摘要

在数字化时代背景下，图像编辑技术已成为人们日常生活中不可或缺的一部分。在艺术创作、影视制作、广告设计等专业领域，设计师们追求能够精确操控图像内容的工具，以实现个性化需求和创意表达。同时，在日常生活、社交网络和旅游等娱乐活动中，广大用户也在寻找低门槛的、自由编辑图像的方法，记录和分享他们的生活点滴。因此，市场上迫切需要一种高效的跨模态条件引导的可控图像编辑模型，一方面把专业人员从重复性劳动中解放出来，提升工作效率；另一方面，让普通用户能够更加便捷地编辑图像，表达他们的想法和情感。

目前，图像可控编辑技术的研究主要围绕三个方面展开探索：（1）能够粗粒度编辑图像色调与整体风格的全局属性编辑模型；（2）能够调整图像实例的轮廓与属性特征的局部实例调整模型；（3）能够支持用户通过多种方式表达编辑意图的多重指令重绘模型。然而，已有研究所提出的模型仍然存在难以实现用户提供指令、难以理解图像实例表征，以及难以解决多条用户指令间的冲突与耦合的问题。针对这些问题，本文研究了跨模态条件引导下的可控图像编辑技术，旨在克服现有模型在用户指令实现、图像实例表征处理，以及多条用户指令间冲突与耦合的优化等方面的局限性。通过设计专门的用户指令理解模块、利用预训练模型的生成先验、Transformer结构的编码功能和跨模态注意力机制的匹配能力，本文提出了精准控制图像多项属性特征的解决方案。并且，本文进一步提出了一种统一的模型架构，能够同时理解并处理多条用户指令，并对图像的具体特征执行独立的编辑操作。本文的研究成果主要包括以下几个方面：

(1) 强化指令理解的图像全局属性编辑：本文针对模型难以实现用户提供指令的问题，提出了两种解决策略。首先，针对自然语言描述作为用户指令时出现的颜色与物体的耦合与不匹配问题，提出了一个基于属性解耦的颜色编辑方法。该方法利用双线性机制建立物体与颜色之间的对应矩阵，通过注意力转移模块有效建立形容词和图像区域之间的对应关系，设计软门控注入模块确保颜色描述被正确应用于指定的图像区域，实现正确的图像颜色编辑。相比于现有的相关技术，该方法将PSNR和SSIM指标分别提高了3.56和6.3%。其次，针对自然语言作为用户指令时情感描述难以理解并可视化的问题，提出了一个基于情感分析的纹理编辑方法。该方法采用多模态Transformer架构，通过统一图像内容与情感描述的特征表示，使用跨模态注意力机制融合跨模态特征，并进一步设计情感描述理解和可视化损失函数，从而有效地将视觉抽象的情感描述投影到到视觉具象的颜色与纹理特征。在与相关方法的用户调研中，超过一半的

用户表现出对该方法生成效果的明显偏好。

（2）基于实例感知的图像局部实例编辑：本文针对模型难以理解图像实例表征的问题，为理解图像实例表征分别设计了显式和隐式的实例感知机制。当用户指令明确描述实例轮廓的调整方向时，提出了基于掩码已知的实例轮廓调整算法。该方法在第一阶段基于用户指令提取并补全输入图像的结构化线索，并在第二阶段设计了结构驱动的卷积神经网络以精准地调整图像中指定实例的轮廓，避免填充区域中不同实例间的混合。该方法在 FID、SSIM 和 PSNR 指标上获得了显著提升。当用户指令仅仅暗示实例属性的编辑方向时，提出了基于自适应感知的实例属性编辑方法，通过将图像分为相似颜色聚合成的多个颜色组，设计分组编码、分组 Transformer 和分组注意力机制等模块，使用无监督学习技术自动识别并区分图像中的不同实例并进行独立的属性调整。根据用户调研的评测，该方法与标准答案之间仅存在 2.64% 的偏差。

（3）跨模态多条件引导的图像重绘制：本文针对多条用户指令间的冲突与耦合问题，提出了两种技术策略来支持用户同时使用颜色、纹理、轮廓等不同模态的指令，从而在图片的指定区域生成与背景无缝衔接、满足用户意图的视觉内容。第一种策略采用两阶段架构，由理解与融合多模态输入条件的生成网络和根据背景区域微调重绘结果色调的拼接网络组成。第二种策略采用一体化架构，通过重新设计条件注入与融合模块，构建背景与其他条件之间的交互和依赖关系，从而避免了两阶段方法的冗余计算，进而提示了图像重绘制的性能。本文还收集了四个不同场景的数据集，并针对图像重绘制任务的需求进行了额外标注，展示了该技术在广泛应用场景中的潜力。实验结果表明，一体化图像重绘制模型显著提升了生成质量（FID）、条件一致性（R-prcn）和拼接效果（M-score）。

本文提出的三种解决方案，为可控图像编辑研究开辟了新的道路，有助于提升图像编辑的性能和降低图像编辑的技术门槛，从而支持用户以更直观和自然的方式参与图像编辑，记录和分享生活的点滴。

关键词：多模态，生成模型，可控图像编辑

Research on Controllable Image Editing Guided by Cross-modality Conditions

Shuchen Weng (Computer Application Technology)
Directed by Boxin Shi

ABSTRACT

In the era of digitalization, image editing has become an indispensable part of people's daily lives. In professional fields such as artistic creation, film production, and advertising design, content creators crave tools to control the image content for personalized needs and creative expression. On the other hand, in daily life, social networking, and entertainment activities such as tourism, users also need handy methods to edit images at ease, to record and share the moments of their lives. Thus, there is an urgent need for an efficient controllable image editing model guided by cross-modality conditions, which not only improves professionals' efficiency by saving them from repetitive work but also helps ordinary users express their ideas and emotions by editing images more conveniently.

Currently, research on controllable image editing primarily explores three areas: (1) global property editing models that can coarsely edit image colors and overall style; (2) local instance adjustment models that can modify the contour and property features of specified instances; and (3) repainting models that supports expressing editing intentions with multiple instructions across various modalities. Despite these advancements, previous works still struggle to implement the instructions provided by users and the instances in the image. They also cannot resolve the conflicts and dependencies in the instructions. To address these issues, this thesis delves into controllable image editing guided by cross-modality conditions, aiming to overcome the limitations of existing models in understanding user instructions, processing instances in the image, and resolving conflicts and dependencies in the instructions. By designing a specialized user instruction implementing module, leveraging the generative priors of pre-trained models, the powerful encoding capability of the Transformer, and the matching ability of the cross-attention mechanism, this thesis proposes solutions for precisely controlling multiple property features of images. Furthermore, this thesis proposes a unified model architecture that can understand and process multiple user instructions and execute independent editing operations on specific image features. The contributions include the following aspects:

(1) Image global property editing enhanced by instruction understanding: This thesis proposes two solutions to the problem of difficulty in implementing user instructions. First, for the coupling and mismatch between colors and objects when natural language descriptions are given as user instructions, a color editing method based on color-object decoupling is proposed. This method uses a bilinear mechanism to establish a correspondence matrix between objects and colors, effectively establishes relationships between adjectives and image regions through an attention transfer module, and ensures correctly assign color descriptions to specified image regions with a soft-gate injection module, achieving accurate image color editing. Compared to previous state-of-the-art methods, the proposed method has improved PSNR and SSIM by 3.56 and 6.3%, respectively. Moreover, for the difficulty in understanding and visualizing emotional descriptions when natural language is used as user instructions, an emotion analysis-based texture editing method is proposed. This method adopts a multi-modal Transformer architecture, unifies the feature representations of image content and emotional descriptions, fuses cross-modality features through a cross-attention mechanism, and further designs emotional description understanding and visualization loss functions, effectively projecting visually abstract emotional descriptions onto visually concrete color and texture features. In the user study, more than half of the users show a clear preference for the proposed method on the synthesis results.

(2) Instance-aware local image instance editing: This thesis designs an explicit instance perception mechanism and an implicit one to understand the instances in the image, depending on whether user instructions specify the instance. When user instructions explicitly describe the direction of instance contour adjustment, a two-phase instance contour adjustment algorithm is proposed. This method extracts and completes structured cues from the input image based on user instructions in the first stage, and designs a structure-driven convolutional neural network to precisely adjust the contours of specified instances in the image in the second stage, avoiding the blending and mixture of different instances in the areas to be inpainted. This method significantly improves the quantitative results in terms of FID, SSIM, and PSNR. An adaptive instance-aware property adjustment method is proposed, in the case of the user instructions which only imply the direction of instance property adjustment. It aggregates the color in the image into multiple color groups, applies customized group embeddings and tailored modules such as the grouping Transformer and the grouping attention mechanism, and uses unsupervised learning techniques to automatically identify different instances in the image for independent property adjustments. In the user study experiment, this method performs

only 2.64% lower than the ground truth.

(3) Cross-modality multi-condition guided image repainting: This thesis proposes two technical strategies to support users' instructions in different modalities such as color, texture, and geometry simultaneously, thus generating visual content in specified regions of the picture that seamlessly integrates with the background and meets user intentions. The first strategy adopts a two-phase architecture, consisting of a generative network that understands and integrates multi-modal input conditions and a joining network that refines the tone of the repainted result according to the background regions. The second strategy adopts an unified architecture, by redesigning the condition injection and integration module, constructing the interaction and dependence between the background and other conditions, thus avoiding the redundant computation of the two-phase method and thereby enhancing the performance of image repainting. We also collect datasets from four different scenarios and conduct additional annotations tailored to the needs of the image repainting task, demonstrating the potential of this technology in a wide range of scenarios. Experimental results show that the unified image repainting model achieves the best synthetic quality (FID), condition consistency (R-prcn), and compositing effect (M-score).

The three solutions proposed in the thesis pave novel avenue for the research in controllable image editing, expected to improve the performances of image editing methods and lower the skill barriers, enabling a more intuitive and natural way to edit the image for the majority of users.

KEY WORDS: Multi-modal, generative model, controllable image editing