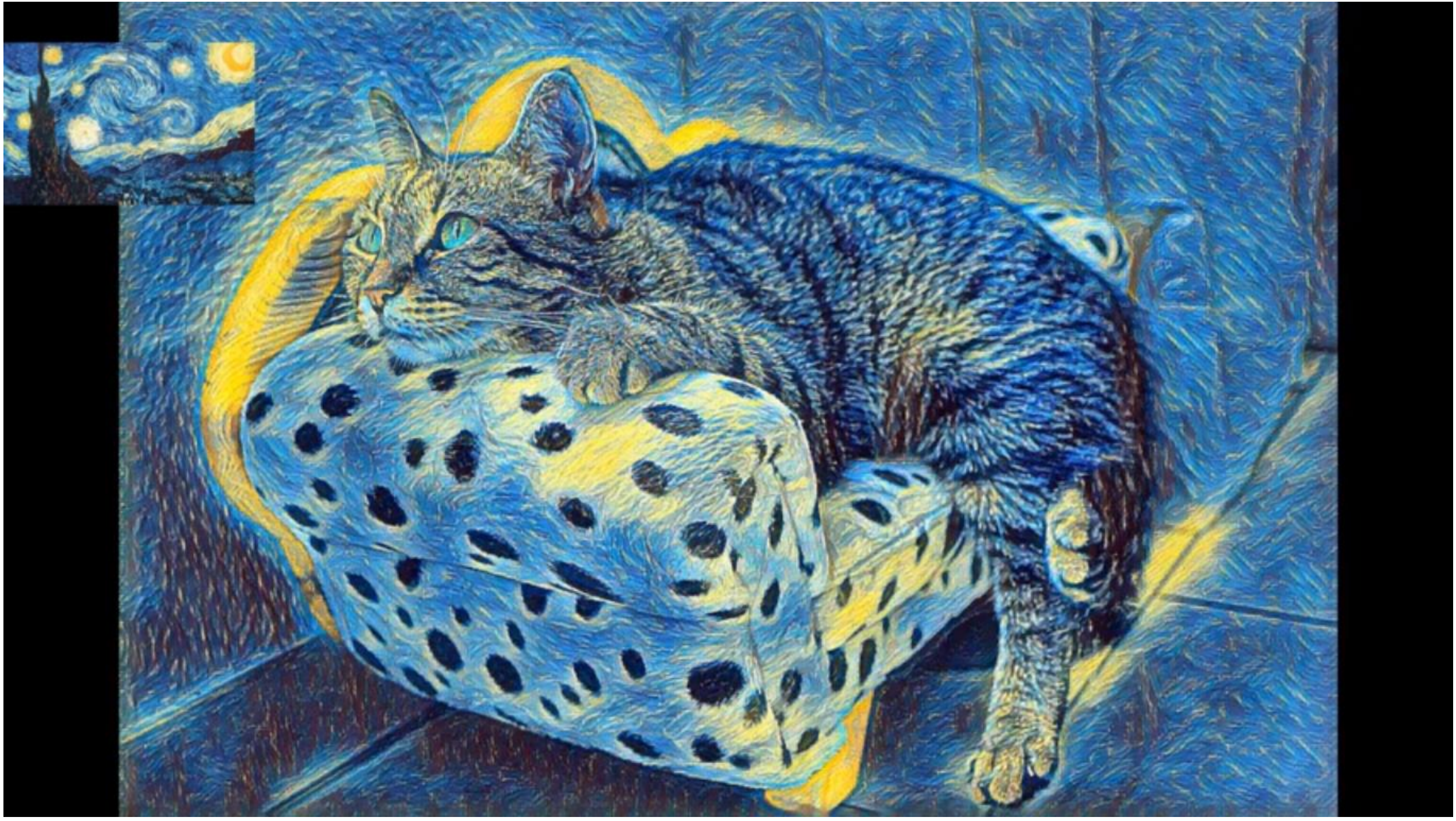
Perceptual Losses for Real-Time Style Transfer and Super-Resolution

Stanford, 2016

발표 최성욱



이미지 변환의 기존 방법들

	Per pixel loss function	Perceptual loss function
최적화	Low-level pixel 정보 간의 차이 (Output vs ground-truth)	High-level feature 간의 차이
구조	Feed-forward CNN	Forward CNN + Backward optimization
활용	Super-resolution, Segmentation	Image generation
단점	Perceptual difference 인식 못함	Computationally expensive

두 방법론의 장점을 결합

- Per-pixel loss 아닌 perceptual loss function 이용

&

- Feed-forward 방식으로 quick approximation 가능

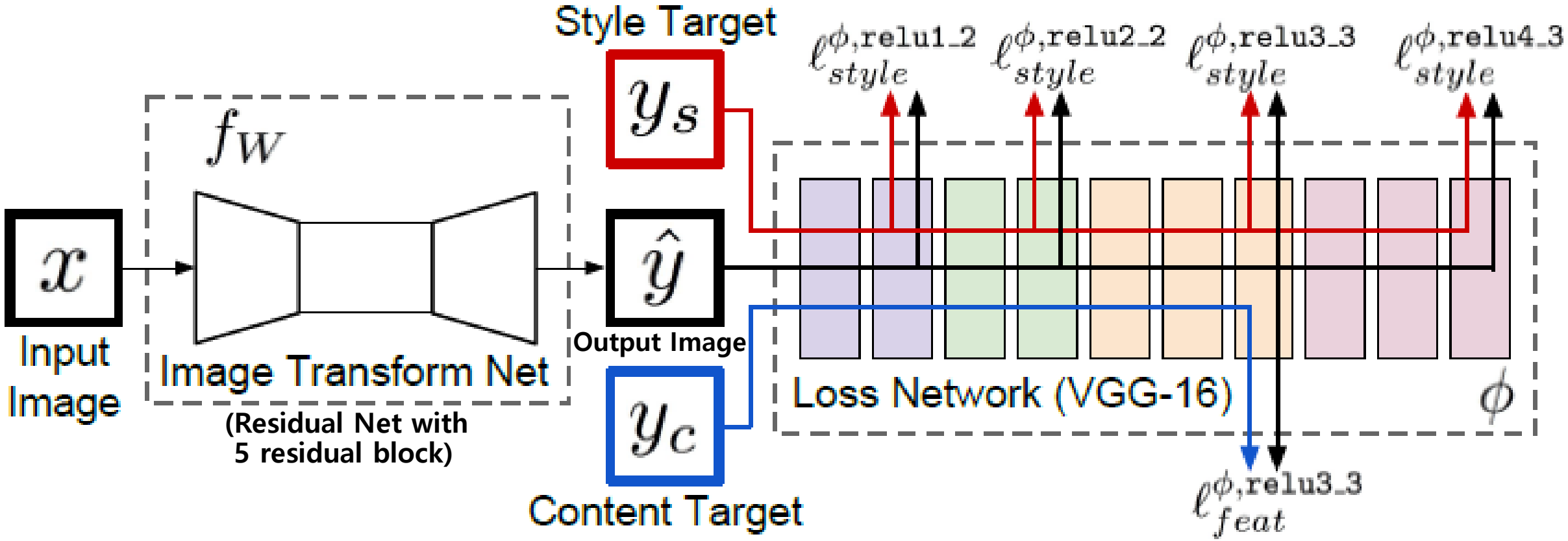
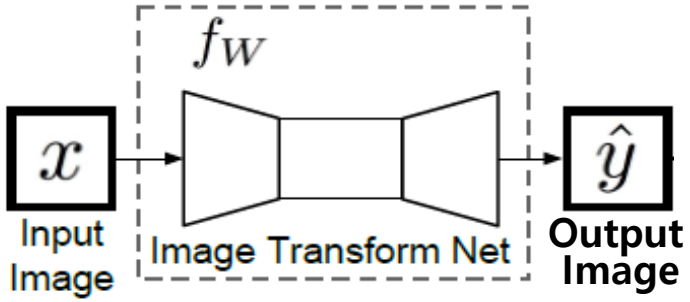

Style Target
y_s
$\ell^{\phi,\text{relu1\_2}}_{style}$
$\ell^{\phi,\text{relu2\_2}}_{style}$
$\ell^{\phi,\text{relu3\_3}}_{style}$
$\ell^{\phi,\text{relu4\_3}}_{style}$
f_W
x
Input Image
Image Transform Net
(Residual Net with 5 residual block)
$\hat{y}$
Output Image
Loss Network (VGG-16)
ϕ
y_c
Content Target
$\ell^{\phi,\text{relu3\_3}}_{feat}$

Image transform network (f_W)



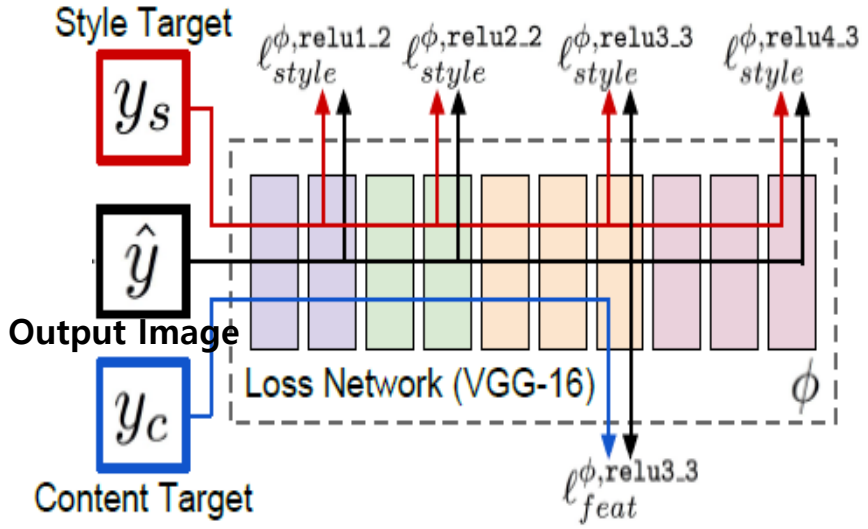
Deep residual CNN (parameter ; W)

Input image (x) 를 output image ($\hat{y}$) 로 변환시킴 ($\hat{y} = f_W(x)$)

$$W^* = \arg\min_W \mathbf{E}_{x,\{y_i\}} \left[\sum_{i=1} \lambda_i \ell_i(f_W(x), y_i) \right]$$

* loss function $l_i(\hat{y},y_i)$ = output image $\hat{y}$ 과 target image y_i 의 feature vector(또는 matrix) 의 차이

Loss network (Φ)



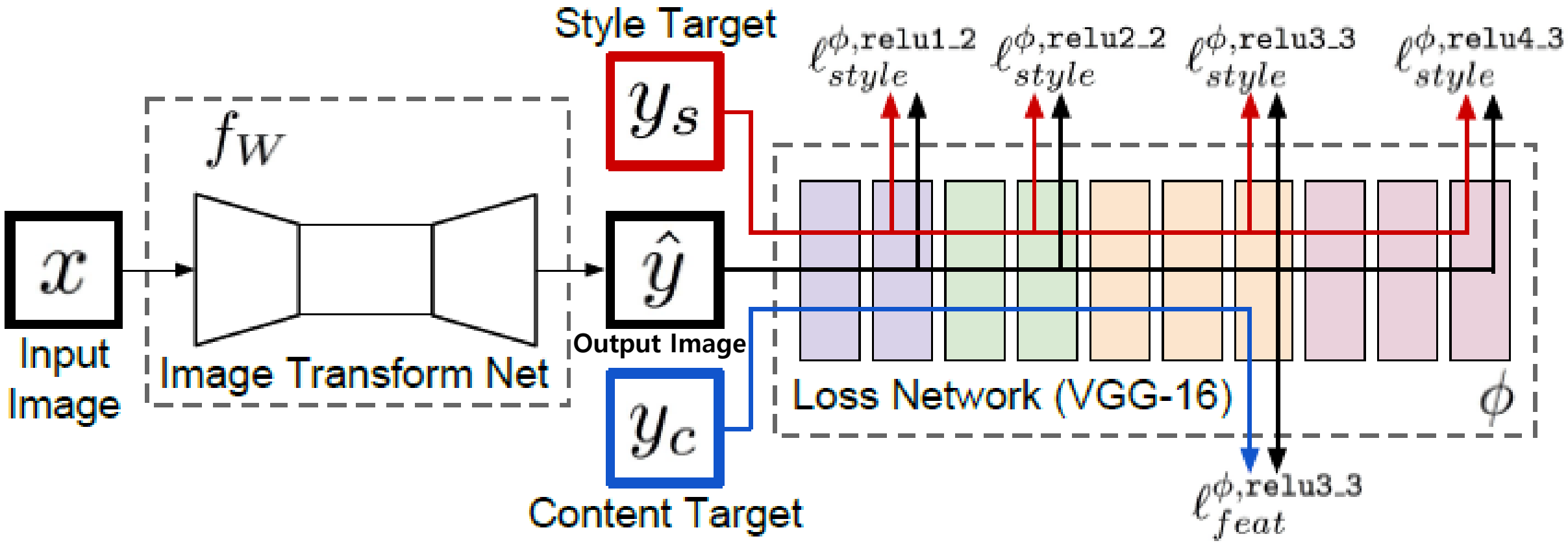
CNN(VGG-16), pre-trained for image classification
→ Perceptual & semantic information 을 측정

$\Phi_j(y)$ = Loss network Φ 가 image y 를 process 할 때,
j-th layer 에서 발생하는 feature vector

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

$$\ell^{\phi,j}_{style}(\hat{y}, y) = \|G^{\phi}_j(\hat{y}) - G^{\phi}_j(y)\|_F^2$$

$$\left(\quad G^{\phi}_j(x)_{c,c'} = \frac{1}{C_j H_j W_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} \phi_j(x)_{h,w,c} \phi_j(x)_{h,w,c'} \right)$$



$$W^* = \arg\min_W \mathbf{E}_{x,\{y_i\}} \left[\sum_{i=1} \lambda_i \ell_i(\underbrace{f_W(x)}_{=\hat{y}}, y_i) \right]$$

적용

- Style transfer & Super-resolution

  - 정답이 있는 문제가 아니기 때문에 (ill-posed),
    input image 에 대한 semantic reasoning 이 필요한 task

  - Perceptual loss function 쓰면 semantic knowledge 전달 가능!

Task 에 따른 target 설정

- Style transfer 의 경우

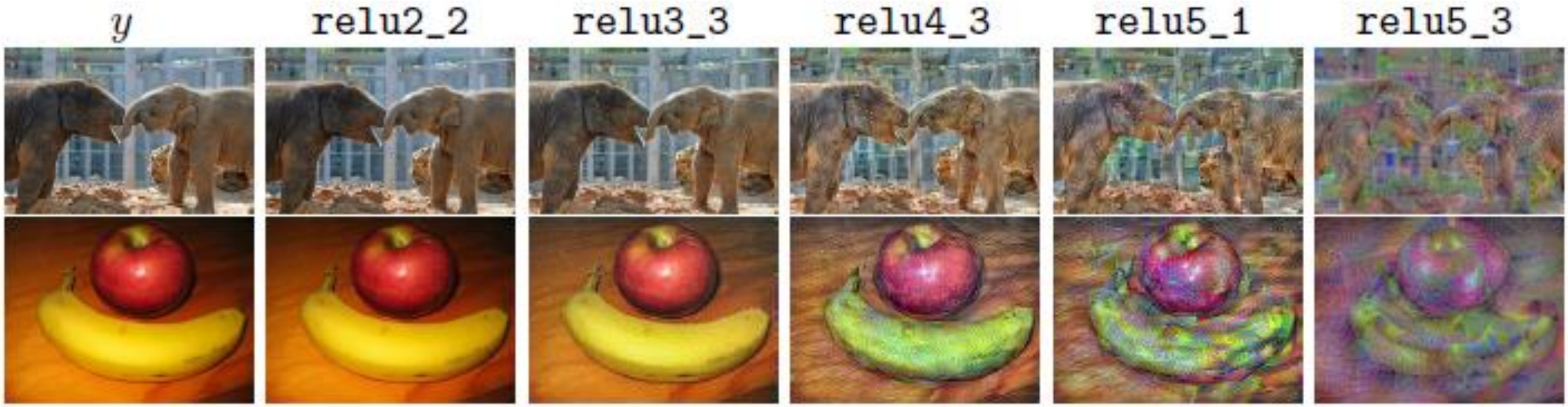
  ① content target y_c = input image x ② style target = y_s

  (output image $\hat{y}$ 은 그 둘을 combine 한 결과)

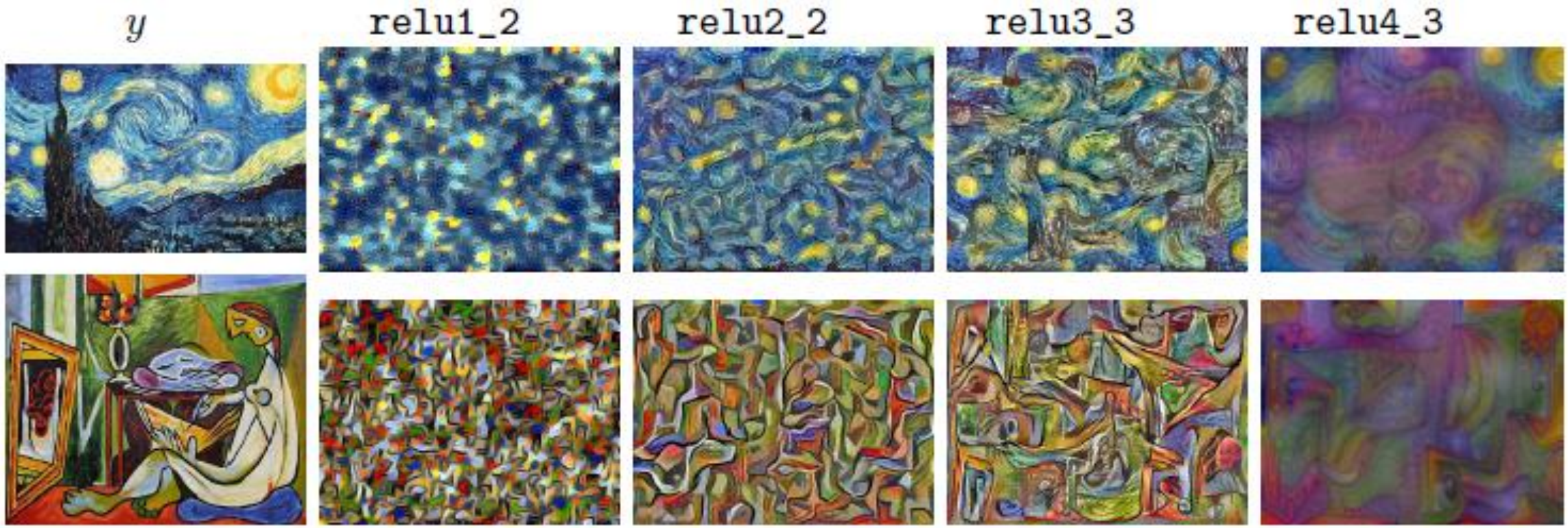
- Super-resolution 의 경우

  ① input image x 는 low-resolution image ,

  ② y_c 는 ground-truth high-resolution image (y_s 는 안 쓰임)



Higher layer 에서 reconstruct 한 것일수록 (오른쪽)
원래의 pixel 정보 많이 일어버리기 때문에 blurring 심해진다


y
relu1_2
relu2_2
relu3_3
relu4_3

Experiments

1. Style transfer

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

output image $\hat{y}$ (=fw(x))를 white noise image 로 초기화 후 optimize
→ loss 줄이는 방향으로 W 형성되며 image 나타남

- Training details

  - MS-COCO dataset
  - 80k training images resized to 256x256 patches.
  - Batch size: 4
  - 40k iterations (~2 epochs)
  - Optimizer used: Adam
  - Learning rate: 1 x 10^{-3}
  - Training takes ~4 hours on Titan X GPU

Style
The Muse,
Pablo Picasso,
1935

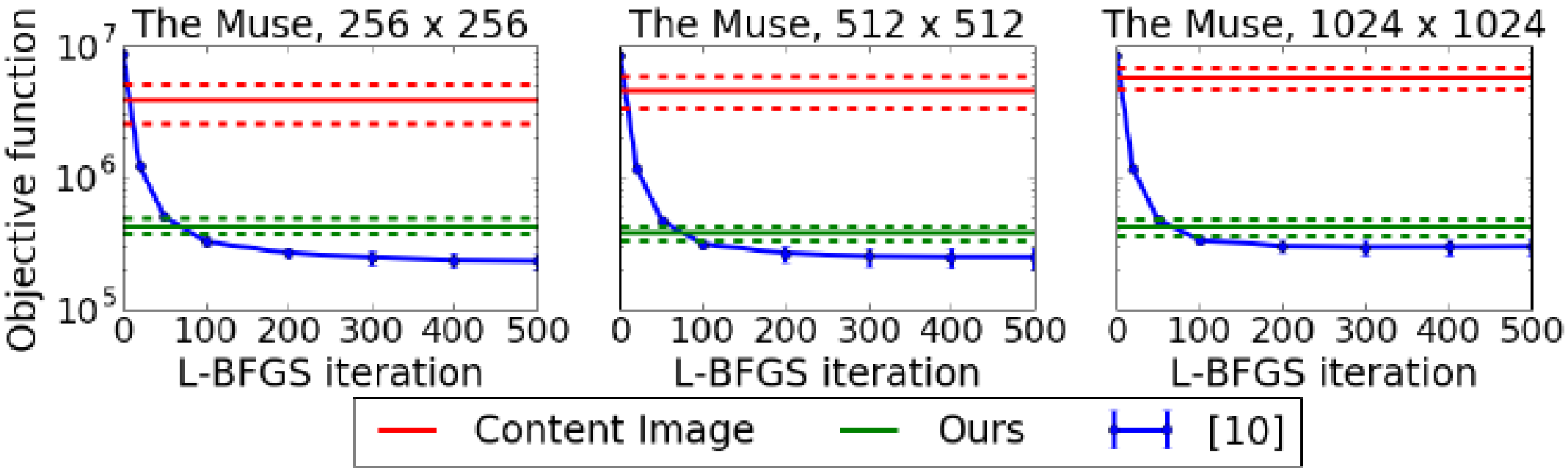

The Muse, 256 x 256
The Muse, 512 x 512
The Muse, 1024 x 1024
Objective function
10^7
10^6
10^5
0 100 200 300 400 500
L-BFGS iteration
Content Image
Ours
[10]

Image Size	Gatys *et al.* [11] 100	300	500	Ours	Speedup 100	300	500
256×256	3.17	9.52s	15.86s	**0.015s**	212x	636x	**1060x**
512×512	10.97	32.91s	54.85s	**0.05s**	205x	615x	**1026x**
1024×1024	42.89	128.66s	214.44s	**0.21s**	208x	625x	**1042x**

2. Single-Image Super-Resolution

Issues to be defined

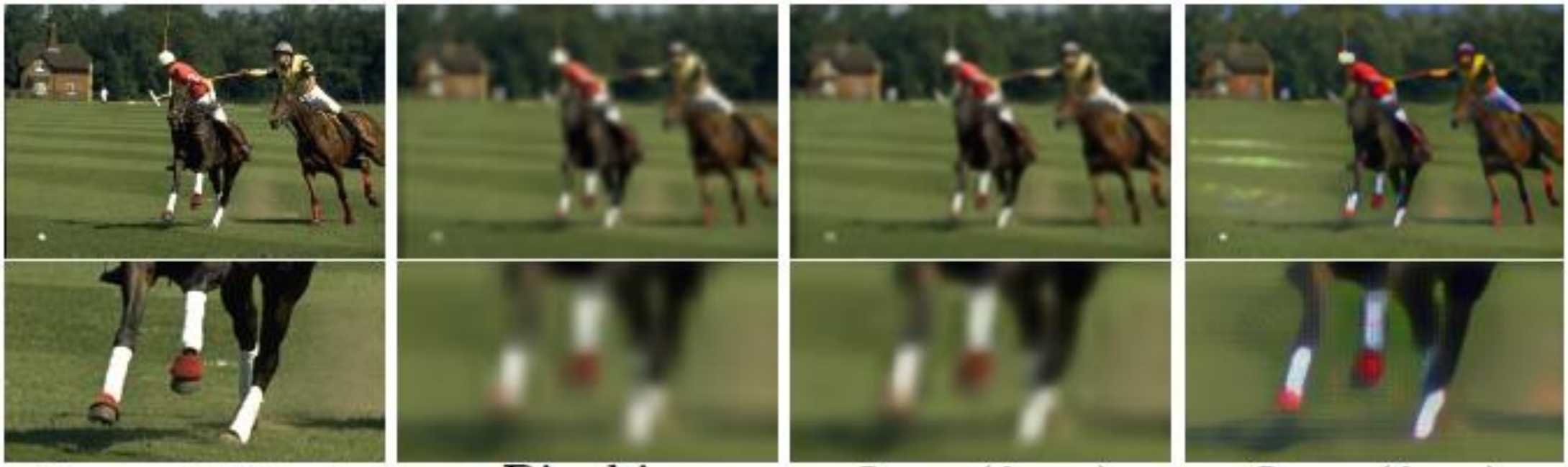
① ill-posed problem 이므로 semantic information 이 중요
∴ per-pixel loss 대신 Feature reconstruction loss 사용

② PSNR or SSIM 대신 Qualitative metrics 이용해서 평가

- Training details

  - 288x288 patches from 10k images from the MS-COCO
  - Prepared low-resolution inputs by blurring with a Gaussian kernel (width σ=1.0) and downsampling with bicubic interpolation.
  - Batch size: 4
  - Iterations: 200k
  - Optimizer: Adam
  - Learning rate: 1 x 10^{-3}
  - Compared against Super-resolution CNN

Ground Truth	Bicubic	Ours (ℓ_{pixel})	SRCNN [13]	Ours (ℓ_{feat})
This image	31.78 / 0.8577	31.47 / 0.8573	32.99 / 0.8784	29.24 / 0.7841
Set5 mean	28.43 / 0.8114	28.40 / 0.8205	30.48 / 0.8628	27.09 / 0.7680



Ground Truth	Bicubic	Ours (ℓ_{pixel})	Ours (ℓ_{feat})
This image	22.75 / 0.5946	23.42 / 0.6168	21.90 / 0.6083
Set5 mean	23.80 / 0.6455	24.77 / 0.6864	23.26 / 0.7058
Set14 mean	22.37 / 0.5518	23.02 / 0.5787	21.64 / 0.5837
BSD100 mean	22.11 / 0.5322	22.54 / 0.5526	21.35 / 0.5474

결론

- Feed-forward image transformation network combined with Perceptual loss function

  - Style transfer) Similar quality, but much faster speed

  - Super-resolution) Better reconstruction (edge, details)