Halftone Based Adversarial Example

蕭擎軒

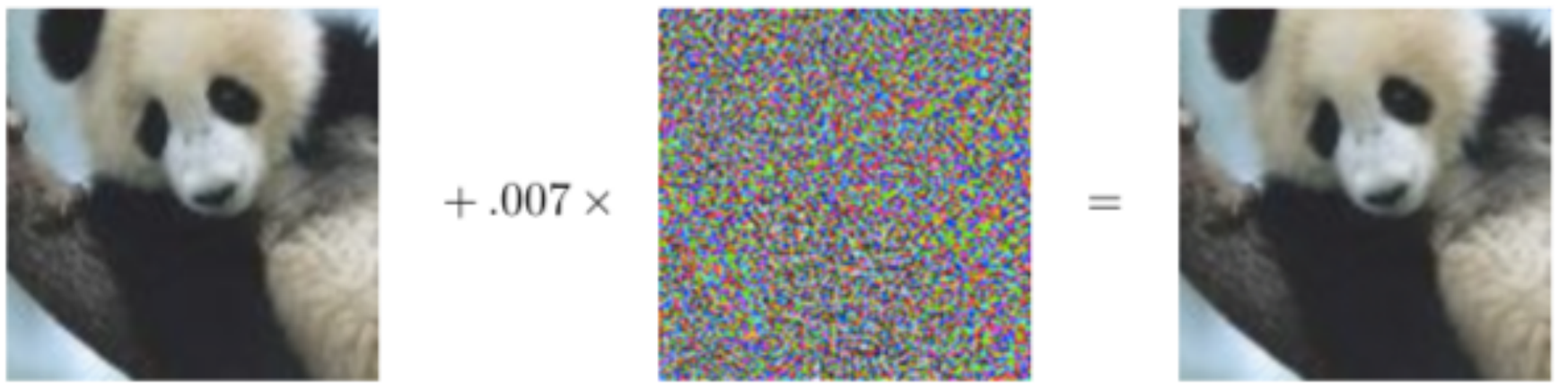
2018/08/21

Outline

1. Recall: What is Adversarial Example?
2. Introduction: Halftone Based Adversarial Example
3. CVPR 2016: Image Style Transfer Using Convolutional Neural Networks
4. Experiment: Halftone MNIST
5. Application: Reinforce CAPTCHA
6. Conclusion and Future Work

What is Adversarial Example?

- Adversarial examples are **recognizable by human** but can **fool deep neural networks** in the testing stage.



$\boldsymbol{x}$

"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

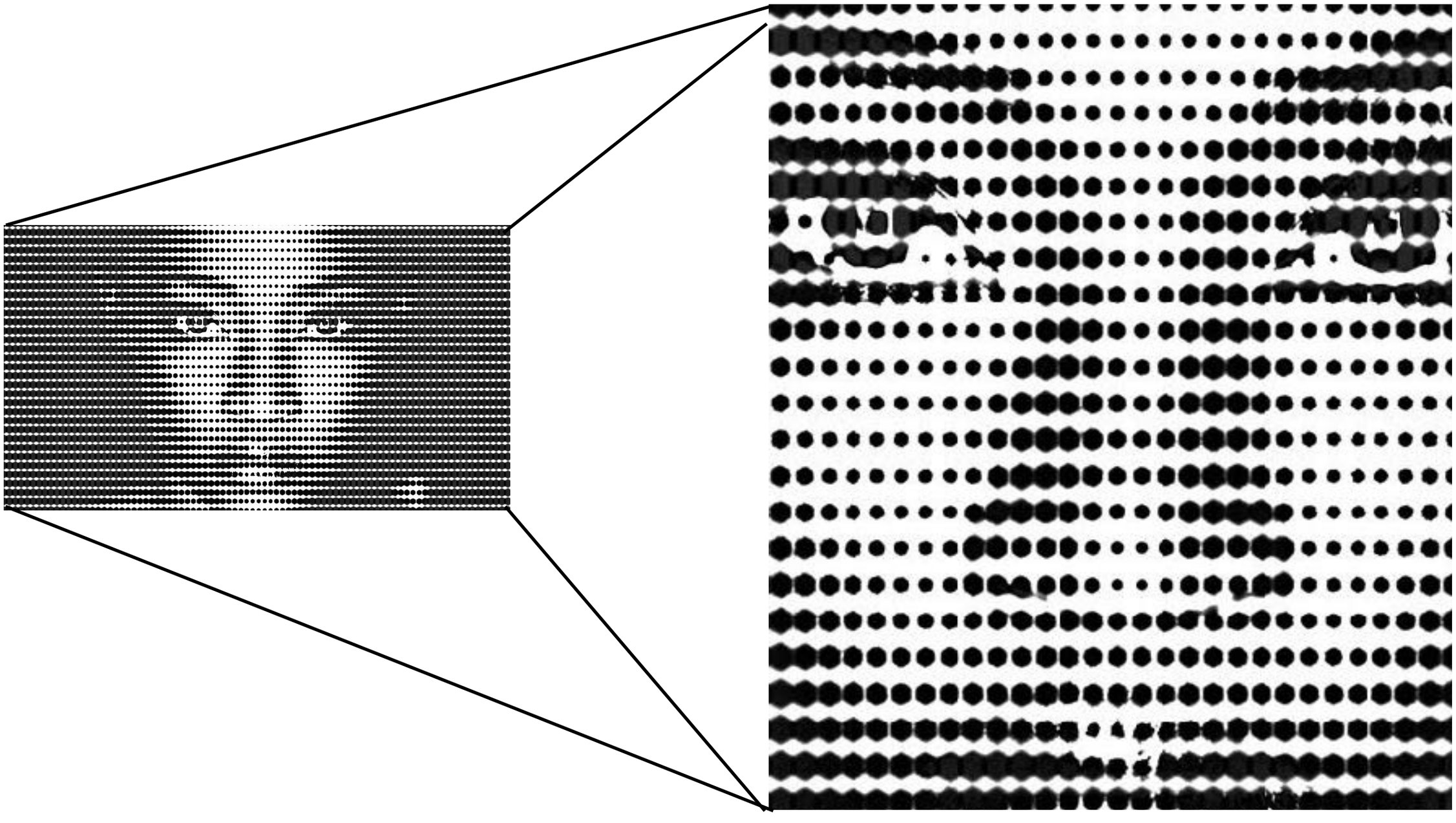
What is Adversarial Example?

$f(x) = 3$

$g(x) = x'$

$f(x') = 2$

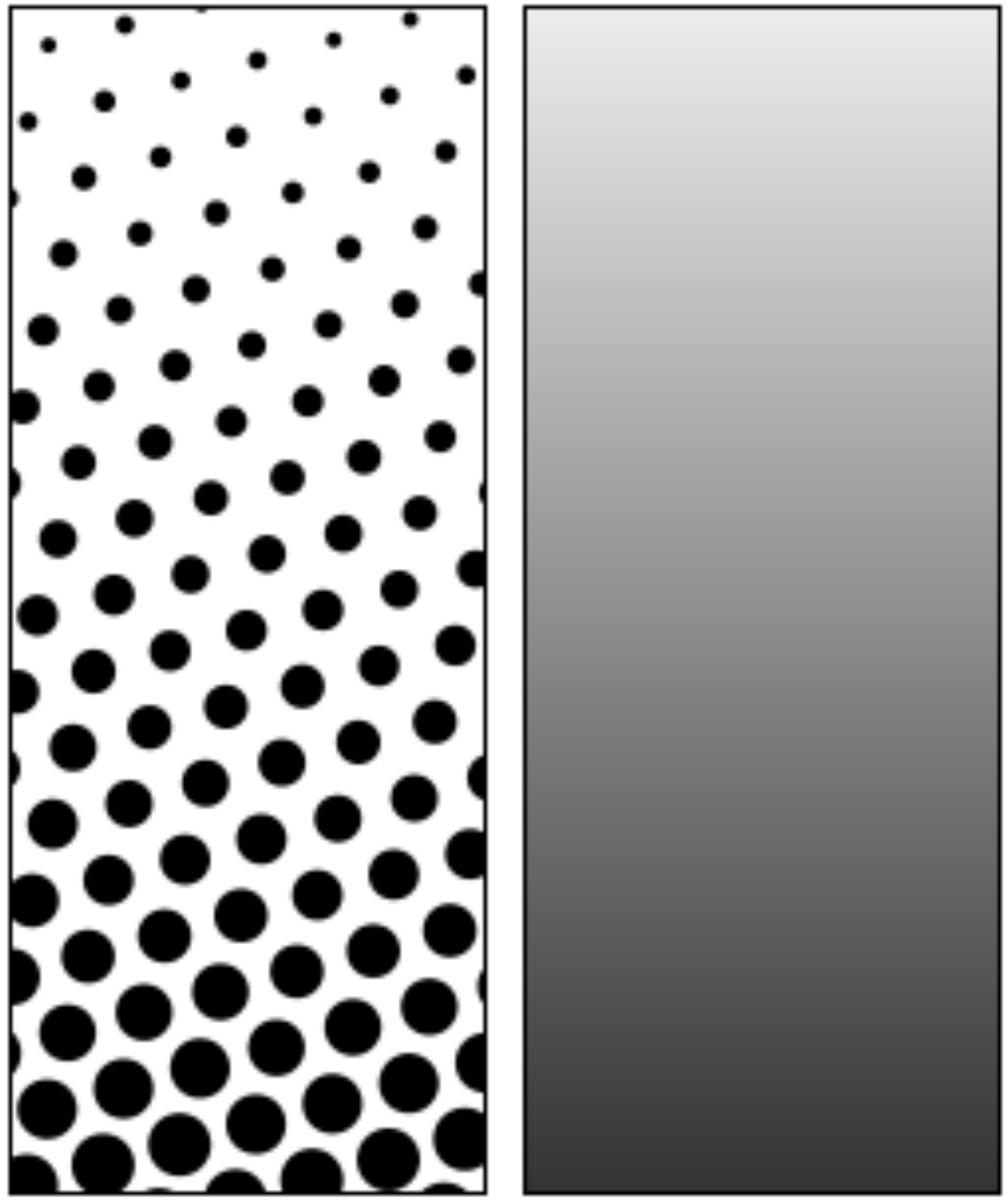
Halftone Based Adversarial Example



What is Halftone?

- Halftone is the reprographic technique that simulates continuous tone imagery through the use of dots, varying either in size or in spacing, thus generating a **gradient-like effect**.

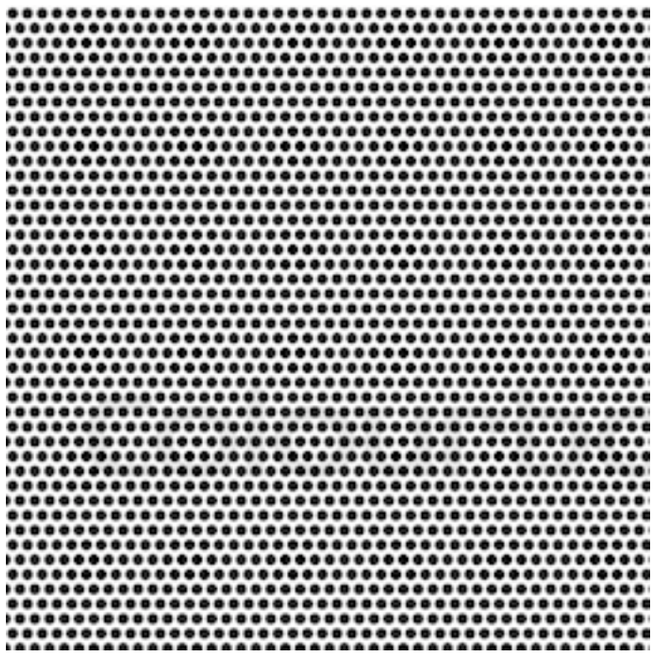
- This reproduction relies on a basic **optical illusion**: the tiny halftone dots are blended into smooth tones by the human eye.**（視覺積分）**



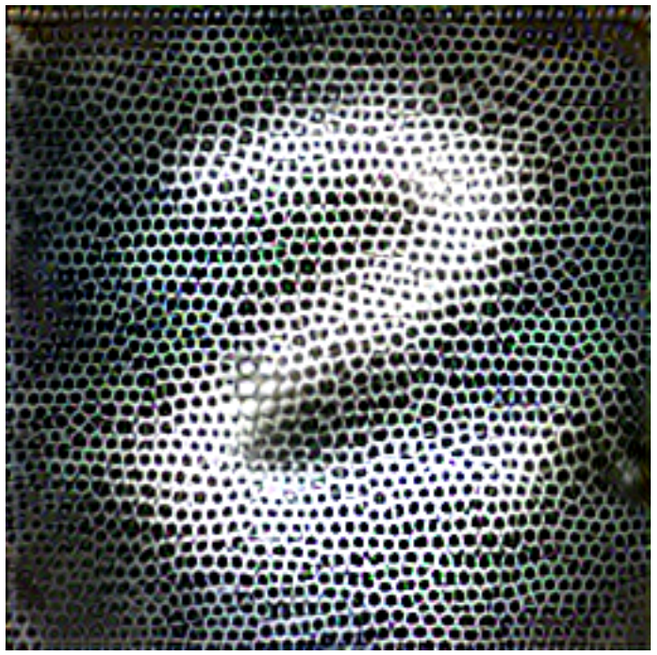
離散 **連續**

Does CNN Have Optical Illusion?

+



=



$f(x) = 2$

$g(x) = x'$

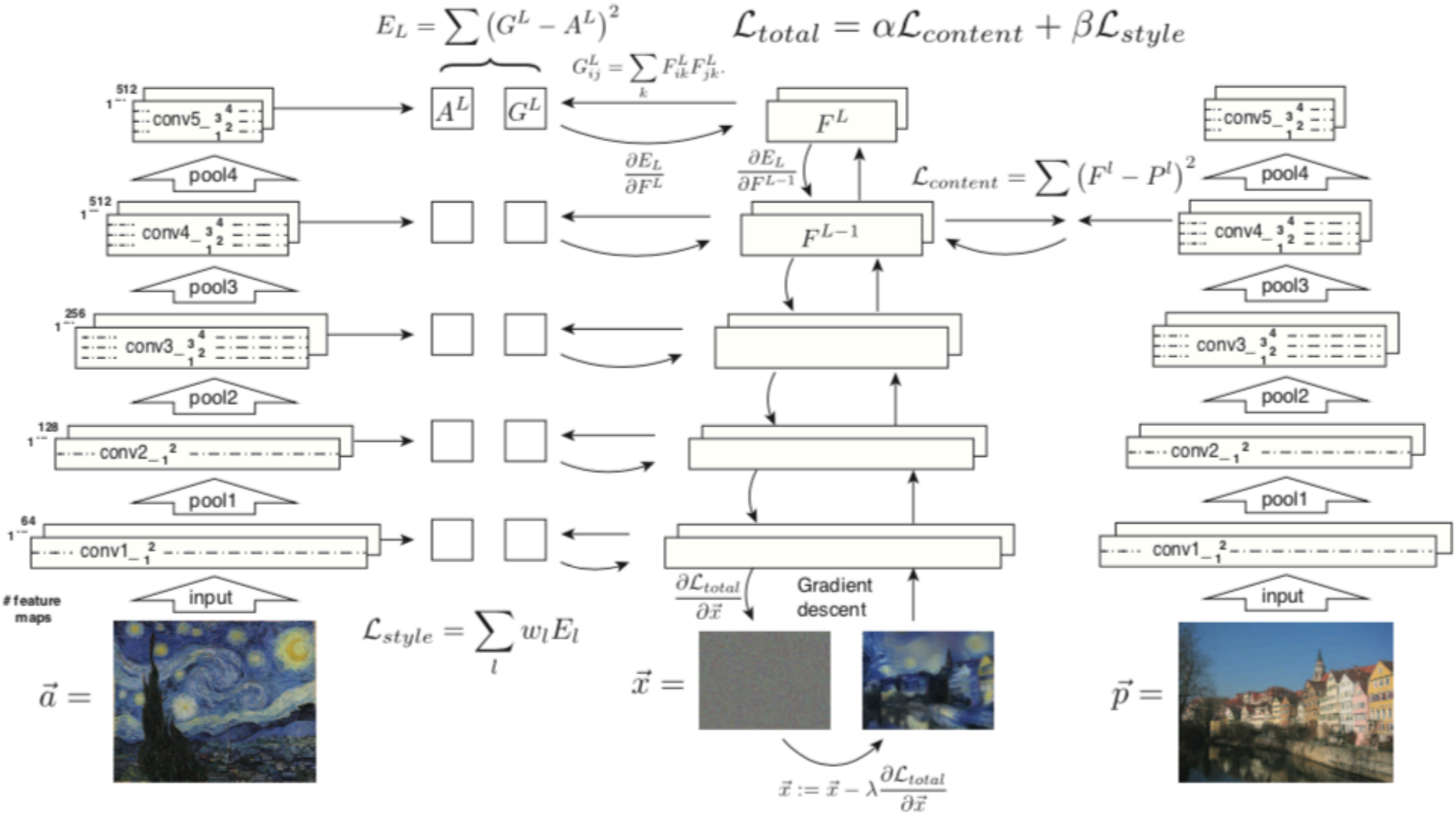
$f(x') = ?$

Image Style Transfer Using Convolutional Neural Networks

Leon A. Gatys[1]
Alexander S. Ecker[2]
Matthias Bethge[3]

CVPR 2016 (ORAL)

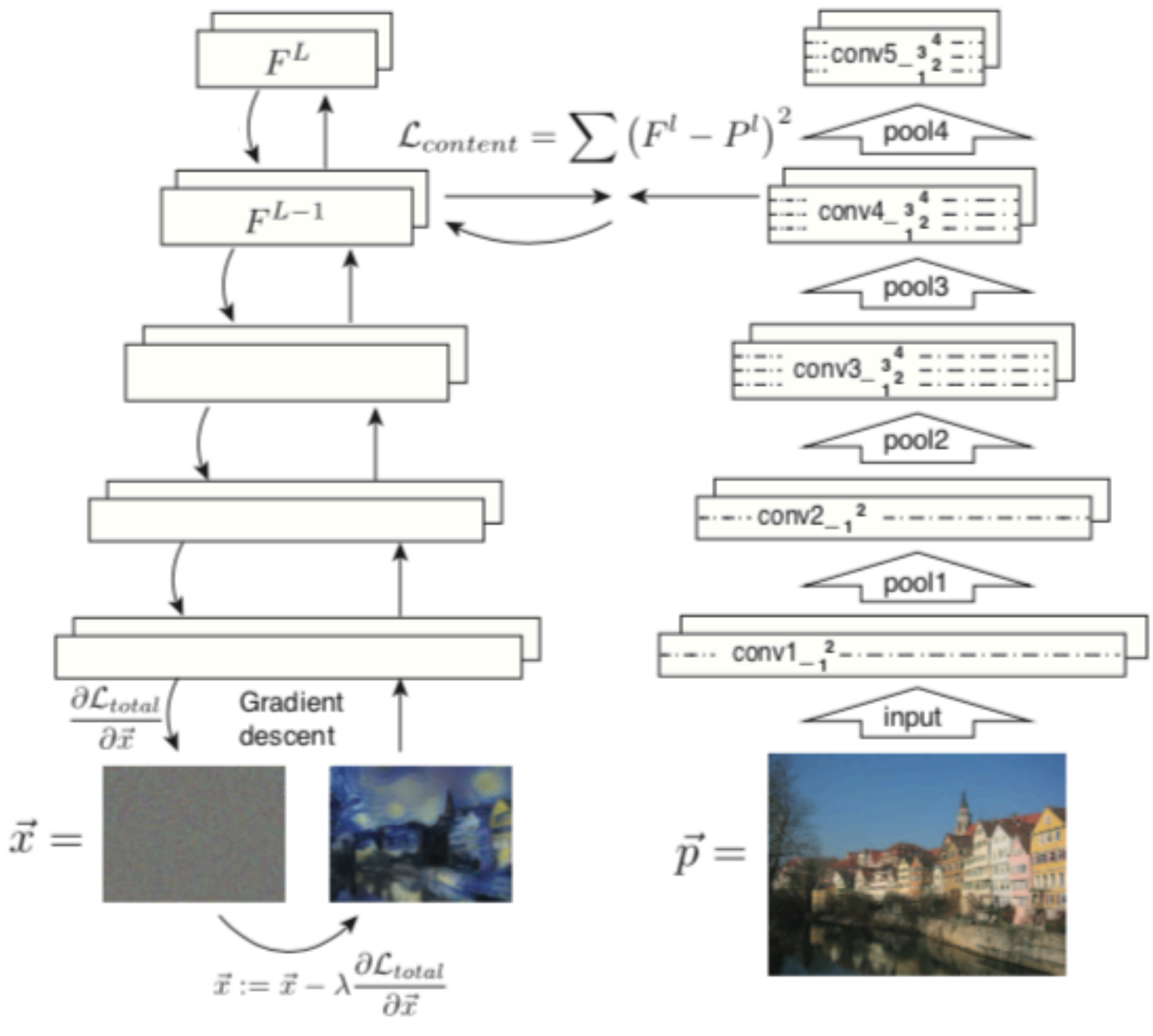
Architecture



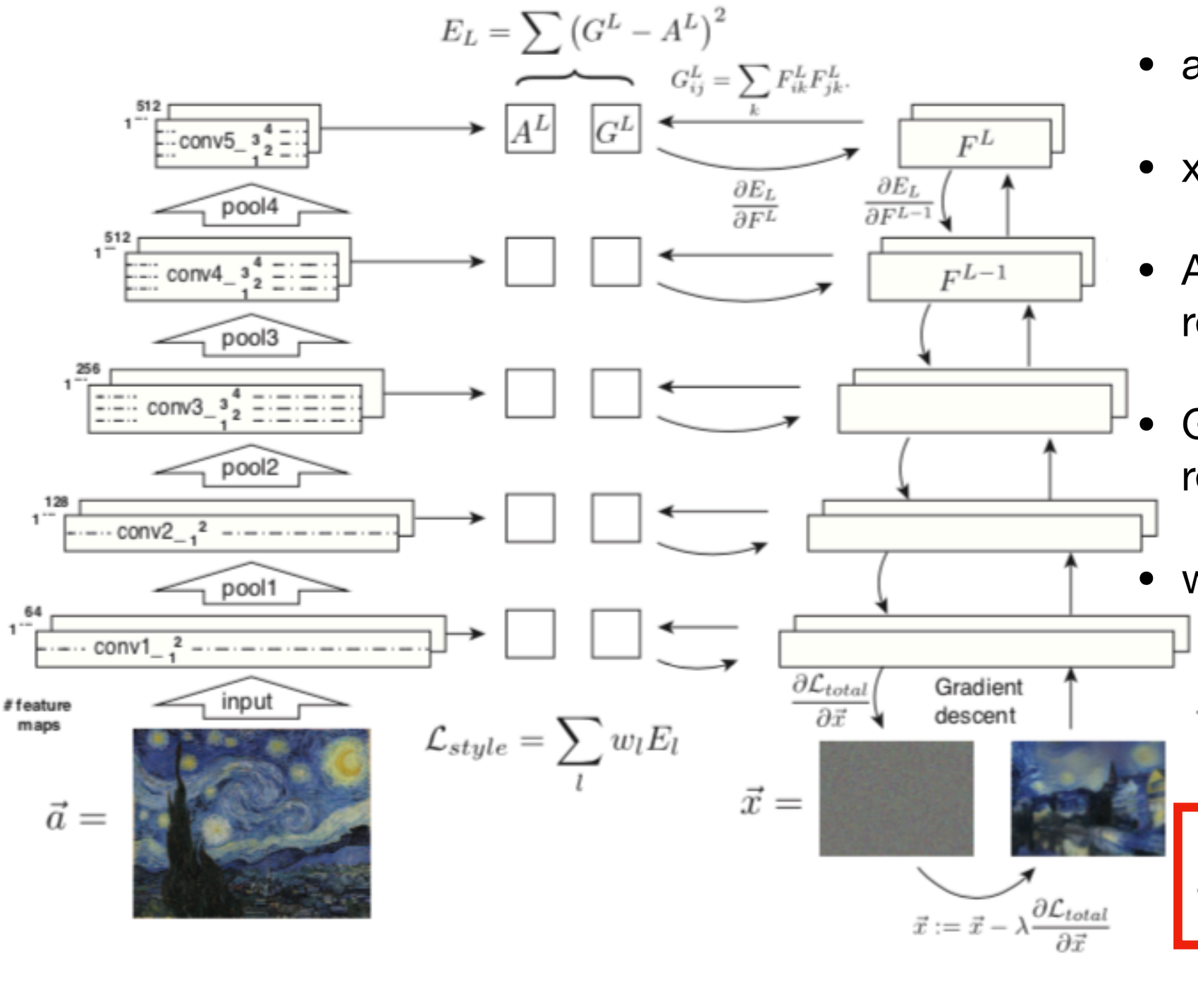
1. Square Loss of Content

- p: original image
- x: generated image
- P^L: original image's feature representation in layer i
- F^L: generated image's feature representation in layer i

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



2. Square Loss of Style



- a: original image
- x: generated image
- A^L: original image's feature representation in layer i
- G^L: generated image's feature representation in layer i
- w_L: weighting factors in layer i

$$E_l = \frac{1}{4N_l^2 M_l^2} \sum_{i,j} \left(G_{ij}^l - A_{ij}^l\right)^2$$

$$\mathcal{L}_{\text{style}}(\vec{a}, \vec{x}) = \sum_{l=0}^{L} w_l E_l$$

3. Square Loss of Total

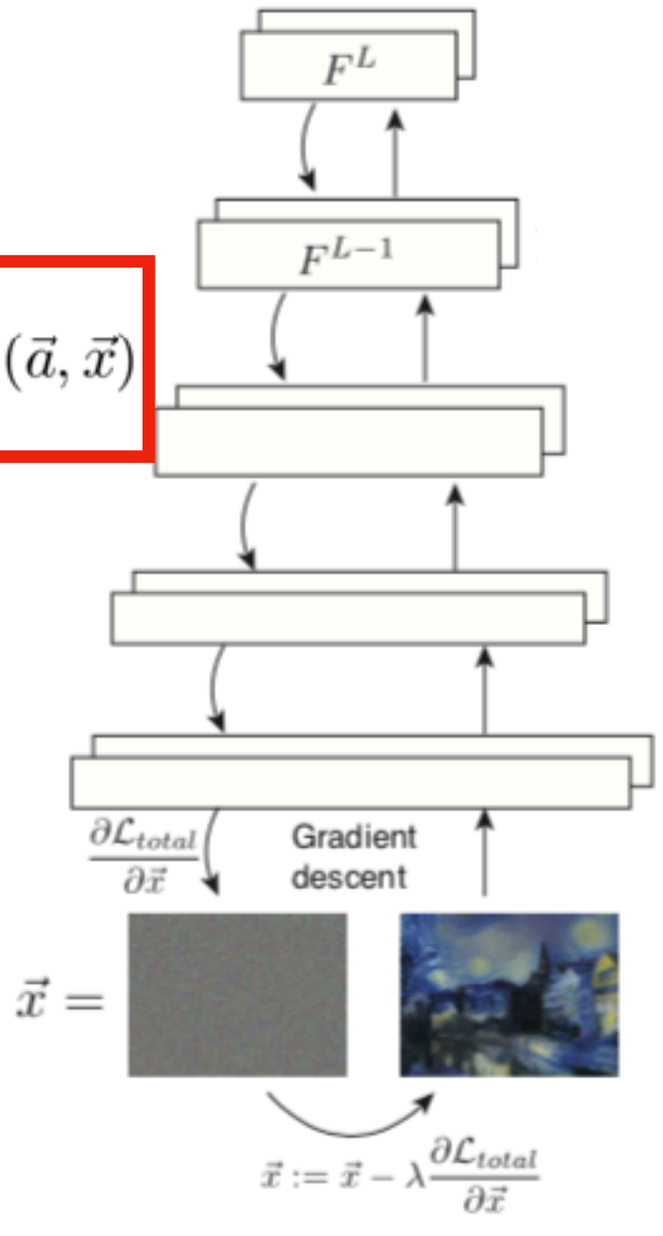
α : **weighting factors for content**

β : **weighting factors for style**

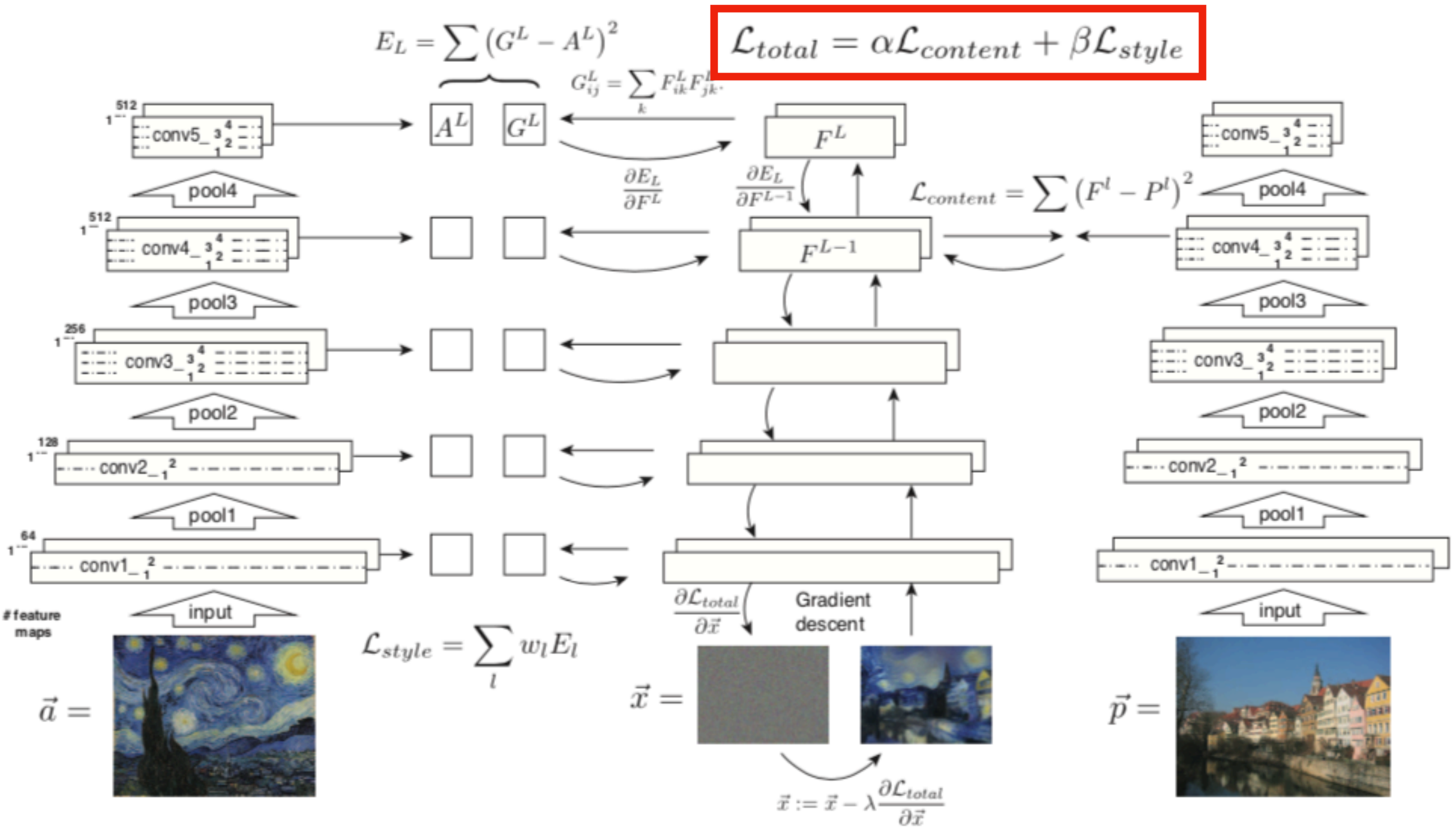
$$\mathcal{L}_{total} = \alpha\mathcal{L}_{content} + \beta\mathcal{L}_{style}$$

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

$$\vec{x} := \vec{x} - \lambda\frac{\partial\mathcal{L}_{total}}{\partial\vec{x}}$$



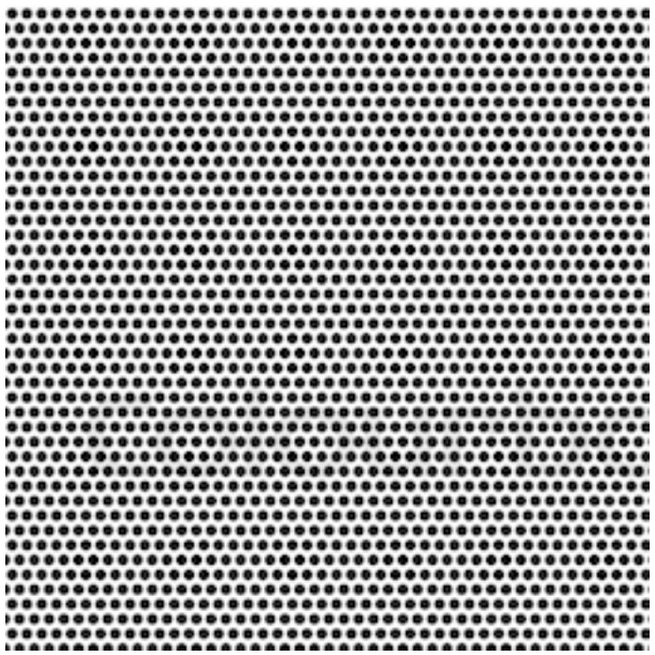
Architecture



Content / Style Weight

Content

α **Content** + β **Style**

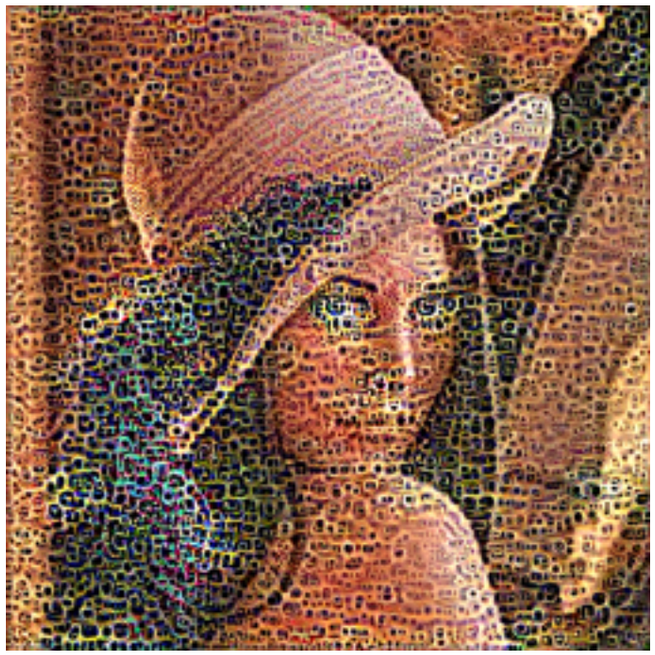


Style

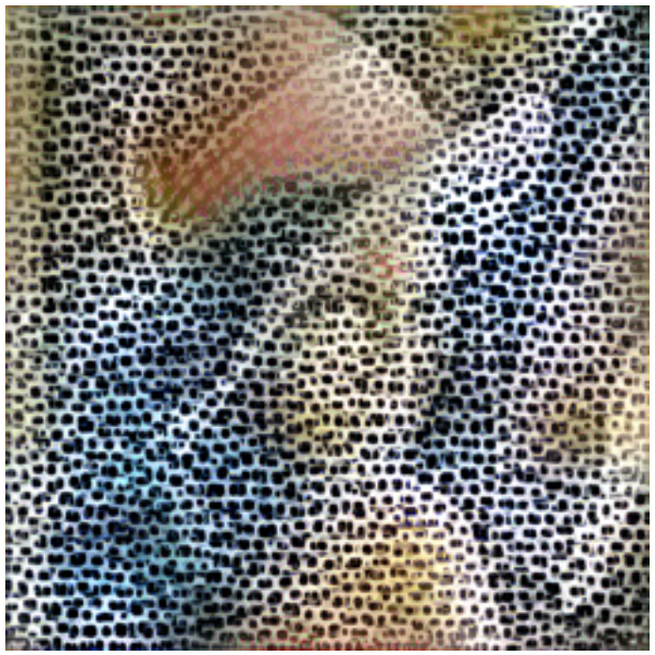
α ⟷ β

Iteration

Iteration: 0



Iteration: 1



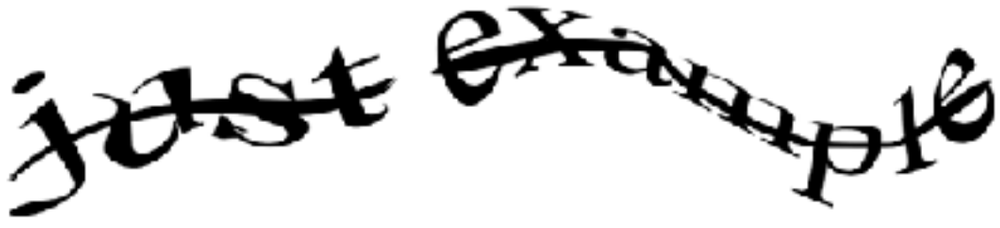
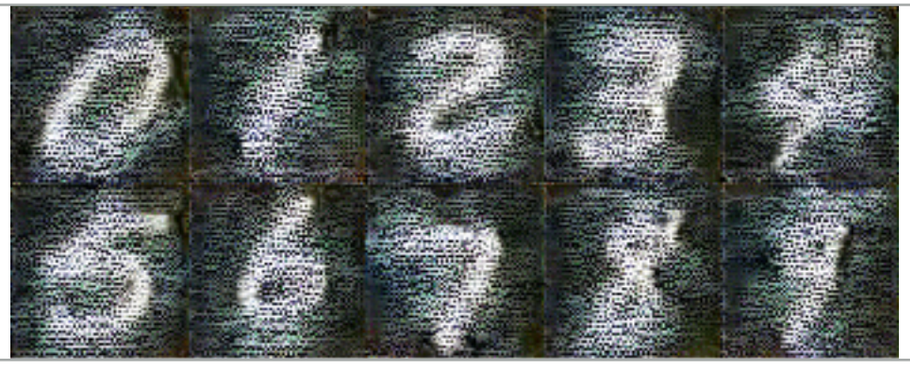
Iteration: 10

Halftone MNIST

Experiment

MNIST	Training	Testing	LeNet	AlexNet	VGG16	VGG19	ResNet18	ResNet50
32 x 32	Original	Original	98.89%	98.99%				
	Original	Halftone	**81.46%**	**21.46%**				
	Halftone	Halftone	96.88%	97.94%				
256 x 256	Original	Original						
	Original	Halftone						
	Halftone	Halftone						

Application

CAPTCHA	AI	Human
	Easy	Easy
	Hard	Hard
	**Hard**	**Easy**

Conclusion

- Progress of Adversarial Example
- Generate Halftone Effect with Style Transfer
- Discuss the impact of Halftone on Convolutional Neural Networks Models, eg: LeNet, AlexNet…

Future Work

1. Adversarial Retraining

    - Train on Halftone
    - Test on Halftone

2. Resize to 256 x 256 (Enhance the Halftone Effect)

3. Testing on more models, eg: VGG, ResNet…

Reference

- Adversarial Examples: Attacks and Defenses for Deep Learning: https://arxiv.org/abs/1712.07107

- Halftone: https://en.wikipedia.org/wiki/Halftone

- Image Style Transfer Using Convolutional Neural Networks: https://ieeexplore.ieee.org/document/7780634/