Sketch-Based Scene Image Generation with Two-Stage Latent Diffusion Model

2110414 Zhang Tianyu

Image generation is currently in a stage of rapid development with new methods constantly emerging. Researchers aim to improve generated images' quality, diversity, and controllability, and explore broader application areas, such as face generation, image restoration, and image style transfer. The diffusion model is undoubtedly one of the most revolutionary technologies that have surfaced in the past few years. In computer vision and computer graphics, new text-to-image generation methods have demonstrated remarkable image quality. Such as Stable Diffusion, DALL-E, and GLIDE. Scene image generation is a challenging task in the field of image generation. Scene images typically involve the combination of multiple objects and elements, such as humans, animals, and backgrounds, with contextual relationships between them. In scene image generation, the model needs to capture various elements and details in the scene while maintaining the realism of the generated image. Additionally, the model must understand and preserve semantic connections between objects during the generation process, ensuring overall coherence and consistency in the generated images.

Despite the successes, the powerful pre-trained diffusion models still lack a high level of control that can guide the spatial properties of the scene images. The current diffusion models face the following issues: 1) the text prompts are difficult to describe the semantic information, especially in scene images; 2) text-to-image generation models lack position control of generated results; 3) diffusion models may lose the objects that depicted in text prompts.

To solve these issues, we propose a sketch-based method to control the position of corresponding objects in image generation and solve the issue of object disappearance in state-of-the-art diffusion models. The pre-trained text-to-image latent diffusion model was utilized by the proposed method as the image generator without additional fine-tuning or training. We manipulate the cross-attention layers used by the model to connect textual and visual information via the input sketch and guide the image reconstruction with the given desired layout. Specifically, we partition the model into two stages. In the feature extraction stage, the sketches are segmented into individual objects using the image segmentation approach, and the obtained bounding boxes and labels are then used as position-guided inputs to the attention layers of the diffusion models. In the image generation stage, the proposed model utilizes Latent Diffusion Model (LDM) as the generator to generate corresponding images. In the early stages of the diffusion process,

the object's position is influenced by the sketch's object position guidance in generating the attention maps.

We conduct experiments quantitatively and qualitatively to evaluate the proposed model's effectiveness. In qualitative experiments, we show the position control of the proposed method. Compared with LDM, the proposed model can control the precise positions of objects in the generated image and effectively solve the issue of object loss. We also explore that the positions of objects in the generated image are determined early in the diffusion process. In addition, We further demonstrate its versatility by changing the position relationships and relative scales in sketches. We demonstrated that after altering the spatial relationships and relative scales among sketch objects, the generated objects in the image could change accordingly, even if it contradicts the content of the text. In quantitative experiments, we compared the proposed model with LDM and Generative Adversarial Networks (GANs). We show excellent image quality (21.04 FID value) than GANs (143.1 for pix2pix, 141.5 for SketchyGAN, 87.6 for SketchyCOCO).