摘要

当通过玻璃窗等透明介质拍摄照片时，被污染的混合图像（记为 **M**）可以被视为背景层（记为 **T**）和反射层（记为 **R**）的组合，这样的图像不仅不利于人眼观测，还会损害下游计算机视觉任务的性能。因此，反射消除——旨在去除不需要的反射并从受污染的混合图像中恢复清晰的背景层——已成为计算摄影领域的一个热门话题。最先进的单图像反射消除方法主要通过从合成和真实数据的混合数据集中学习先验来消除反射。然而，由于缺乏解决此问题所需的背景和反射场景的充分的知识，它们在处理复杂或低透射反射的场景时常常遇到限制。多图像反射消除方法通过利用从专门设备或特殊拍摄设置获得的额外的辅助场景信息，实现了比单图像方法更稳健的反射消除效果。然而，特殊的数据采集要求限制了它们的应用范围，特别是对于移动设备和互联网上的图像。因此，需要找到一个用户友好的辅助输入，既能帮助缓解反射消除的病态性，又能保持单图像方法的适用性和可访问性。

近期，通过结合自然语言描述的文本提示和预训练扩散模型的生成先验，基于语言的扩散模型在图像着色和超分辨率等图像恢复任务中生成了视觉效果令人满意的结果。它在反射消除除中也显示出提供语义先验以消除反射的潜力，其中预训练扩散模型中的生成先验可以巧妙地处理复杂场景，并使在低透射反射下恢复背景层成为可能。然而，直接将现有的基于语言的扩散模型用于反射消除将遇到控制条件不准确和恢复图像保真度不足的挑战。

基于此，本文提出了一种新的用于单图像反射消除的基于语言的扩散模型，旨在在强反射的干扰下通过语言引导实现高保真度的背景图像恢复。利用背景层的描述作为正面提示，指导模型保留必要的图像内容，而反射层的描述作为负面提示，抑制不需要的反射。同时，本文还在反向扩散过程中提出了一种迭代条件优化策略，确保随着去躁过程，颜色和结构条件的表示越来越准确。此外，本文引入了一种多条件约束机制，旨在保护指定条件的保真度，通过让恢复的背景图像与所需颜色和结构条件对齐，有效地解决了潜在的颜色偏移和结构失真问题。

进一步地，针对方法在反射区域定位不精确和生成结果色彩失真等局限，本文引入了空间感知与色彩可控机制。通过允许用户交互式提供区域掩码，对具体反射区域进行定向处理，显著提升了反射消除的空间精准性和用户可控性。同时，提出的色彩矫正模块有效保证了恢复图像的色彩一致性，增强了模型在极端光照和复杂场景下的表现能力。为支撑方法评估与泛化，本研究还构建了高质量的半合成训练集及覆盖多场景的真实反射消除测试集。

本文的主要贡献概括如下：

1. 提出了一种用于反射消除的基于语言的扩散模型，利用文本提示强化对背景层和反射层的区分。
2. 创新性地设计了迭代条件优化策略和多条件约束机制，为反射消除提供更准确的颜色与结构指导，提升了颜色与结构恢复保真度。
3. 引入空间感知和色彩矫正机制，实现了区域精准的可控反射消除，满足灵活输入的实际需求，并且保证了背景层的色彩自然和真实。

综上，本文提出了一套专门针对单图像反射消除任务的融合文本提示与空间感知控制的扩散模型框架。新方法以多模态条件和灵活控制实现高质量反射消除，兼顾目标图像的保真性、一致性与可控性。大规模实验验证了所提出方法在定量和定性结果上较现有方法的显著优势，在复杂反射和真实场景下展示了优良的泛化能力。

关键词：反射消除，图像恢复，基于语言的扩散模型

Research on Reflection Removal Algorithm Integrating Textual Prompts and Generative Priors

Haofeng Zhong (Computer Science and Technology (Intelligent Science and Technology))
Directed by Boxin Shi

ABSTRACT

When photographing through transparent media such as glass windows, the resulting mixture image (denoted as **M**) can be regarded as a combination of a transmission layer (**T**) and a reflection layer (**R**). Such images often degrade the performance of downstream computer vision tasks. Consequently, reflection removal—aimed at eliminating unwanted reflections and recovering a clean background from contaminated mixtures images has become a widely studied problem in computational photography. The state-of-the-art single-image reflection removal methods generally learn priors from hybrid datasets that combine synthetic and real-world examples. However, due to insufficient knowledge about the intricacies of real background and reflection composition, these methods frequently struggle in complex or low-transmission scenarios. While multi-image reflection removal approaches leverage additional scene cues collected via specialized hardware or controlled acquisition setups for more robust separation, their data collection requirements inherently restrict practical applicability, especially on mobile devices or Internet images. As such, there is an urgent need for a user-friendly form of auxiliary input that can mitigate the ill-posedness of reflection removal while maintaining the accessibility and applicability of single-image approaches.

Recently, language-driven diffusion models, which combine natural language prompts with generative priors from pre-trained diffusion models, have achieved visually impressive results in image restoration tasks such as colorization and super-resolution. They have also demonstrated potential in providing semantic priors for reflection removal, with the generative prior enabling the handling of complex scenarios and the recovery of backgrounds even in severe reflection conditions. However, directly applying existing language-driven diffusion models to reflection removal faces challenges in terms of insufficient control over conditions and suboptimal restoration fidelity.

To address these issues, this thesis proposes a novel language-guided diffusion framework

for single-image reflection removal, designed to restore high-fidelity background images in the presence of strong reflections via language prompting. Positive prompts, describing the background layer, guide the model to retain essential image content, while negative prompts, describing the reflection layer, suppress unwanted reflections. Additionally, our work introduces an iterative conditional optimization strategy within the reverse diffusion process to progressively refine the accuracy of color and structure constraints. A multi-constraint framework is further employed to ensure fidelity to specified conditions—effectively mitigating color shifts and structural distortions in the recovered background.

Furthermore, to overcome limitations related to inaccurate reflection region localization and color distortion in generated results, we introduce spatial-aware and color-controllable mechanisms. By allowing users to interactively provide spatial masks for targeted reflection removal, our method improves spatial precision and controllability. The developed color correction module ensures consistency in color restoration, significantly enhancing performance in extreme lighting and complex scenarios. To support evaluation and promote generalization, we also construct a high-quality semi-synthetic training set and a comprehensive real-world benchmark for reflection removal.

The main contributions of this thesis can be summarized as follows:

1. This thesis proposes a language-driven diffusion model for reflection removal, which utilizes text prompts to enhance the distinction between the background and reflection layers.
2. This thesis designs an iterative conditional optimization strategy and a multi-condition constraint mechanism, providing more accurate color and structure guidance for reflection removal, and improving the fidelity of color and structure restoration.
3. This thesis introduces spatial awareness and color correction mechanisms to achieve region-specific and controllable reflection removal, meeting the practical needs of flexible input while ensuring the naturalness and authenticity of the recovered background colors.

In summary, this thesis presents a diffusion-based framework for single-image reflection removal, integrating language prompts and spatial awareness. The proposed method achieves high-quality, controllable reflection removal by leveraging multi-modal conditional guidance. Extensive experiments, both quantitative and qualitative, demonstrate that our approach significantly outperforms existing methods, and exhibits strong generalization capability in real-world and challenging reflection scenarios.

KEY WORDS: Reflection Removal, Image Restoration, Language-based Diffusion Model