Imagic: Text-Based Real Image Editing with Diffusion Models

Abstract:

Text-conditioned image editing has recently attracted considerable interest. However, most methods are currently either limited to specific editing types (e.g., object overlay, style transfer), or apply to synthetically generated images, or require multiple input images of a common object. In this paper we demonstrate, for the very first time, the ability to apply complex (e.g., non-rigid) text-guided semantic edits to a single real image. For example, we can change the posture and composition of one or multiple objects inside an image, while preserving its original characteristics. Our method can make a standing dog sit down or jump, cause a bird to spread its wings, etc. Each within its single high-resolution natural image provided by the user. Contrary to previous work, our proposed method requires only a single input image and a target text (the desired edit). It operates on real images, and does not require any additional inputs (such as image masks or additional views of the object). Our method, which we call "Imagic", leverages a pre-trained text-to-image diffusion model for this task. It produces a text embedding that aligns with both the input image and the target text, while fine-tuning the diffusion model to capture the image-specific appearance. We demonstrate the quality and versatility of our method on numerous inputs from various domains, showcasing a plethora of high quality complex semantic image edits, all within a single unified framework.