Text-Driven Image Editing via Learnable Regions

Reporter: Fatemeh Sadat Masoumi

Department of Computer Science, University of Texas at San Antonio. December, 2024

Abstract

The paper explores language as a natural interface for image editing and introduces a novel method for region-based image editing driven by textual prompts. Unlike traditional approaches, this method eliminates the need for user-provided masks or sketches. It leverages a pre-trained text-to-image model and integrates a bounding box generator to automatically identify editing regions aligned with textual descriptions. The paper is authored by Yuanze Lin, Yi-Wen Chen, Yi-Hsuan Tsai, Lu Jiang, and Ming-Hsuan Yang, representing a collaboration between Google, UC Merced, and the University of Oxford, and was published in April 2024. For more details, the project page is available at Learnable Regions Project Page.

Contents

1 Introduction

The paper focuses on the advancements in text-driven image editing, leveraging large-scale vision-language models and text-image paired data. Two major paradigms for image editing are discussed:

- **Mask-Based Editing:**Requires manual masks for precise local editing but is labor-intensive and user-restrictive.
- **Mask-Free Editing:**Eliminates masks but struggles with precision, particularly in local modifications due to reliance on pixel-level accuracy.

The study identifies bounding boxes as a promising alternative to pixel masks, offering a more intuitive, user-friendly approach for local editing. Unlike pixel-level masking, bounding boxes are quicker to adjust and compatible with models like Muse, which lack pixel-level editing support.

Key Contributions: **Region-Based Editing Network:** Introduces mask-free editing using bounding boxes guided by a CLIP-based loss function, enhancing pretrained text-to-image models. **Versatility:** Integrates with diverse models like MaskGIT, Muse, and Stable Diffusion. **Experimental Results:** Produces high-quality, realistic image edits aligned with text prompts, validated by a user study outperforming five state-of-the-art methods.

2 Related Work

Text-to-Image Synthesis

Recent advancements in text-to-image synthesis have transitioned from early Generative Adversarial Network (GAN) approaches to state-of-the-art diffusion and transformer-based models. Notable developments include:

- **Diffusion Models:** DALL·E 2 and Imagen condition textual prompts on diffusion frameworks to generate realistic images.
- **Transformer Models:** Muse employs masked generative transformers to produce images from textual descriptions.

- **CLIP Integration:** Models like Stable Diffusion leverage pre-trained CLIP models for guiding image generation based on text prompts.

Stable Diffusion, trained on large-scale image-text datasets, has become foundational for numerous image generation and editing tasks. Innovations like ControlNet further enhance Stable Diffusion by enabling spatial control for synthesis.

The proposed method introduces a new region generation model to enable mask-free local image editing. Using self-supervised learning (SSL) (e.g., DINO) for anchor initialization and a region generation network (RGN) for selecting optimal regions, the framework integrates with pre-trained text-to-image models. The CLIP model guides training by scoring the alignment between textual prompts and edited results. This approach enhances the capabilities of existing models for localized editing without requiring masks.

Text-driven Image Manipulation

Recent advancements in text-driven image manipulation leverage pre-trained generator models and CLIP for high-quality and controllable edits. Notable approaches include:

StyleCLIP: Combines StyleGAN and CLIP to manipulate latent codes for diverse edits. **VQGAN-CLIP:**Guides VQ-GAN with CLIP to produce high-quality image generation and editing. **Diffusion Models:**

- **Imagic:**Aligns textual embeddings with images for editing using fine-tuned diffusion models.
- **InstructPix2Pix:**Integrates GPT-3 and Stable Diffusion for human-instruction-based edits.
- **DiffEdit:**Uses DDIM inversion and auto-generated masks for local editing.
- **MasaCtrl:**Employs mutual self-attention to learn editing masks via cross-attention maps.

Building on these advancements, the proposed method introduces a bounding box-based approach for local edits, offering greater flexibility compared to mask-dependent methods like DiffEdit and MasaCtrl. By utilizing diffusion models and CLIP guidance, this approach enhances adaptability to diverse text prompts, addressing the limitations of mask sensitivity in existing methods.

3 Proposed Method

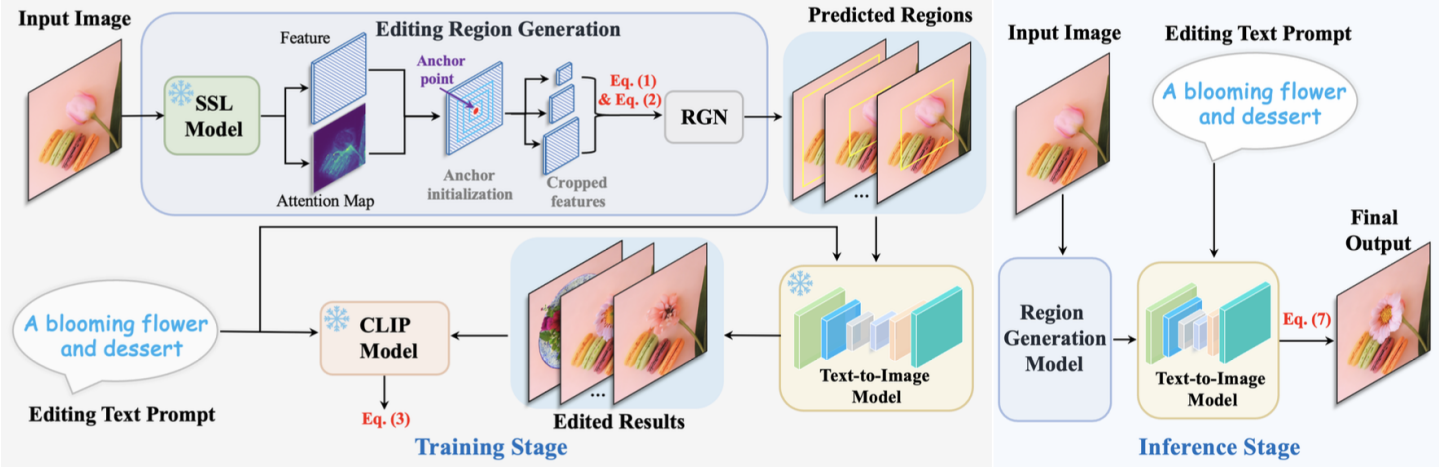

Figure 1: Framework of the proposed method.

Edit-Region Generation

Given an input image $X \in R^{3\times H\times W}$ and text $T \in Z^p$, they utilize a pre-trained Vision Transformer (ViT-B/16) trained with the DINO self-supervised objective. The extracted feature $F \in R^{d\times h\times w}$ provides semantic segmentation as a prior for region

generation.

They initialize K anchor points $\{C_i\}_{i=1}^{K}$ from the top-K scoring patches in the DINO self-attention map. For each anchor point C_i, square bounding box proposals $\{B_j\}_{j=1}^{M}$ are defined and evaluated using a Region Generation Network (RGN). Features f_j are pooled using ROI-pooling:

$$f_j = ROI - pool(F, B_j)$$

The RGN assigns scores via softmax over logits $[\pi_1, \ldots, \pi_M]$, identifying the best region:

$$Softmax([\pi_1, \ldots, \pi_M])$$

Training employs the Gumbel-Softmax trick to ensure differentiability. The highest-scoring region generates a mask, which, combined with the input image and prompt, produces the edited image. This process generates K edited images for all anchor points.

Training Objective

The proposed method leverages the CLIP model to guide image editing based on the similarity between text prompts and generated images. It introduces a composite editing loss with three components:

CLIP Guidance Loss (L_{Clip}): Measures cosine similarity between the text prompt and the edited image. **Directional Loss (L_{Dir})**: Aligns edits in the CLIP space with the intended modifications. **Structural Loss (L_{Str})**: Maintains spatial structure and layout of the source image.

The total loss is defined as:

$$L = \lambda_C L_{Clip} + \lambda_S L_{Str} + \lambda_D L_{Dir}$$

where weights $\lambda_C = 1$, $\lambda_S = 1$, and $\lambda_D = 1$ balance the loss components.

During inference, a *quality score* ranks the edited images to select the best result:

$$S = \alpha \cdot S_{t2i} + \beta \cdot S_{i2i}$$

The method ensures high-quality edits aligned with the text prompts while preserving the source image's structure and appearance.

Compatibility with Pretrained Editing Models

The proposed region generator demonstrates its versatility by integrating with various image editing models to modify source images based on textual prompts. To evaluate its applicability, we apply it to two distinct image synthesis models:

- **Non-Autoregressive Transformers:** Examples include *MaskGIT* and *Muse*, which operate on discrete tokens generated by a VQ autoencoder. These models are compatible with box-like masks but lack precision for pixel-level editing.
- **Diffusion U-Nets:** Examples include *Stable Diffusion*, which operates in continuous latent space, enabling pixel-level editing.

Key Observations: Transformers like MaskGIT and Muse work well with box-like masks but lack precision for fine edits. Diffusion models like Stable Diffusion excel at pixel-level precision for image editing tasks.

Experimental Setup:

- The *official MaskGIT model* is used instead of Muse, which is not publicly available.
- Text prompts are constrained to the class vocabulary that the model was trained on.

This setup validates the compatibility of the proposed region generator with both transformer and diffusion-based image editing frameworks, demonstrating its adaptability across different editing paradigms.

4 Experimental Results

Implementation Details

They evaluated the method using high-resolution images from Unsplash (https://unsplash.com/). For edit-region generation, 7 bounding box proposals ($M = 7$) are used with CLIP initialized by ViT-B/16.

Stable Diffusion-v-1-2 is the default editing model, with experiments conducted on two A5000 GPUs for 5 epochs using the Adam optimizer and a learning rate of 0.003.

Qualitative Evaluation

Key observations include(Figure 4 in paper):

Single-object descriptions: Handles prompts focused on one category of objects. **Multiple objects:** Successfully edits images based on prompts involving multiple objects in the scene. **Geometric relations:** Demonstrates the ability to interpret and edit based on spatial relationships between objects. **Long paragraphs:** Effectively handles detailed and lengthy descriptions for image editing.

Comparisons with Prior Work

- **InstructPix2Pix:** Alters global appearance, including background.
- **DiffEdit & MasaCtrl:** Struggle with complex prompts involving multiple objects.
- **Plug-and-Play & Next-text Inversion:** Generate unrealistic or text-inconsistent results.

User Study

A user study with 203 participants evaluated 60 images and prompts. The proposed method was preferred in 84.9% of comparisons, outperforming five baselines. Failures were due to anchor initialization in background areas.

Ablation Study

Region Generation Methods: The proposed method outperformed two baselines:

- **Random-anchor-random-size:** Editing regions sampled uniformly from the image.
- **DINO-anchor-random-size:** Regions centered at anchor points from the DINO self-attention map with random sizes.

User Preference:

- Preferred over Random-anchor-random-size in **83.9%** of comparisons.
- Preferred over DINO-anchor-random-size in **71.0%** of comparisons.

Loss Component Analysis:

- Removing the *Directional Loss* (L_{Dir}) resulted in edits that failed to fully align with text prompts.
- Omitting the *Structural Loss* (L_{Str}) led to the loss of object shape and posture from the source image.
- Using all loss components ensures adherence to text prompts while preserving source image concepts.

Limitations

The method's performance depends on the self-supervised model used for anchor initialization. Predicted regions may unintentionally include background areas, causing undesired edits.

5 Conclusions

This paper introduces a method for image editing using language descriptions, removing the need for user-defined edit regions. A region generation network and text-driven losses ensure high-quality results, validated through experiments and user studies.