Implicit Image Inpainting

Yunkyu Lee, POSTECH

February 17, 2022

Related work

LIIF[1]

Superresolution thorough an implicit representation. Core ideas:

- Relative coordinate based implicit representation
- Self-supervised encoder encodes into latent space (feature map), LIIF decodes latent space
- Only nearest few latent codes to a coordinate are considered

Possibly extendable to inpainting, but receptive field is too small

[1] Chen, Liu, and Wang, *Learning Continuous Image Representation with Local Implicit Image Function*.

Problem formulation

Given an image $I \in \mathbb{R}^{C \times W \times H}$ and a binary mask $M \in \{0, 1\}^{W \times H}$, the masked image can be expressed as $I_M = I \odot M$.

With I_M as input, our objective is to generate an inpainted image $\hat{I}$ which is the closest to I.

Method

Implicit representation

For an implicit representation, we sample at points X from I to obtain a signal set S.

$$X = \{x_k : x_k \in \{0, \ldots, W-1\} \times \{0, \ldots, H-1\}\}_{k=1}^{n}$$
$$S = \{s_k : I_{i_k j_k}\}_{k=1}^{n}, \quad (i_k, j_k) = x_k$$

The inpainting problem can be expressed as finding a mapping from X to S given knowledge of I_M.

Architecture

Overview

We use an autoencoder architecture.
Trainable params $\approx 4.7 \times 10^5$

Encoder E_ϕ: Conv2D ResBlock $\times 3$

- Input: masked image I_M
- Output: Latent feature map $Z \in \mathbb{R}^{C \times W \times H}$

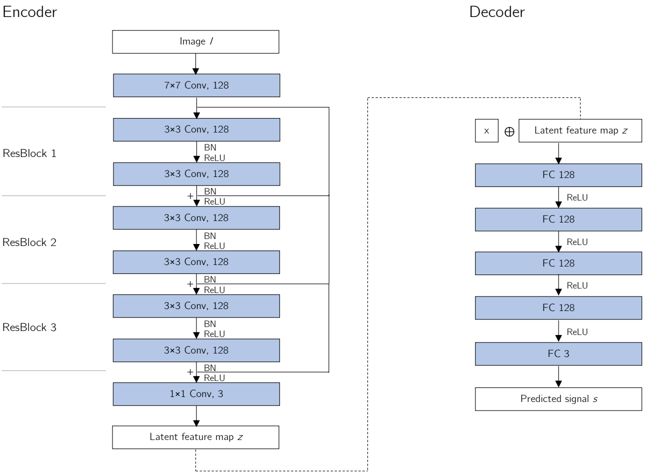
Decoder f_θ: FC 128×4

- Input: $x_i \oplus Z$
- Output: Predicted signal s_i

Results in a implicit representation in respect to x for fixed parameters ϕ, θ.

Architecture

Diagram



The encoder and decoder are end-to-end trained on samples. A weighted combination of MSE loss and TV loss is used.

$$\mathcal{L} = \alpha \mathcal{L}_{MSE} + \beta \mathcal{L}_{TV}$$

$\mathcal{L}_{MSE}$ is calculated pixel-wise, and $\mathcal{L}_{TV}$ is calculated over the entire inpainted image.

$$\mathcal{L}_{MSE} = \frac{1}{nC} \sum_{C} \sum_{k} (s_k - \hat{s}_k)^2$$

$$\mathcal{L}_{TV} = \frac{1}{nCWH} \sum_{C} \sum_{i,j} \left((\hat{I}_{i+1,j} - \hat{I}_{i,j})^2 - (\hat{I}_{i,j+1} - \hat{I}_{i,j})^2 \right)$$

Training

Masks and coordinates

Masks used in training has a large influence in inpainting performance.[2] We train over a wide range of randomly sampled masks of filling rate $p \sim \mathcal{U}(0.1, 0.9)$.

We convert a mask M to X by taking the coordinates of zero elements: $X = \{(i, j) : M_{ij} = 0\}$.

X is normalized to $[-1, 1]$.

---

[2] Suvorov et al., *Resolution-robust Large Mask Inpainting with Fourier Convolutions.*

Method

Inference

During inference, the image is reconstructed from $\hat{s}$, x.
Only the empty parts of the image are inpainted:

$$\hat{I} = I_M + \hat{I}_S \odot (\neg M)$$

Where $\hat{I}_S$ is $\hat{S}$ rearranged into a matrix:

$$\hat{I}_{S_{i_k j_k}} = \begin{cases} \hat{s}_k & \text{if } (i_k, j_k) \in X, \\ 0 & \text{otherwise} \end{cases}$$

Dataset

We concentrate on facial inpainting with the CelebA dataset. RGB Images were centered and cropped to

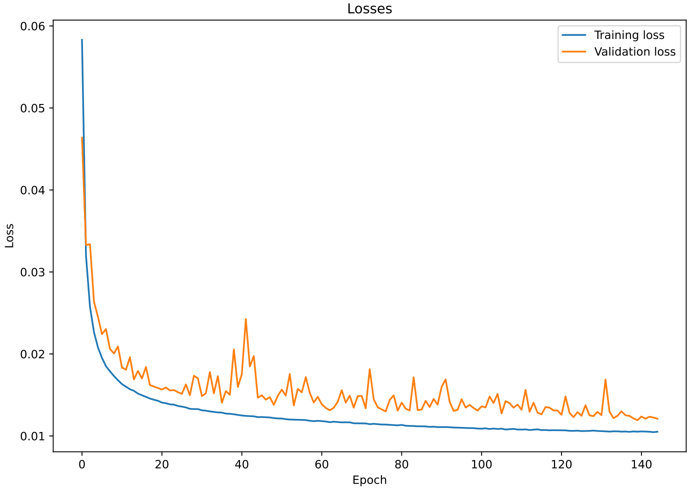
$$C \times W \times H = 3 \times 32 \times 32$$

Dataset split:

- Train: 162770
- Validation: 19687
- Test: 19962

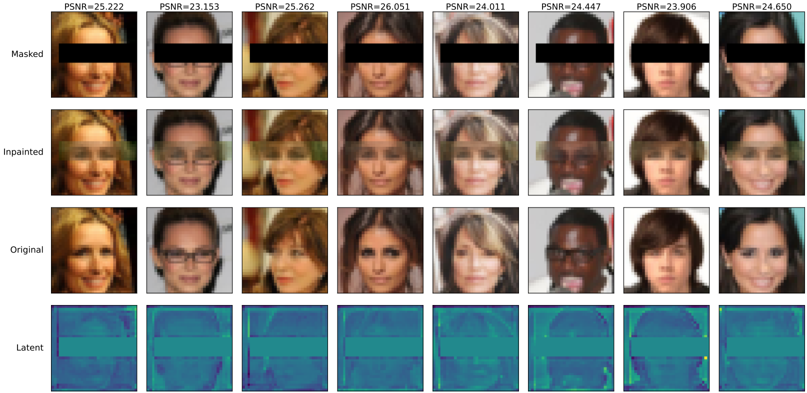
Experiments

Training curve



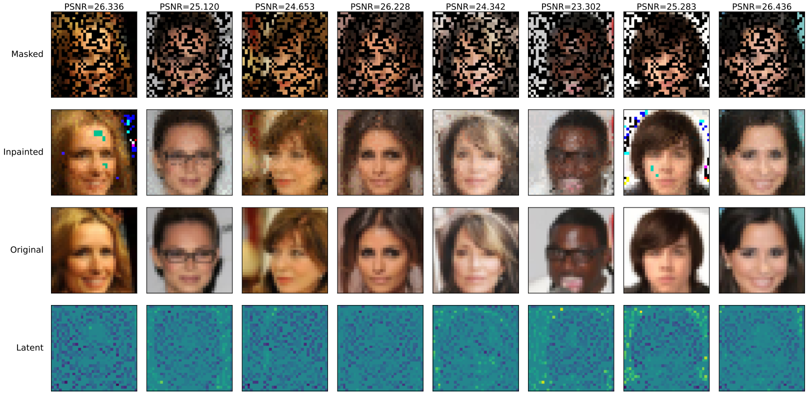
Experiments

Rectangular mask $p = 0.75$



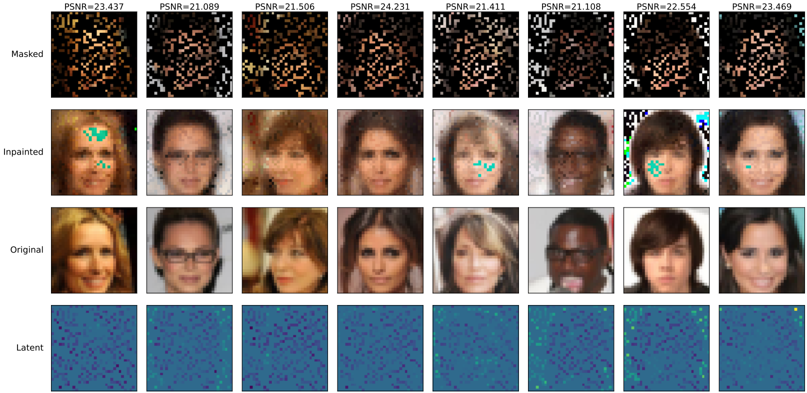
Experiments

Random mask, $p = 0.5$



Experiments

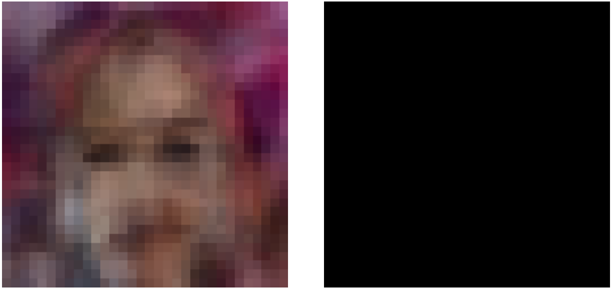
Random mask, $p = 0.3$



Experiments

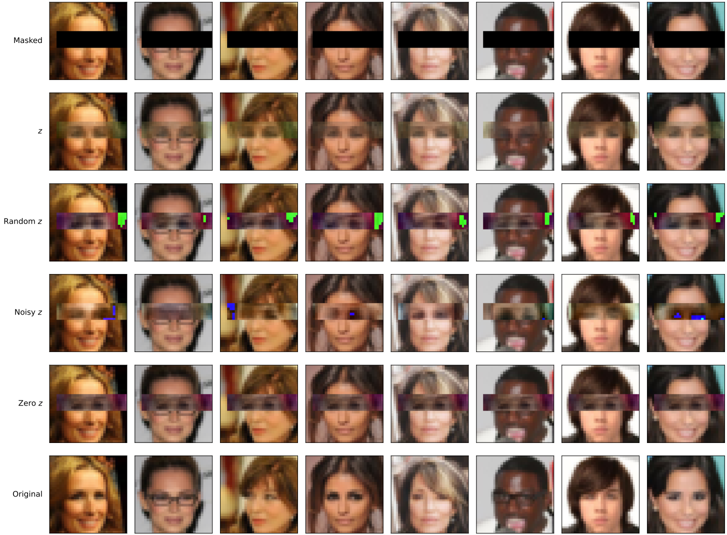
Totally masked input

Totally masked input



Experiments

Exploring the latent space



Remarks

+ Future tasks

- Does applying implicit neural representations make sense?
- A way of sampling latent codes (with inductive bias)
- GAN, discriminative loss
- Probabilistic approach
- Larger images, more diverse datasets
- SOTA baselines, better empirical control
- Ablation study

- Chen, Yinbo, Sifei Liu, and Xiaolong Wang. *Learning Continuous Image Representation with Local Implicit Image Function*. 2021. arXiv: 2012.09161 [cs.CV].
- Suvorov, Roman et al. *Resolution-robust Large Mask Inpainting with Fourier Convolutions*. 2021. arXiv: 2109.07161 [cs.CV].