# SAM-DiffSR: Structure-Modulated Diffusion Model for Image Super-Resolution

#### ソース

- https://arxiv.org/pdf/2402.17133
- https://github.com/lose4578/SAM-DiffSR

### 概要

- SAM(Segment Anythin Model)をdiffusionに取り入れてSuper Resolution
- 低画質画像のsegmentation maskを学習時(拡散過程)のノイズに加える

## 提案手法

- $m{x}_0$ をSAMに入力して得られるsegmentation maskを $m{M}_{\mathrm{SAM}}$ とする
- $m{M}_{
  m SAM}$ に対してencoder, embeddingを加えたものを $E_{
  m SAM}$ とする

#### diffusion modelの変更

拡散過程を以下のように定義する

$$q\left(\boldsymbol{x}_{t} \mid \boldsymbol{x}_{t-1}, \boldsymbol{E}_{\mathrm{SAM}}\right) = \mathcal{N}\left(\boldsymbol{x}_{t}; \sqrt{1-\beta_{t}}\boldsymbol{x}_{t-1} + \sqrt{\beta_{t}}\boldsymbol{E}_{\mathrm{SAM}}, \beta_{t}\boldsymbol{I}\right)$$
(1)

このとき、 $q\left(\boldsymbol{x}_{t} \mid \boldsymbol{x}_{0}, \boldsymbol{E}_{\mathrm{SAM}}\right)$ を計算すると

$$q\left(\boldsymbol{x}_{t} \mid \boldsymbol{x}_{0}, \boldsymbol{E}_{\mathrm{SAM}}\right) = \mathcal{N}\left(\boldsymbol{x}_{t}; \sqrt{\bar{\alpha}_{t}}\boldsymbol{x}_{0} + \phi_{t}\boldsymbol{E}_{\mathrm{SAM}}, (1 - \bar{\alpha}_{t})\boldsymbol{I}\right)$$
(2)

$$\alpha_t = \beta_t \tag{3}$$

$$\bar{\alpha}_t = \prod_{i=1}^t \alpha_i \tag{4}$$

$$\phi_t = \sum_{i=1}^t \sqrt{\bar{\alpha}_t \frac{\beta_i}{\bar{\alpha}_i}} \tag{5}$$

さらに、 $p\left(oldsymbol{x}_{t-1} \mid oldsymbol{x}_t, oldsymbol{x}_0, oldsymbol{E}_{ ext{SAM}}
ight)$ については以下の式になる

$$\tilde{\boldsymbol{\mu}}_t\left(\boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{E}_{\mathrm{SAM}}\right) = \frac{1}{\sqrt{lpha_t}} \left(\boldsymbol{x}_t - \frac{eta_t}{\sqrt{1 - ar{lpha}_t}} \left(\frac{\sqrt{1 - ar{lpha}_t}}{\sqrt{eta_t}} \boldsymbol{E}_{\mathrm{SAM}} + \boldsymbol{\epsilon}\right)\right)$$
 (6)

$$\tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t \tag{7}$$

$$p\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_{t}, \boldsymbol{x}_{0}, \boldsymbol{E}_{\mathrm{SAM}}\right) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \tilde{\mu}_{t}\left(\boldsymbol{x}_{t}, \boldsymbol{x}_{0}, \boldsymbol{E}_{\mathrm{SAM}}\right), \tilde{\beta}_{t} \boldsymbol{I}\right)$$
 (8)

$$\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}\right)$$
 (9)

そこで、損失関数を以下で定義する

$$\mathcal{L}\left(oldsymbol{ heta}
ight) = \mathbb{E}_{t,oldsymbol{x}_0,oldsymbol{\epsilon}}\left[\|rac{\sqrt{1-ar{lpha}}}{\sqrt{eta_t}}oldsymbol{E}_{ ext{SAM}} + oldsymbol{\epsilon} - oldsymbol{\epsilon}_{oldsymbol{ heta}}\left(oldsymbol{x}_t,t
ight)\|_2^2
ight]$$

# segmentation maskに対するencoding,embedding

画像 $m{x}\,(C imes H imes W)$ に対してposition embedding(RoPE)により $m{x}_{\mathrm{RoPE}}\in\mathbb{R}^{1 imes H imes W}$ を得る。

一方で $m{x}$  (C imes H imes W)に対してSAMによりK個のsegmentation mask  $M_{\mathrm{SAM},i}\in\{0,1\}^{1 imes H imes W}$ を得る。 $(i=1,2,\cdots,K)$ 

これらに対して以下の計算式でStructural position encodingを行い $m{E}_{ ext{SAM}}$ を得る

$$m{E}_{\mathrm{SAM}} = \sum_{i} m{M}_{\mathrm{SAM},i} \cdot \mathrm{mean}\left(m{x}_{\mathrm{RoPE},i}\right)$$
 (10)