Generative Image Inpainting with Contextual Attention

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, Thomas Huang
University of Illinois at Urbana–Champaign, Adobe Research
In CVPR, 2018

Sharer: Yan Zhao
2018.06.20

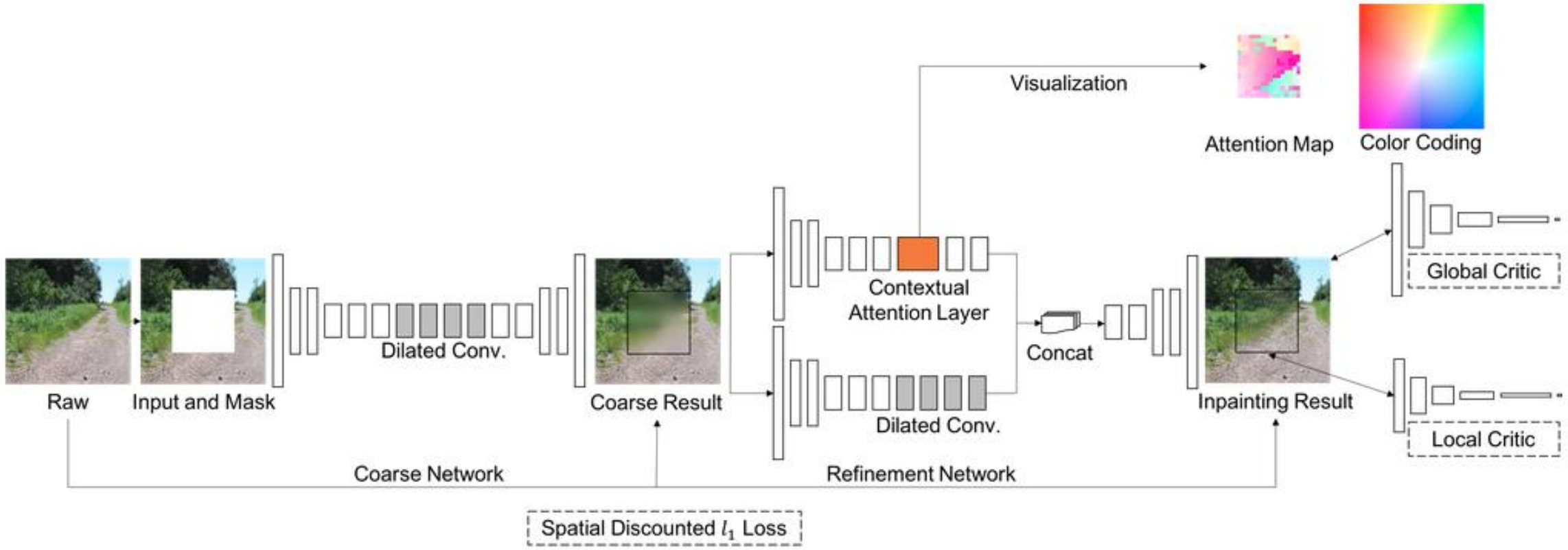
Motivation

Example inpainting results of our method on images of natural scene (Places2), face (CelebA) and object (ImageNet).

Contributions

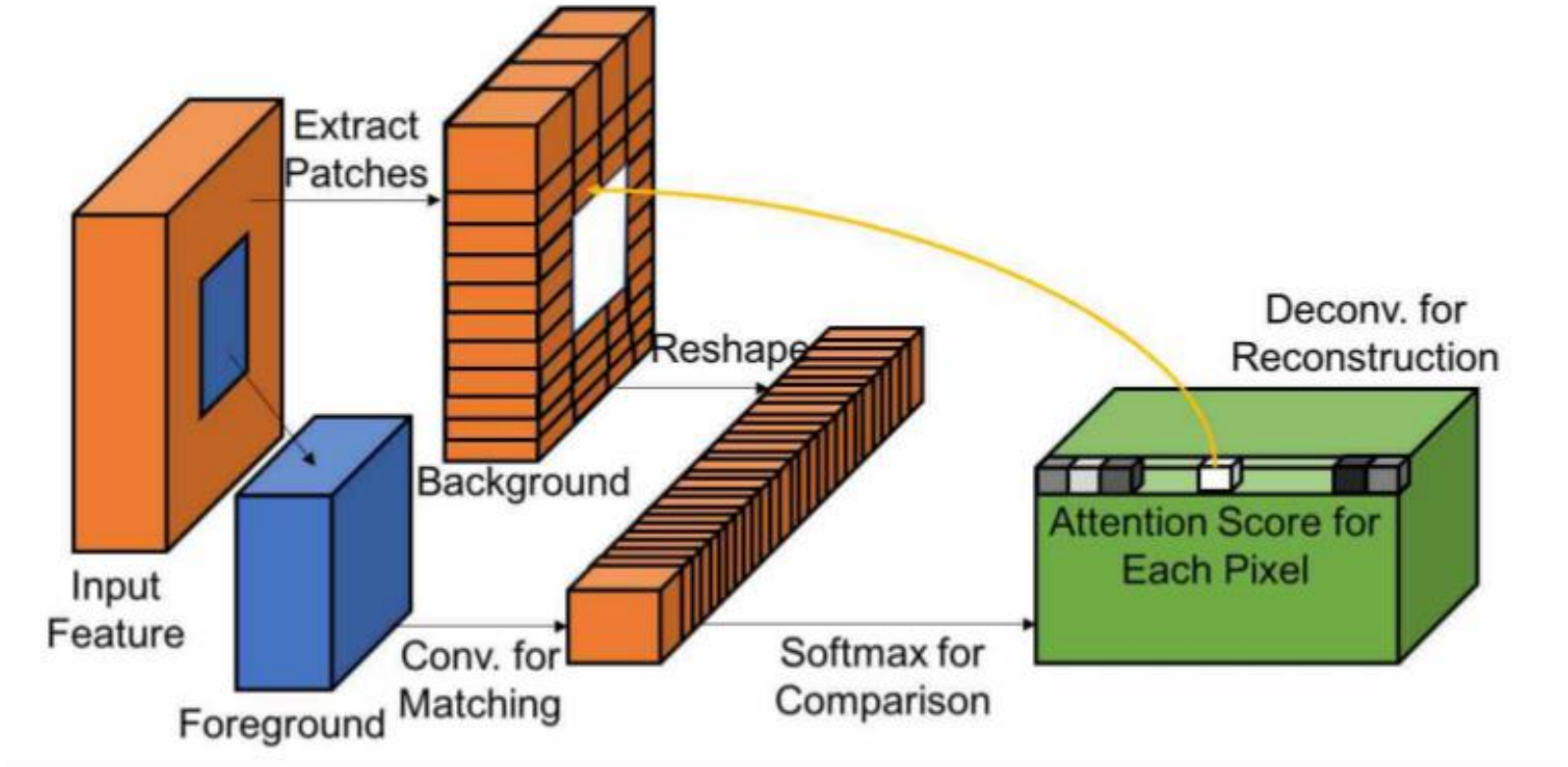
- We propose a novel contextual attention layer to explicitly attend on related feature patches at distant spatial locations
- We introduce several techniques including inpainting network enhancements, global and local WGANs and spatially discounted reconstruction loss to improve the training stability and speed based on the current the state-of-the-art generative image inpainting network .
- Our unified feed-forward generative network achieves high-quality inpainting results on a variety of challenging datasets including CelebA faces, CelebA-HQ faces, DTD textures, ImageNet and Places2.

Approach



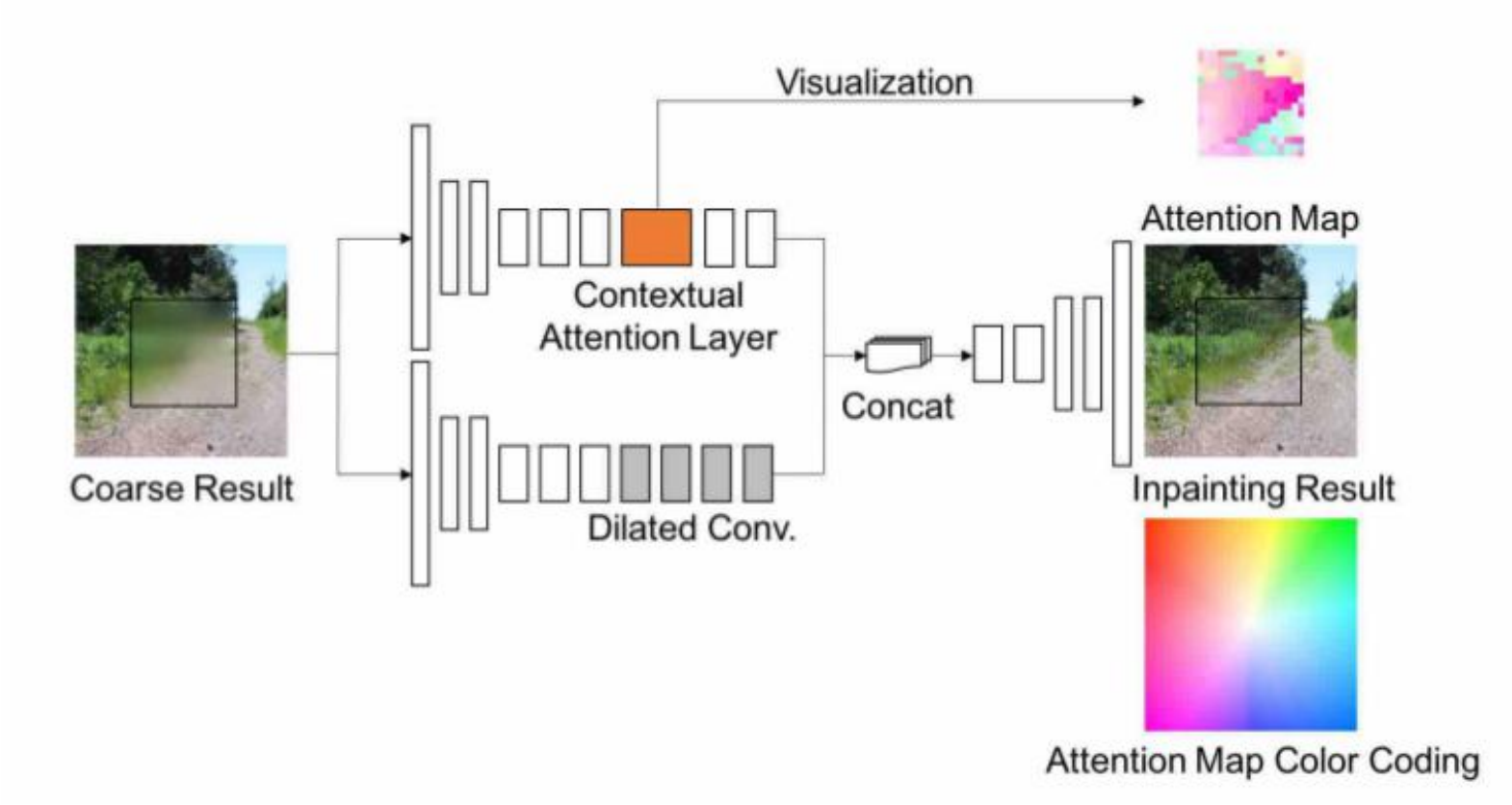
The coarse network is trained with reconstruction loss explicitly, while the refinement network is trained with reconstruction loss, global and local WGAN-GP adversarial loss.

Contextual Attention



Firstly we use convolution to compute matching score of foreground patches with background patches (as convolutional filters). Then we apply softmax to compare and get attention score for each pixel. Finally we reconstruct foreground patches with background patches by performing deconvolution on attention score.

Unified Inpainting Network



For visualization of attention map, color indicates relative location of the most interested background patch for each pixel in foreground. For examples, white (center of color coding map) means the pixel attends on itself, pink on bottom-left, green means on top-right.

Experiments

Datasets

- Places2
- CelebA faces
- CelebA-HQ faces
- DTD textures
- ImageNet

Quantitative comparisons

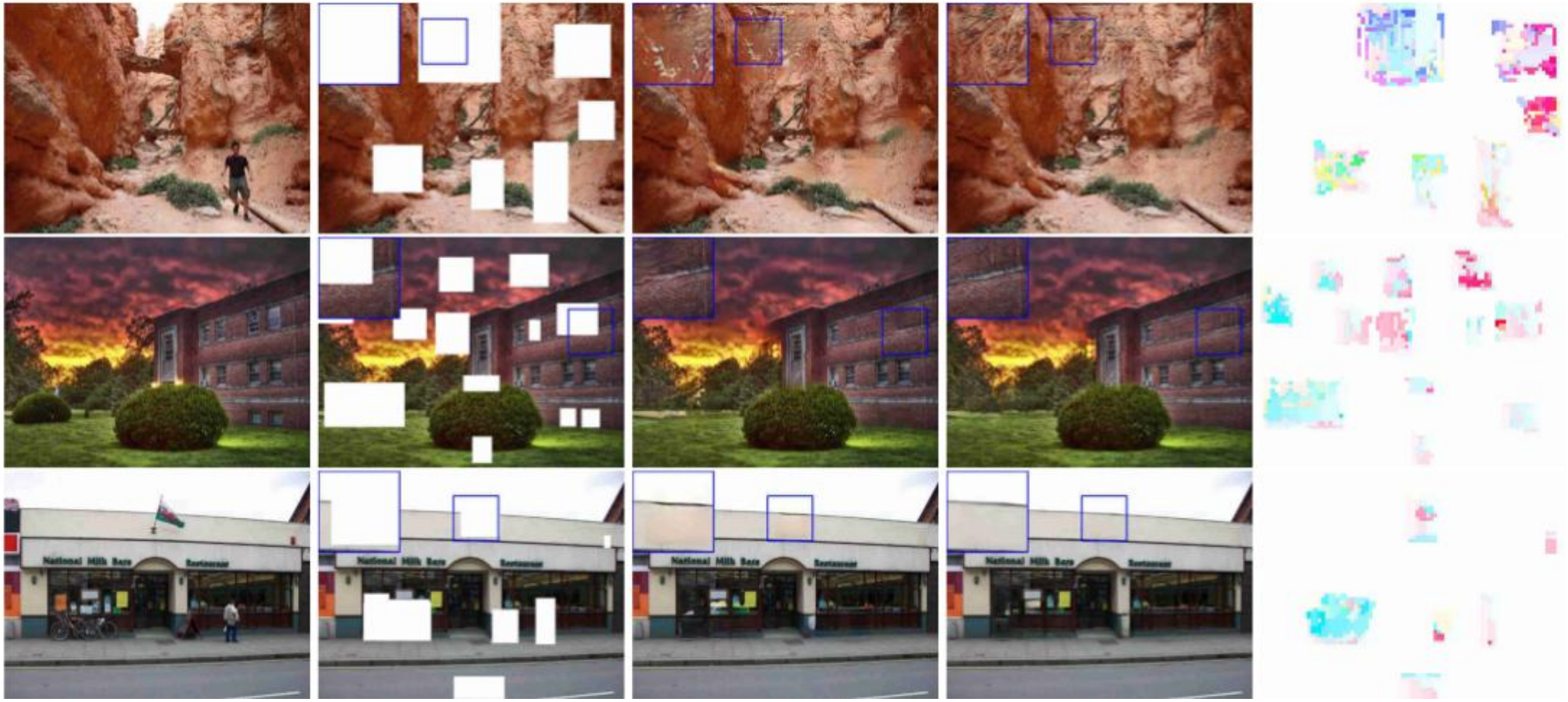
Method	ℓ_1 loss	ℓ_2 loss	PSNR	TV loss
PatchMatch [3]	16.1%	3.9%	16.62	**25.0%**
Baseline model	9.4%	2.4%	18.15	25.7%
Our method	**8.6%**	**2.1%**	**18.91**	25.3%

Table 1: Results of mean ℓ_1 eror, mean ℓ_2 error, PSNR and TV loss on validation set on Places2 for reference.

Comparisons

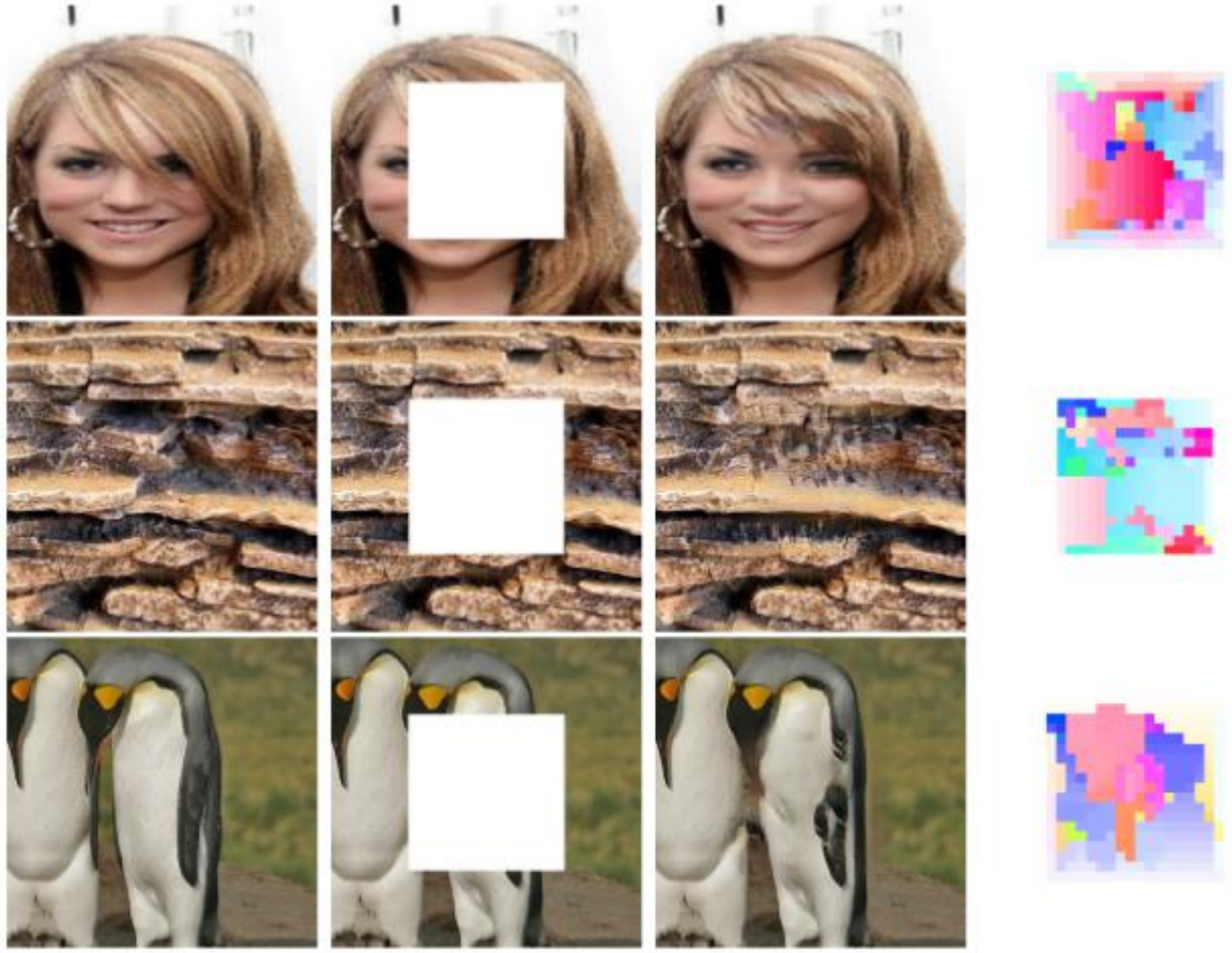
Figure 5: Comparison of our baseline model with Iizuka et al. [17]. From left to right, we show the input image, result copied from main paper of work [17], and result of our baseline model. Note that no post-processing step is performed for our baseline model, while image blending is applied for the result of [17]. Best viewed with zoom-in.

Results



We show from left to right the original image, input image, result of our baseline model, result and attention map (upscaled 4×) of our full model.

Results



Q & A