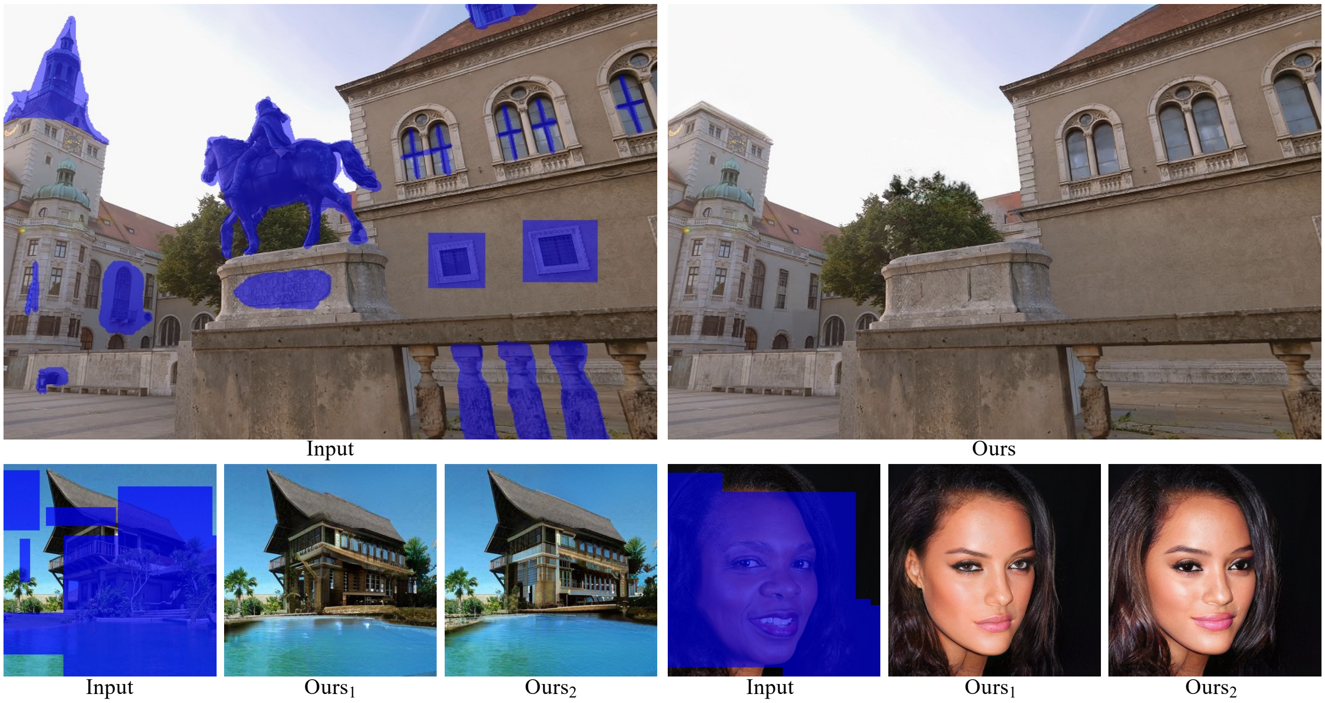
MAT: Mask-Aware Transformer for Large Hole Image Inpainting

Wenbo Li[1] Zhe Lin[2] Kun Zhou[3] Lu Qi[1] Yi Wang[4*] Jiaya Jia[1]
[1]The Chinese University of Hong Kong [2]Adobe Inc.
[3]The Chinese University of Hong Kong (Shenzhen) [4]Shanghai AI Laboratory
{wenboli,luqi,leojia}@cse.cuhk.edu.hk
zlin@adobe.com kunzhou@link.cuhk.edu.cn wangyi@pjlab.org.cn



Input Ours

Input Ours$_1$ Ours$_2$ Input Ours$_1$ Ours$_2$

Figure 1. The proposed MAT supports photo-realistic and pluralistic large hole image inpainting. The first example is a real-world high-resolution image and the other two examples (512 × 512) are from Places [78] and FFHQ [26] datasets.

Abstract

Recent studies have shown the importance of modeling long-range interactions in the inpainting problem. To achieve this goal, existing approaches exploit either standalone attention techniques or transformers, but usually under a low resolution in consideration of computational cost. In this paper, we present a novel transformer-based model for large hole inpainting, which unifies the merits of transformers and convolutions to efficiently process high-resolution images. We carefully design each component of our framework to guarantee the high fidelity and diversity of recovered images. Specifically, we customize an inpainting-oriented transformer block, where the attention module aggregates non-local information only from partial valid tokens, indicated by a dynamic mask. Extensive experiments demonstrate the state-of-the-art performance of the new model on multiple benchmark datasets. Code is released at https://github.com/fenglinglwb/MAT.

1. Introduction

Image completion (a.k.a. inpainting) is a fundamental problem in computer vision, which aims to fill missing regions with plausible contents. It has many applications including image editing [23], image re-targeting [9], photo

[*] Corresponding author