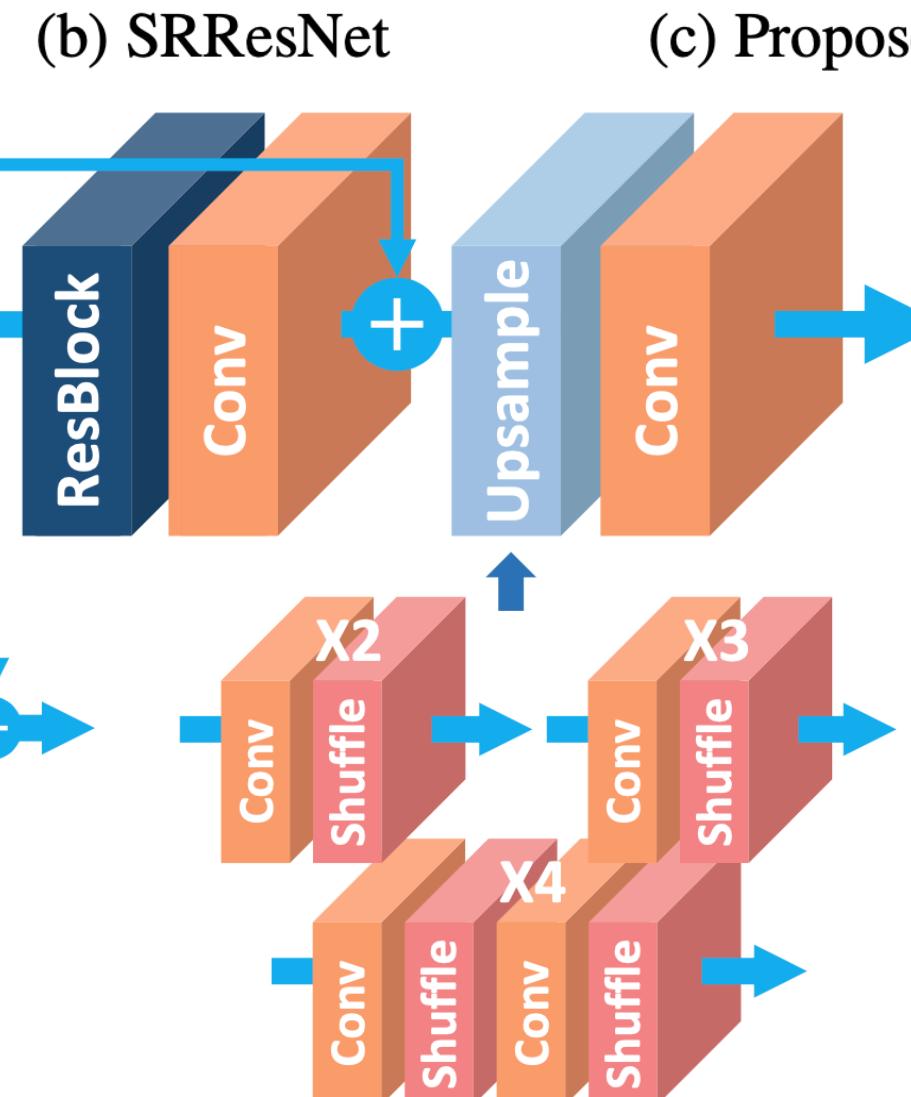
PSNR(dB) on DIV2K validation set (x4)



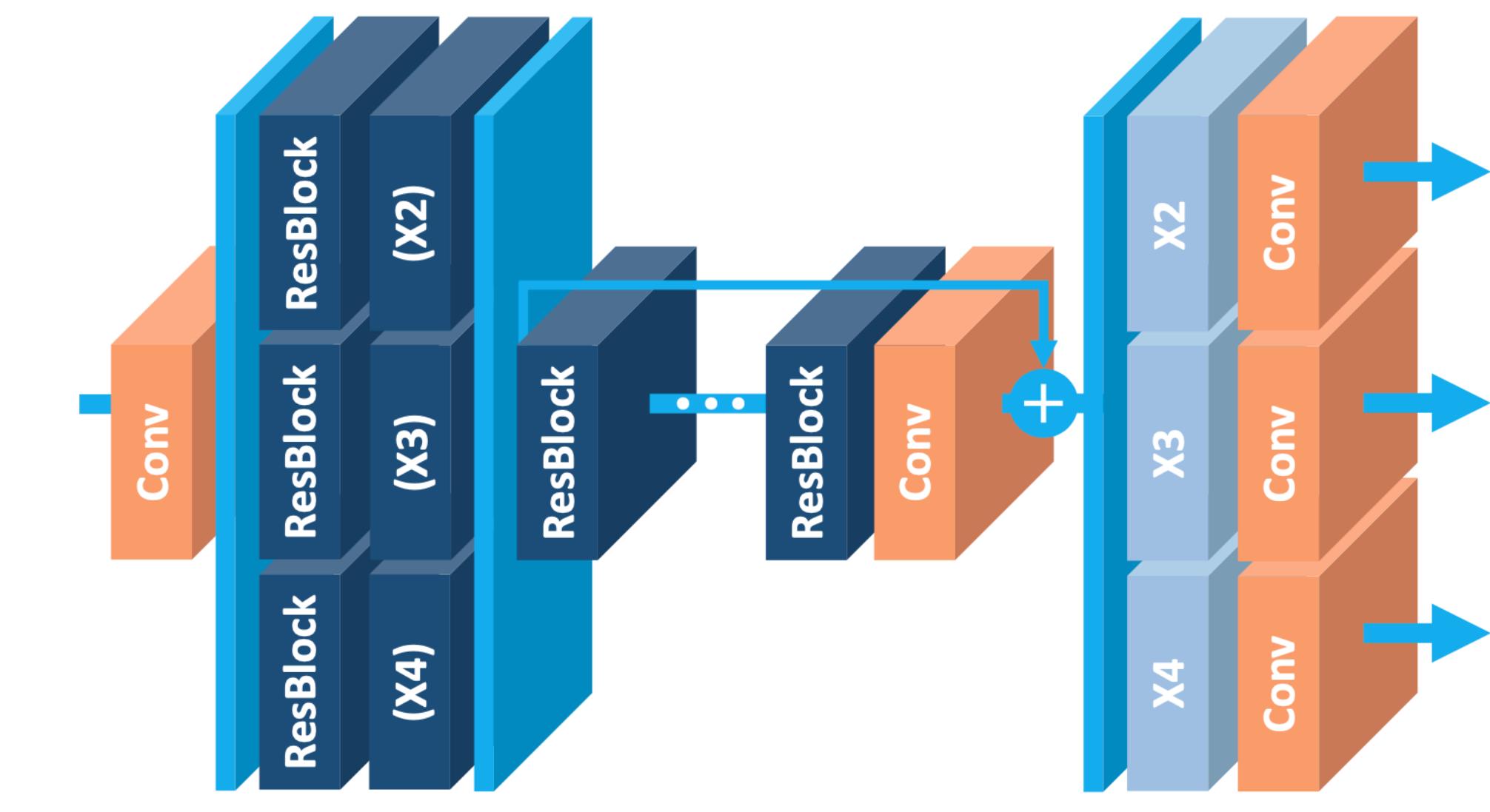
(a) Original



(b) SRResNet



(c) Proposed





Boulder

Deep Image Prior

Zero training data, zero validation data, and one test data!

$$x^* = \arg \min_x \|d(x) - x_0\|^2 + R(x)$$

$x^* \in \mathbb{R}^{rH \times rW \times C}$ → super-resolved image

$x_0 \in \mathbb{R}^{H \times W \times C}$ → low resolution image

$x \in \mathbb{R}^{rH \times rW \times C}$

$d : \mathbb{R}^{rH \times rW \times C} \rightarrow \mathbb{R}^{H \times W \times C}$

└ downsampling operator

$R(x)$ → regularizer (e.g., total variation (TV) of the image)

Use a CNN as the regularizer!

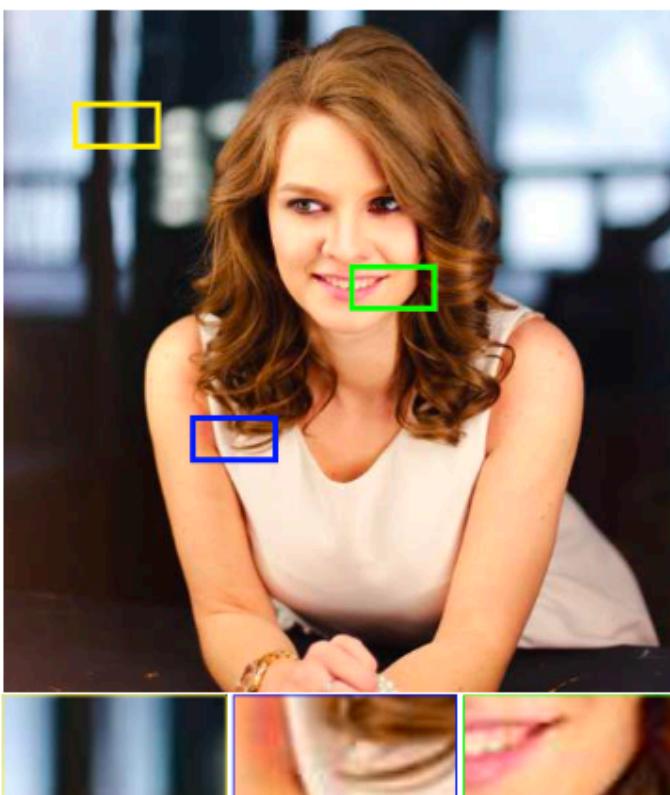
$R(x) = 0$ when $x = f_\theta(z)$ and $R(x) = +\infty$ otherwise

z → a vector filled with uniform noise (fixed)

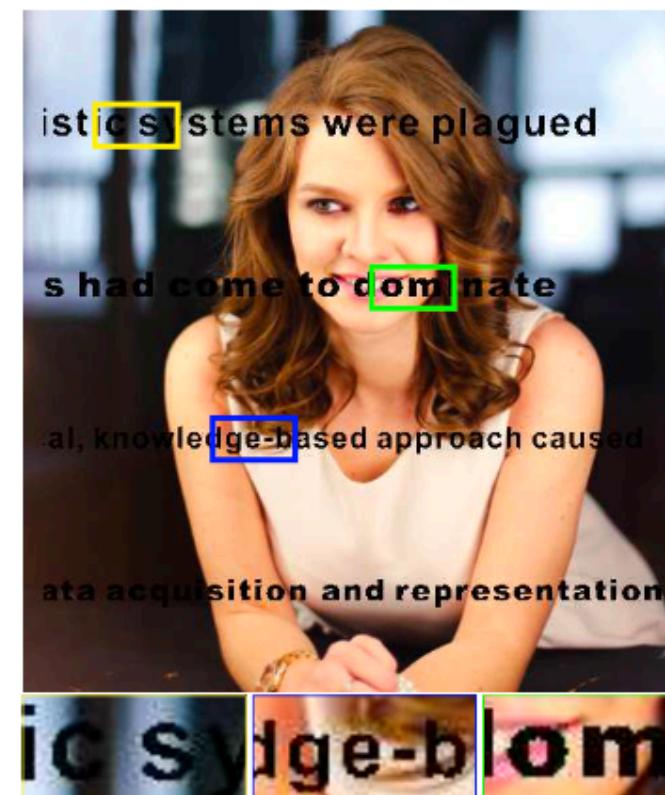
f_θ → CNN

$$\theta^* = \arg \min_\theta \|d(f_\theta(z)) - x_0\|^2$$

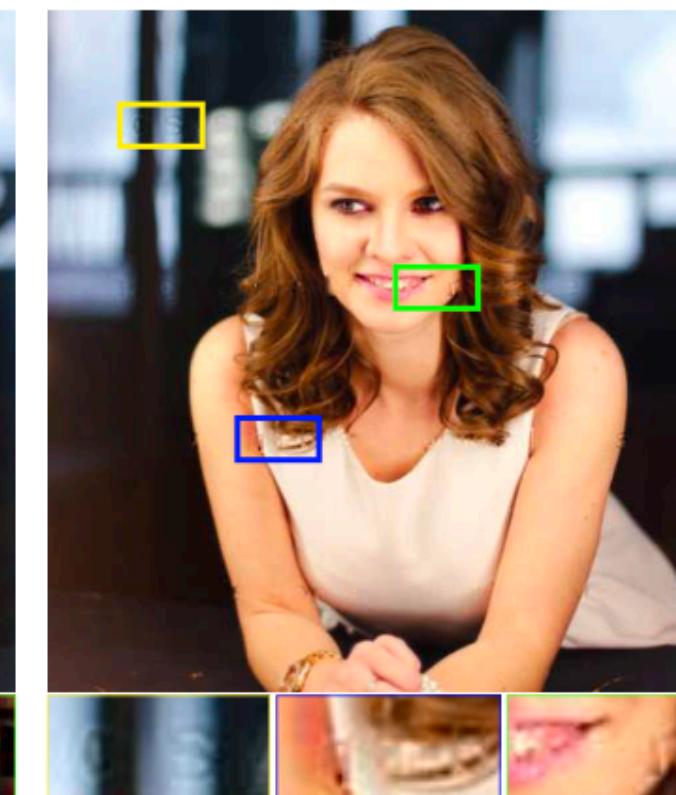
$$x^* = f_{\theta^*}(z)$$



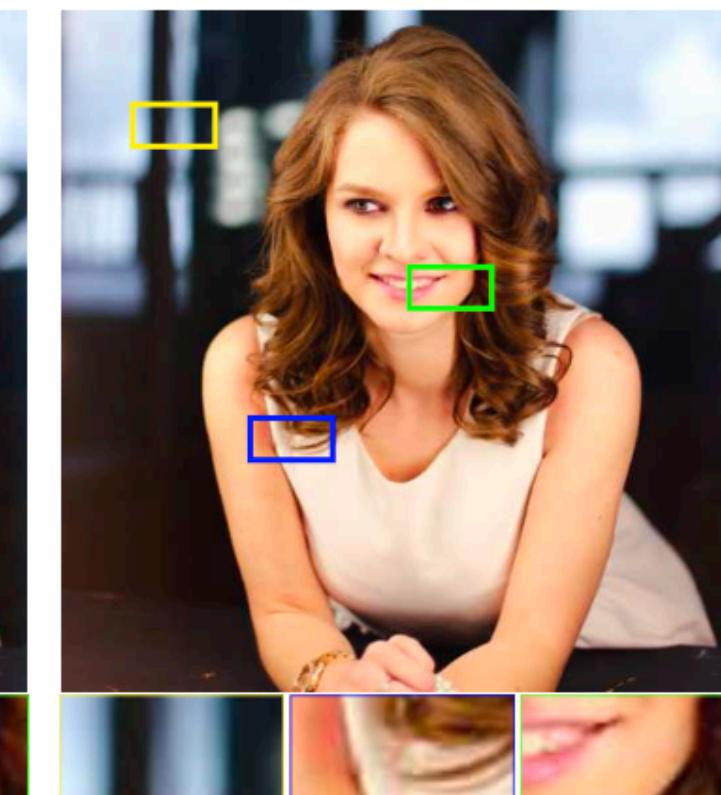
(a) Original image



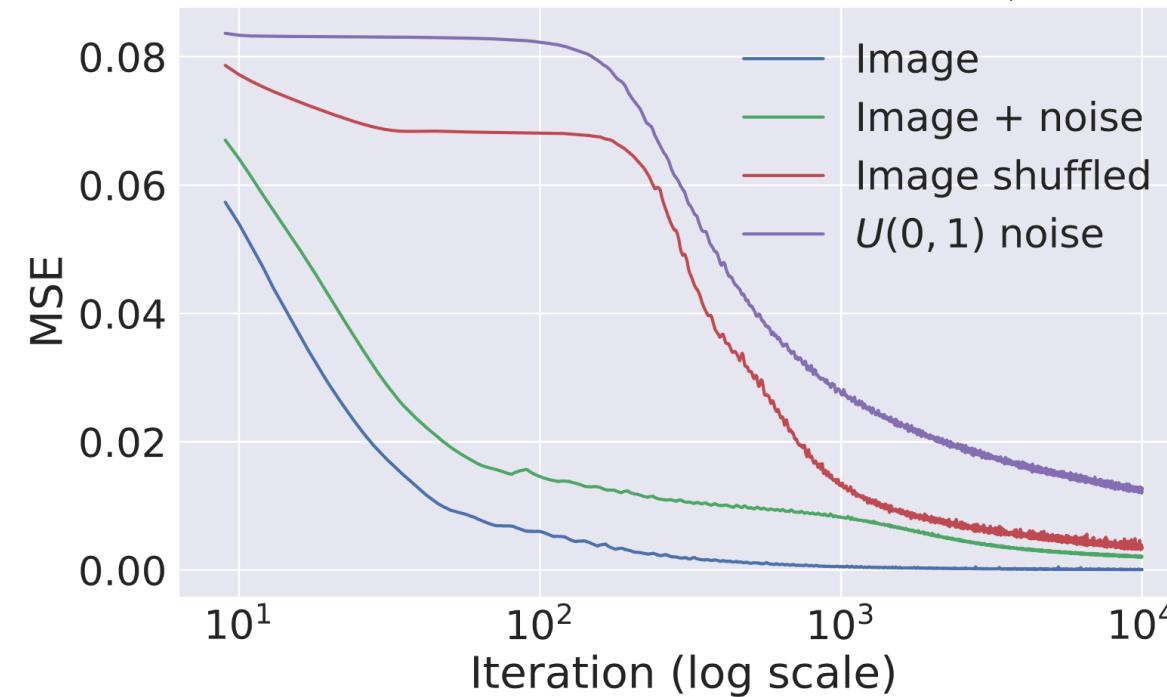
(b) Corrupted image



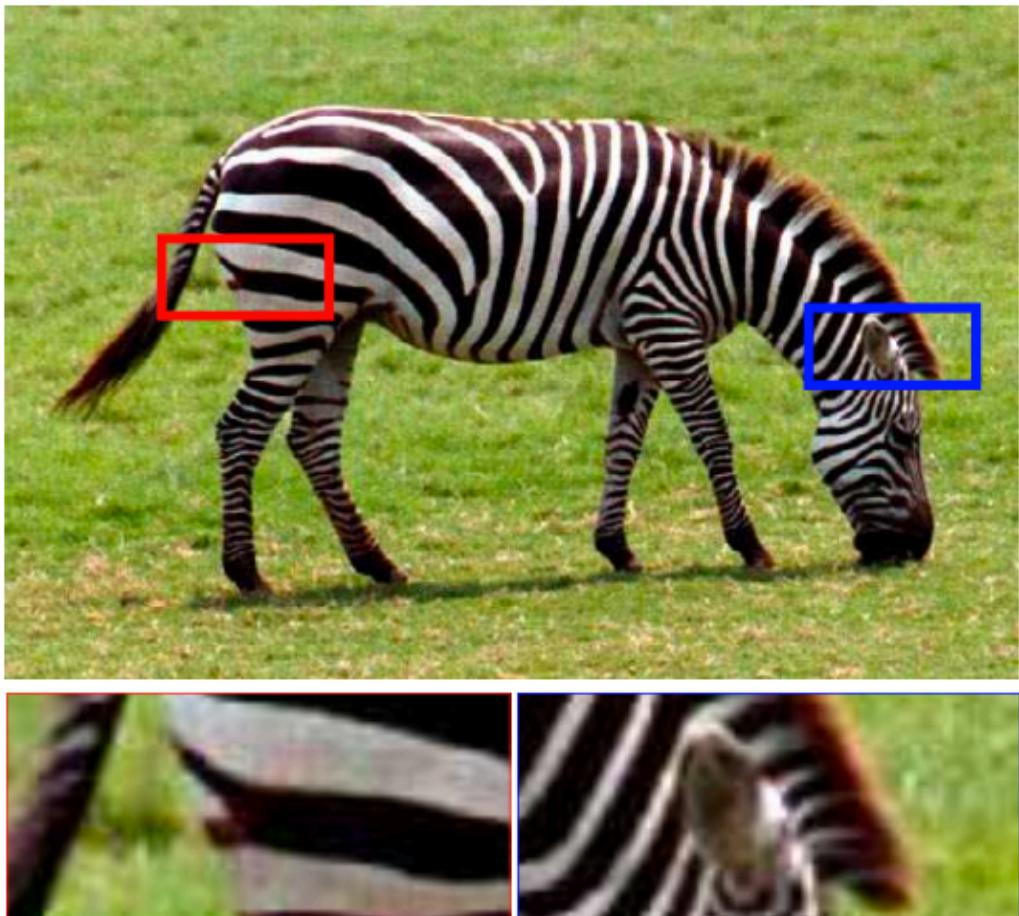
(c) Sheppard networks [27]



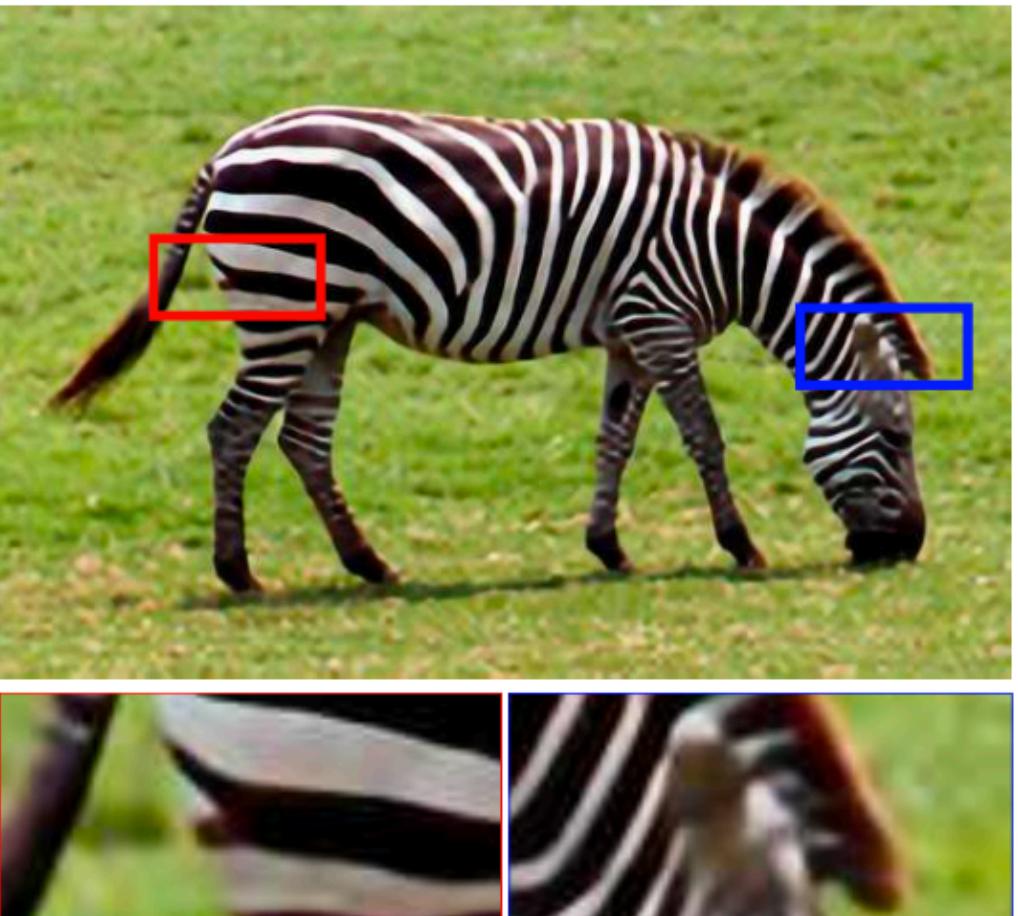
(d) Deep Image Prior



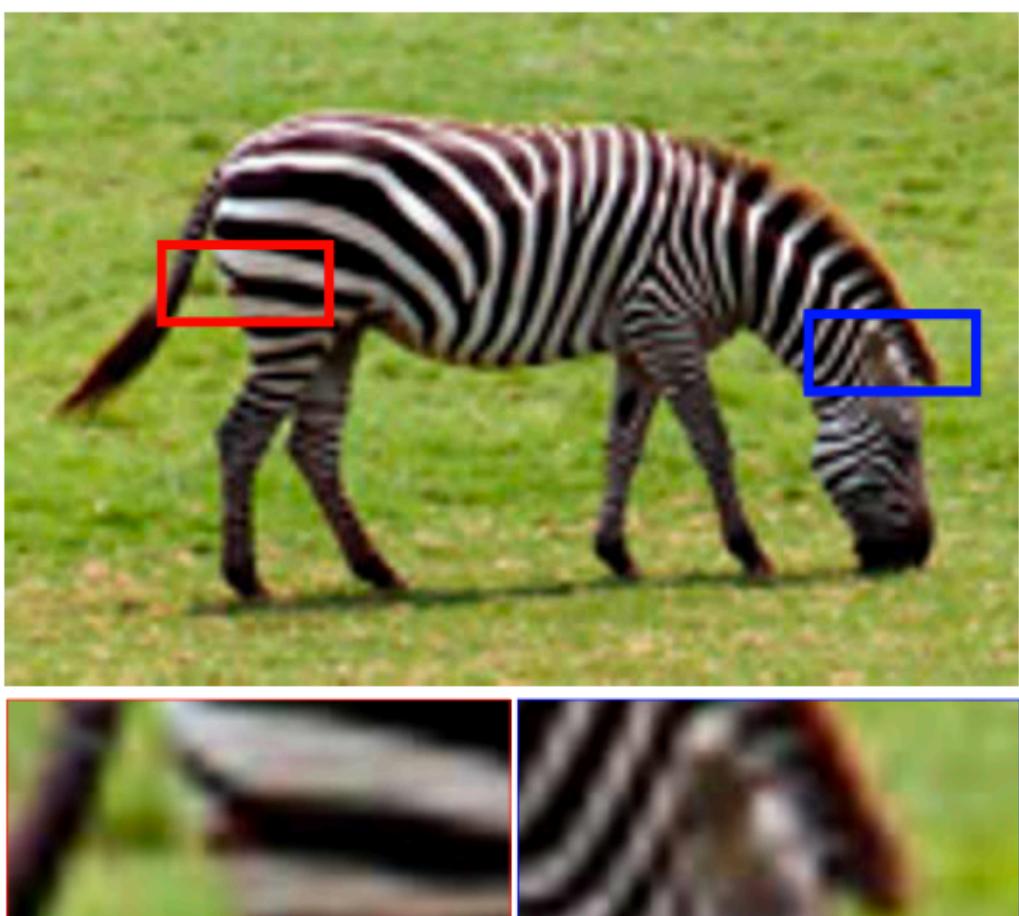
The network has a harder time overfitting noise than overfitting to natural images!



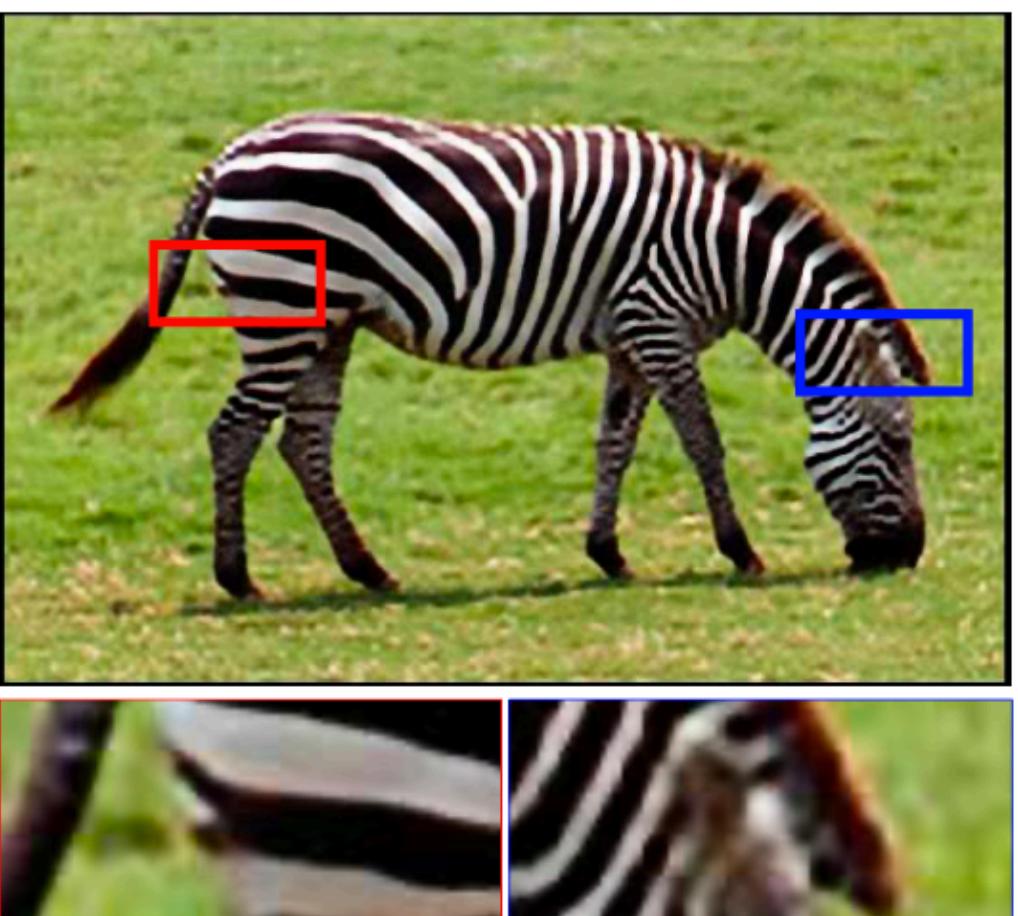
(a) Ground truth



(b) SRResNet [19], Trained



(c) Bicubic, Not trained



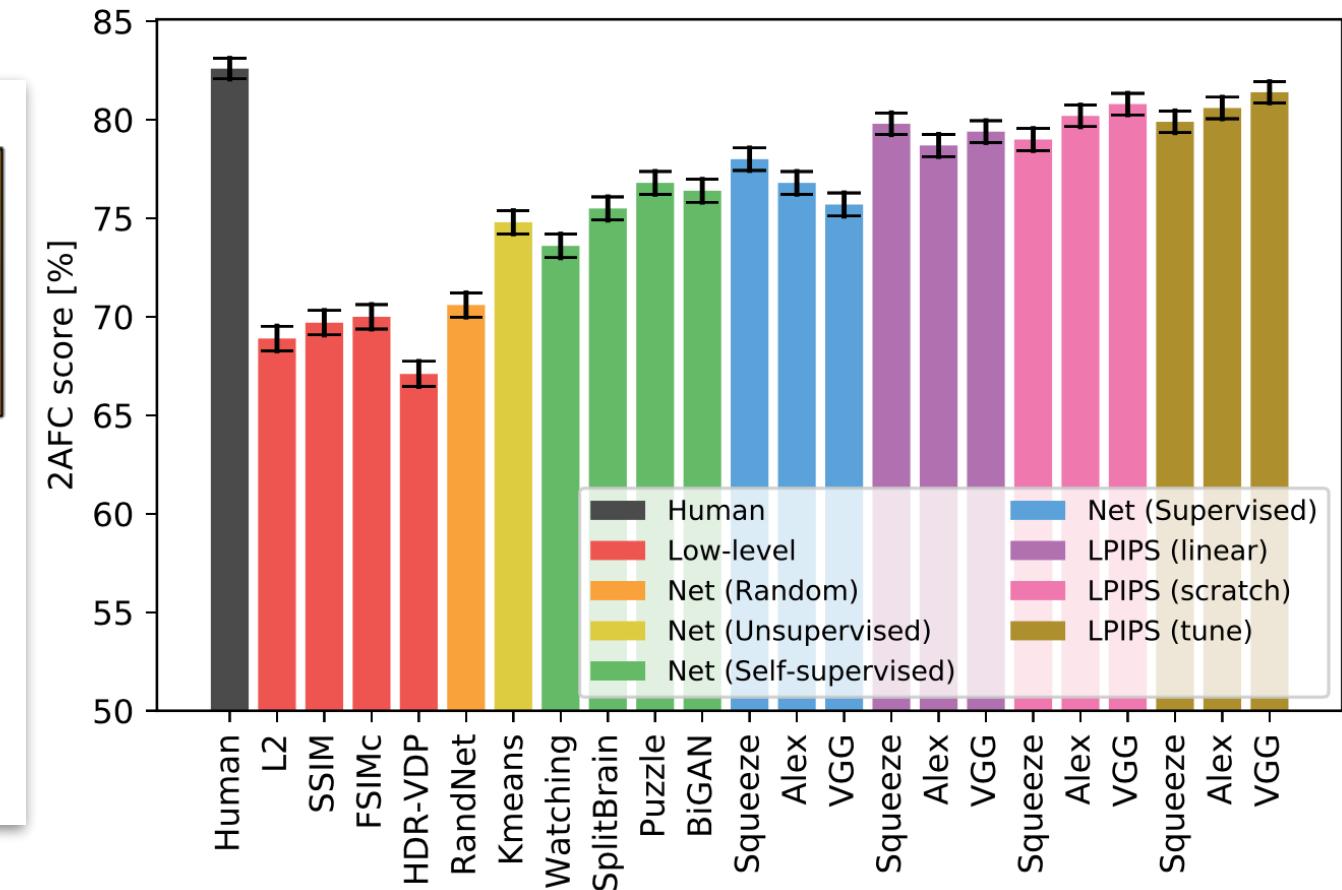
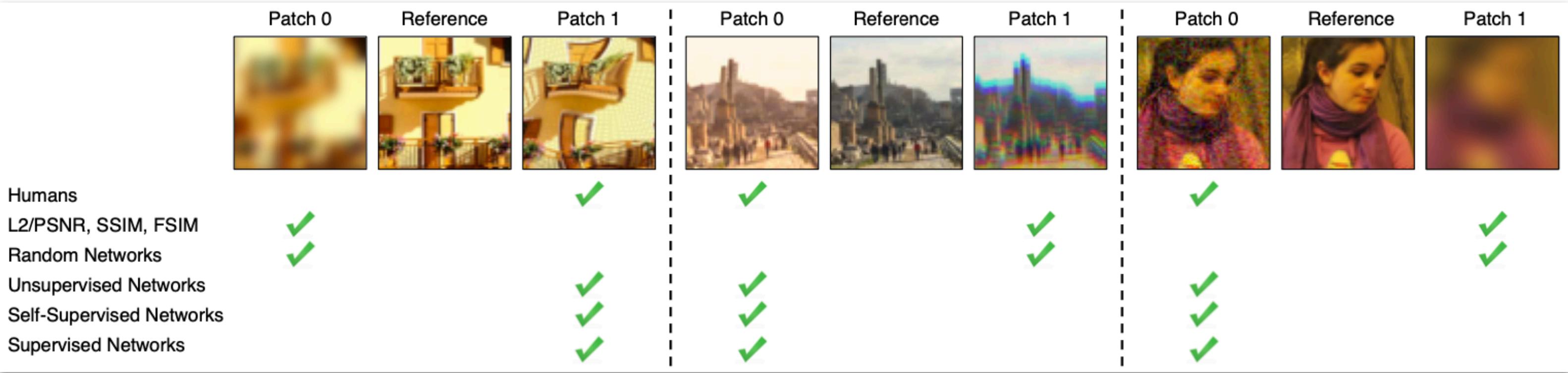
(d) Deep prior, Not trained



Boulder

The Unreasonable Effectiveness of Deep Features as a Perceptual Metric

Which patch (left or right) is “closer” to the middle patch in these examples?



Perceptual Distance: Is a red circle more similar to a red square or to a blue circle?

Berkeley-Adobe Perceptual Patch Similarity (BAPPS) Dataset

– Two Alternative Forced Choice (2AFC): Which of two distortions is more similar to a reference?

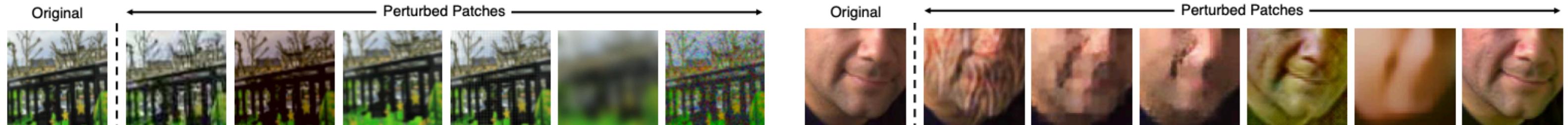
$x \rightarrow$ randomly selected image patch

$x_0, x_1 \rightarrow$ distorted patches

$h \in \{0, 1\} \rightarrow$ response of a human (which one is closer to the original patch?)

$\mathcal{T} \rightarrow$ dataset of patch triplets (x, x_0, x_1, h)

– Just Noticeable Difference (JND): Are two patches (one reference and one distorted) the same?

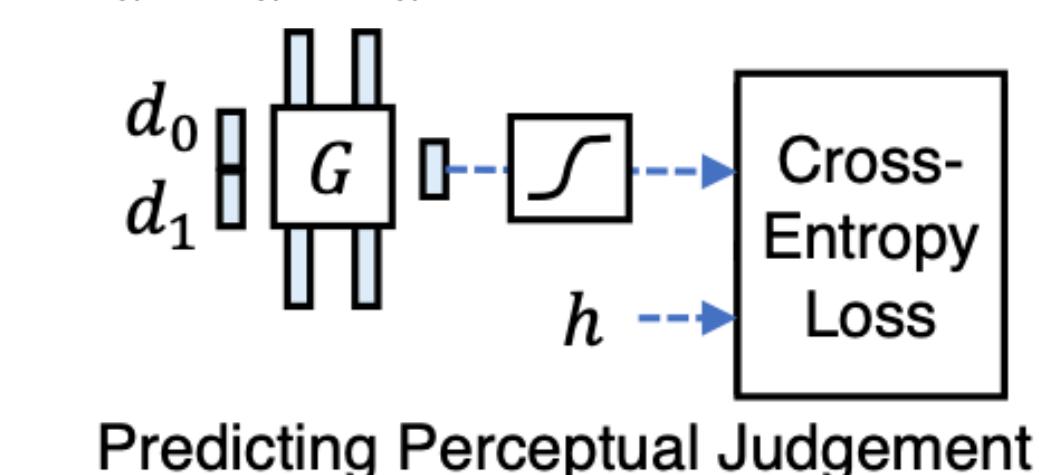
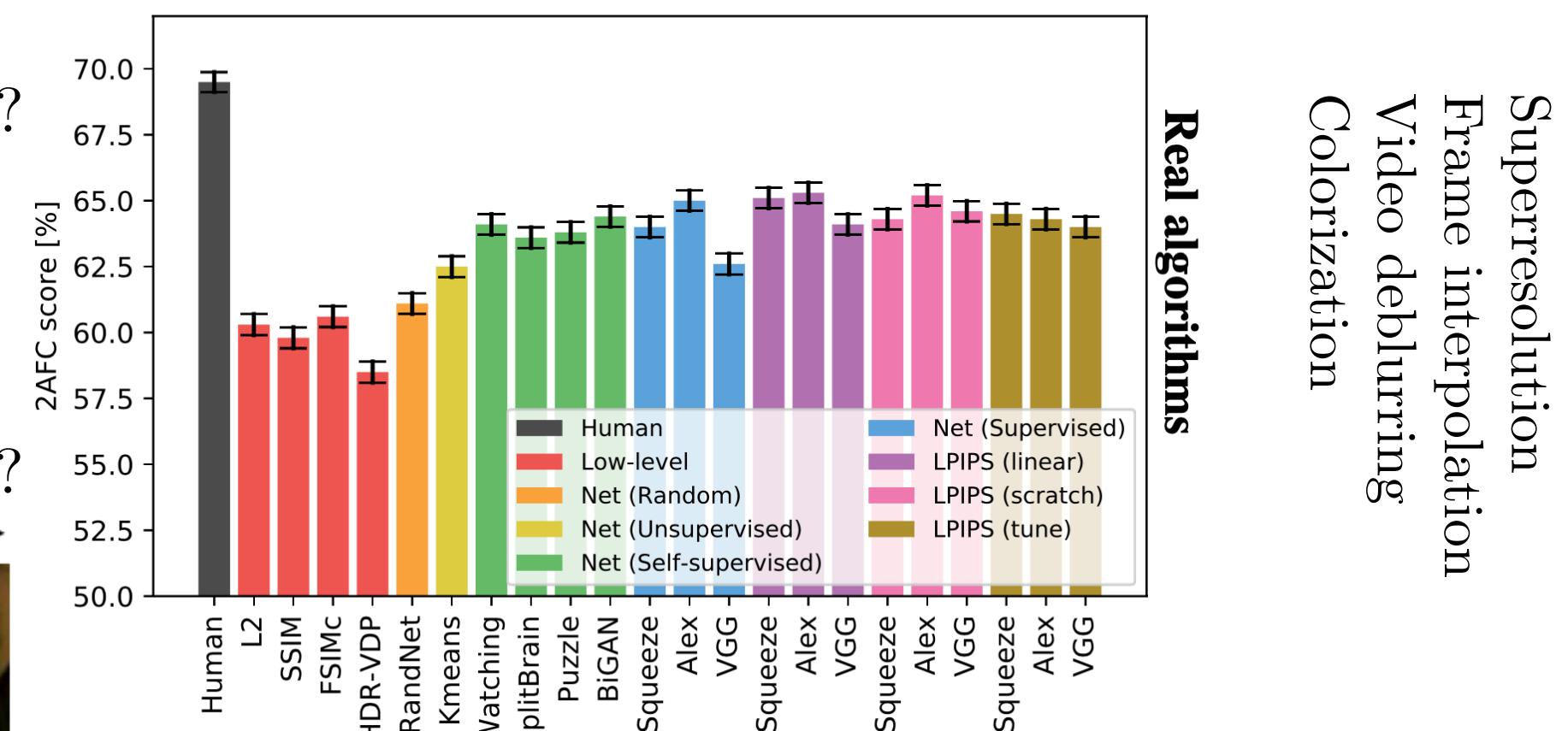


Learned Perceptual Image Patch Similarity (LPIPS) metric
Distance between reference and distorted patches x, x_0 with network \mathcal{F} :

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|w_l \odot (\hat{y}_{hw}^l - \hat{y}_{0hw}^l)\|_2^2 \quad \hat{y}^\ell, \hat{y}_0^\ell \in \mathbb{R}^{H_\ell \times W_\ell \times C_\ell} \text{ for layer } \ell \quad w^\ell \in \mathbb{R}^C \rightarrow \text{scaling factor}$$

$$\|\hat{y}^\ell\| = \|\hat{y}_0^\ell\| = 1 \rightarrow \text{unit normalized in the channel dimension}$$

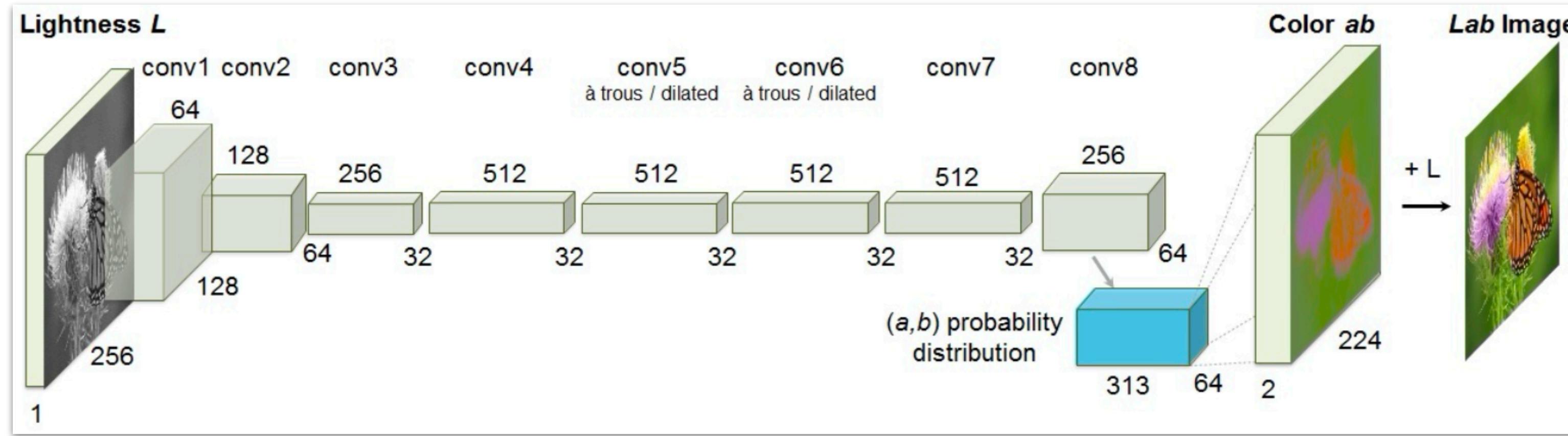
Zhang, Richard, et al. "The unreasonable effectiveness of deep features as a perceptual metric." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.





Boulder

Colorful Image Colorization



$X \in \mathbb{R}^{H \times W \times 1}$ → input lightness channel

$Y \in \mathbb{R}^{H \times W \times 2}$ → two associated color channels

$\hat{Y} = F(X) \in \mathbb{R}^{H \times W \times 2}$ → mapping (to be learned)

Distances in CIE Lab color space model perceptual distances

$$L_2(\hat{Y}, Y) = \frac{1}{2} \sum_{h,w} \|Y_{h,w} - \hat{Y}_{h,w}\|_2^2$$

Not robust to the inherent ambiguity and multimodal nature of the colorization problem.

For example, an apple is typically red, green, or yellow, but unlikely to be blue or orange.

Quantize the $ab \in [-110, 110]^2$ output space into bins with grid size 10 and keep the $Q = 313$ values which are in-gamut.

$\hat{Z} = G(X) \in [0, 1]^{H \times W \times Q}$ → probability distribution over colors

$Z = H_{gt}^{-1}(Y)$ → converts ground truth color Y to vector Z , using a soft-encoding scheme

Find the 5-nearest neighbors to $Y_{h,w}$ in the output space and weight them proportionally to their distance from the ground truth using a Gaussian kernel with $\sigma = 5$.

$$L_{cl}(\hat{Z}, Z) = - \sum_{h,w} v(Z_{h,w}) \sum_q Z_{h,w,q} \log(\hat{Z}_{h,w,q})$$

$v(Z_{h,w})$ → weighting term used to rebalance the loss based on color-class rarity

$$\hat{Y} = H(\hat{Z}) \rightarrow \text{class probabilities to point estimates}$$

$$H(Z_{h,w}) = \mathbb{E}[f_T(Z_{h,w})]$$

$$f_T(z) = \frac{\exp(\log(z)/T)}{\sum_q \exp(\log(z_q)/T)}$$

$$T = 0.38$$



Class Rebalancing

$p \in \Delta^Q$ → empirical probability of colors in the quantized ab space

$\tilde{p} \in \Delta^Q$ → smoothed empirical distribution

(smoothed version of p using Gaussian kernel G_σ)

$w \propto ((1 - \lambda)\tilde{p} + \lambda/Q)^{-1} \in \mathbb{R}^Q$ → weight factor

$1/Q$ → uniform distribution

$$\mathbb{E}[w] = \sum_q \tilde{p}_q w_q = 1 \rightarrow \text{normalization constraint}$$

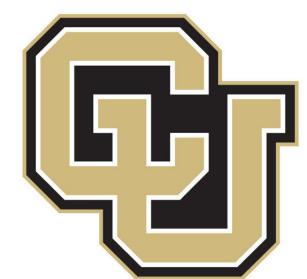
$$q^* = \arg \max_q Z_{h,w,q} \rightarrow \text{closest } ab \text{ bin}$$

$$v(Z_{h,w,q}) = w_{q^*}$$

Colorization as a pretext task for representation learning

Acting as a cross-channel encoder!

PASCAL classification, detection, and segmentation



Boulder



Questions?

[YouTube Playlist](#)
