Learning to Incorporate Structure Knowledge for Image Inpainting

Abstract

This paper develops a multi-task learning framework that attempts to incorporate the image structure knowledge to assist image inpainting, which is not well explored in previous works. The primary idea is to train a shared generator to simultaneously complete the corrupted image and corresponding structures - edge and gradient, thus encourages the generator to exploit relevant structure knowledge while inpainting. Besides, we also introduce a structure embedding scheme to explicitly embed the learned structure features into the inpainting process serving as preconditions for image completion. Specifically, a novel pyramid structure loss is proposed to supervise structure learning and embedding. Moreover, an attention mechanism is developed to exploit the recurrent patterns in the image to refine the generated structures and contents. Through multi-task learning, structure embedding besides with attention, our framework takes advantage of the structure knowledge and outperforms several state-of-the-art methods on benchmark datasets quantitatively and qualitatively.

Introduction

Image inpainting targets at filling corrupted or replacing unwanted regions of images with plausible and fine-detailed contents, which is widely applied in fields of restoring damaged photographs, retouching pictures, et al.

Existing inpainting approaches can be roughly divided into two groups: conventional and deep learning based approaches. Conventional inpainting approaches usually make use of low-level features (e.g. color and texture descriptors) hand-crafted from the incomplete input image and resort to priors (e.g. smoothness and image statistics) or auxiliary data (e.g. external image databases). They either propagate low-level features from surroundings to the missing regions following a diffusive process or fill holes by searching and fusing similar patches from the same image or external image databases. Without a high-level understanding of the image contents and structures, conventional approaches usually struggle to generate semantically meaningful content, especially when a large portion of an image is missing or corrupted.

- We propose a multi-task learning framework to incorporate the image structure knowledge to assist image inpainting.
- We introduce a structure embedding scheme which can explicitly provide structure preconditions for image completion, and an attention mechanism to exploit the similar patterns in the image to refine the generated structures and contents.

Related Work

Numerous image inpainting approaches have been proposed; here, we focus to review the representative deep learning based methods.

Method

Our multi-task framework leverages the structure knowledge with multi-tasking learning (simultaneous image and structure generation), structure embedding and attention mechanism. As a future work, we plan to investigate adapting the proposed multi-task framework to other specific inpainting architectures to leverage the structure knowledge.

Experiments

In this section, we present our experimental comparisons with several state-of-the-art image inpainting approaches and ablation studies of the effectiveness of our multi-task framework. More results can reference our supplementary material.

Conclusion

We have primarily presented a framework for incorporating image structure knowledge for image inpainting. We propose to utilize the multi-task learning strategy, explicit structure embedding besides with an an attention mechanism to make use of the image structure knowledge for inpainting. The experiments results demonstrate that the proposed approach shows superior performance compared with several state-of-the-art inpainting methods which either ignore or not well exploit the structure knowledge. Besides, each component for incorporating structure knowledge is verified by ablation studies.

Supplementary Material

In this supplementary material, we present more details of the network architectures and training, additional qualitative comparisons and visual results.

A. Network Architectures

The detailed architectures of our generator and discriminator are shown in Table 1 and Table 2 respectively. Our model is trained end-to-end using Adam optimizer with $\beta_1 = 0$ and $\beta_2 = 0.9$. We set the initial learning rate 10^{-4} then lower it to 10^{-5} until metrics converge. The code will be made public in the future.

B. More Experimental Results

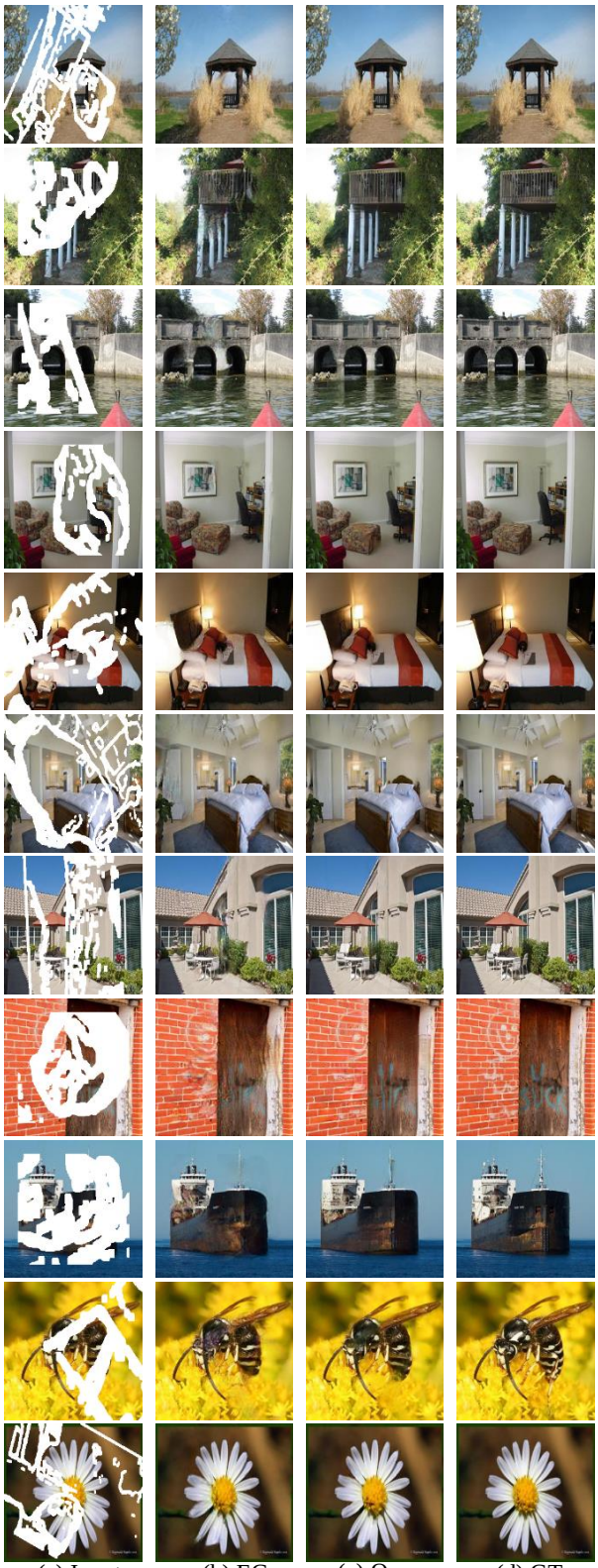
For the experiments, we show more qualitative comparison results on Places2 in Figure 1 and CelebA in Figure 2. More visual results are shown in Figure 4 and Figure 5. Additional ablation study results are shown in Figure 3.

Table 1: The architecture of our generator. $\oplus$ denotes feature concatenation, ϕ^1 the feature maps in the encoder, A^1 attention maps, S^1 structure feature maps, φ^i the features maps in the decoder. [IN: Instance Normalization; RBs: Residual Blocks (Nazeri et al. 2019); AT: Attention Layer; SE: Structure Embedding Layer.]

Input: $\hat{\mathbf{I}} \oplus \mathbf{M} \oplus \hat{\mathbf{E}} \oplus \hat{\mathbf{C}}$ $(256 \times 256 \times 11)$
ϕ^1: Conv. (7, 7, 64), stride=1; IN; ReLU;
ϕ^2: Conv. (4, 4, 128), stride=2; IN; ReLU;
ϕ^3: Conv. (4, 4, 256), stride=2; IN; ReLU;
ϕ^4: Eight RBs(ϕ^3)
A^1: AT(ϕ^4)
S^1: SE(A^1)
Structure Output: Conv. (1, 1, 6), stride=1;
φ^1: $A^1 \oplus S^1$; Deconv. (3, 3, 128), stride=2; IN; ReLU;
A^2: AT(φ^1)
S^2: SE(A^2)
Structure Output: Conv. (1, 1, 6), stride=1;
φ^2: $A^2 \oplus S^2$; Deconv. (3, 3, 64), stride=2; IN; ReLU;
φ^3: AT(φ^2); Conv. (5, 5, 64), stride=1; IN; ReLU;
Structure Output: Conv. (1, 1, 6), stride=1; **Image Output**: Conv. (1, 1, 3), stride=1;

Table 2: The architecture of our discriminator. SNConv. denotes the convolutions with spectral normalization.

Input: $\mathbf{I}_{comp}$ $(256 \times 256 \times 3)$
[layer 1]: SNConv. (5,5,64), stride=2; LReLU;
[layer 2]: SNConv. (5,5,128), stride=2; LReLU;
[layer 3]: SNConv. (5,5,256), stride=2; LReLU;
[layer 4]: SNConv. (5,5,512), stride=2; LReLU;
[layer 5]: SNConv. (5,5,1), stride=1;

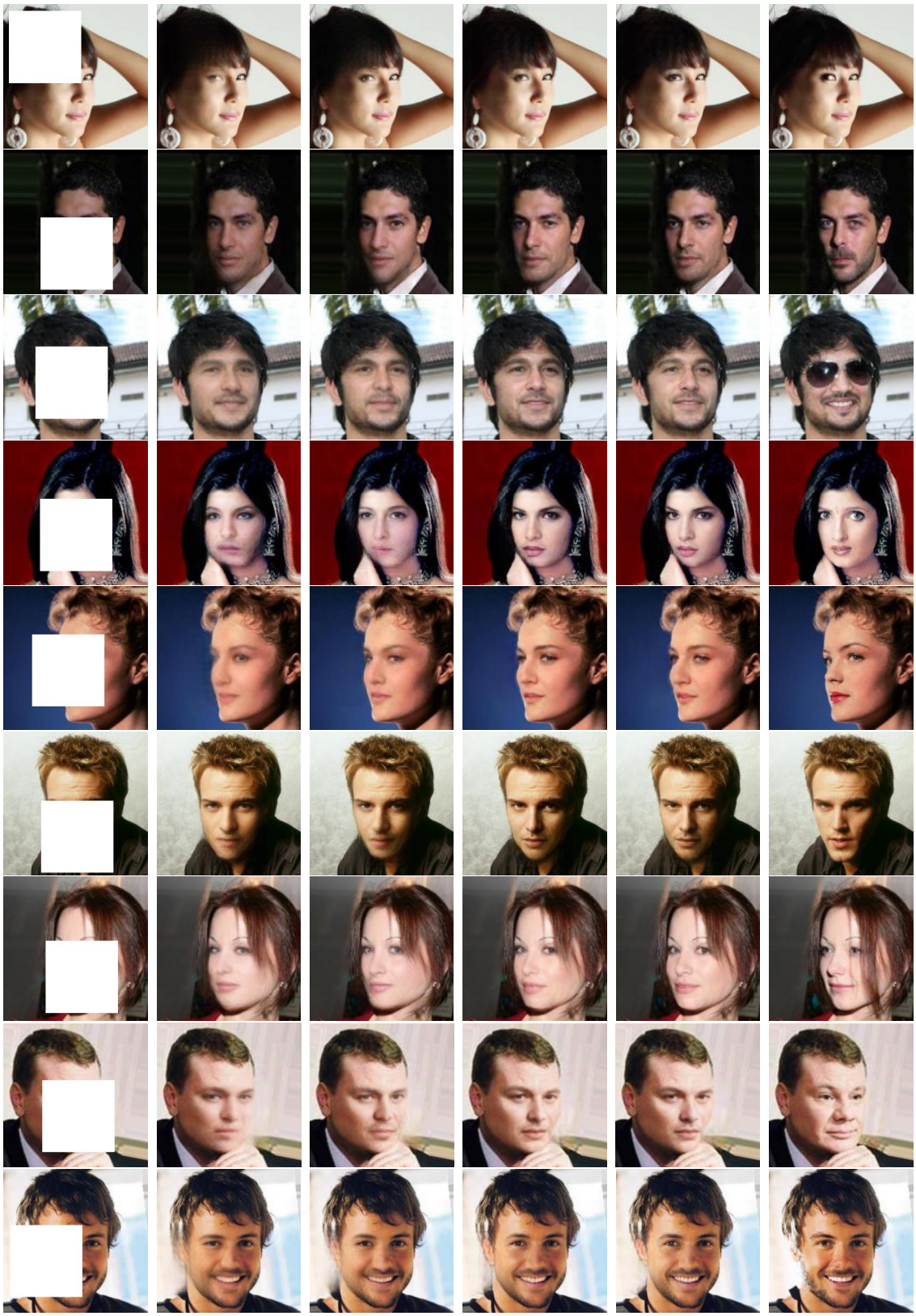


(a) Input (b) EG (c) Ours (d) GT

Figure 1: Ours compared with EG (Nazeri et al. 2019) and the ground truth (GT) on Places2. [Best viewed with zoom-in.]

(a) Input (b) CA (c) PEN-Net (d) EG (e) Ours (f) GT

Figure 2: Ours compared with baselines and the ground truth (GT) on CelebA. [Best viewed with zoom-in.]

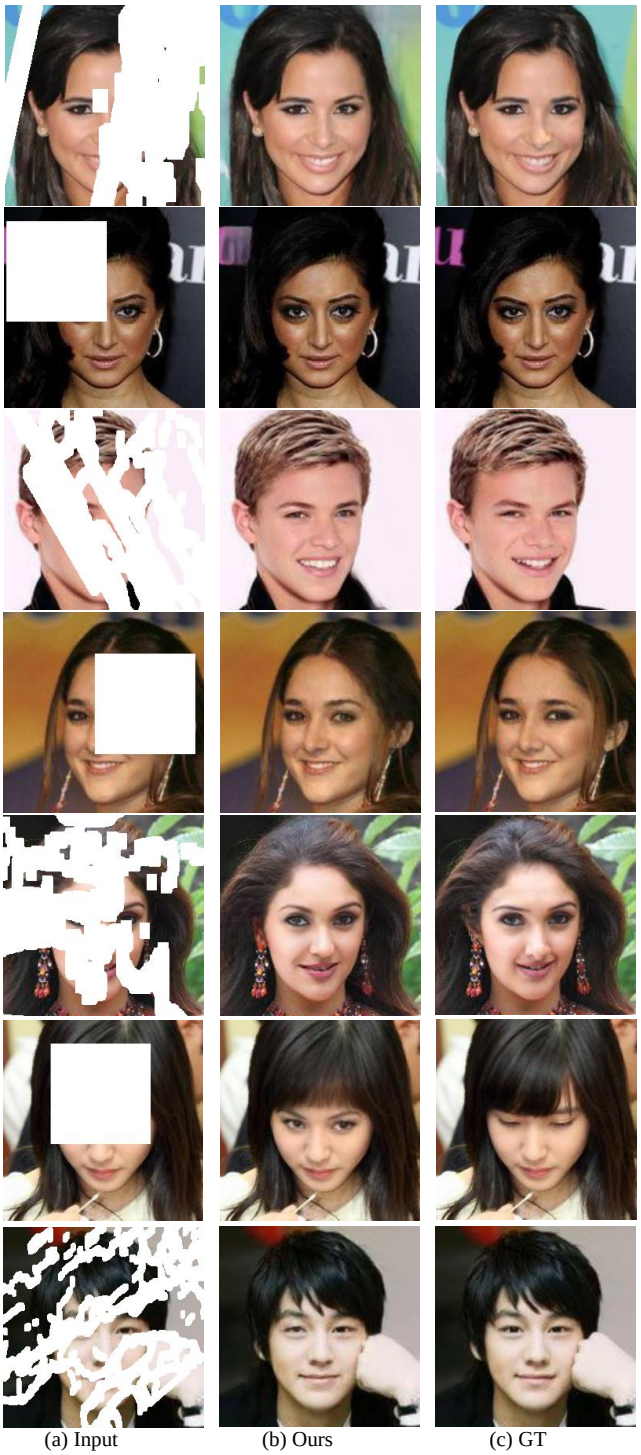


(a) Input (b) Baseline (c) MT (d) MT, SE (e) MT, SE, AT (f) GT

Figure 3: Qualitative results of the ablation study. [Best viewed with zoom-in.]

(a) Input (b) Ours (c) GT

Figure 4: Example inpainting results on Facade. [Best viewed with zoom-in.]



(a) Input (b) Ours (c) GT

Figure 5: Example inpainting results on CelebA. [Best viewed with zoom-in.]