中图法分类号:TP18; TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2026)06-2103-22

论文引用格式:Shu Y, Zhao F M, Chen Z Y, Zhao T Q, Wang Y Z, Li K C, Zhou Y, Wang D L, Peng L R, Gao L C and Yin X C. 2026. Survey on visual text generation techniques. Journal of Image and Graphics, 31(6):2103-2124(舒言, 赵方敏, 陈泽宇, 赵天齐, 王逸竹, 李焜炽, 周宇, 王大寒, 彭良瑞, 高良才, 殷绪成. 2026. 可视文本生成技术综述. 中国图象图形学报, 31(6):2103-2124)[DOI:10.11834/jig.260047]

可视文本生成技术综述

舒言[1],赵方敏[2],陈泽宇[1],赵天齐[3],王逸竹[3],李焜炽[4],周宇[1],
王大寒[4],彭良瑞[3],高良才[5],殷绪成[6*]

1. 南开大学计算机/密码与网络空间完全学院,天津　300350; 2. 中国科学院信息工程研究所,北京　100085;
3. 清华大学电子工程系,北京　100084; 4. 厦门理工学院计算机与信息工程学院,厦门　361024;
5. 北京大学王选计算机研究所,北京　100871; 6. 北京科技大学计算机与通信工程学院,北京　100083

摘　要: 可视文本图像生成与编辑是计算机视觉与自然语言处理交叉领域的重要研究方向,旨在实现图像中文本内容的无痕擦除、精准编辑与智能生成。不同于一般图像生成任务,可视文本兼具语义信息与视觉特征的双重属性,在字形结构、笔画细节、颜色纹理和排版布局等方面对模型的多模态表征能力和生成精度提出了更高的要求。随着生成对抗网络(generative adversarial network,GAN)、扩散模型以及多模态大模型的快速发展,该领域在技术范式与应用场景上取得了显著突破。本综述系统梳理了可视文本擦除(visual text removal)、可视文本编辑(visual text editing)与可视文本生成(visual text generation)三大核心任务的研究进展。在可视文本擦除方面,知识迁移、多任务学习与渐进式学习三大范式推动了文本检测与背景修复能力的协同优化,在保留背景完整性的前提下实现了文本的彻底消除;在可视文本编辑方面,从基于GAN的分步处理到端到端的条件生成,研究聚焦于文本风格特征、笔画特征与语义特征的精准提取与迁移,实现了风格保持与内容替换的统一建模;在可视文本生成方面,研究已从早期基于图形学的渲染合成演进到数据驱动的神经生成,通过引入字符感知编码、字形条件控制与多模态对齐机制,显著提升了文本拼写准确性、场景融合度与多语言泛化能力。本综述进一步分析了该领域面临的核心挑战:多语言复杂字符的精准渲染、跨场景跨风格的泛化能力、生成内容与人类意图的精确对齐,以及实时交互所需的计算效率。展望未来,随着多模态大模型能力的持续增强、扩散模型架构的不断优化,以及高质量基准数据集的完善,可视文本图像生成与编辑技术将在智能媒体创作、信息可视化、文化遗产保护以及无障碍阅读等领域发挥更加重要的作用,成为推动人机交互与视觉智能发展的关键技术。

关键词: 可视文本擦除(VTR);可视文本编辑(VTE);可视文本生成(VTG);扩散模型;多模态学习;图像生成

Survey on visual text generation techniques

Shu Yan[1], Zhao Fangmin[2], Chen Zeyu[1], Zhao Tianqi[3], Wang Yizhu[3], Li Kunchi[4],
Zhou Yu[1], Wang Dahan[4], Peng Liangrui[3], Gao Liangcai[5], Yin Xucheng[6*]

1. College of Computer Science/Cyberspace Security, Nankai University, Tianjin　300350, China; 2. Institute of Information Engineering, Chinese Academy of Sciences, Beijing　100085, China; 3. Department of Electronic Engineering, Tsinghua University, Beijing　100084, China; 4. School of Computer and Information Engineering, Xiamen University of Technology, Xiamen　361024, China; 5. Wangxuan Institute of Computer Technology, Peking University, Beijing　100871, China; 6. School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing　100083, China

收稿日期:2026-01-21;**修回日期**:2026-02-18;**预印本日期**:2026-02-25

*通信作者:殷绪成　xuchengyin@ustb.edu.cn

基金项目:国家自然科学基金项目(62376266,62406318,62576301);多模态人工智能系统全国重点实验室开放课题项目(MAIS2024101)

Supported by: National Natural Science Foundation of China (62376266,62406318,62576301); Open Project of the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS2024101)

Abstract: Visual text image generation and editing are important research directions at the intersection of computer vision and natural language processing, aiming to achieve seamless text removal, precise text editing, and intelligent text generation in images. In contrast with general image generation tasks, visual text possesses the dual attributes of context semantic information and visual graph features, imposing higher requirements for models' multimodal representation capability and generation precision in terms of glyph structure, stroke details, color texture, and layout composition. With the rapid development of generative adversarial networks (GANs), diffusion models, flow matching, and multimodal large models (e. g. , CLIP and Flamingo), this field has achieved significant breakthroughs in technical paradigms and practical application scenarios over the past decade. This survey systematically reviews research progress in three core tasks: visual text removal (VTR), visual text editing (VTE), and visual text generation (VTG), which constitute the technical system of visual text image processing when they are combined. The three tasks are mutually complementary and form a closed loop of visual text processing: VTR lays the foundation for clean image background, VTE realizes flexible modification of existing text information, and VTG completes the active creation of text content in images, jointly supporting the full chain application of visual text technology. In VTR, three major paradigms, namely, knowledge transfer, multitask learning, and progressive learning, have continuously advanced the synergistic optimization of text detection and background inpainting capabilities. Knowledge transfer paradigms leverage pretrained image inpainting models (e. g. , contextual residual aggregation networks) to enhance background restoration quality, effectively addressing the problem of texture inconsistency caused by the independent training of detection and inpainting modules. Multitask learning frameworks integrate text detection and inpainting into a unified model architecture, sharing feature extraction backbones to reduce cumulative errors between upstream and downstream tasks, and improving end-to-end processing efficiency. By contrast, progressive learning approaches adopt a coarse-to-fine strategy. First, text regions based on semantic features are roughly located and eliminated. Then, background details are iteratively refined by fusing local texture and global structure information, achieving thorough text elimination while maximally preserving the integrity and consistency of the original background texture, lighting, and spatial structure. In VTE, technical evolution shifts from GAN-based stepwise processing (including text detection, removal, and regeneration) to end-to-end conditional generation models, considerably improving the efficiency and fidelity of editing results. Early stepwise methods suffer from evident seams between edited text and the background due to the separation of each module. Meanwhile, modern end-to-end models focus on the precise extraction and cross-domain transfer of text style features (e. g. , font, size, color, and transparency), stroke features (e. g. , thickness, smoothness, texture, and wear degree), and semantic features. By introducing attention mechanisms to focus on the correlation between text and background regions, and style embedding modules to encode scene-specific visual attributes, these models realize a unified modeling of style preservation and content replacement. Such modeling enables seamless editing effects wherein the edited text is consistent with the surrounding scene in terms of visual appearance, layout logic, and lighting conditions, effectively avoiding disharmony of the edited content. In VTG, research has undergone a fundamental transformation from early graphics-based rendering synthesis (relying on predefined font libraries and layout rules) to data-driven neural generation, breaking through the limitations of fixed styles and single scenes in traditional methods. Recent advances in character-aware encoding (which captures fine-grained glyph features), glyph-conditioned control (which regulates text shape and layout), and multimodal alignment mechanisms (which align text content with image context) have significantly improved the performance of text generation. Specifically, these improvements are reflected in three aspects: text spelling accuracy (decreasing missing characters, distortions, and errors), scene coherence (matching the lighting, perspective, texture, and noise of the target image), and multilingual generalization (supporting complex scripts, such as Chinese, Arabic, Sanskrit, and other languages with irregular stroke structures). These advancements have made neural text generation more adaptable to real-world application scenarios, such as advertising design, paper poster, scene customization, and intelligent annotation. This survey further analyzes the core challenges that confront the field, restricting the large-scale practical application of visual text image technologies. First, accurate rendering of multilingual complex characters remains difficult due to the diverse glyph structures, stroke variations, and semantic associations across languages. For example, Chinese calligraphy with freehand brushwork and Arabic cursive script pose considerable challenges to model feature extraction. Second, models lack strong generalization capability across diverse scenes (e. g. , indoor, outdoor, low-light, and motion-

blurred environments) and text styles (handwritten, printed, artistic fonts, and worn ancient text), frequently exhibiting a sharp performance decline when facing unseen scenarios. Third, achieving precise alignment between generated content and human intention requires more effective interaction mechanisms to capture fine-grained user requirements, because current models experience difficulty in accurately understanding ambiguous editing instructions. Finally, the high computational cost of current models (especially diffusion and large multimodal models) hinders their deployment in real-time interaction scenarios, such as mobile devices and edge computing platforms. In the future, with the continuous enhancement of multimodal large model capabilities, the ongoing optimization of diffusion model architectures (e. g. , efficiency-oriented lightweight designs, distillation strategies, and fast sampling algorithms), and the refinement of high-quality benchmark datasets (covering more diverse languages, scenes, and text styles), visual text image generation and editing technologies will play increasingly important roles in multiple fields. In intelligent media creation, they can assist designers in quickly generating and editing text elements in posters, videos, and animations. In information visualization, they can dynamically generate scene-adaptive text labels for data charts. In cultural heritage preservation, they can restore blurred or damaged text in ancient artifacts and manuscripts without damaging the original cultural relics. In accessible reading, they can generate large-font, high-contrast accessible text for visually impaired users to improve reading experience. Ultimately, these technologies will become key enablers for advancing human-computer interaction and visual intelligence in the digital era, bridging the gap between text information and visual scenes.

Key words: visual text removal(VTR); visual text editing(VTE); visual text generation(VTG); diffusion models; multimodal learning; image generation

0 引　言

随着人工智能技术的快速发展，计算机视觉与自然语言处理的交叉融合催生了一系列创新应用场景。其中，图像中的文本信息作为连接视觉与语言的关键桥梁，承载着丰富的语义内容与独特的视觉表现形式。从街景照片中的店铺招牌、社交媒体上的图文海报，到文档扫描件中的印刷文字，可视文本无处不在，其智能化处理需求日益迫切。在此背景下，可视文本图像生成与编辑技术应运而生，成为计算机视觉领域的重要研究方向。

不同于一般的跨模态生成任务，嵌入于图像中的文本具有语义与视觉的双重属性。一方面，文本承载着明确的语言信息，需要保证字符拼写准确、语义表达清晰；另一方面，文本作为视觉元素，在字体字形、笔画结构、颜色纹理和排版布局等方面呈现出高度的多样性与复杂性。这种双重属性使得可视文本的生成与操控在技术上更具挑战性，对模型的多模态表征能力、生成精度与场景适应性提出了更高的要求。

随着生成对抗网络(generative adversarial network，GAN)(Goodfellow 等，2020)、扩散模型(diffusion model)以及多模态大模型的突破性进展，可视文本处理技术经历了从传统方法到深度学习的跨越式发展，在文本擦除、编辑与生成三大核心任务上取得显著进展。在可视文本擦除(visual text removal，VTR)方面，该任务旨在保留图像背景完整性与视觉一致性的前提下消除文本区域，应用于隐私保护与图像编辑等场景。早期传统方法依赖手工特征与分步处理，难以应对复杂场景挑战(Telea，2004；Khodadadi 和 Behrad，2012)。2017 年，Nakamura 等人(2017)提出首个基于神经网络的 VTR 方法，明确了“文本消除—背景修复”的核心任务方向。近年来研究围绕“精准捕捉文本笔画特征、平衡擦除彻底性与背景完整性”形成了三大技术范式：知识迁移范式借助预训练检测或分割模型获取文本位置信息作为修复指导(Conrad 和 Chen，2021；Qin 等，2018)；多任务学习范式通过统一框架联合训练文本特征学习与背景修复(Zhang 等，2019；Liu 等，2022)；渐进式学习范式模拟“粗修—精修”过程逐步细化掩码与修复结果(Du 等，2023b)。此外，基于扩散模型的 TextDestroyer 等方法展现出优异泛化能力(Li 和 Chao，2024)。

在可视文本编辑(visual text editing，VTE)方面，该任务旨在保留原始背景风格、位置和尺寸的基础上对文本进行精细化调控，涵盖风格编辑与内容编辑两大功能，应用于图像翻译、数字媒体编辑与增强

现实等场景。研究从早期基于传统图像处理的局部编辑演进到深度生成模型。STEFANN（scene text editor using font adaptive neural network）首次实现字符形状与纹理特征的解耦处理（Roy 等，2020）；SRNet（end-to-end trainable style retention network）引入骨架结构约束文本形态，成为后续模型的基础架构（Wu 等，2019）；SwapText 通过注意力模块与内容形状转换网络提升弯曲文本生成能力（Yang 等，2020）；RewriteNet 增加文本识别模块并设计针对性训练策略增强真实场景泛化（Lee 等，2021）；Text-StyleBrush 基于 StyleGAN 构建多层次融合架构并采用无监督训练方式（Krishnan 等，2023）。随着扩散模型兴起，越来越多的工作将 VTE 建模为条件生成任务。研究聚焦于 3 类特征提取：文本风格特征（字体、颜色和布局）可通过隐式学习或显式定义（Wu 等，2019；Zeng 等，2024；Chen 等，2023a）；文本笔画特征通过模板表示或字符感知编码提供字形渲染条件（Liu 等，2023a）；文本语义特征通过识别损失辅助监督提升渲染准确性（Lee 等，2021；Zhang 等，2024a）。

在可视文本生成（visual text generation，VTG）方面，该任务旨在自然图像或合成场景中自动生成可读文本，使其在字体、颜色、光照、透视与材质等方面与背景高度融合，应用于场景文本合成、广告排版、Logo 生成与跨语言替换等场景。研究目标概括为 3 个维度：保真度（全局外观与局部细节真实感）、合理性（文本位置与场景几何语义协调）和实用性（提升下游任务性能）（Gupta 等，2016；Zhan 等，2018）。早期 VTG 采用基于图形学的渲染合成方法，如 SynthText 管线通过深度估计与光照建模生成大规模合成数据（Jaderberg 等，2014），但在复杂字体、多语言文本与语义一致性方面存在不足（Liao 等，2020；Long 和 Yao，2020）。随着 GAN、扩散模型与多模态大模型发展，VTG 从“规则驱动渲染”演进到“数据驱动的神经生成”。一方面，通过引入字符字形、位置掩码等显式结构先验增强可控性，如 Spatial Fusion GAN、SynthTIGER 显式控制文本姿态与风格（Zhan 等，2019；Yim 等，2021），Character-Aware Models、GlyphControl 通过字符感知编码显著提升拼写准确率（Liu 等，2023a；Yang 等，2023b）；另一方面，结合视觉编码器、语言模型与扩散生成器，如 TextDiffuser、AnyText、TextFlux 等在布局控制、多语言生成与 OCR-free 架构方面推动技术发展（Chen 等，2023b；Tuo 等，2024b；Xie 等，2025）。

该研究方向不仅推动了多模态智能的理论与方法创新，更在多个实际场景中展现出广阔的应用前景。在隐私保护领域，文本擦除技术可用于遮挡敏感信息如身份证号、车牌号等；在广告与海报设计中，智能文本生成与编辑可大幅提高创作效率；在跨语言传播场景中，文本替换技术能够实现场景图像的快速本地化；在文化遗产保护中，相关技术有助于修复受损文献与碑刻中的文字信息；在无障碍阅读与数字媒体编辑领域，也展现出重要的应用价值。图 1 展示了可视文本擦除、可视文本编辑和可视文本生成方法。

1 国际研究现状

1.1 生成式基础模型

近年来，生成式模型已从早期的对抗生成网络逐步发展到扩散模型，成为计算机视觉中数据生成与数据增强的重要技术基础。相关研究已系统总结了生成式方法在视觉数据构建、合成与应用中的发展脉络和关键挑战（马愈卓 等，2025）。

1.1.1 生成对抗网络

生成对抗网络是可视文本生成的重要基础模型之一。原始 GAN 模型（Goodfellow 等，2014）通过生成器 G 与判别器 D 的对抗训练，目标是让生成器 G 生成以假乱真的样本，判别器 D 学会区分真伪。由于原始的 GAN 模型存在训练不稳定、梯度消失、模式崩溃（mode collapse）、生成图像可控性差以及图像质量不高等问题，制约了其应用范围。

为提升训练稳定性，深度卷积 GAN（deep convolutional GAN，DCGAN）（Radford 等，2016）采取了一系列措施，包括在生成器 G 中使用转置卷积（transposed convolution）进行上采样，在判别器 D 中采用具有适当步长的卷积代替池化（pooling）等，成为后续众多 GAN 相关工作的“骨架”模型。

为缓解梯度消失与模式崩溃问题，WGAN（Wasserstein GAN）（Arjovsky 等，2017）利用最优传输的 Wasserstein 距离作为训练优化目标；Gulrajani 等人在此基础上引入梯度惩罚约束以保证 Lipschitz 连续性（Gulrajani 等，2017），提高了模型稳定性。

在提升图像生成可控性和质量方面，条件 GAN（conditional GAN，cGAN）（Mirza 和 Simon，2014）在生

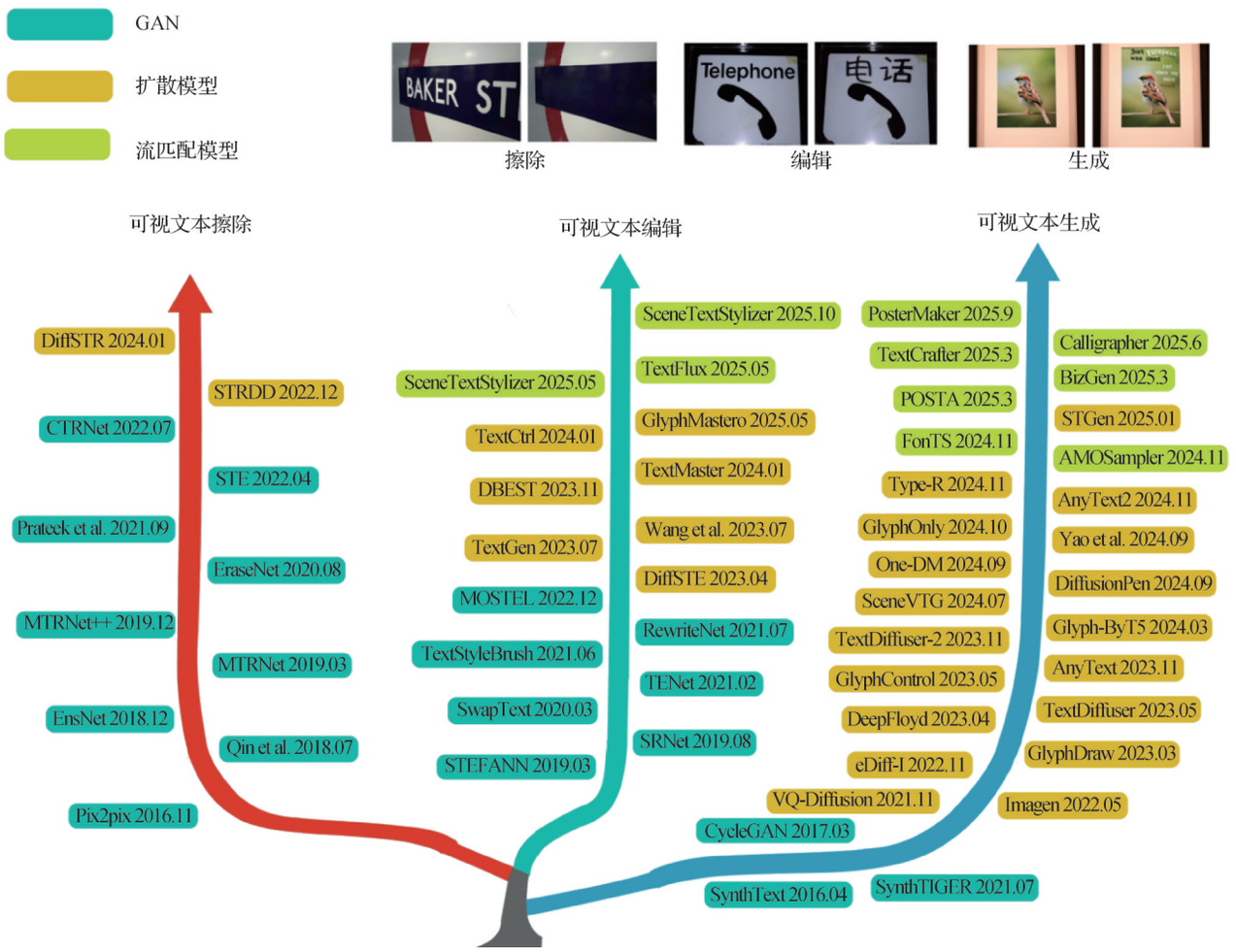

图 1　可视文本擦除、编辑、生成方法概览

Fig. 1　An overview of visual text removal, editing and generation methods

成器与判别器中引入类别或文字条件信息，实现具有条件控制的图像生成。InfoGAN（Chen 等，2016）通过极大化潜变量与生成样本间的互信息，增强潜变量的可解释性。Progressive GAN（Karras，2017）采用逐步生长的分辨率生成策略，可生成分辨率高达 1 024 × 1 024 像素的人脸图像。随后，StyleGAN（Karras 等，2019）引入映射网络和样式块，分离高层语义与随机细节，从而实现对视觉风格的精细控制。BigGAN（Brock 等，2018）通过截断技巧与正交正则化在 ImageNet 上实现高质量生成。为解决非配对数据的生成问题，CycleGAN（Zhu 等，2017）通过循环一致性损失实现无配对图像之间的双向映射。

在上述发展的基础上，近期研究工作进一步推动了 GAN 在高分辨率及多模态领域的应用。例如，StyleGAN-XL（Sauer 等，2022）将 Projected GAN 的高效判别器设计与渐进式训练策略相结合，使得 StyleGAN3 首次能够在 ImageNet 等大规模数据集上实现稳定的高分辨率图像生成。

在文本相关的图像生成任务（尤其是手写文本图像生成）中，也出现了专门面向可变长度文字和风格迁移的模型。ScrabbleGAN（Fogel 等，2020）提出以“字母拼贴”方式进行生成：将每个字符视为可独立建模的图像块，通过可变长度的特征拼接实现任意长度单词的合成，从而显著提升了手写文本生成的灵活性。JokerGAN（Zdenek 和 Nakayama，2020）进一步解决了 ScrabbleGAN 在大字符集场景中的模型膨胀问题。JokerGAN 引入多类别条件批归一化，使得模型参数规模几乎与字符集大小无关，能够支持如日语等拥有大量字符的语言。同时，该模型加入“文本基线位置”条件，使生成器能够区分上下延伸字符，提高行内字符的对齐与版面稳定性，在手写文本生成质量和对手写文本识别（handwritten text recognition，HTR）的数据增强效果方面均表现优异。

1.1.2 扩散模型

国际扩散模型研究主要围绕“从概率图模型到高保真图像生成”、“从像素空间到潜空间”以及“从无条件到多模态条件控制”等方向系统展开，代表性研究机构包括美国加州大学伯克利分校、美国斯坦福大学、OpenAI、Google Research、德国慕尼黑工业大学等。Ho等人(2020)提出的DDPM(denoising diffusion probabilistic model)奠定了离散时间扩散模型的基本框架，通过前向高斯加噪与反向去噪马尔可夫链建模，将去噪得分匹配与变分推断联系起来，在CIFAR-10 (Canadian Institute for Advanced Research)、ImageNet等数据集上首次展示了与GAN可比甚至更优的无条件图像生成质量，但采样步骤较多、推理开销较大。随后，美国斯坦福大学Song等人(2021)提出DDIM(denoising diffusion implicit model)，在保持与DDPM相同训练目标的前提下，引入非马尔可夫反向过程与确定性采样路径，将采样步数在感知质量基本不降的情况下大幅压缩，为后续高效扩散采样方法提供了基础。OpenAI的Nichol和Dhariwal在Improved DDPM与“Diffusion Models Beat GANs”中进一步优化噪声调度、方差建模与网络结构，并结合分类器引导(classifier guidance)，在ImageNet上实现FID(Fréchet inception distance)指标全面超越BigGAN等主流GAN模型，标志着DDPM系列在图像合成领域首次系统性“反超”对抗生成范式(Nichol和Dhariwal, 2021; Dhariwal和Nichol, 2021)。

在条件扩散与文本驱动图像生成方面，国际研究重点集中在条件编码机制与引导策略的设计。OpenAI的GLIDE模型系统比较了基于CLIP(contrastive language-image pre-training)的引导与无分类器引导(classifier-free guidance)，并在大规模文本—图像数据上训练35亿参数的文本条件扩散模型，在人类主观评价中超过DALL-E，同时展示了通过微调实现图像修复与局部重绘等下游能力(Nichol等, 2022; Ho和Salimans, 2022)。Google Research提出的Imagen将大规模预训练语言模型(如T5)作为文本编码器，配合级联扩散解码器，在COCO(common objects in context)等基准上实现了同时兼顾图文一致性与视觉保真度的SOTA(state-of-the-art)表现，并提出DrawBench评测基准系统性比较多种文本到图像模型，为后续研究提供了标准参考(Saharia等, 2022)。德国慕尼黑工业大学与Runway团队提出的LDM(latent diffusion model)则将DDPM从像素空间迁移到自编码器的潜空间，通过卷积自编码器压缩图像，再在低维潜空间执行扩散与去噪，并在U-Net主干中引入跨模态交叉注意力以支持文本、框等通用条件，显著降低训练与采样成本的同时保持高分辨率细节，为后续Stable Diffusion等开源模型奠定了技术基础(Rombach等, 2022)。NVIDIA的eDiff-I进一步提出“多专家去噪器”思想，将不同时间段的去噪过程交由不同的专用子模型完成，在计算量不变的前提下提升文本对齐度和视觉质量(Balaji等, 2022)。

在模型结构与可扩展性方面，扩散模型从早期的U-Net主干逐步演进到基于Transformer的通用架构。Peebles和Xie提出的DiT(diffusion Transformers)用Vision Transformer式的patch-Transformer替代LDM中常用的U-Net，在ImageNet上系统分析了算力GFLOPs(giga floating-point operations per second)与FID之间的缩放规律，表明在潜空间上堆叠更深/更宽的Transformer可以稳定提升扩散模型性能，并为后续大规模DiT系列模型提供了结构参考(Peebles和Xie, 2023)。同时，围绕DDPM框架也出现了一系列基于扩散先验的图像引导与重建方法，如利用预训练扩散模型作为先验进行结构保持的图像生成与编辑(SDEdit等)，在草图上色、图像修复和条件重绘等任务中展现出良好的通用性(Meng等, 2022)。总体而言，基于DDPM的离散时间扩散模型已在国际上形成从理论建模、条件控制以及潜空间压缩到Transformer主干与多模态扩展的完整技术谱系，为后续在可视文本生成、文本擦除和图像编辑等下游视觉任务中的应用提供了坚实的基础。

1.1.3 流匹配模型

流匹配(flow matching)是一种新型生成模型，其核心思想是通过求解常微分方程(ordinary differential equation, ODE)直接建模样本从简单分布(如高斯噪声)到复杂数据分布(如自然图像)的连续流动过程。与扩散模型通过多步去噪生成不同，流匹配学习依赖于一个时间的速度场$v_t(x_t, t)$，该速度场描述了样本在时间t时的瞬时移动方向。给定初始分布x_0(通常为标准高斯分布)和目标分布x_1，流匹配采用线性插值构建中间轨迹：$x_t = (1-t) + tx_1$。速度场的真实值为$v_t(x_t, t) = x_t' = x_1 - x_0$，训练目标是

使神经网络预测的速度场$f(x_t, t)$尽可能接近真实速度。在采样阶段，从x_0开始，通过求解ODE逐步将样本从噪声空间移动到数据空间$x_1 = x_0 + \int_0^1 f(x_u, u)\mathrm{d}u$。与扩散模型相比，流匹配无需进行数百步的迭代去噪，通常只需几十步即可生成高质量图像，显著提升了生成效率。同时，流匹配通过直接建模连续路径，避免了扩散模型中噪声渐变的离散假设，为生成过程提供了更精细的控制。

国际上，流匹配研究已取得突破性进展。美国Meta研究院的Lipman等人(2023)提出首个系统性流匹配框架，引入条件概率路径和矢量场回归的训练目标，证明了该范式能同时包含常见扩散路径并在训练、效率与样本质量上更优，从而明确了流匹配的核心任务框架与研究方向。美国得克萨斯大学奥斯汀分校的Liu等人(2023b)提出了修正流(recitified flow)，让模型学习的传输路径尽可能沿着连接两个分布的直线路径，降低传输成本，并避免了路径交叉的现象。由于概念直观、架构简单，流匹配的思想已经广泛应用于实践当中。Stability AI基于流匹配的思想提出了基于全新MMDiT(multimodal denoising Transformer)架构的SD3模型(Esser等，2024)，在此基础上，德国黑森林实验室提出了参数量更大、MMDiT和DiT结合的一系列Flux模型，美国纽约大学的谢赛宁团队提出了SiT(Ma等，2024)，更容易扩展到更高的模型量级，Meta团队(Polyak等，2024)提出了30 B量级的视频生成基础模型Movie-Gen，促进了流匹配模型的应用。

1.2 可视文本擦除

国际VTR研究主要有美国加州大学伯克利分校、澳大利亚昆士兰科技大学、日本东北大学、日本九州大学等研究机构，围绕“精准捕捉文本特征、平衡擦除与背景修复”构建技术体系。早期国际VTR以传统方法为主(如纹理分析分割、轮廓提取填充)，受限于手工特征，难以应对复杂场景。日本九州大学Nakamura等人(2017)提出首个神经网络VTR方法，通过滑动窗口与编码器—解码器实现英文文本移除，虽有局限，但明确核心任务框架。此后，美国加州大学伯克利分校Isola等人(2017)提出Pix2pix模型，将条件生成对抗网络(cGAN)与U-Net结合，开创图像翻译范式，其PatchGAN判别器与L1损失设计，成为GAN类擦除模型的基础。

知识迁移范式借助预训练模型降低学习成本，分为串行与并行迁移：串行方向，日本东京大学Zdenek和Nakayama(2020)提出弱监督方案，将预训练检测器与GAN修复网络分离训练，无需配对数据，降低标注依赖；荷兰阿姆斯特丹大学Conrad和Chen(2021)利用预训练检测器获取粗掩码，经空洞空间金字塔池化(atrous spatial pyramid pooling，ASPP)模块细化，提升复杂纹理背景擦除精度。并行方向，美国加州大学圣克鲁兹分校Qin等人(2018)提出双解码器cGAN架构，一个输出笔画掩码，一个修复背景，擅长不规则手写文本处理。

多任务学习范式通过统一框架联合训练“文本特征学习”与“背景修复”：GAN架构方向，印度理工学院Keserwani和Roy(2022)通过字符级对称线与双判别器，无需笔画标注即可精准隐藏文本；专用模块方向，韩国研究团队Lee和Choi(2022)设计门控注意力(gated attention，GA)与感兴趣区域生成(region of interest generation，RoIG)机制，GA模块动态调整文本笔画与周边区域特征权重，RoIG模块聚焦文本区域提升训练效率，在SCUT-EnsText数据集上取得峰值信噪比(peak signal to noise ratio，PSNR)为41.37 dB、SSIM为0.984 6的优异性能。

渐进式学习范式模拟“粗修—精修”过程，迭代优化掩码与修复结果：掩码细化方向，澳大利亚昆士兰科技大学Tursun等人(2019)提出MTRNet(mask-based text removal network)模型，以cGAN为基础引入辅助文本掩码，无需重新训练即可适配多语言、弯曲文本场景，实现部分文本选择性移除，在ICDAR 2013等数据集上表现突出；其后续升级的MTRNet++设计“掩码细化→粗修复→精修复”三分支架构，引入Tversky损失提升掩码召回率，参数仅18.7 M，兼顾轻量化与效率(Tursun等，2020)；日本东北大学Tang等人(2021)提出笔画级场景文本擦除方法，基于改进的合成文本引擎生成训练数据，通过笔画掩码预测模块与背景修复模块提取文本笔画作为小尺寸孔洞，保留更多背景内容，仅依赖合成数据训练即在真实数据集上表现优异。

特定任务创新方面，日本九州大学Mitani等人(2023)提出选择性场景文本移除新任务，设计包含背景提取、文本提取、选择性移除以及重建的四模块U-Net架构，通过条件U-Net与FiLM层实现目标词

指定移除，无需显式文字识别模块，为隐私敏感词定向移除提供解决方案。

鉴于扩散模型的VTR是近年热点，印度理工学院提出的DiffSTR模型，以预训练PBE（paint-by-diffusion）扩散模型为基础，引入ControlNet实现结构与纹理的双重控制，结合DINO-V2编码器提取的细粒度特征增强像素级可控性，在复杂纹理与透明反射场景中表现出色（Pathak等，2024）。

1.3 可视文本编辑

STEFANN（scene text editor using font adaptive neural network）（Roy等，2020）首先通过文本检测算法选中图像中需要修改的文本区域，随后利用字体自适应网络和颜色网络生成与原始字体风格匹配的目标字符并保留颜色一致性。STEFANN作为第一个提出VTE的深度学习工作，引发了后续对VTE的研究热潮。国外的可视文本编辑的研究主要涉及了Meta等互联网公司以及日本的九州大学团队等高校。

在基于文本风格特征的编辑方法中，日本九州大学团队（Shimoda等，2021）提出De-Rendering的方法，认为像素域直接风格迁移易产生伪影和分辨率依赖，难以高质量重渲染与放缩，而将像素中的文本编辑首先反渲染为结构化的参数，包含字体、风格、字号、颜色、填充、描边、阴影和位置等属性，从而更有针对性地对文本元素进行内容和风格的修改，最后用可微渲染器重绘目标文本。Fast（Das等，2023）针对先前的文本风格编码器多样字体属性、不同词长泛化差，渲染一致性不足的问题，通过整合文本检测、字符移除、新文本生成与渲染等任务，结合目标掩码生成、风格提取与级联自注意力实现任意位置/样式的文本生成与替换。

除了将编辑任务分为单个小任务之外，端到端的编辑方法减少了错误累积。随着生成模型的提出，端到端的编辑方法也逐渐被建模为条件生成模型。RewriteNet（Lee等，2021）分别引入风格编码器和文本内容编码器各自提取风格与文本特征，并引入场景文本识别器和判别器等优化生成器和风格编码器的性能。TextStyleBrush（Krishnan等，2023）基于StyleGAN2改造，通过引入字体分类器计算风格损失、文字识别模型计算内容损失、对抗损失以及自监督的重建损失，通过特征编码与条件约束解决StyleGAN2在文本生成上的缺陷。DBEST（diffusion-based scene text manipulation network）（Santoso等，2024）致力于通过提升文本提示的方法，利用Text-Inversion的方式，借助扩散模型的生成能力提升编辑效果，并改善了原有扩散模型在文生渲染领域的缺陷。SceneTextStylizer（Yuan和Yanai，2025）引入了文本风格特征注入模块和渐进式的以距离为基础的控制掩膜，从而实现了精确而无需训练的VTE模型。

基于文本笔画特征的编辑方法，在提升VTE任务的字符准确性方面更有成效。Google团队（Liu等，2023a）提出对于VTE任务，文本编码器的粒度限制了文字生成的准确性，字符级粒度的编码器更加实用，并实现了从盲字符编码到字符感知型的编码。

在基于文本语义特征的编辑方法中，先前提到的RewriteNet（Lee等，2021）、TextStyleBrush（Krishnan等，2023）和Fast（Das等，2023）等模型通过引入光学字符识别（optical character recognition，OCR）的识别损失或者OCR的主干特征，引导VTE的文字生成质量。随着生成模型能力的提升，OCR与生成模型之间的关系正由“结果校验”向“过程约束”转变，其在多模态大模型体系中的定位与挑战已得到系统性讨论（李鸿亮 等，2025）。

1.4 可视文本生成

国际VTG研究主要围绕从传统渲染合成向扩散模型驱动的可控生成演进，代表性研究机构包括英国牛津大学、新加坡南洋理工大学、以色列Amazon Rekognition团队与Cornell Tech、韩国NAVER Clova AI，以及微软研究院、Google Research等。整体来看，研究重点集中在构建大规模文本图像数据集、设计字符感知的条件编码器，以及将布局规划与图像生成解耦的两阶段框架3个方面（Gupta等，2016；Zhan等，2018；Zhan等，2019；Liu等，2023a；Chen等，2023a）。

在渲染合成范式方面，英国牛津大学Gupta等人（2016）提出SynthText场景文本合成管线，通过几何与光照建模，将二维文字通过透视变换与分割结果对齐后“贴附”到三维场景表面，并叠加模糊与噪声以提升真实感，为后续场景文本检测与识别预训练提供了大规模基础数据集。新加坡南洋理工大学Zhan等人（2018，2019）提出Verisimilar Image Synthesis与SF-GAN（spatial fusion GAN），前者通过语义一致性与显著性约束选择文本嵌入位置，后者在统一框架中显式建模几何合成与外观合成两个子网

络，使前景文本在透视、尺度以及颜色、纹理上同时与背景匹配，合成图像可直接用于训练深度文本检测与识别模型。随后，SynthText3D 与 UnrealText 将合成从 2D“贴图”扩展到基于 3D 引擎的场景级渲染，通过同时渲染场景与文本实例，精确建模复杂透视、遮挡和光照变化，生成更接近真实世界的场景文本数据（Liao 等，2020；Long 和 Yao，2020）。

在神经渲染与多任务学习范式方面，以色列 Amazon Rekognition 团队与美国 Cornell Tech 的 Fogel 等人（2020）提出 ScrabbleGAN，将判别器与手写文本识别网络联合训练，在保证对抗生成逼真度的同时，以识别损失约束生成结果的可读性与内容正确性，实现可变长度单词级手写文本图像生成，并显著提升下游手写识别性能。韩国 NAVER Clova AI 团队 Yim 等人（2021）提出 SynthTIGER，通过模块化的文本图像合成管线整合多种背景、字体与形变建模策略，并针对文本长度和字符分布的长尾问题设计重采样机制，在 ICDAR 基准上证明其生成数据优于传统合成数据组合，为通用场景文本识别提供高质量训练语料。此外，Google Research 的 Character-Aware Models 工作从文本编码层面出发，引入字符感知编码器并构建 DrawText 基准，系统分析了字符级信息对视觉文本渲染质量的影响，为后续基于 ByT5/Glyph-ByT5 的字符感知 VTG 提供了重要启示（Liu 等，2023a）。

在扩散模型驱动的可控生成范式方面，微软研究院与香港科技大学 Chen 等人（2023b）提出 Text-Diffuser 框架，将 VTG 分解为“文本布局先行、再进行图像生成”的两阶段流程：首先利用 Transformer 从输入提示中抽取关键词并生成包含位置与大小的文本布局图，然后将布局图与文本提示共同作为条件，引导扩散模型在指定区域书写目标字符串，并保证文本与背景的一致性。与此同时，GlyphControl 利用字形条件控制 Stable Diffusion，在无需重训主干模型的前提下提升了字符级控制能力（Yang 等，2023b），Glyph-ByT5 通过定制化字符感知文本编码器与 SDXL 结合，实现长段落、高密度文本的高精度渲染（Liu 等，2024）。这些工作共同推动国际 VTG 研究从传统渲染迈向“字符感知 + 布局可控 + 多语言扩展”的扩散模型新阶段。

2 国内研究现状

2.1 生成式基础模型

2.1.1 生成对抗网络

近年来，生成对抗网络在字体生成与文本图像合成方面取得了一系列进展。

Auto-Encoder Guided GAN（Lyu 等，2023）将书法生成视为图像到图像的翻译问题，模型结合自编码器与迁移子网，可从标准字体生成多风格书法图像。在文本描述到图像生成领域，DF-GAN（deep fusion generative adversarial network）（Tao 等，2020）利用单阶段生成器与目标感知判别器，提升了语义一致性并简化了训练流程。TH-GAN（generative adversarial network based transfer learning for historical Chinese character recognition）（Cai 等，2019）在 GAN 模型中引入字形边缘和骨架加权的图像重构目标函数，提升了字形生成的效果，生成的古籍汉字图像可用于古籍汉字识别训练集的扩充。TET-GAN（texture effects transfer GAN）（Yang 等，2019）实现文字图像风格化与去风格化一体化训练，并构建新数据集以支持单样本学习。同时，基于生成式对抗网络的中文字体风格迁移方法也不断推动相关技术发展，例如滕少华和孔棱睿利用残差网络结构与对抗训练，在提高字体细节生成质量与风格迁移灵活性方面取得显著效果。此外，字体生成技术的应用也在拓展，如面向文档水印的自动字库生成方法（孙杉 等，2022）通过端到端的编码—解码结构与可导噪声建模，实现了在视觉质量与抗打印扫描鲁棒性方面均优于传统手工字库的中文水印字体生成，为字体生成技术在信息安全领域的应用提供了新的可能性。

VLMGAN（Cheng 等，2022）中引入视觉—语言匹配监督，通过双向匹配机制强化图像与文本的对应关系，并提出视觉语言匹配分数（vision-language matching score，VLMS）作为新评价指标。

2.1.2 扩散模型

国内在扩散模型方向的研究与应用同样走在前列，整体呈现出“面向中文与多语言场景的大模型预训练 + 垂直领域可控生成框架”的格局。一方面，百度、智源研究院、IDEA 研究院以及字节跳动等机构面向中文和多语言文本到图像生成提出了一系列具

有代表性的扩散模型;另一方面,高校与科研机构在遥感等专业视觉领域构建了面向特定场景的可控扩散框架,推动扩散模型从通用生成走向行业级应用。

在中文与多语言文本到图像生成方面,百度提出的 ERNIE-ViLG 2.0 将知识增强机制与“混合去噪专家(mixture-of-denoising-experts)”相结合,在不同时间步使用不同去噪子网络,并显式注入精细的语言与视觉知识,在 MS-COCO 上取得了新的零样本 FID-30k 最优结果,同时显著提升了中文语义对齐与细粒度可控性,是国内最早系统探索大规模中文扩散文生图的工作之一(Feng 等,2023)。北京智源研究院推出的 AltDiffusion 及 AltDiffusion-m18 通过训练多语言文本编码器并插入预训练 Stable Diffusion,将支持语言扩展到 18 种,并在多语言与文化特定概念生成上优于原始 Stable Diffusion,代表了国内在多语言扩散建模与知识蒸馏训练范式上的实践(Ye 等,2024)。IDEA 研究院的 Taiyi-Diffusion-XL 在 CLIP 和 Stable Diffusion XL 基础上,通过连续双语预训练扩展词表与位置编码,并借助视觉—语言大模型生成高质量图文对,使模型在 COCO / COCO-CN 上的中英双语生成能力均优于以往开源模型,成为中文场景可控生成的重要开源基线(Wu 等,2024)。字节跳动的 Seedream 2.0 进一步面向中英双语图像生成基础模型,在数据系统、双语文本编码器和多阶段后训练(含 RLHF (reinforcement learning from human feedback))方面系统优化,并引入 Glyph-Aligned ByT5 以增强字符级文本渲染能力,在提示跟随、美学质量、文本渲染和结构正确性等多项指标上取得领先表现,体现了国内工业界在大规模扩散基础模型与中文文化要素建模方面的最新进展(Gong 等,2025a)。

在垂直领域与可控生成方面,国内高校与研究机构探索了将扩散模型用于遥感等专业场景的可控图像生成。西北工业大学等单位提出的 CRS-Diff 框架,针对遥感图像的地理与时间信息,设计了同时支持文本、元数据与图像多条件输入的可控扩散模型,并通过多尺度特征融合机制整合控制信号,在单条件与多条件生成任务中均明显优于传统方法,并证明生成数据可有效提升道路提取等下游任务性能(Tang 等,2024)。这类工作体现了国内在“扩散模型 + 遥感/工程场景”方向从图像质量到下游任务实用性的系统探索。

总体而言,国内扩散模型研究一端以 ERNIE-ViLG2.0、AltDiffusion、Taiyi-Diffusion-XL 与 Seedream 2.0 等大规模中文/多语言文生图模型为代表,在语义对齐、多语言支持与中文文化要素建模等方面形成了具有国际竞争力的技术体系;另一端则以 CRS-Diff 等面向遥感的可控扩散框架为代表,将扩散模型从通用内容生成拓展到高价值行业场景,兼顾生成质量、可控性与下游任务增益,为扩散模型在国内的产业化落地奠定了基础。

2.1.3 流匹配模型

在国内,基于流匹配模型的研究机构主要包括智源实验室、阿里通义团队、腾讯混元团队等。腾讯混元团队提出了 HunYuanVideo(Kong 等,2024),作为国内首个开源的 13 B 量级的视频生成模型。阿里的通义实验室提出了 Wan2.1、Wan2.2 (Wan 等,2025)等一系列参数量级,最大参数量达到 14 B 的视频生成模型,以及 Qwen-Image、Qwen-Image-Edit (Wu 等,2025)等 20 B 量级的图像生成与编辑模型。这些模型通过将 DiT 与流匹配模型结合,引入多阶段大数据量预训练过程,实现了在通用领域的大幅度提升。智源实验室则将自回归架构与流匹配进行结合,提出 OmniGen(Song 等,2021),同时支持图像的生成和编辑任务,支持文本输入和任意图像组合的输入。不仅如此,自回归模型的上下文能力,使得在未训练过的编辑任务上,通过少样本的示例实现精确处理。

2.2 可视文本擦除

国内 VTR 研究早期以国际技术框架为基础,针对中文笔画复杂、多语言混合等特性展开适配优化。华南理工大学金连文团队借鉴 Pix2pix 的 cGAN 架构提出 EnsNet (ensconce network) 模型 (Zhang 等,2019),创新多尺度中文特征融合模块与横向连接结构,设计 4 项精细化损失函数,成为国内首个在真实中文场景数据集上实现端到端 VTR 的成果,推理速度达 333 帧/s。团队后续提出的 EraseNet (end-to-end text removal in the wild) 模型(Liu 等,2020) 新增中文笔画轮廓校正模块与分割头,通过 Dice 损失优化掩码,同时引入 SN-Patch-GAN 与光谱归一化技术,进一步提升中文场景适配性。

在知识迁移范式下,国内研究聚焦中文场景迁移误差抑制。复旦大学金城团队的 SAEN (stroke-aware erasing network) 模型(Du 等,2023a)创新文本

笔画掩码建模模块，通过结合ASPP技术（Chen等，2018）的两阶段架构避免文本框掩码导致的背景冗余替换问题；武汉理工大学朱安娜团队的DINet（deformation inpainting network）模型（Gong等，2025b）设计双向交互增强与傅里叶注意力模块，提升中英双语混合及弯曲文本场景适配性，同团队的FETNet（feature erasing and transferring）模型（Lyu等，2023）创新特征擦除与迁移机制，以轻量化参数实现高效文本擦除，兼顾精度与效率平衡。

多任务学习范式的国内研究重点强化协同机制。华南理工大学金连文团队的CTRNet（contextual-guided text removal network）模型（Liu等，2022）融合CNN与Transformer优势，创新局部—全局内容建模模块，减少复杂场景下的修复伪影；复旦大学于海洋团队的EAFormer模型（Yu等，2024）创新边缘感知Transformer架构，通过文本边缘提取器过滤非文本区域干扰，为“分割—擦除”链路提供高可靠性支撑。

渐进式学习范式的国内研究持续优化中文文本擦除流程。中国科学技术大学团队的PERT（a progressively region-based network for scene text removal）模型（Wang等，2021）创新显式擦除引导与平衡多阶段擦除机制，通过区域修改策略仅调整文本区域，参数复杂度较现有多阶段模型降低25%；同团队的DeepEraser模型（Feng等，2024）以极简参数支持指定文本区域自适应擦除；中山大学的HiCIR（hierarchical context-aware interaction reconstruction）模型（Dai等，2025）创新分层掩码引导与通道级上下文交互机制，在SCUT-EnsText数据集实现PSNR 36.82 dB，展现复杂形状与多尺度中文文本处理能力；复旦大学金城团队的PEN（progressive scene text erasing with self-supervision）模型（Du等，2023b）参考自监督思路，通过图像变体约束与三阶段迭代，提升真实中文场景泛化能力；中国科学技术大学的DARLING（disentangled representation learning framework）框架（Zhang等，2024a）创新风格与内容特征解耦策略，构建多任务解码器与门控注入机制，在SCUT-EnsText数据集实现当前SOTA性能的PSNR 38.85 dB，兼顾多任务适配性与擦除精度。

国内在VTR新兴方向的研究也持续拓展技术边界。特征解耦与文本迁移方向，国内团队创新显式解耦文本迁移框架，使中文笔画重叠场景背景重建一致性提升15%以上；视觉上下文学习与ViT（vision Transformer）适配方向，上海交通大学的ConText模型（Zhang等，2024c）设计“输入—擦除—分割”任务链提示机制，华南理工大学的ViTEraser模型（Peng等，2024）成为首个全ViT架构VTR模型，通过SegMIM（the encoder on text box segmentation and the decoder on masked image modeling）预训练策略增强泛化能力；评估体系与扩散模型方向，中国科学技术大学团队提出“背景完整性—擦除彻底性”双核心评估体系（Wang等，2023b），为中文VTR方法评估提供精细化标准；厦门大学团队的TextDestroyer模型基于预训练扩散模型（Li和Chao，2024），设计3级分层文本定位机制，无训练无标注即可适配中文弯曲、小字体文本擦除，显著降低应用门槛。

2.3 可视文本编辑

国内的可视文本编辑的研究主要有清华大学、北京大学和中国科学院等科研机构与院校，以及阿里、腾讯等企业。国内在VTE方向的研究呈现出非常清晰的阶段性特点：从以OCR和视觉分割为基础的“替换式编辑”，发展到风格与内容解耦的“生成式编辑”，再进入扩散模型驱动的“自由风格文本生成”，最终走向多模态大模型参与的“统一理解与编辑”。

在基于文本风格编辑的范式中，SRNet（editing text in the wild）（Wu等，2019）模型作为国内的先驱性工作，提出了风格保留网络，实现词级别文本替换的同时，兼顾文本风格与背景纹理的一致性。阿里团队的SwapText（Yang等，2020）和深圳大学的TENet（text editing module）（Zhao等，2021）在SRNet基础上，进一步引入多阶段的替换，即首先利用现有的文本检测或识别模块检测文本位置和内容，随后引入文本转换模块用于将源图像中的前景文本风格迁移到目标模板图像上，最后将生成后的前景文本图像融入原目标背景中。具体而言，SwapText集成了形状变换网络以控制文本形状，TENet采用硬编码组件提取中文字符的骨架结构。此类显式迁移网络的设计最初使用图像到图像翻译模型中进行字体生成。然而，文本风格不仅限于字体本身，预先定义好的字体具有局限性，难以应用于艺术字和手写等更灵活的风格范式，其复杂性使得迁移学习更具挑战性。不仅如此，这种多阶段级联的操作，降低了每一步的难度，但也使得不理想的迁移结果或擦除效果在后续融合过程中引发误差累积。端到端的方法

使得编辑流程得以简化，MOSTEL（modifying scene text image at stroke level）（Qu等，2023）提出编辑引导机制，通过预测引导图标记需要编辑的区域，显式分离了文本区域的修改与背景区域的保留，使网络能够专注于学习文本区域的编辑规则，降低了编辑规则的学习难度，并确保生成文本的风格与原始文本保持一致。同时利用无配对场景文本数据进行半监督学习，结合增强风格参考和识别损失优化训练。TextCtrl（Zeng等，2024）设计风格解耦预训练策略以捕获属性特征，通过独立模块分别提取字形结构和细粒度风格特征，再将其作为扩散模型的生成条件，从根源上提升字形准确性和风格一致性。DiffUTE（Chen等，2023a）、TextDiffuser（Chen等，2023b）以及UDiffText（Zhao和Lian，2024）等利用Inpainting范式针对掩膜区域实现了精确的编辑，其中，DiffUTE（Chen等，2023a）还利用基于OCR的图像编码器替代CLIP文本编码器，提升文本—风格保真度，Zhang等人（2024b）利用ControlNet范式（Zhang等，2023）将带有位置信息的文本渲染图像送入主干模型中，TextMaster（Wang等，2024）则通过自适应间距与掩码控制增强文本布局能力，对字体风格进行解耦并通过IP-Adapter（Ye等，2023）条件化风格。这些工作利用通用场景下已被证明有效的图像编辑范式，借助了预训练扩散模型，实现了初步探索。

基于文本笔画特征的编辑方法，绝大多数方法选择提供带有固定字体文本渲染的二值图像作为条件，从而实现对目标文本的精确生成。早期基于GAN的编辑方法依赖转换网络，以模板文本图像作为模型输入。模板表示提供显式的笔画引导，避免生成字符的随机性，包括SRNet（Wu等，2019）和SwapText（Yang等，2020）等，基于修复或者ControlNet的扩散模型方法，例如TextDiffuser（Chen等，2023b）、Glyphdraw（Ma等，2023）和DiffUTE（Chen等，2023a）同样在模型设计中利用文本模板图像，或将其与潜在属性拼接，或编码后作为交叉注意力机制的条件。随着细粒度字符级文本编码器的引入以及基础文生图模型在文本渲染能力上的提升，文本笔画也可以通过文本提示表示，例如TextCtrl（Zeng等，2024）等。为更好适配文本引导的扩散模型，近期方法，例如AnyText（Tuo等，2024b）以及DiffSTE（Ji等，2023）采用多编码器处理文本提示，字符编码器保障渲染精度，指令编码器实现风格控制，实现渲染文本的细粒度表示。阿里提出的FLUXText则采用更丰富的输入模态提升了渲染精度：先将待编辑文本按对应位置渲染成字形图像，用预训练变分自编码器（variational autoencoder，VAE）对该图像编码提取字形特征，还采用双编码器融合策略，使用T5等标准文本编码器处理输入提示获取语义嵌入，通过Glyph-ByT5编码器（Liu等，2024）提取待编辑文本的细粒度字形与语义信息，解决了多语言场景下的VTE任务，模态对齐的进一步探索有望提升文本提示表示的有效性。随着Qwen-Image-Edit（Wu等，2025）和SeeDream（Gao等，2025a，b）等文生图模型在控制能力和基础生成能力上的提升，基于文本提示的方法逐渐占据了主导，见证了随着模型参数量与训练数据量的提升，文本提示表示能力进一步加强的希望。

在基于文本语义特征的方法中，文本识别损失得到广泛应用，为文字渲染的准确性提供辅助监督。早期基于GAN的方法MOSTEL（Qu等，2023）在自监督学习中利用识别损失；基于扩散的方法AnyText（Tuo等，2024b）以及DiffUTE（Chen等，2023a）进一步对解码图像施加语义损失。此外，Zhang等人（2024a）将文本编辑与文本识别训练集成于单一模块，以实现视觉与内容表征的更好解耦，这进一步启发了语义信息如何辅助编辑的探索。同时，具有文本理解和推理能力的文本编码器逐渐取代CLIP文本编辑器，越来越常用于基础文生图模型中，这也驱动了语义能力的发展。

2.4 可视文本生成

国内在可视文本生成方向的研究与应用实践同样活跃，既在多语言、多字体复杂场景下提出了具有国际影响力的基础模型，也在电商海报、品牌广告以及新媒体内容生产等产业场景中进行了探索，形成“多语言支持 + 布局语义理解 + 产业级应用”的鲜明特色（Tuo等，2024a，b；Zhu等，2024；Wang等，2025b；Gao等，2025b；Peng等，2025；Du等，2025；Xie等，2025）。

在多语言与多字体生成方面，国内团队提出的AnyText系列是具有代表性的工作。该系列模型以预训练文本到图像扩散模型为基础，引入编码文本字形、位置与掩码的隐空间条件模块，并结合OCR提取的笔画级信息与文本语义嵌入，通过专门设计的文本控制损失与文本一致性损失，显著提升了多

语言(尤其是中文、日文、韩文等复杂字符)场景下的书写准确率与背景融合度;其后续版本 AnyText2 进一步在字体、颜色等文本属性上实现可控生成,并构建了包含多语言、多字体和多场景的专用数据集与评测基准,推动了面向中文场景的 VTG 系统化研究(Tuo 等,2024a,b)。TextFlux 在场景级多语言文本合成方面提出 OCR-free DiT 架构,系统性提升了中文及多语言场景下的渲染精度与可控性(Liu 等,2024;Xie 等,2025)。

针对复杂字形与艺术字生成问题,国内研究者提出了以字形为中心的可控扩散框架,如 GlyphDraw 系列通过显式引入字形图像与位置信号作为条件,解决了传统文本到图像模型难以精细描绘笔画结构的问题,并在此基础上拓展到海报与广告设计场景,将文案生成、布局规划与文本渲染整合为一体化流程(Ma 等,2023)。后续的 GlyphDraw2、FonTS 与 POSTA 等工作在字形先验、版式控制与艺术风格建模上进一步提升,实现了复杂艺术字和海报版式的一体化生成,为国内海报设计与品牌视觉系统提供了可行方案(GlyphDraw2 2025;FonTS 2025;POSTA 2025)。在面向商品营销的应用方面,PosterMaker 框架结合字符判别式特征建模与双分支生成网络,在电商海报场景中实现高精度中文字符渲染与商品主体保持(Gao 等,2025b),BizGen 则面向长文案信息图设计,提出分层检索增强与布局引导的可视文本渲染机制,支撑多语言、多段落的高密度排版(Peng 等,2025)。

在场景级可视文本生成与数据集构建方面,国内高校与工业界合作提出了多种面向“野外场景”的 VTG 框架。“Visual Text Generation in the Wild/SceneVTG” 系列工作将自然场景可视文本生成拆解为“多模态大模型驱动的布局与内容推荐 + 条件扩散图像生成”两阶段:第 1 阶段利用多模态大模型对输入图像进行语义解析与版式评估,在多尺度上推荐合理的文本位置与内容;第 2 阶段则以推荐结果作为条件进行扩散生成,并构建了覆盖多类自然场景的 SceneVTG 数据集与统一评测基准(Zhu 等,2024)。后续的 SceneVTG++在此基础上引入可控多语言属性与更细粒度的文本外观控制模块,使得在同一场景中可以针对不同文本区域独立控制语言、字体与颜色,进一步提升了真实场景中 VTG 的可控性与实用性(Liu 等,2025)。同时,TextFlux 在场景级多语言文本合成方面提出 OCR-free DiT 架构与高效训练策略,显著降低了多语言扩展的标注与算力成本(Xie 等,2025)。

在通用文本编辑与统一框架方面,国内研究提出了统一支持“场景文本编辑、任意图像文本生成与高精度文本到图像生成”的字符感知扩散模型。一类代表性工作如 DreamText 通过重构扩散训练过程,在去噪过程中显式引入字符级注意力引导与潜在字符掩码估计,在多字体、多风格场景中显著缓解字符重复、缺失与扭曲问题,提升了场景文本合成的细节保真度与字形一致性(Wang 等,2025b);另一类工作如 TextCrafter 则面向复杂多区域视觉文本生成,提出“实例融合—区域隔离—文本聚焦”的三阶段渐进式框架,通过布局优化与区域级扩散过程,有效支持多块、小字号以及混合字体文本在复杂视觉载体上的高精度呈现(Du 等,2025)。这些方法通常配合字符级分割监督与识别一致性损失,在保持模型参数规模可控的同时,实现对复杂中文场景中小字号、弯曲与密集文本的高质量生成与编辑,并与 AnyText/AnyText2、SceneVTG、TextFlux 等多语言与场景级框架形成互补(Tuo 等,2024a;Zhu 等,2024;Xie 等,2025;Wang 等,2025b;Du 等,2025)。

总体来看,国内 VTG 研究在多语言复杂字形、场景语义理解与工业级应用方面形成了较为完整的技术链条:一端是面向通用场景的基础模型与大规模多语言数据集,如 AnyWord-3M、SceneVTG 与 TextFlux 系列数据集,为训练与评测提供了系统支撑(Zhu 等,2024;Tuo 等,2024a;Xie 等,2025);另一端是与电商设计、短视频封面以及新媒体运营等垂直场景深度结合的生产系统与原型框架,例如面向商品图文海报生成的 PosterMaker、面向信息图与幻灯片内容生成的 BizGen,以及面向复杂多区域视觉文本生成的 TextCrafter 等模型,均在公开数据集和基准上验证了其在文本渲染精度与布局控制方面的优势,为电商营销、新媒体内容生产等业务场景提供了可落地的技术方案(Gao 等,2025b;Peng 等,2025;Du 等,2025)。

3 国内外研究进展比较

3.1 可视文本擦除

国内外 VTR 研究呈现差异化发展特征,技术路

线与核心目标各有侧重。国内以“轻量化适配 + 中文场景深耕”为核心，围绕中文笔画复杂、多语言混合等特性优化，如华南理工大学EnsNet、EraseNet针对中文场景设计特征融合与掩码优化模块，中国科学技术大学PERT、DeepEraser通过架构创新将参数量控制在1.4～18.7 M级，兼顾效率与中文场景落地；生成式模型应用聚焦中文产业需求，厦门大学TextDestroyer基于扩散模型优化中文弯曲、小字体文本擦除，上海交通大学ConText、华南理工大学ViTEraser则在ViT架构与上下文学习上突破，同时国内数据集如SCUT-EnsText（3 562 幅，中文占比45%）以多层级标注（像素级、笔画级）覆盖古籍、车牌等中文特色场景，真实数据占比超60%，产业端已形成“技术—场景—产品”闭环，服务千万级用户。

国外则以“通用质量提升 + 架构创新”为导向，日本Nakamura团队奠定神经网络VTR基础，美国Isola团队Pix2pix开创GAN类模型框架，荷兰Conrad团队、美国Qin团队在知识迁移范式的掩码细化与双解码器架构上突破，韩国Lee团队模型虽然在Flickr-ST等通用数据集客观指标（PSNR 36.87 dB+）上占优，但参数量多为65～120 M级，推理速度仅0.8～1.2帧/s，且国外数据集（SynthText等，超70万幅）中文样本占比不足5%，中文场景研究空白，产业落地多为通用工具（如Adobe Content-Aware Fill），中文处理准确率不足90%，缺乏规模化适配方案。

3.2 可视文本编辑

国内外的可视文字编辑发展相比，国内的研究多样化更高，且在维持拉丁语言编辑能力的情况下，更注重中文场景以及跨语言场景的编辑，例如AnyText、AnyText2等针对多语言场景的统一生成和编辑模型，而非局限于拉丁语言场景，且专门提出了用于中文编辑评测的统一基准Anytext-Chinese。除此之外，相比于国外的基础模型GPT-4o等，国内的模型如Qwen-Image-Edit等使用了更多的中文语料用于预训练，在中文文本的生成和编辑上显著优于国外模型。

除此之外，国内的研究范式，在提升字符准确率方面更加直接，相比国际上使用更多的识别结果损失引导或文字分类引导，国内的模型更倾向于提供更丰富的信息输入，例如多编码器的结合、文字分割结果或渲染图像的直接注入等。

总体而言，国内外的可视文字编辑发展范式与阶段相一致，从原来的基于分而治之的深度学习编辑方法，逐步演化到基于生成对抗网络、扩散模型和流匹配模型的方法，一直在顺应当前视觉领域前沿的技术发展，也随着基础模型的能力不断改进技术路线。模型流程逐渐从多阶段发展到端到端，模型新增组件逐渐更轻量化，使得文生图基础生成和编辑模型在VTE领域焕发了新的活力，拓宽了应用场景。

3.3 可视文本生成

总体来看，国际VTG研究形成了从传统渲染合成到神经网络生成再到扩散模型可控生成的完整技术演进脉络，国内工作则在此基础上，围绕多语言复杂字形与产业场景深度定制，两者在目标侧重点和技术路径上形成互补格局。

在国际方面，以英国牛津大学SynthText系列及其后续Verisimilar Image Synthesis、SynthText3D、UnrealText为代表的渲染合成方法，率先通过语义分割、几何与光照建模以及大规模2D/3D渲染等手段，构建了服务于文本检测和识别的合成数据基础（Jaderberg等，2014；Gupta等，2016；Zhan等，2018；Liao等，2020；Long和Yao，2020）。随后，Scrabble-GAN、SynthTIGER与Character-Aware Models等神经生成方法将识别损失与合成过程联动，在保持视觉真实感的同时显式优化可读性和可识别性（Fogel等，2020；Yim等，2021；Liu等，2023a）。新近的Text-Diffuser、GlyphControl、Glyph-ByT5等扩散/DiT框架则进一步引入布局图、字形与字符感知文本编码器，实现精细的字符级控制和多语言扩展，并配套MARIO、DrawText等评测基准，推动国际VTG在通用性与可控性上持续演进（Chen等，2023b；Yang等，2023b；Liu等，2024；Xie等，2025a）。

在国内方面，研究重点更多聚焦于中文为主的多语言复杂字形、场景语义理解与产业级应用。AnyText/AnyText2通过字形、位置与属性嵌入，将预训练扩散模型扩展为多语言、多字体可控文本生成框架，并构建多语言专用数据集与评测体系（Tuo等，2024a，b）。SceneVTG/Visual Text Generation in the Wild将多模态大模型引入“布局—内容推荐”阶段，再以条件扩散在自然图像上渲染文本，形成面向真实场景的完整VTG管线及统一评测基准（Zhu等，2024；Liu等，2025）。在应用方面，DreamText面向高保真场景文本合成，GlyphDraw/GlyphDraw2、Poster-

Maker、BizGen 与 TextCrafter 等工作则面向复杂艺术字、商品海报、信息图与多区域文本生成任务，提出了集文案、布局与渲染为一体的多种系统方案，为电商设计和新媒体内容生产提供了可直接迁移的技术基础（Ma 等，2023；Wang 等，2025b；Gao 等，2025b；Peng 等，2025；Du 等，2025）。整体来看，国外侧重通用框架与英文场景基准建设，国内则在多语言复杂字形与产业场景适配方面形成了更突出的优势。

4 发展趋势与展望

4.1 可视文字擦除

结合国内外研究现状及现存技术短板，未来 VTR 研究将推动技术从通用适配向精准定制、图像擦除向视频擦除、实验室向产业化深度转型，而大模型将成为贯穿多方向的核心驱动力。一方面，针对超小、透明以及强反光等极端文本场景，需依托大模型强化“语义 + 深度 + 视觉”多模态融合能力——基于 DiT 等预训练大模型的强泛化基底，可更精准捕捉文本与背景的关联特征，搭配字体级分割技术实现单个字符的独立背景填充，有效减少传统方案的伪影问题；同时聚焦扩散模型效率瓶颈，通过大模型轻量化改造突破部署限制，如移除冗余的 Text Prompt 与 Cross-Attention 模块、结合 4 bit 量化与知识蒸馏技术，在保证 PSNR ≥ 38 dB 的前提下降低计算量，配合端侧 NPU（neural network processing unit）硬件优化，推动推理速度提升至 10 帧/s 以上，满足手机、车载设备等实时需求。

另一方面，大模型为降低标注依赖与拓展应用边界提供新路径：借助大模型半自动化标注能力生成高质量伪掩码，优化“检测—伪掩码—修复”无监督框架；通过细粒度概念擦除技术，可精准移除目标文本同时保留关联背景知识，解决传统模型的“邻域知识丢失”问题。此外，深化大模型与中文场景的适配创新，基于端云协同架构部署定制化模型，在古籍修复中融合大模型的纸张老化与墨迹扩散规律学习能力，在智能驾驶、新媒体领域实现“文本擦除—内容替换—风格统一”一体化服务，同时依托多语言大模型的自适应模块，支撑中、英、日、韩等多语言场景的高效处理，构建大模型驱动的中文 VTR 产业生态。

4.2 可视文字编辑

结合国内外的研究现状和现存技术短板，未来 VTE 仍然存在三大技术研究方向。1）跨语言的 VTE，现存的基础模型或专有模型，通过扩大训练数据而显著提升在多语言编辑上的效果，然而，这仍然只局限于中文或其他的拉丁语系，因为这类语言在网络上的海量数据更易获取，但对于稀有训练数据的语言，如何通过模型的迁移能力和零样本或少样本泛化能力，实现编辑效果，仍然是一个值得研究和深耕的课题。2）对于极端情况下的文本场景，例如极小文本、背景昏暗等极端场景，如何提升模型的鲁棒性，以及对生成文字的准确性，是一个极大的挑战。3）一个条件更统一和全面的 VTE 模型更值得研究，由于当前的研究范式多样，仍然无法存在一个统一的架构，能够同时支持多图参考、指定掩码形状或多边形边界框输入、文本描述输入等多种输入形式，限制了模型的广泛使用。

除此之外，如何更公平地评价模型的编辑效果也值得思考。当前的评测指标，例如 FID score 等，仍然需要目标图像，然而编辑的目标图像往往更难获取。更准确地说，似乎不存在绝对标准的参考图像，而 CLIP score 则对文本更不敏感，无法有较好的评测效果和区分度。能够更好地反映人类偏好与图像保真度、文字风格相似度、文字字符正确性以及文字背景融合程度的评测指标或评测集亟待提出。

最后，随着单步扩散模型等加速方法的提出，以及基础模型在分辨率和文字密集场景上的优化，进一步拓宽了可视文本编辑的使用场景和边界，对于高分辨率文档图像的编辑、批量化高效可视文本编辑等起到了极大的促进作用。

4.3 可视文字渲染

结合国内外研究现状及现存技术短板，未来 VTG 研究将主要围绕“更强字符感知与多语言统一建模、更复杂载体上的布局可控生成以及更完备的评估与数据体系”三大方向推进，推动技术从单语种向多语言、多脚本扩展，从静态图像向文档级与视频级多区域生成演进，而扩散/ DiT 结构与多模态大模型将成为贯穿多方向的核心驱动力。一方面，针对中文等复杂字形以及超小字号、高密度排版等极端文本场景，需依托字符感知文本编码器与字形先验，强化语义理解、字形建模与空间布局的联合表征能力——在 Glyph-ByT5、GlyphControl、TextFlux 等工作

验证的基础上，通过字形图像、笔画结构与多语言嵌入的协同建模，可以更精确地捕捉字符级轮廓与局部纹理，配合布局图或层级结构提示实现整段、多行文本的一致书写（Liu 等，2024；Yang 等，2023a；Xie 等，2025b）。

同时，面向多页信息图、长文档与视频封面序列，有必要在 DiT 等预训练结构之上引入版式层级与时序建模机制，将“区块级布局规划—区域级文本生成—跨页/跨帧风格一致性约束”统一到单一框架中，缓解当前方法在长文本、跨场景迁移时易出现的风格漂移与语义割裂问题（Zhu 等，2024；Peng 等，2025；Du 等，2025）。

另外，多模态大模型为降低标注依赖与拓展应用边界提供了新路径：借助视觉—语言对齐能力与“看图写字”能力，可以半自动构建大规模多语言 VTG 数据集与布局标注，优化从背景合成到版式规划再到文本渲染的端到端训练流程，在电商海报、信息图和短视频封面等高价值场景中更贴近真实分布（Tuo 等，2024a；Gao 等，2025b；Peng 等，2025）。

与之相配套，构建同时衡量“保真度、合理性与实用性”的统一评估体系将愈发重要——既需要字符级可读性与拼写准确率，也需要版式合理性、跨语言一致性以及对下游检测、识别等任务的增益指标，形成类似 MARIO、SceneVTG、DrawText 这类覆盖多任务、多语言的综合评测基准（Chen 等，2023b；Zhu 等，2024；Liu 等，2023a；Shu 等，2025）。在这一过程中，面向中文及多语言生态的细粒度评价标准与公开数据集，将成为推动 VTG 从方法研究走向规模化应用的关键支撑。

5 结 语

本文系统综述了可视文本图像生成与编辑领域的研究进展，围绕可视文本擦除、可视文本编辑与可视文本生成三大核心任务，全面梳理了从传统方法到深度学习时代的技术演进脉络，深入分析了各技术范式的核心思想、代表性方法及其优缺点，并对该领域面临的挑战与未来发展趋势进行了展望。

在可视文本擦除方面，本文总结了知识迁移、多任务学习与渐进式学习三大技术范式，揭示了从“借助预训练模型特征”到“端到端联合优化”再到“多阶段迭代细化”的发展路径，并分析了基于扩散模型的新兴方法在泛化能力上的突破；在可视文本编辑方面，本文梳理了从字符级解耦、前背景分离融合以及注意力增强到端到端条件生成的技术演进，重点阐述了文本风格特征、笔画特征与语义特征 3 类关键表征的提取与应用策略；在可视文本生成方面，本文对比了基于图形学渲染与数据驱动神经生成两大范式的优劣，系统介绍了字符感知编码、字形条件控制以及多模态对齐等核心技术，展现了从合成数据生成工具到智能文本创作系统的跨越。

本综述的主要贡献在于：1）构建了可视文本处理领域的完整技术谱系，为研究者提供了清晰的知识框架与方法论指导；2）通过对比分析不同技术范式的适用场景与局限性，为算法选择与改进提供了参考依据；3）总结了该领域在数据集构建、评测指标设计等基础设施方面的成果与不足，为后续研究指明了方向；4）从多语言支持、人类意图对齐以及实时交互效率等维度提出了未来研究的关键挑战，为学术界与产业界的进一步探索提供了思路。

展望未来，可视文本图像生成与编辑技术将在以下方向持续发展：1）多模态大模型与扩散模型的深度融合，通过更强的语义理解与生成能力实现复杂场景下的高质量文本处理；2）从单一任务向统一框架演进，构建擦除—编辑—生成一体化的通用模型；3）增强跨语言、跨字体和跨场景的泛化能力，特别是对中文、阿拉伯文等复杂书写系统的支持；4）提升模型的可解释性与可控性，实现更精准的人机协同创作；5）优化计算效率，支持移动端与边缘设备的实时应用。随着技术的不断成熟，可视文本处理将在智能媒体创作、文化遗产保护以及信息无障碍传播等领域发挥日益重要的作用，成为推动人工智能与人类社会深度融合的关键技术之一。

致谢：本文由中国图象图形学学会文档图像分析与识别专业委员会组织撰写，该专委会链接为：https://www.csig.org.cn/16/201801/49333.html。

参考文献（References）

Arjovsky M, Chintala S and Bottou L. 2017. Wasserstein generative adversarial networks//Proceedings of the International Conference on Machine Learning. [s.l.]: [s.n.]: 214-223

Balaji Y, Nah S, Huang X, Vahdat A, Song J M, Zhang Q S, et al. 2022. eDiff-I: text-to-image diffusion models with an ensemble of expert denoisers [EB/OL]. [2026-01-10].

https://arxiv.org/pdf/2211.01324.pdf

Brock A, Donahue J and Simonyan K. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis//Proceedings of the International Conference on Learning Representations.[s.l.]:[s.n.]

Cai J, Peng L, Tang Y, Liu C and Li P. 2019. TH-GAN: generative adversarial network based transfer learning for historical Chinese character recognition//Proceedings of 2019 International Conference on Document Analysis and Recognition (ICDAR). [s.l.]: IEEE: 178-183

Chen H X, Xu Z E, Gu Z X, Lan J, Zhneg X, Li Y H, et al. 2023a. DiffUTE: universal text editing diffusion model//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #2753

Chen J Y, Huang Y P, Lyu T C, Cui L, Chen Q F and Wei F R. 2023b. TextDiffuser: diffusion models as text painters//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #410

Chen L C, Papandreou G, Kokkinos I, Murphy K and Yuille A L. 2018. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4): 834-848 [DOI: 10.1109/TPAMI.2017.2699184]

Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I and Abbeel P. 2016. InfoGAN: interpretable representation learning by information maximizing generative adversarial nets//Proceedings of the 30th International Conference on Neural Information Processing Systems.[s.l.]:[s.n.] : 2180-2188

Cheng Q, Wen K and Gu X. 2022. Vision-language matching for text-to-image synthesis via generative adversarial networks. IEEE Transactions on Multimedia, 25: 7062-7075

Conrad B and Chen P I. 2021. Two-stage seamless text erasing on real-world scene images//Proceedings of 2021 IEEE International Conference on Image Processing (ICIP). Anchorage, USA: IEEE: 1309-1313 [DOI: 10.1109/ICIP42928.2021.9506394]

Dai P W, Li J Y, Wu D Y, Zheng P J and Cao X C. 2025. TextSafety: visual text vanishing via hierarchical context-aware interaction reconstruction. IEEE Transactions on Information Forensics and Security, 20: 1421-1433 [DOI: 10.1109/TIFS.2025.3528249]

Das A, Biswas S, Roy P, Ghosh S, Pal U, Blumenstein M, et al. 2023. FASTER: a font-agnostic scene text editing and rendering framework [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2308.02905.pdf

Dhariwal P and Nichol A. 2021. Diffusion models beat GANs on image synthesis//Proceedings of the 35th International Conference on Neural Information Processing Systems. [s.l.]: Curran Associates Inc.: #672

Du N K, Chen Z N, Chen Z Z, Gao S, Chen X, Jiang Z K, et al. 2025. TextCrafter: accurately rendering multiple texts in complex visual scenes [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2503.23461.pdf

Du X C, Zhou Z, Zheng Y B, Ma T L, Wu X J and Jin C. 2023a. Modeling stroke mask for end-to-end text erasing//Proceedings of 2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 6140-6148 [DOI: 10.1109/WACV56688.2023.00609]

Du X C, Zhou Z, Zheng Y B, Wu X J, Ma T L and Jin C. 2023b. Progressive scene text erasing with self-supervision. Computer Vision and Image Understanding, 233: #103712 [DOI: 10.1016/j.cviu.2023.103712]

Esser P, Kulal S, Blattmann A, Entezari R, Müller J, Saini H, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis//Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: JMLR.org: #503

Feng H, Wang W D, Liu S K, Deng J J, Zhou W G and Li H Q. 2024. DeepEraser: deep iterative context mining for generic text eraser. IEEE Transactions on Multimedia, 27: 1914-1925 [DOI: 10.1109/TMM.2024.3521809]

Feng Z D, Zhang Z Y, Yu X T, Fang Y W, Li L X, Chen X Y, et al. 2023. ERNIE-ViLG 2.0: improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023). Vancouver, Canada: IEEE: 10135-10145 [DOI: 10.1109/CVPR52729.2023.00977]

Fogel S, Averbuch-Elor H, Cohen S, Mazor S and Litman R. 2020. ScrabbleGAN: semi-supervised varying length handwritten text generation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020). Seattle, USA: IEEE: 4323-4332 [DOI: 10.1109/CVPR42600.2020.00438]

Gao Y F, Lin Z H, Liu C B, Zhou M, Ge T Z, Zheng B, et al. 2025b. PosterMaker: towards high-quality product poster generation with accurate text rendering//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025). Nashville, USA: IEEE: 8083-8093 [DOI: 10.1109/CVPR52734.2025.00757]

Gao Y, Gong L X, Guo Q S, Hou X X, Lai Z C, Li F S, et al. 2025a. Seedream 3.0 technical report [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2504.11346.pdf

Gong L X, Hou X X, Li F S, Li L, Lian X C, Liu F, et al. 2025a. Seedream 2.0: a native Chinese-English bilingual image generation foundation model [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2503.07703.pdf

Gong R, Zhu A N and Liu K. 2025b. Edge guided and Fourier attention-based dual interaction network for scene text erasing. Image and Vision Computing, 154: #105406 [DOI: 10.1016/j.imavis.2024.105406]

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. 2020. Generative adversarial networks. Communications of the ACM, 63(11): 139-144

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. 2014. Generative adversarial nets. Advances in Neural Information Processing Systems, #27

Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A C. Improved training of wasserstein gans. Advances in neural information processing systems, 30(2017): 5767-5777.

Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V and Courville A C. 2017. Improved training of wasserstein GANs//Proceedings of the 31st International Conference on Neural Information Processing Systems. [s.l.]:[s.n.] : 5769-5779

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models//Proceedings of the 34th International Conference on Neural Information Processing System. Vancouver, Canada: Curran Associates Inc.: #574

Ho J and Salimans T. 2020. Classifier-free diffusion guidance [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2207.12598.pdf

Isola P, Zhu J Y, Zhou T H and Efros A A. 2017. Image-to-image translation with conditional adversarial networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5967-5976 [DOI: 10.1109/CVPR.2017.632]

Jaderberg M, Simonyan K, Vedaldi A and Zisserman A. 2014. Synthetic data and artificial neural networks for natural scene text recognition [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/1406.2227.pdf

Ji J B, Zhang G H, Wang Z W, Hou B R, Zhang Z F, Price B, et al. 2023. Improving diffusion models for scene text editing with dual encoders [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2304.05568.pdf

Karras T, Laine S and Aila T. 2019. A style-based generator architecture for generative adversarial networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4401-4410

Karras T, Aila T, Laine S and Lehtinen J. 2017. Progressive growing of gans for improved quality, stability, and variation [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/1710.10196.pdf

Keserwani P and Roy P P. 2022. Text region conditional generative adversarial network for text concealment in the wild. IEEE Transactions on Circuits and Systems for Video Technology, 32(5): 3152-3163 [DOI: 10.1109/TCSVT.2021.3103922]

Khodadadi M and Behrad A. 2012. Text localization, extraction and inpainting in color images//Proceedings of the 20th Iranian Conference on Electrical Engineering. Tehran, Iran: IEEE: 1035-1040. [DOI: 10.1109/IranianCEE.2012.6292505]

Kong W J, Tian Q, Zhang Z J, Min R, Dai Z Z, Zhou J, et al. 2024. HunyuanVideo: a systematic framework for large video generative models [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2412.03603.pdf

Krishnan P, Kovvuri R, Pang G, Vassilev B and Hassner T. 2023. TextStyleBrush: transfer of text aesthetics from a single example. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(7): 9122-9134 [DOI: 10.1109/TPAMI.2023.3239736]

Lee H and Choi C. 2022. The surprisingly straightforward scene text removal method with gated attention and region of interest generation: a comprehensive prominent model analysis//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 457-472 [DOI: 10.1007/978-3-031-19787-1_26]

Lee J, Kim Y, Kim S, Yim M, Shin S, Lee G, et al. 2021. RewriteNet: reliable scene text editing with implicit decomposition of text contents and styles [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2107.11041.pdf

Li H L, Liu Y L, Liao W H, Huang M X, Zhang S, Jin L W. 2025. OCR in the era of large models: current status and prospects. Journal of Image and Graphics, 30(6): 2023-2050 (李鸿亮, 刘禹良, 廖文辉, 黄明鑫, 张朔, 金连文. 2025. 大模型时代的光学文字识别: 现状及展望. 中国图象图形学报, 30(6): 2023-2050) [DOI: 10.11834/jig.250098]

Li M C and Chao F. 2024. TextDestroyer: a training- and annotation-free diffusion method for destroying anomal text from images [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2411.00355.pdf

Liao M H, Song B Y, Long S B, He M H, Yao C and Bai X. 2020. SynthText3D: synthesizing scene text images from 3D virtual worlds. Science China Information Sciences, 63(2): #120105 [DOI: 10.1007/s11432-019-2737-0]

Lipman Y, Chen R T Q, Hamu H B, Nickel M and Le M. 2023. Flow matching for generative modeling//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: ICLR

Liu C Y, Jin L W, Liu Y L, Luo C J, Chen B D, Guo F J, et al. 2022. Don't forget me: accurate background recovery for text removal via modeling local-global context//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 409-426 [DOI: 10.1007/978-3-031-19815-1_24]

Liu C Y, Liu Y L, Jin L W, Zhang S T, Luo C J and Wang Y P. 2020. EraseNet: end-to-end text removal in the wild. IEEE Transactions on Image Processing, 29: 8760-8775 [DOI: 10.1109/TIP.2020.3018859]

Liu J, Zhu Y, Gao F, Yang Z, Wang P, Lin J, et al. 2025. SceneVTG++: Controllable multilingual visual text generation in the wild [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2501.02962.pdf

Liu R, Garrette D, Saharia C, Chan W, Roberts A, Narang S, et al. 2023a. Character-aware models improve visual text rendering//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Canada: ACL: 16270-16297 [DOI: 10.18653/v1/2023.acl-long.900]

Liu X C, Gong C Y and Liu Q. 2023b. Flow straight and fast: learning to generate and transfer data with rectified flow//Proceedings of the 11th International Conference on Learning Representations. Kigali,

Rwanda: OpenReview.net

Liu Z Y, Liang W C, Liang Z H, Luo C, Li J, Huang G, et al. 2024. Glyph-ByT5: a customized text encoder for accurate visual text rendering//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 361-377 [DOI: 10.1007/978-3-031-73226-3_21]

Long S B and Yao C. 2020. UnrealText: synthesizing realistic scene text images from the unreal world//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE: 5488-5497

Lyu G T, Liu K, Zhu A N, Uchida S and Iwana B K. 2023. FETNet: feature erasing and transferring network for scene text removal. Pattern Recognition, 140: #109531 [DOI: 10.1016/j. patcog. 2023. 109531]

Ma J, Deng Y L, Chen C, Du N Y, Lu H N, Yang Z Y, et al. 2025. GlyphDraw2: automatic generation of complex glyph posters with diffusion models and large language models//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, Pennsylvania: AAAI Press: 5955-5963 [DOI: 10.1609/aaai.v39i6. 32636]

Ma J, Zhao M J, Chen C, Wang R C, Niu D, Lu H N, et al. 2023. GlyphDraw: seamlessly rendering text with intricate spatial structures in text-to-image generation [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2303.17870.pdf

Ma N Y, Goldstein M, Albergo M S, Boffi N M, Vanden-Eijnden E and Xie S N. 2024. SiT: exploring flow and diffusion-based generative models with scalable interpolant transformers//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 23-40 [DOI: 10.1007/978-3-031-72980-5_2]

Ma Y Z, Zhang Y F, Jia W, Liu J Y, Gan T, Yang W H, et al. 2025. Recent advances in data generation and its applications in computer vision. Journal of Image and Graphics, 30(6): 1872-1952 (马愈卓, 张永飞, 贾伟, 刘家瑛, 甘甜, 杨文瀚, 等. 2025. 面向计算机视觉的数据生成与应用研究进展. 中国图象图形学报, 30(6): 1872-1952) [DOI: 10.11834/jig.250085]

Meng C L, He Y T, Song Y, Song J M, Wu J J, Zhu J Y, et al. 2022. SDEdit: guided image synthesis and editing with stochastic differential equations//Proceedings of the 10th International Conference on Learning Representations. [s.l.]: OpenReview.net

Mirza M and Osindero S. 2014. Conditional generative adversarial nets [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/1411.1784.pdf

Mitani H, Kimura A and Uchida S. 2023. Selective scene text removal [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2309.00410.pdf

Nakamura T, Zhu A N, Yanai K and Uchida S. 2017. Scene text eraser//Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition. Kyoto, Japan: IEEE: 832-837 [DOI: 10.1109/ICDAR.2017.141]

Nichol A and Dhariwal P. 2021. Improved denoising diffusion probabilistic models//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: PMLR: 8162-8171

Nichol A Q, Dhariwal P, Ramesh A, Shyam P, Mishkin P, McGrew B, et al. 2022. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models//Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR: 16784-16804

Pathak S, Kaushik V and Lall B. 2024. DiffSTR: controlled diffusion models for scene text removal [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2410.21721.pdf

Peebles W and Xie S N. 2023. Scalable diffusion models with transformers//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 4172-4182 [DOI: 10.1109/ICCV51070.2023.00387]

Peng D Z, Liu C Y, Liu Y L and Jin L W. 2024. ViTEraser: harnessing the power of vision transformers for scene text removal with segMIM pretraining//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press: 4468-4477 [DOI: 10.1609/aaai.v38i5.28245]

Peng Y Y, Xiao S S, Wu K M, Liao Q S, Chen B H, Lin K, et al. 2025. BizGen: advancing article-level visual text rendering for infographics generation//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 23615-23624

Polyak A, Zohar A, Brown A, Tjandra A, Sinha A, Lee A, et al. 2024. Movie gen: a cast of media foundation models [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2410.13720.pdf

Qin S Y, Wei J H and Manduchi R. 2018. Automatic semantic content removal by learning to neglect//Proceedings of 2018 British Machine Vision Conference. Newcastle, UK: IEEE: #157

Qu Y D, Tan Q F, Xie H T, Xu J J, Wang Y X and Zhang Y D. 2023. Exploring stroke-level modifications for scene text editing//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press: 2119-2127 [DOI: 10.1609/aaai.v37i2. 25305]

Radford A, Metz L and Chintala S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/1511.06434.pdf

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 10674-10685 [DOI: 10.1109/CVPR52688.2022.01042]

Roy P, Bhattacharya S, Ghosh S and Pal U. 2020. STEFANN: scene text editor using font adaptive neural network//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 13225-13234 [DOI: 10.1109/CVPR42600.2020.01324]

Saharia C, Chan W, Saxena S, Li L, Whang J, Denton E, et al. 2022. Photorealistic text-to-image diffusion models with deep language

understanding//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #2643

Santoso J, Simon C and Williem. 2024. On manipulating scene text in the wild with diffusion models//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 5190-5199 [DOI: 10.1109/WACV57701.2024.00512]

Sauer A, Schwarz K and Geiger A. 2022. Stylegan-xl: Scaling stylegan to large diverse datasets//Proceedings of ACM SIGGRAPH 2022 Conference New York, USA: ACM: 1-10

Sun S, Zhang W, Fang H and Yu N. 2022. Automatic generation of Chinese document watermarking fonts. Journal of Image and Graphics, 27(1): 262-276(孙杉, 张卫明, 方涵, 俞能海. 中文水印字库的自动生成方法. 2022. 中国图象图形学报, 27(1):262-276) [DOI: 10.11834/jig.200695]

Shimoda W, Haraguchi D, Uchida S and Yamaguchi K. 2021. Derendering stylized texts//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 1056-1065 [DOI: 10.1109/ICCV48922.2021.00111]

Shu Y, Zeng W C, Zhao F M, Chen Z Y, Li Z H, Yang X M, et al. 2025. Visual text processing: a comprehensive review and unified evaluation [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2504.21682.pdf

Song J M, Meng C L and Ermon S. 2021. Denoising diffusion implicit models//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: OpenReview.net

Sun S, Zhang W M, Fang H and Yu N H. 2022. Automatic generation of Chinese document watermarking fonts [J]. Journal of Image and Graphics, 27 (1): 262-276 (孙杉,张卫明,方涵,俞能海. 2022. 中文水印字库的自动生成方法. 中国图象图形学报,27 (1): 262-276)[DOI:10.11834/jig.200695]

Tang D T, Cao X Y, Hou X S, Jiang Z Y, Liu J M and Meng D Y. 2024. CRS-Diff: controllable remote sensing image generation with diffusion model. IEEE Transactions on Geoscience and Remote Sensing, 62: #5638714 [DOI: 10.1109/TGRS.2024.3453414]

Tang Z M, Miyazaki T, Sugaya Y and Omachi S. 2021. Stroke-based scene text erasing using synthetic data for training. IEEE Transactions on Image Processing, 30: 9306-9320 [DOI: 10.1109/TIP.2021.3125260]

Tao M, Tang H, Wu S, Sebe N, Jing X Y, Wu F, et al. 2022. Df-gan: deep fusion generative adversarial networks for text-to-image synthesis [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2020.200805865.pdf

Telea A. 2004. An image inpainting technique based on the fast marching method. Journal of Graphics Tools, 9(1): 23-34 [DOI: 10.1080/10867651.2004.10487596]

Teng S H and Kong L R. 2019. Chinese fonts style transfer based on generative adversarial networks. Computer Application Research, 10: 3164-3167 (滕少华,孔棱睿. 2019. 基于生成式对抗网络的中文字体风格迁移. 计算机应用研究,(10): 3164-3167)

Tao M, Tang H, Wu S, Sebe N, Jing X Y, Wu F, et al. 2020. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis [EB/OL].[2026-01-10]. https://arxiv.org/pdf/2008.05865.pdf

Tuo Y X, Geng Y F and Bo L F. 2024a. AnyText2: visual text generation and editing with customizable attributes [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2411.15245.pdf

Tuo Y X, Xiang W M, He J Y, Geng Y F and Xie X S. 2024b. AnyText: multilingual visual text generation and editing//Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net

Tursun O, Denman S, Zeng R, Sivapalan S, Sridharan S and Fookes C. 2020. MTRNet++: one-stage mask-based scene text eraser. Computer Vision and Image Understanding, 201: #103066 [DOI: 10.1016/j.cviu.2020.103066]

Tursun O, Zeng R, Denman S, Sivapalan S, Sridharan S and Fookes C. 2019. MTRNet: a generic scene text eraser//Proceedings of 2019 International Conference on Document Analysis and Recognition. Sydney, Australia: IEEE: 39-44 [DOI: 10.1109/icdar.2019.00016]

Wan T, Wang A, Ai B L, Wen B, Mao C J, Xie C W, et al. 2025. Wan: open and advanced large-scale video generative models [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2503.20314.pdf

Wang A Q, Wang J, Yan Z Y, Shang W X, Lin R and Zhang Z. 2024. TextMaster: universal controllable text edit [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2410.09879.pdf

Wang C S, Wu L, Chen X, Li X, Meng L and Meng X X. 2023a. Letter embedding guidance diffusion model for scene text editing//Proceedings of 2023 IEEE International Conference on Multimedia and Expo. Brisbane, Australia: IEEE: 588-593 [DOI: 10.1109/ICME55011.2023.00107]

Wang T, Liu T, Qu X C, Wu C J, Liu L Q and Hu X L. 2025a. GlyphMastero: a glyph encoder for high-fidelity scene text editing//Proceedings of 2025 Computer Vision and Pattern Recognition Conference. Nashville, USA: IEEE: 28523-28532 [DOI: 10.1109/CVPR52734.2025.02656]

Wang Y B, Zhang W Z, Xu H H and Jin C. 2025b. DreamText: high fidelity scene text synthesis//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 28555-28563

Wang Y X, Xie H T, Fang S C, Qu Y D and Zhang Y D. 2021. PERT: a progressively region-based network for scene text removal [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2106.13029.pdf

Wang Y X, Xie H T, Wang Z X, Qu Y D and Zhang Y D. 2023b. What is the real need for scene text removal? Exploring the background integrity and erasure exhaustivity properties. IEEE Transactions on Image Processing, 32: 4567-4580 [DOI: 10.1109/TIP.2023.3290517]

Wu C F, Li J H, Zhou J R, Lin J Y, Gao K Y, Yan K, et al. 2025. Qwen-image technical report [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2508.02324.pdf

Wu H J, Zhang D X, Gan R Y, Lu J Y, Wu Z W, Sun R L, et al. 2024. Taiyi-Diffusion-XL: advancing bilingual text-to-image generation with large vision-language model support [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2401.14688.pdf

Wu L, Zhang C Q, Liu J M, Han J M, Liu J M, Ding E R, et al. 2019. Editing text in the wild//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: ACM: 1500-1508 [DOI: 10.1145/3343031.3350929]

Xie Y, Zhang J L, Chen P Y, Wang W H, Gao L W, Li P Y, et al. 2025. TextFlux: an OCR-free DiT model for high-fidelity multilingual scene text synthesis [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2505.17778.pdf

Yang F X, Su T H, Zhou X, Di D L, Wang Z J and Li S Z. 2023a. Self-supervised cross-language scene text editing//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM: 4546-4554 [DOI: 10.1145/3581783.3612174]

Yang Q P, Huang J and Lin W. 2020. SwapText: image based texts transfer in scenes//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 14700-14709 [DOI: 10.1109/CVPR42600.2020.01471]

Yang S, Liu J, Wang W and Guo Z. 2019. TET-GAN: text effects transfer via stylization and destylization//Proceedings of the AAAI Conference on Artificial Intelligence. [s.l.]: [s.n.]: 1238-1245

Yang Y K, Gui D N, Yuan Y H, Liang W C, Ding H S, Hu H, et al. 2023b. GlyphControl: glyph conditional control for visual text generation//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #1912

Ye F L, Liu G, Wu X Y and Wu L. 2024. AltDiffusion: a multilingual text-to-image diffusion model//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press: 6648-6656 [DOI: 10.1609/aaai.v38i7.28487]

Ye H, Zhang J, Liu S B, Han X and Yang W. 2023. IP-adapter: text compatible image prompt adapter for text-to-image diffusion models [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2308.06721.pdf

Yim M, Kim Y, Cho H C and Park S. 2021. SynthTIGER: synthetic text image generator towards better text recognition models//Proceedings of the 16th International Conference on Document Analysis and Recognition. Lausanne, Switzerland: Springer: 109-124 [DOI: 10.1007/978-3-030-86337-1_8]

Yu H Y, Fu T, Li B and Xue X Y. 2024. EAFormer: scene text segmentation with edge-aware transformers//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 410-427 [DOI: 10.1007/978-3-031-72698-9_24]

Yuan H H and Yanai K. 2025. SceneTextStylizer: a training-free scene text style transfer framework with diffusion model [EB/OL]. [2026-01-10]. https://arxiv.org/pdf/2510.10910.pdf

Zdenek J and Nakayama H. 2020. Erasing scene text with weak supervision//Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). Snowmass, USA: IEEE: 2227-2235 [DOI: 10.1109/WACV45572.2020.9093544]

Zeng W C, Shu Y, Li Z H, Yang D B and Zhou Y. 2024. TextCtrl: diffusion-based scene text editing with prior guidance control//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #4396

Zhan F N, Lu S J and Xue C H. 2018. Verisimilar image synthesis for accurate detection and recognition of texts in scenes//Proceedings of the 15th European Conference on Computer Vision. Milan, Italy: Springer: 527-273 [DOI: 10.1007/978-3-030-01237-3_16]

Zhan F N, Zhu H Y and Lu S J. 2019. Spatial fusion GAN for image synthesis//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3648-3657 [DOI: 10.1109/CVPR.2019.00377]

Zhang B Q, Gao Z, Qu Y D and Xie H T. 2024b. How control information influences multilingual text image generation and editing?//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #221

Zhang B Q, Xie H T, Gao Z and Wang Y X. 2024a. Choose what you need: disentangled representation learning for scene text recognition, removal and editing//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 28358-28368 [DOI: 10.1109/cvpr52733.2024.02679]

Zhang F, Zhang P, Yang B S, Huang F, Wang Y F and Zhang Y. 2024c. ConText: driving in-context learning for text removal and segmentation//Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada: OpenReview.net

Zhang L M, Rao A Y and Agrawala M. 2023. Adding conditional control to text-to-image diffusion models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 3813-3824

Zhang S T, Liu Y L, Jin L W, Huang Y X and Lai S X. 2019. EnsNet: ensconce text in the wild//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI Press: 801-808 [DOI: 10.1609/aaai.v33i01.3301801]

Zhao L, Chen C S and Huang J W. 2021. Deep learning-based forgery attack on document images. IEEE Transactions on Image Processing, 30: 7964-7979 [DOI: 10.1109/TIP.2021.3112048]

Zhao L, Dong S, Liu J, Zhang X, Gao X and Wu X. 2025. Skeleton-guided diffusion for font generation. Electronics, 14(19): #3932

Zhao Y M and Lian Z H. 2024. UDiffText: a unified framework for high-quality text synthesis in arbitrary images via character-aware diffusion models//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 217-233 [DOI: 10.1007/978-

3-031-72751-1_13]

Zhu Y Z, Liu J W, Gao F Y, Liu W Y, Wang X G, Wang P, et al. 2024. Visual text generation in the wild//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 89-106 [DOI: 10.1007/978-3-031-73668-1_6]

Zhu J Y, Park T, Isola P and Efros A A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks//Proceedings of the IEEE International Conference on Computer Vision. [s.l.]: [s.n.]: 2223-2232

作者简介

舒言，男，博士研究生，主要研究方向为多模态大模型与文档智能。E-mail: shuyan9812@gmail.com

殷绪成，通信作者，男，教授，博士生导师，主要研究方向为模式识别、文字识别、计算机视觉、工业智能与工业软件技术及应用。E-mail: xuchengyin@ustb.edu.cn

赵方敏，女，硕士研究生，主要研究方向为计算机视觉、文档图像几何校正、多对一文档图像增强。

E-mail: zhaofangmin@iie.ac.cn

陈泽宇，男，硕士研究生，主要研究方向为图像生成与编辑。

E-mail: chenzeyu@mail.nankai.edu.cn

赵天齐，男，博士研究生，主要研究方向为多模态大模型、计算机视觉、图像生成。E-mail: ztq24@mails.tsinghua.edu.cn

王逸竹，女，本科生，主要研究方向为人工智能安全、对抗机器学习。E-mail: yizhu-wa21@mails.tsinghua.edu.cn

李焜炽，男，讲师，主要研究方向为机器学习与深度学习。

E-mail: likunchi@xmut.edu.cn

周宇，男，教授，博士生导师，主要研究方向为计算机视觉、多模态人工智能、具身智能、文档智能（OCR）、多模态大模型、多模态智能体和终身学习。E-mail: yzhou@nankai.edu.cn

王大寒，男，教授，主要研究方向为模式识别、人工智能及在智慧医疗和智慧交通等领域的应用。

E-mail: wangdh@xmut.edu.cn

彭良瑞，女，副研究员，博士生导师，主要研究方向为机器学习与计算机视觉、智能图文信息处理。

E-mail: penglr@tsinghua.edu.cn

高良才，男，副教授，博士生导师，主要研究方向为人工智能、模式识别和数字出版。E-mail: glc@pku.edu.cn