Face Inpainting Restoration with Multiple Conditions Diffusion Model

Team 14

Introduction

Face inpainting aims to restore missing or corrupted facial regions with visually plausible content. Despite recent progress, restoring severely degraded faces remains a challenging problem due to the loss of semantic and structural information. To address this, we explore a novel conditional face inpainting framework built upon the pretrained Stable Diffusion v1.5 inpainting model. Our method **incorporates both high-level semantic and low-level structural signals as multi-modal conditions**, enabling precise guidance for high-fidelity face restoration while minimizing modifications to the pretrained diffusion architecture.

Problem Formulation

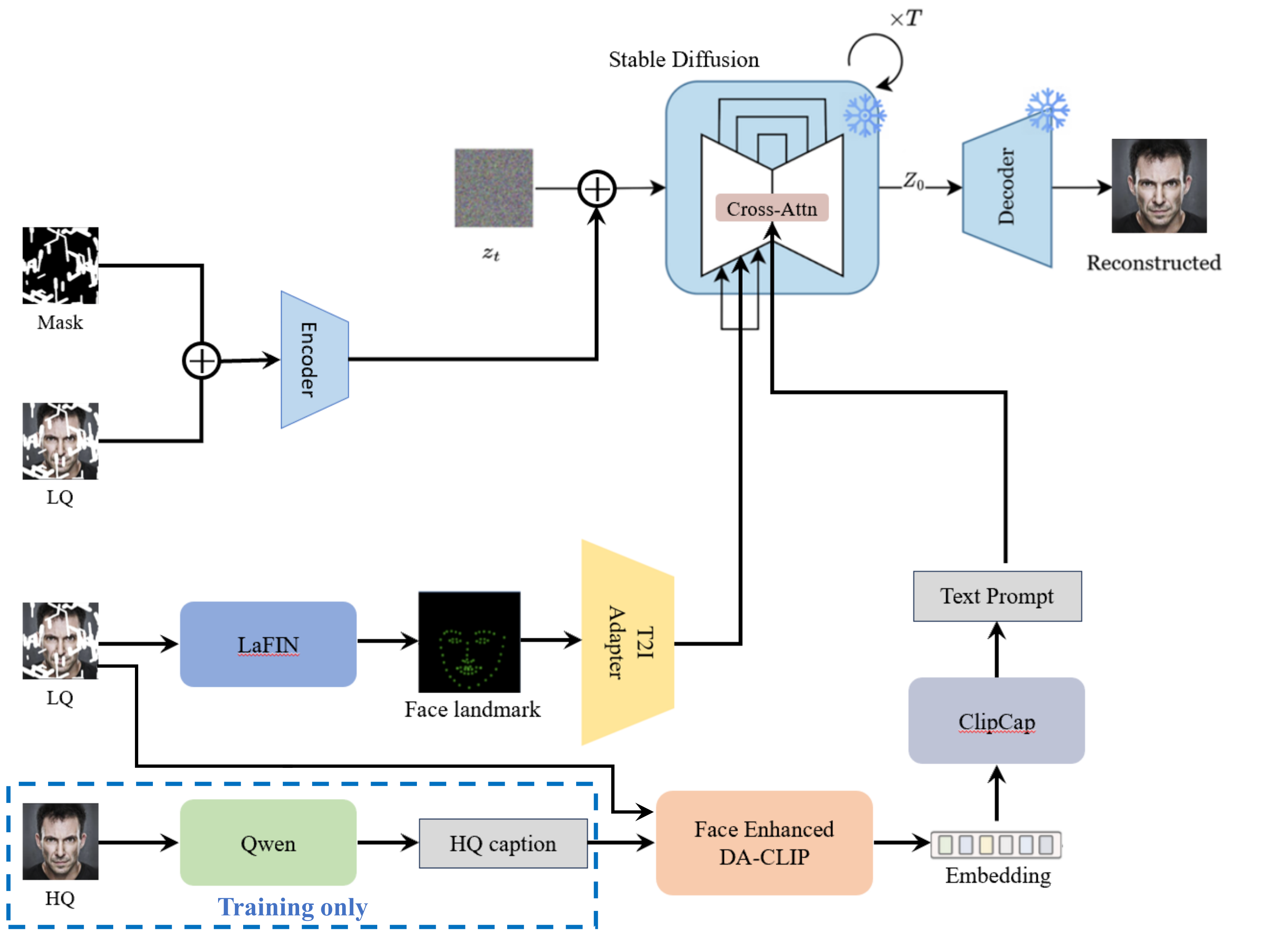
Given a low-quality (LQ) masked face image, the goal is to generate a high-quality (HQ) restored image such that it is visually realistic, semantically consistent, and structurally aligned with the original identity. Formally, we aim to learn a mapping: $x_{HQ} = D(z, z_{mask}, c_{sem}, c_{str})$

where:

- z: latent representation obtained the from VAE encoder within Stable Diffusion
- z_{mask}: latent feature derived from masked image content
- c_{sem}: text prompt derived from semantic embeddings
- c_{str}: structural guidance derived from facial landmarks

We aim to learn the adapter that inject c_{str} into the denoising process, while maintaining a frozen pretrained backbone D. The overall generation process operates in latent space, following the framework of Stable Diffusion.

Methodology



Semantic Prompt Extraction

- HQ are used to generate descriptive captions via Qwen
- Fine-tune DA-CLIP with CLIP-SMU for face attributes
- Use pre-trained ClipCap model to generate text prompt

Structural Guidance via T2I Adapter

- Extract facial landmarks from the LQ image using LaFIN
- T2I Adapter encodes landmarks into spatial feature maps
- Inject to UNet intermediate layers via residual connections

Latent Guidance from Masked Input

- Passed through encoder to produce a latent z_{mask}
- Concat with noise latent to guide the denoising process

The Stable Diffusion v1.5 inpainting model is used as the frozen backbone, with a guidance scale of 7.5 and 50 sampling steps. Conditioning is applied without retraining the backbone, enabling efficient generation of high-quality faces.

Experiments

Experiment Setup

- Model : Stable Diffusion v1.5 – inpainting
- Guidance Scale : 7.5 (classifier free)
- Resolution: 512x512
- Steps: 50
- Prompt: A detailed, natural, ultra-realistic photograph of a face close-up, an analog photo, Associated Press photo, National Geographic photo, Polaroid photo, portrait, detailed, high resolution, 4K, 8K, idyllic, scenic, Canon EOS R3, nikon, f/1.4, ISO 200, 1/160s, 8K, RAW, unedited, in-frame
- Negative prompt: unnatural, ugly, deformed, unreal eyes, deformed eyes , mutations, drawing, render, CGI, childish, painting, blurred, low-resolution, monochrome photo, masked