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

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

diffusion modelの変更

拡散過程を以下のように定義する

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

このとき、$q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_0, \boldsymbol{E}_{\mathrm{SAM}}\right)$を計算すると

$$q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_0, \boldsymbol{E}_{\mathrm{SAM}}\right)=\mathcal{N}\left(\boldsymbol{x}_t ; \sqrt{\bar{\alpha}_t} \boldsymbol{x}_0+\phi_t \boldsymbol{E}_{\mathrm{SAM}},\left(1-\bar{\alpha}_t\right) \boldsymbol{I}\right) \tag{2}$$

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

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

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

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

$$\tilde{\boldsymbol{\mu}}_t\left(\boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{E}_{\mathrm{SAM}}\right)=\frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\left(\frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\beta_t}} \boldsymbol{E}_{\mathrm{SAM}}+\boldsymbol{\epsilon}\right)\right) \tag{6}$$

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

$$p\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{E}_{\mathrm{SAM}}\right)=\mathcal{N}\left(\boldsymbol{x}_{t-1} ; \tilde{\mu}_t\left(\boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{E}_{\mathrm{SAM}}\right), \tilde{\beta}_t \boldsymbol{I}\right) \tag{8}$$

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

そこで、損失関数を以下で定義する

$$\mathcal{L}(\boldsymbol{\theta})=\mathbb{E}_{t, \boldsymbol{x}_0, \boldsymbol{\epsilon}}\left[\| \frac{\sqrt{1-\bar{\alpha}}}{\sqrt{\beta_t}} \boldsymbol{E}_{\mathrm{SAM}}+\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_{\boldsymbol{\theta}}\left(\boldsymbol{x}_t, t\right) \|_2^2\right]$$

segmentation maskに対するencoding,embedding

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

一方で$\boldsymbol{x}\left(C \times H \times W\right)$に対してSAMにより$K$個のsegmentation mask $M_{\mathrm{SAM},i} \in\{0,1\}^{1 \times H \times W}$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$\boldsymbol{E}_{\mathrm{SAM}}$を得る

$$\boldsymbol{E}_{\mathrm{SAM}}=\sum_i \boldsymbol{M}_{\mathrm{SAM},i} \cdot \operatorname{mean}\left(\boldsymbol{x}_{\mathrm{RoPE},i}\right) \tag{10}$$