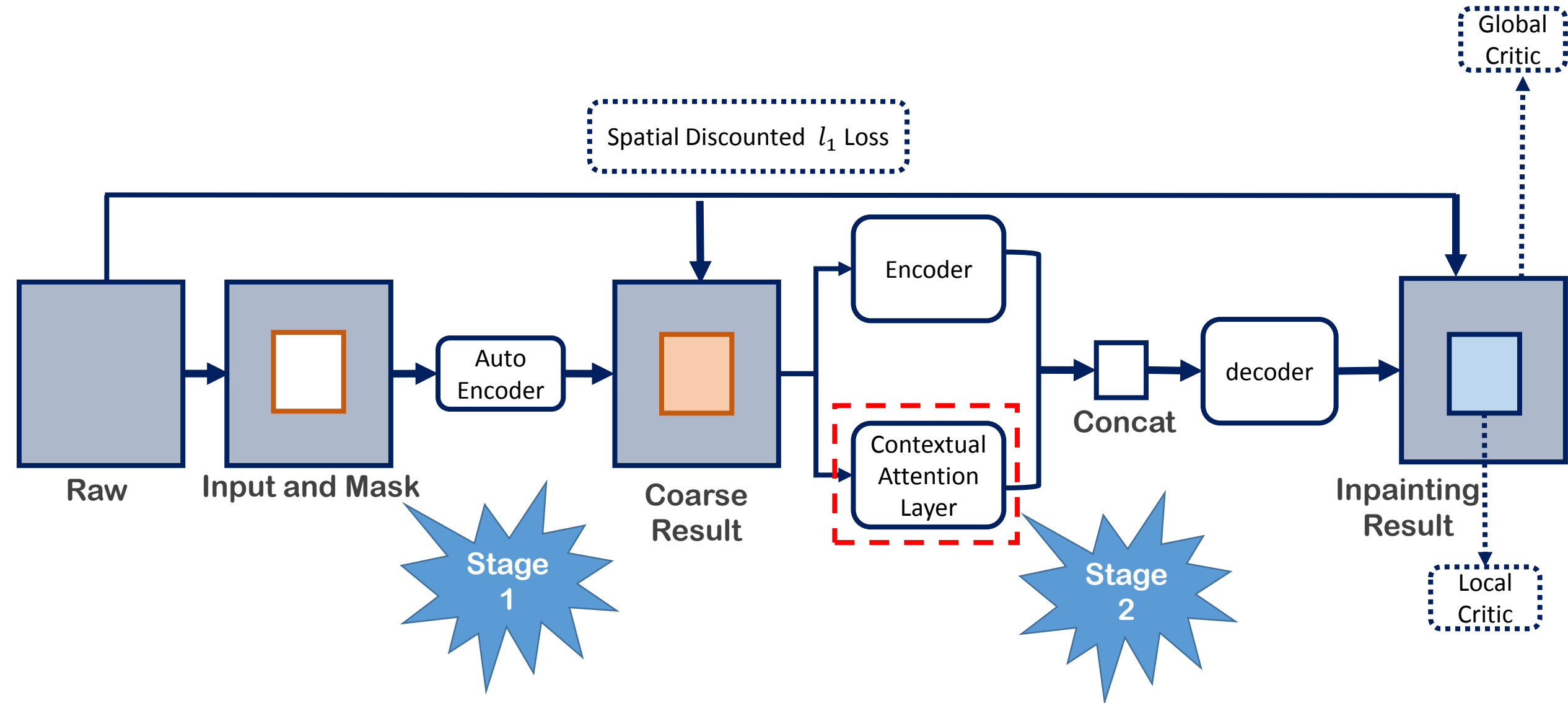# Generative Image Inpainting with Contextual Attention

7107053114 沈佳詠

Improved generative inpainting Network
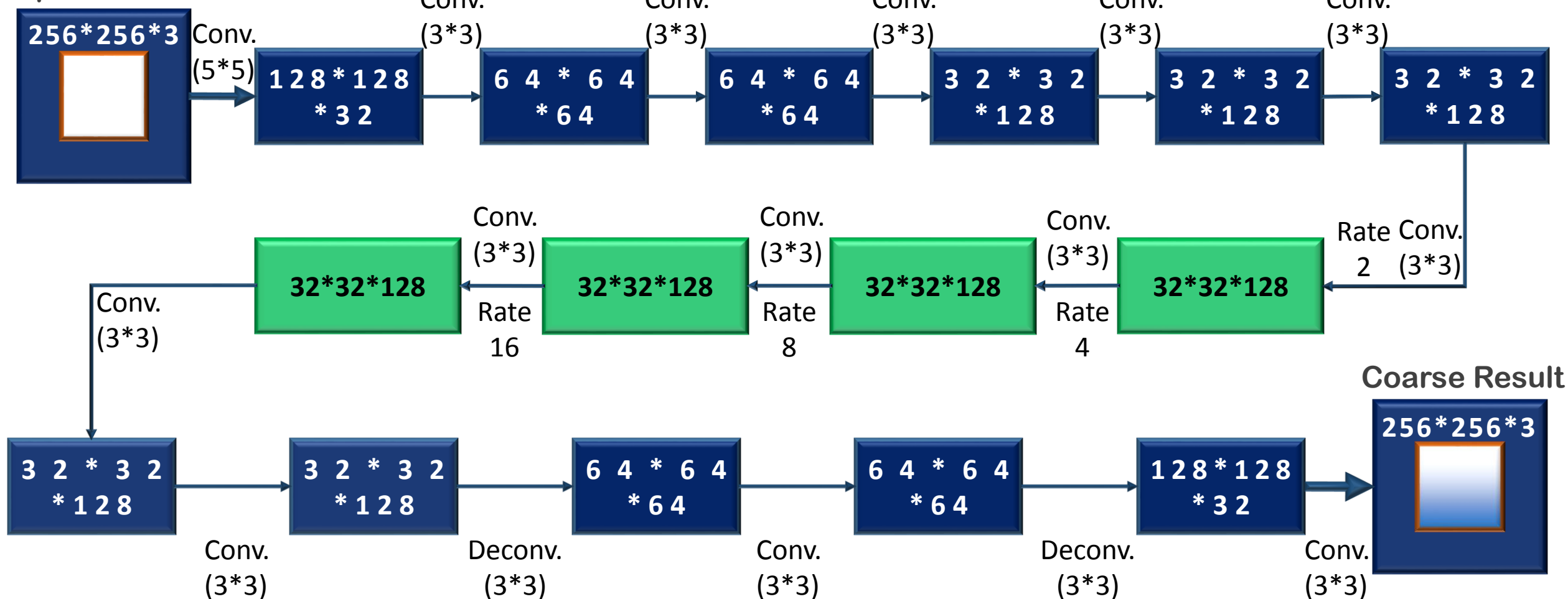
Generator stage 1

L1 Loss

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \prod(\mathbb{P}_r, \mathbb{P}_g)} E_{(\mathbf{x}, \mathbf{y}) \sim \gamma}[\|\mathbf{x} - \mathbf{y}\|]$$

Rate

Input and Mask

256*256*3

Conv. (5*5)

Conv. (3*3) | 128*128*32

Conv. (3*3) | 64*64*64

Conv. (3*3) | 64*64*64

Conv. (3*3) | 32*32*128

Conv. (3*3) | 32*32*128

32*32*128

Rate Conv. 2 (3*3)

Conv. (3*3) | 32*32*128 | Rate 16

Conv. (3*3) | 32*32*128 | Rate 8

Conv. (3*3) | 32*32*128 | Rate 4

32*32*128

Conv. (3*3)

Coarse Result

256*256*3

32*32*128

Conv. (3*3) | 32*32*128

Deconv. (3*3) | 64*64*64

Conv. (3*3) | 64*64*64

Deconv. (3*3) | 128*128*32

Conv. (3*3)

Generator stage 2

Attention

Dilated Convolution

Input and Mask

256*256*3 → Conv. (5*5) → 128*128*32 → Conv. (3*3) → 64*64*32 → Conv. (3*3) → 64*64*64 → Conv. (3*3) → 32*32*64 → Conv. (3*3) → 32*32*128 → Conv. (3*3) → 32*32*128

32*32*128 ← Conv. (3*3) Rate 16 ← 32*32*128 ← Conv. (3*3) Rate 8 ← 32*32*128 ← Conv. (3*3) Rate 4 ← 32*32*128 ← Rate 2 Conv. (3*3)

Contextual Attention Layer

Input and Mask

256*256*3 → Conv. (5*5) → 128*128*32 → Conv. (3*3) → 64*64*32 → Conv. (3*3) → 64*64*64 → Conv. (3*3) → 32*32*128 → Conv. (3*3) → 32*32*128 → Conv. (3*3) → 32*32*128

32*32*128 ← Conv. (3*3) Rate 16 ← 32*32*128 ← Conv. (3*3) Rate 8 ← 32*32*128 ← Contextual Attention

Generator stage 2

L1 Loss

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \prod(\mathbb{P}_r, \mathbb{P}_g)} E_{(\mathbf{x}, \mathbf{y}) \sim \gamma}[\|\mathbf{x} - \mathbf{y}\|]$$
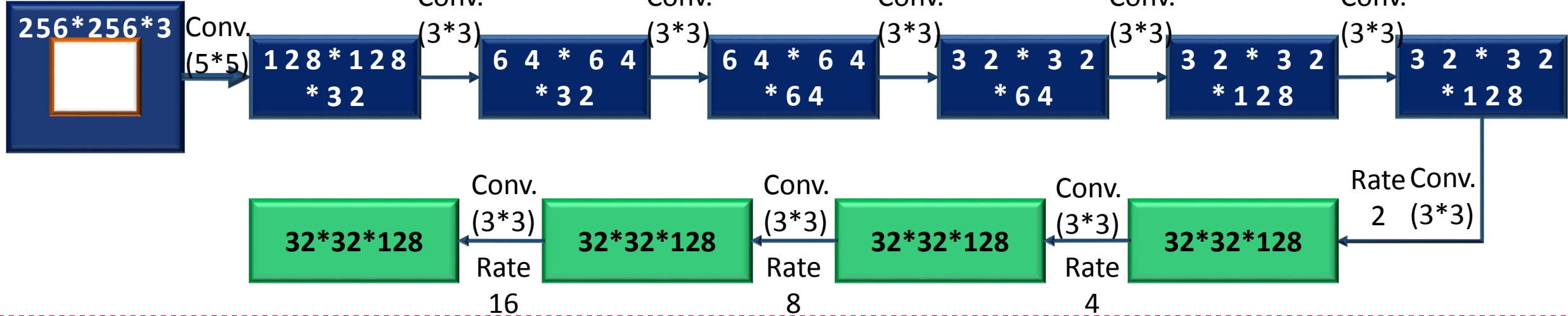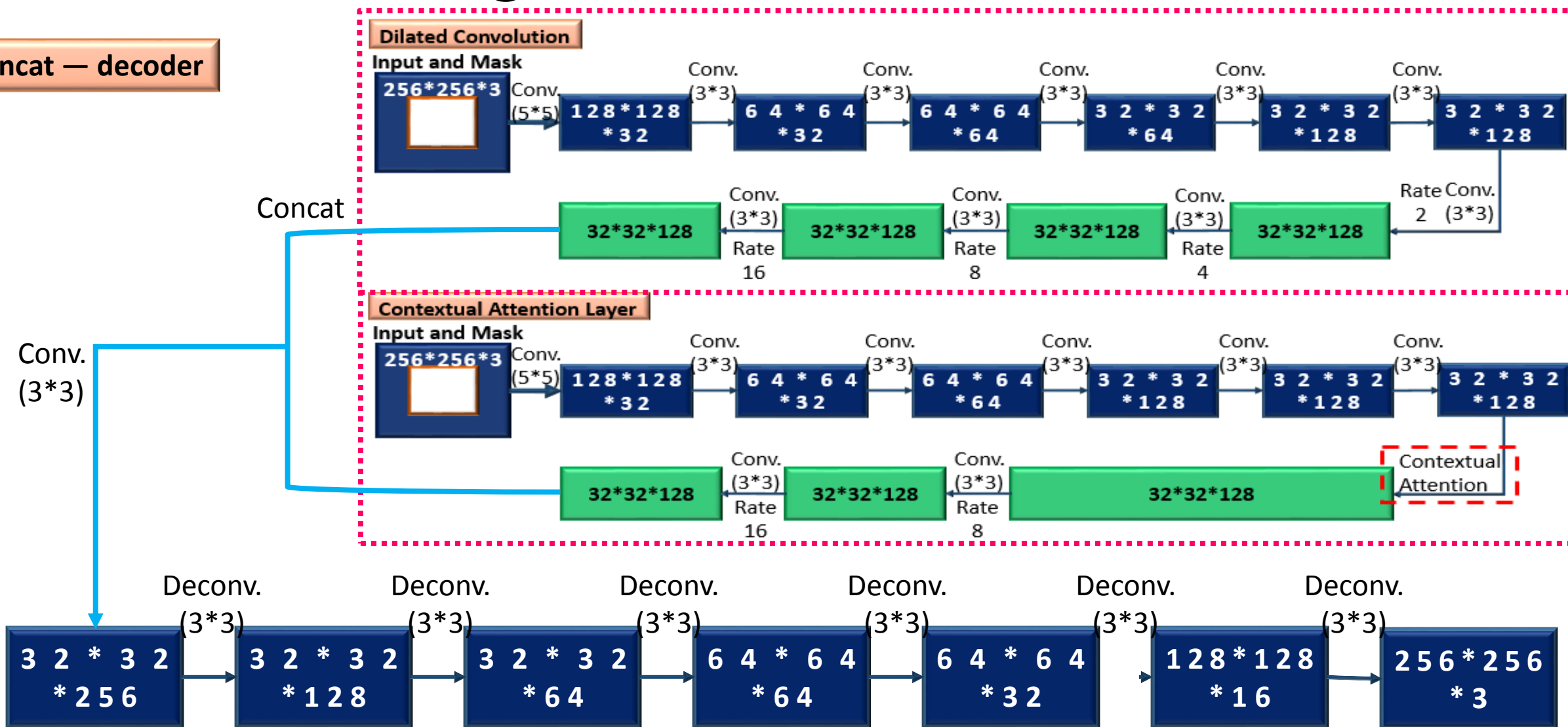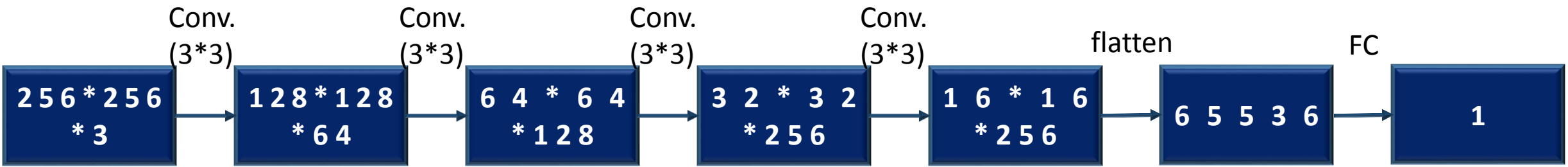
Concat — decoder

Dilated Convolution
Input and Mask

256*256*3

Conv. (5*5)

128*128*32

Conv. (3*3)

64*64*32

Conv. (3*3)

64*64*64

Conv. (3*3)

32*32*64

Conv. (3*3)

32*32*128

Conv. (3*3)

32*32*128

Rate Conv. 2 (3*3)

32*32*128 — Conv. (3*3) Rate 16 — 32*32*128 — Conv. (3*3) Rate 8 — 32*32*128 — Conv. (3*3) Rate 4 — 32*32*128

Concat

Contextual Attention Layer
Input and Mask

256*256*3

Conv. (5*5)

128*128*32

Conv. (3*3)

64*64*32

Conv. (3*3)

64*64*64

Conv. (3*3)

32*32*128

Conv. (3*3)

32*32*128

Conv. (3*3)

32*32*128

Contextual Attention

32*32*128 — Conv. (3*3) Rate 16 — 32*32*128 — Conv. (3*3) Rate 8 — 32*32*128

Conv. (3*3)

32*32*256

Deconv. (3*3)

32*32*128

Deconv. (3*3)

32*32*64

Deconv. (3*3)

64*64*64

Deconv. (3*3)

64*64*32

Deconv. (3*3)

128*128*16

Deconv. (3*3)

256*256*3

Discriminator
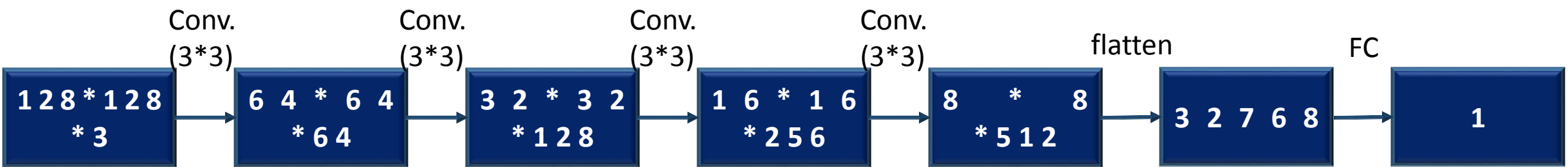
$$\min_{G} \max_{D \in \mathcal{D}} E_{\mathbf{x} \sim \mathbb{P}_r}[D(\mathbf{x})] - \lambda E_{\hat{\mathbf{x}} \sim \mathbb{P}_{\hat{\mathbf{x}}}} (\|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\|_2 - 1)^2]$$

| 256*256 *3 | Conv. (3*3) → | 128*128 *64 | Conv. (3*3) → | 64*64 *128 | Conv. (3*3) → | 32*32 *256 | Conv. (3*3) → | 16*16 *256 | flatten → | 65536 | FC → | 1 |

Local Critic

$$\min_{G} \max_{D \in \mathcal{D}} E_{\mathbf{x} \sim \mathbb{P}_r}[D(\mathbf{x})] - \lambda E_{\hat{\mathbf{x}} \sim \mathbb{P}_{\hat{\mathbf{x}}}} (\|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}}) \odot (\mathbf{1} - \mathbf{m})\|_2 - 1)^2]$$

| 128*128 *3 | Conv. (3*3) → | 64*64 *64 | Conv. (3*3) → | 32*32 *128 | Conv. (3*3) → | 16*16 *256 | Conv. (3*3) → | 8*8 *512 | flatten → | 32768 | FC → | 1 |

Parameter

- $x =$ real image
- $z =$ input image, $x \odot m$
- $\tilde{x} =$ inpainting output, $G(\text{z})$
- $\hat{x} = (1 - \text{t})x + \text{t}\tilde{x}, \text{t} \sim \text{U}[0, 1]$
- G ：generator
- D ：discriminator
- $\mathcal{D}$ ：the set of 1-Lipschitz function
- P_r ：the model distribution defined by x
- P_g ：the model distribution implicitly defined by $\tilde{x}$

Parameter

- λ ： set to 10
- m ： input and mask

$$\begin{cases} 0, & \text{for missing pixels} \\ 1, & \text{for elsewhere} \end{cases}$$

- $P_{\hat{x}}$ ： the model distribution defined by $\hat{x}$
- $\nabla_{\hat{x}} D(\hat{x})$ ： the gradient penalty apply to pixels inside the holes