Large Scale Image Completion via Co-Modulated Generative Adversarial Networks

Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I-Chao Chang, and Yan Xu

Microsoft Research

Motivation

- Numerous task-specific architectures have been proposed in order to attempt the task of image completion. However, a significant shortcoming of such methods is that all existing algorithms tend to fail in the face of **large missing regions**. We argue that this is due to the lack of generative capability.
- We notice the lack of plausible evaluation metrics in the field. Commonly adopted metrics either omit the inherent **paired** relationship between the original and the inpainted images, which are not strict enough and tend to suffer from huge variance, or are entirely **pixel-wise**, which favor blurry results and cannot reflect the semantic difference between images.

Contributions

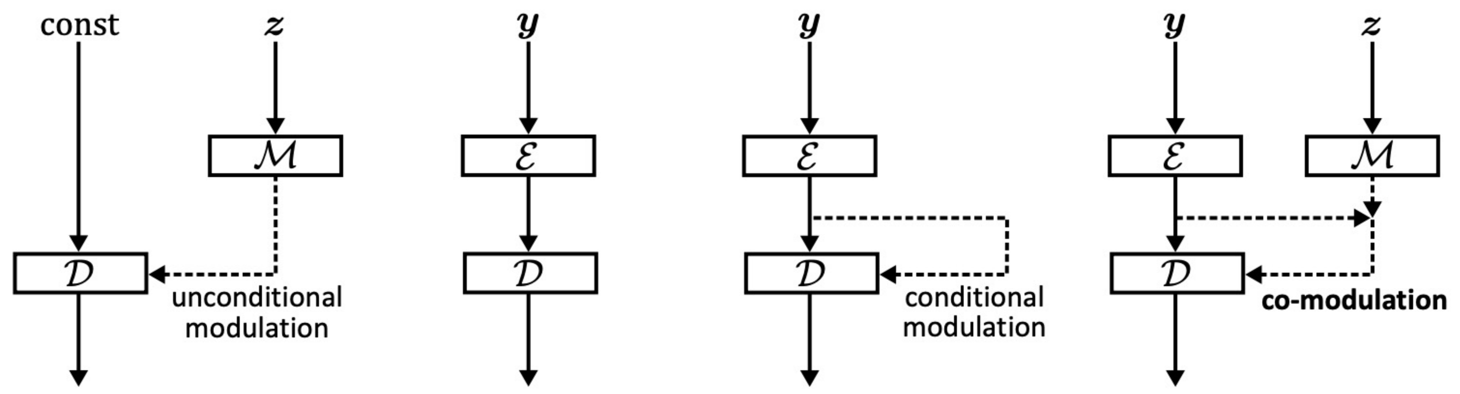
- We propose a new architecture, CoModGAN, that *co-modulates* both the image-conditional information and a stochastic style code.
- We propose P-IDS and U-IDS, which robustly measure the perceptual fidelity of inpainted images compared to real images via linear separability in the Inception-v3 feature space.

Other Details

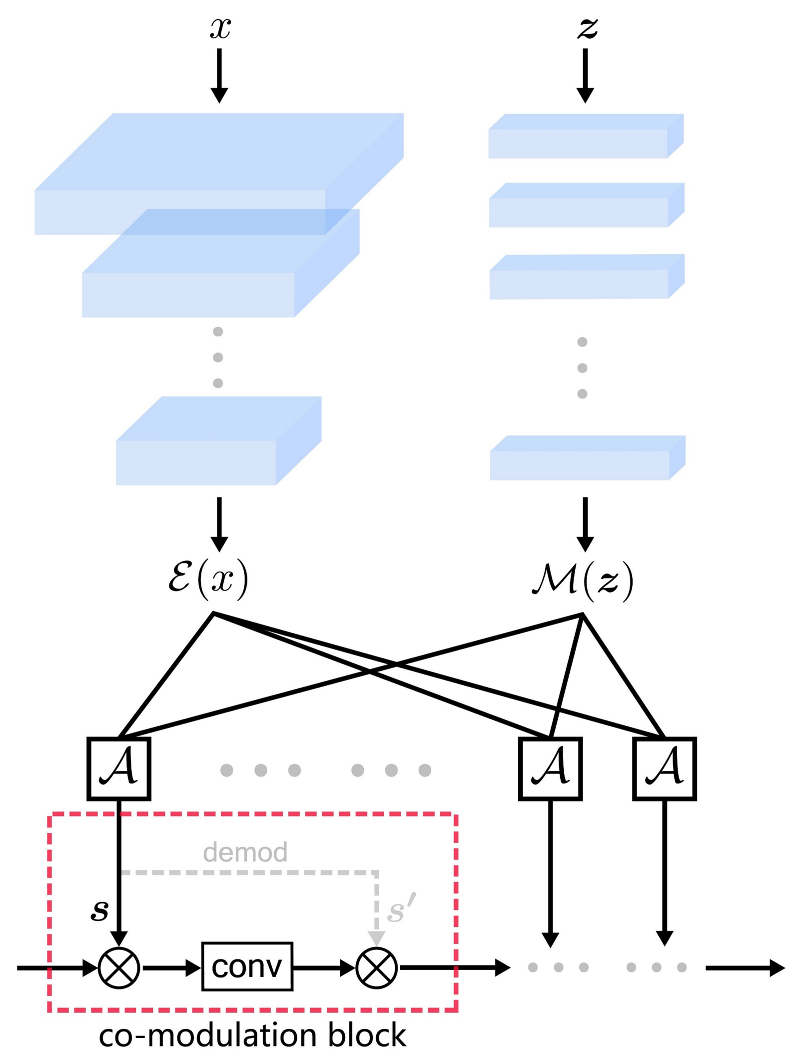
Paper

Code

Network Architecture



From unconditional modulation to co-modulation (left to right): unconditional modulated generator, vanilla image-conditional generator, conditional modulated generator, and *co-modulated* generator.



Given the style vector $\boldsymbol{s} = \mathcal{A}(\mathcal{E}(x), \mathcal{M}(\boldsymbol{z}))$ for modulation, the input feature maps are first channel-wise multiplied by $\boldsymbol{s}$ and then fed into convolution, finally channel-wise multiplied by $\boldsymbol{s}'$, where

$$s'_j = \sqrt{1 \Big/ \sum_{i,k} \left(s_i \cdot w_{ijk}\right)^2}.$$

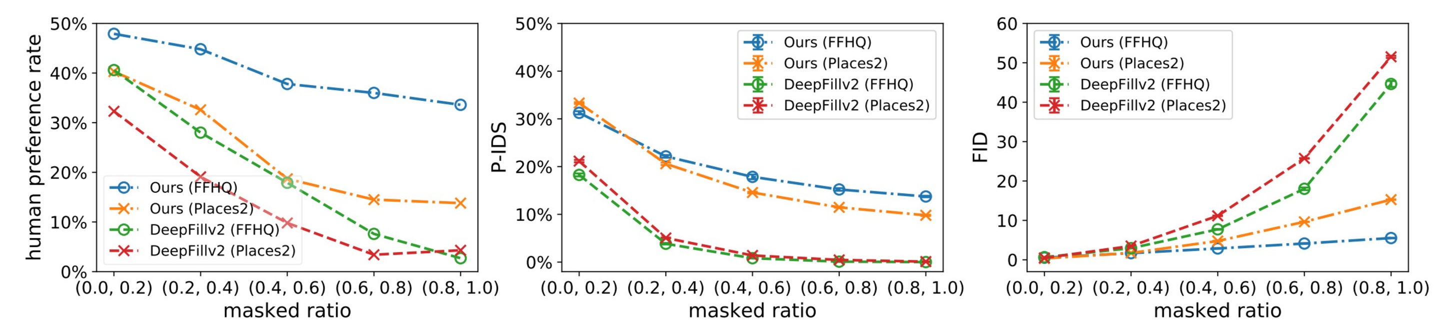
This step acts as weight demodulation, normalizing the feature maps to unit variance.

U-IDS & P-IDS

$$\mathbf{U-IDS}(X, X') := \frac{1}{2} \Pr_{x \in X}\{f(\mathcal{I}(\boldsymbol{x})) < 0\} + \frac{1}{2} \Pr_{x' \in X'}\{f(\mathcal{I}(\boldsymbol{x}')) > 0\}.$$

$$\mathbf{P-IDS}(X) := \Pr_{(x,x') \in X}\{f(\mathcal{I}(\boldsymbol{x}')) > f(\mathcal{I}(\boldsymbol{x}))\}.$$

Model Performance

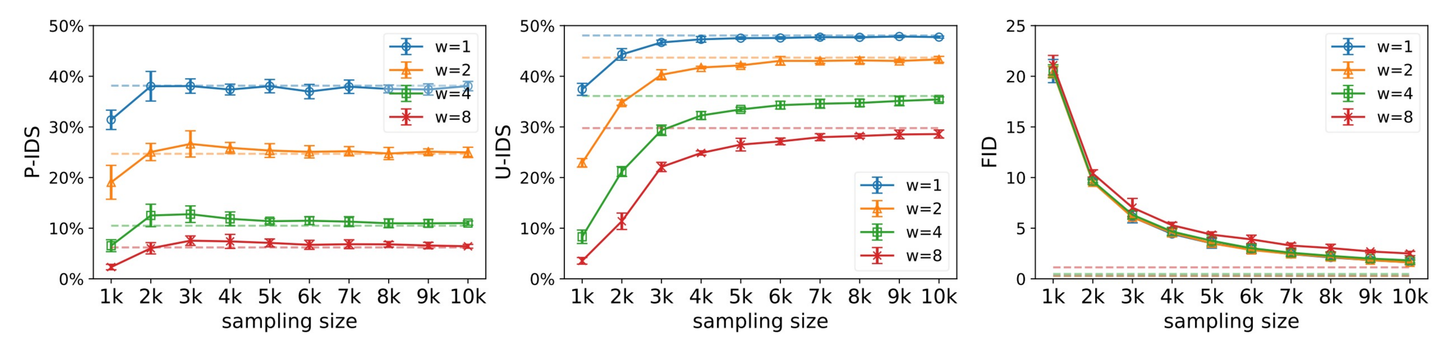


P-IDS correlates well with human preferences, and our model demonstrates superior performance in terms of all metrics, especially in the presence of large masked regions.

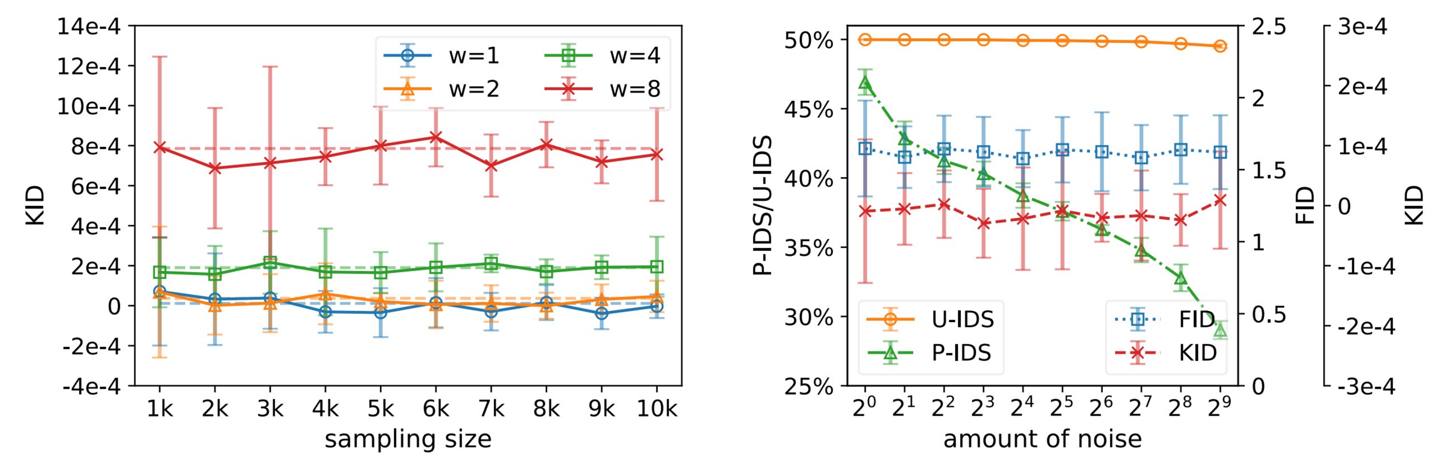
Method	Edges2Shoes		Edges2Handbags	
	FID	LPIPS	FID	LPIPS
Pix2Pix (Isola et al., 2017)	74.2	0.040	95.6	0.042
BicycleGAN (Zhu et al., 2017b)	47.3	0.191	76.0	0.252
MUNIT (Huang et al., 2018)	56.2	0.229	79.1	0.339
BasisGAN (Wang et al., 2019b)	64.2	**0.242**	88.8	0.350
Ours	**38.5**	0.036	**56.9**	0.143
Ours ($\psi = 3$)	**38.5**	0.038	71.1	**0.379**

CoModGAN can be applied to image-to-image translation tasks as well. Our model is capable of generating high-fidelity images when trained on the edges to photos datasets.

Analysis on U-IDS & P-IDS



We randomly mask a square region of size w^2 in images randomly sampled from the FFHQ dataset and calculate the three metrics. Among the three metrics, FID fails to converge within 10k sampling size, while P-IDS and U-IDS converge quickly within a small sampling size.



On the left, we use the same strategy to mask a square region to test KID. Experiments demonstrate that KID is subject to huge variance and fails to distinguish any difference between $w = 2$ and 4.

On the right, we randomly remove a certain number of pixels ("amount of noise") from images randomly sampled from the FFHQ dataset and perform a nearest-neighbor interpolation to fill in the deleted pixels. Results show that P-IDS and U-IDS capture the subtle differences while KID and FID fail to detect them.