Computer Science and Application 计算机科学与应用, 2026, 16(1), 20-27
Published Online January 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.161003

Hans汉斯

文本生成图像技术的控制机制与研究综述

张金虹，罗文秋*，曹　鹏

北京印刷学院信息工程学院，北京

收稿日期：2025年12月2日；录用日期：2025年12月30日；发布日期：2026年1月7日

摘　要

本文综述文本生成图像技术的控制机制与研究进展。文生图技术在计算机视觉与自然语言处理交叉领域意义重大，随GAN、VAE、Transformer等技术发展而进步，且在多行业有广泛应用前景。文中详细阐述图像生成控制机制，包括纯文本控制(GAN、VAE、扩散模型)和多模态控制(草图、语音、布局与文本融合)，介绍了IS、FID、CLIP Score等图像质量评价指标。同时指出当前技术存在语义一致性缺失、多模态控制协同性与易用性失衡等挑战，最后展望未来技术发展方向。

关键词

文本生成图像，扩散模型，多模态控制

Control Mechanisms and Research Review of Text-to-Image Generation Technology

Jinhong Zhang, Wenqiu Luo*, Peng Cao

School of Information Engineering, Beijing Institute of Graphic Communication, Beijing

Received: December 2, 2025; accepted: December 30, 2025; published: January 7, 2026

Abstract

This article reviews the control mechanism and research progress of text generated image technology. This technology is of great significance in the intersection of computer vision and natural language processing, advancing with the development of GAN, VAE, Transformer and other technologies, and has broad application prospects in multiple industries. The article elaborates on the image generation control mechanism in detail, including pure text control (GAN, VAE, diffusion model) and

*通讯作者。

文章引用: 张金虹, 罗文秋, 曹鹏. 文本生成图像技术的控制机制与研究综述[J]. 计算机科学与应用, 2026, 16(1): 20-27.
DOI: 10.12677/csa.2026.161003

multimodal control (sketch, speech, layout and text fusion), and introduces image quality evaluation indicators such as IS, FID, CLIP Score, etc. At the same time, it is pointed out that there are challenges in the current technology, such as the lack of semantic consistency and the imbalance between multimodal control synergy and usability. Finally, the future direction of technological development is discussed.

Keywords

Text-to-Image Generation, Diffusion Models, Multi-Modal Control

1. 引言

文本生成图像(Text-to-Image Generation)是自然语言描述生成对应图像的过程，这一任务在计算机视觉和自然语言处理的交叉领域中具有重要的研究价值。早期的文本生成图像技术主要基于简单的规则或模板，通过手工设计特征和模型，生成较为基础的图像。然而，随着深度学习尤其是生成对抗网络(Generative Adversarial Networks, GAN)、变分自编码器(Variational Autoencoder, VAE)和 Transformer 等先进技术的崛起，文本生成图像技术得到了飞速发展。

文本生成图像的技术在计算机视觉和自然语言处理中的重要性不断增强，它不仅促进了图像生成的多样性，还有效地推进了图像与语言理解的深度融合。在计算机视觉领域，生成高质量图像的技术帮助改善了图像生成、图像修复、图像超分辨率等任务。而在自然语言处理领域，文本生成图像的研究则推动了多模态理解的发展，尤其是跨领域的语义理解能力。这种技术的广泛应用前景使得它在多个行业展现出巨大的潜力。例如，在艺术创作领域，文本生成图像可以为艺术家提供新的创作方式，甚至实现根据简单描述创作出完整的艺术作品。在广告行业，品牌营销中可以通过文本生成图像快速设计创意广告图，从而提高制作效率和创意表达的多样性。此外，游戏设计中可以根据游戏场景的文本描述生成虚拟环境，极大地丰富游戏内容和场景的设计。随着技术的不断成熟，文本生成图像技术预计将在更多领域得到实际应用。

2. 图像生成控制机制

2.1. 生成对抗网络

生成对抗网络是一类深度学习生成模型，由 Ian Goodfellow 等人在 2014 年提出[1]，在文生图生成中有广泛应用，它可以用于各种应用，包括图像合成，语义图像编辑，图像样式转换，图像超分辨率和自动驾驶。GAN 结构独特，与前向传播神经网络、卷积神经网络和循环神经网络等传统网络结构不同，GAN 包括生成网络(Generator, G)和判别网络(Discriminator, D)两个对抗模块[2]，如图 1 所示。生成器通过输入随机噪声生成图像，生成看起来尽可能真实的图像，而判别器的任务是判断图像的真实性，将生成的图像与真实图像区分开来。判别器不断对生成器提出挑战，要求它生成更逼真的图像。经过多轮迭代，生成的结果非常真实，使得鉴别器无法判断出伪造的结果，不断提升图像生成的质量[3]。典型的文生图 GAN 模型如 AttnGAN，通过这种对抗训练，GAN 能够学习到从文本描述到图像生成的映射关系，在细节控制和图像质量上具有一定优势。

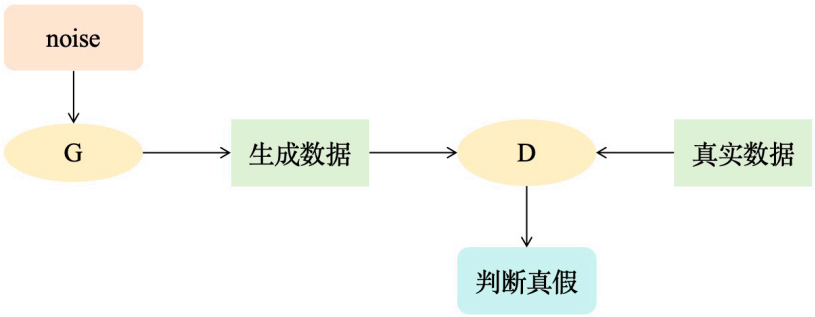

Figure 1. The structure of generative adversarial networks
图 1. 生成对抗网络的结构

总之，GAN 作为生成方法之一，凭借其生成质量和跨模态能力在文生图领域表现优异。尤其是在图像细节、分辨率和复杂场景生成等方面，GAN 结构不断演进并结合其他深度学习技术(如多模态嵌入、注意力机制)，在自动驾驶和其他关键应用中展现了广阔前景。

2.2. 变分自编码器

变分自编码器于 2013 年由 Diederik P. Kingma 和 Max Welling 提出[4]，基于 VAE 的文本生成图像模型主要包括文本编码器、图像编码器、解码器和潜在变量空间四个部分。

与传统的自编码器通过数值的方式描述潜在空间不同，变分自编码器以概率的方式描述对潜在空间的观察，首先将输入图像编码为潜在空间中的低维表示，并支持基于自然语言描述的信息解码出相应的图像，即编码–采样–解码三个过程[5]。如图 2 所示。

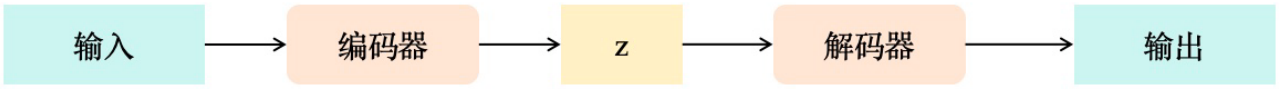

Figure 2. The structure of VAE
图 2. VAE 网络结构

编码器：将输入数据映射到一个高维的潜在空间，并在该空间中建立每个数据点的概率分布。编码器输出的是潜在空间的均值向量和方差向量，用于定义潜在变量的分布。

采样：为了生成新数据，VAE 从潜在空间的概率分布中采样一个潜在变量。这种采样是随机的，确保了生成的数据具有多样性。

解码器：将采样得到的潜在变量解码为生成的数据样本，解码器会尝试使生成的样本尽可能接近真实数据。

在训练过程中，VAE 的损失函数主要包括重构损失和 KL 散度损失两部分[6]。重构损失用于衡量生成图像与真实图像之间的差异，促使解码器能够准确地从潜在变量和文本语义向量中重构出图像；KL 散度损失用于约束潜在变量的分布尽可能接近先验分布(通常为标准正态分布)，保证潜在变量空间的连续性和可解释性，从而使模型具有更好的生成能力和泛化性能。

尽管 VAE 生成的图像质量不如 GAN，但它具有良好的潜在空间结构，可以进行平滑的语义插值[7]。VAE 被广泛应用于图像生成、异常检测、文本生成、图像数据增强等任务中，尤其在生成具有多样性且符合特定数据分布的样本时表现出色。例如，在异常检测中，VAE 可以学习正常样本的分布，当检测到显著偏离该分布的样本时，将其标记为异常。

2.3. 扩散模型

扩散模型是一类基于概率生成的模型，主要用于图像生成、音频生成、文本生成等任务，它们通过逐步降低噪声来生成高质量的数据样本。扩散模型的工作原理可以分为两个主要过程，前向扩散过程：逐步向数据样本(如图像)添加噪声，直到得到一个接近于纯高斯噪声的分布。通过重复多次这个过程，模型将数据分布映射到噪声分布。反向扩散过程：模型从高斯噪声开始逐步去噪，通过估计每一步的噪声并减去它，将噪声数据逐步还原为清晰的样本。这一过程通过学习前向扩散的逆过程实现，生成逼真的样本[8]。

扩散模型的训练通常通过最大化似然估计(MLE)来实现[9]。由于直接优化反向过程的似然通常是不可行的，扩散模型采用变分推断方法，使用下界(ELBO)来近似目标似然函数。通过对正向过程的每个时间步骤的噪声逐步去噪，训练的目标是最小化数据分布与生成分布之间的差距。这类模型中的代表作包DALL-E 2 和 Stable Diffusion，它们通过多步迭代将高质量图像还原自噪声，使得生成的图像具有丰富的细节和高分辨率。

2.4. 条件控制信号与基础模型的融合机制

2.4.1. 基于草图的多模态控制

草图作为一种直观的视觉表达方式，能够快速勾勒出物体的形状、结构和轮廓等关键信息，将草图与文本信息相结合，能够为图像生成提供更精确的结构约束，有效提高生成图像的形状准确性和结构合理性[10]。

草图与文本的融合方式：在基于草图的多模态控制方法中，草图和文本信息的融合主要通过以下几种方式实现：

特征级融合：首先对草图进行特征提取，通常采用 CNN 等模型提取草图的形状、边缘、结构等特征信息，得到草图特征向量；同时，对文本描述进行编码，得到文本语义向量。然后，将草图特征向量与文本语义向量进行拼接、元素相加或注意力加权等操作，得到融合后的多模态特征向量，将其输入到图像生成模型中，指导图像的生成。这种融合方式能够充分利用草图的结构特征和文本的语义信息，实现对图像生成过程的有效控制。

模态交互注意力机制：为了更好地实现草图和文本信息的交互与融合，一些模型引入了模态交互注意力机制。该机制使模型能够根据文本描述中的关键词，关注草图中对应的区域，同时也能够根据草图的结构特征，强化文本中相关语义信息的表达。例如，当文本描述为“一只站在树上的猫”，草图勾勒出树和猫的大致轮廓时，注意力机制能够引导模型在生成猫的区域时，结合文本中“猫”的语义信息，生成具有猫的特征的图像，同时在生成树的区域时，结合文本中“树”的语义信息，确保树的结构和特征与文本描述一致。

阶段式融合：在分阶段生成图像的模型中(如一些基于 GAN 或扩散模型的多尺度生成模型)，草图和文本信息的融合可以采用阶段式的方式。在生成图像的初始阶段，主要利用草图的结构信息，生成符合草图轮廓的低分辨率图像雏形，确定图像的整体结构；在后续的细节生成阶段，逐渐融入文本的语义信息，对图像的细节进行优化和丰富，如添加物体的颜色、纹理、纹理等特征，使生成的图像既符合草图的结构要求，又与文本描述的语义一致。

2.4.2. 基于语音的多模态控制

语音作为人类最自然、最便捷的交互方式之一，将语音与文本信息相结合，能够实现更自然、更高效的图像生成控制，尤其适用于不方便进行文字输入的场景(如移动设备端、hands-free 场景等)。

语音与文本的融合方式：基于语音的多模态控制方法首先需要将语音信号转化为文本信息(即语音识别)，然后再将识别得到的文本与用户可能提供的额外文本描述(或语音中提取的关键语义信息)进行融合，共同指导图像的生成。具体的融合方式主要包括：

语音转文本后融合：首先利用语音识别技术(如基于 Transformer 的语音识别模型)将用户的语音指令转化为文本信息，然后对该文本信息进行语义分析和编码，得到语音文本语义向量。同时，对用户提供的额外文本描述(如果有)进行编码，得到额外文本语义向量。将语音文本语义向量与额外文本语义向量进行融合，得到最终的文本语义向量，输入到图像生成模型中，控制图像的生成。这种融合方式简单直接，充分利用了现有的文本生成图像技术，只需在前端增加语音识别模块即可实现。

语音特征与文本语义融合：除了将语音转化为文本外，一些研究工作还尝试直接利用语音的声学特征(如梅尔频率倒谱系数 MFCC、语音的节奏、语调等)与文本语义信息进行融合。首先提取语音的声学特征，得到语音特征向量；然后对文本描述进行编码，得到文本语义向量；通过注意力机制或多模态融合网络将语音特征向量与文本语义向量进行融合，得到多模态融合特征，用于指导图像生成。这种融合方式能够利用语音中蕴含的情感、语气等额外信息，例如，欢快的语音语调可能对应生成色彩鲜艳、风格活泼的图像，从而使生成的图像更符合用户的情感需求和意图。

2.4.3. 基于布局的多模态控制

布局信息能够明确指定图像中各个物体的位置、大小、姿态等空间分布信息，将布局与文本信息相结合，能够为图像生成提供精确的空间约束，有效解决生成图像中物体位置混乱、遮挡不合理等问题，提高生成图像的空间合理性和语义一致性。

布局与文本的融合方式：基于布局的多模态控制方法中，布局信息通常以边界框、掩码或关键点等形式表示，与文本信息的融合主要通过以下方式实现：

布局引导的生成过程：在图像生成模型中，首先将布局信息作为先验知识输入到模型中，模型根据布局信息确定图像中各个物体的位置和大致范围。然后，将文本描述编码为文本语义向量，模型在布局信息的约束下，根据文本语义向量在各个物体的位置范围内生成对应的物体图像，确保物体的位置、大小与布局信息一致，同时物体的特征与文本描述相符。例如，在生成“客厅场景” 的图像时，布局信息指定了沙发、茶几、电视等物体的位置和大小，模型在生成过程中，会在布局指定的区域内，根据文本描述生成具有相应特征的沙发、茶几、电视等物体。

布局与文本的语义匹配：为了确保布局信息与文本描述的语义一致性，一些模型在训练过程中引入了布局与文本的语义匹配损失函数。通过计算布局信息(如物体的类别、位置等)与文本语义向量之间的相似度，对模型进行约束，促使模型生成的布局信息与文本描述相符。在生成过程中，模型会先根据文本描述生成合理的布局信息(如果用户未提供布局)，或者对用户提供的布局信息进行优化调整，使其与文本描述的语义更匹配，然后再根据优化后的布局信息和文本语义向量生成图像。

多尺度布局融合：对于复杂场景的图像生成，一些模型采用多尺度布局融合的方式。在低尺度层面，布局信息主要确定图像中各个物体的整体位置和大致结构；在高尺度层面，布局信息进一步细化物体的局部位置、细节结构和相互关系。同时，结合文本语义信息在不同尺度层面对图像生成进行控制，使生成的图像在整体结构和局部细节上都能够满足布局和文本的要求，提高生成图像的质量和空间合理性。

2.5. 模型对比分析

为更清晰地揭示 GAN、VAE 与扩散模型在文本生成图像任务中的优劣，本文从生成质量、多样性、训练成本、推理速度、可控性五个维度进行横向对比，如表 1 所示。

Figure 1. Cross-comparative analysis of major generation frameworks
表 1. 主流生成框架的横向对比

	GAN	VAE	扩散模型
生成质量	高	中	极高
多样性	中	高	高
训练成本	中	低	极高
推理速度	快	快	慢

GAN 虽推理快，但其判别器易过拟合文本–图像映射，导致“文本–实体属性错位”问题(如“红色水杯”生成蓝色)。VAE 的潜在空间虽连续，但文本编码器仅作为“条件先验”，缺乏跨模态注意力，导致“文本–图像语义漂移”。扩散模型通过交叉注意力层显式建模文本 token 与图像 patch 的对应关系，显著缓解上述问题，但其推理链长带来实时性瓶颈，在移动端部署时需蒸馏或步数压缩(如 DDIM、DPM-Solver)。

3. 图像质量评价指标

3.1. Inception Score (IS)

IS 指标基于预训练的 Inception-v3 图像分类模型，主要从生成图像的多样性和类别可区分性两个方面对图像质量进行评价。其计算过程如下：首先，将生成的图像输入到 Inception-v3 模型中，得到图像在各个类别上的概率分布(即类别预测概率)；然后，计算所有生成图像类别预测概率的边缘分布的熵(用于衡量图像的多样性，熵值越大，说明生成图像的类别越丰富，多样性越高)，以及每个生成图像类别预测概率的条件熵的平均值(用于衡量图像的类别可区分性，条件熵越小，说明模型对图像类别的预测越确定，图像的类别特征越清晰) [11]；最后，IS 值定义为边缘分布熵与条件熵平均值的差值的指数，即

$$IS = \exp\left(E_x\left[D_{KL}\left(p\left(y|x\right)\|p\left(y\right)\right)\right]\right) \tag{1}$$

其中 x 表示生成图像，y 表示图像类别，$p\left(y|x\right)$ 是给定图像 x 时类别 y 的条件概率，$p\left(y\right)$ 是类别 y 的边缘概率，D_{KL} 表示 KL 散度，衡量 $p\left(y|x\right)$ 和 $p\left(y\right)$ 的差异。

3.2. Fréchet Inception Distance (FID)

FID 指标同样基于预训练的 Inception-v3 模型，通过比较生成图像集和真实图像集在 Inception-v3 模型特征空间中的分布差异来评价生成图像的质量[12]。其计算步骤如下：首先，分别将生成图像集和真实图像集输入到 Inception-v3 模型中，提取模型某一中间层(通常为全连接层前的特征层)的特征向量；然后，计算生成图像集特征向量的均值和协方差矩阵，以及真实图像集特征向量的均值和协方差矩阵；最后，FID 值定义为生成图像集和真实图像集特征分布之间的 Fréchet 距离，计算公式为

$$FID=\left\|\mu_r-\mu_g\right\|^2+T_r\left(\sum r+\sum g-2\left(\sum r\sum g\right)^{1/2}\right) \tag{2}$$

其中 μ_g 和 μ_r 分别表示生成图像集和真实图像集特征向量的均值，$\sum g$ 和 $\sum r$ 分别表示对应的协方差矩阵，$\left\|\mu_r-\mu_g\right\|^2$ 表示欧几里得范数的平方，T_r 表示矩阵的迹。FID 值越小，说明生成图像集的特征分布与真实图像集的特征分布越接近，生成图像的质量越高。

3.3. CLIP Score

CLIP (Contrastive Language-Image Pre-training)是由 OpenAI 提出的一种多模态预训练模型，能够同时

理解图像和文本信息。CLIP Score 指标利用 CLIP 模型来评价生成图像与文本描述之间的语义一致性，同时也能在一定程度上反映生成图像的质量。其计算方法如下：首先，将生成图像输入到 CLIP 的图像编码器中，得到图像特征向量；将对应的文本描述输入到 CLIP 的文本编码器中，得到文本特征向量；然后，计算图像特征向量与文本特征向量之间的余弦相似度；最后，将余弦相似度的值作为 CLIP Score，Score 值越大，说明生成图像与文本描述的语义一致性越高，同时也在一定程度上表明生成图像的质量较好(因为语义一致的图像通常更符合用户需求，被认为质量更高)。

4. 挑战与展望

4.1. 挑战

理想状态下，生成图像应与文本描述实现完全语义匹配，但现有模型在处理多实体关联描述时，常出现语义偏差问题。例如，针对“红色水杯置于蓝色桌面，杯内含银色勺子”的文本输入，模型可能产生物体属性混淆(水杯呈蓝色)、空间关系错位(勺子位于桌面)或关键实体缺失(无勺子生成)等问题。导致该问题的核心原因在于：一是文本编码器对复杂句式中实体关系的解析能力有限，难以构建完整的语义关联图谱；二是多模态信息转换过程中存在语义损耗，即便模型具备细节生成能力，也无法精准区分“哑光材质”与“亮面材质”的视觉差异，或还原“夕阳窗景光斑”等细粒度场景特征。

当前文本生成图像技术虽支持草图、语音、布局等多模态控制方式，但存在协同效率低与用户门槛高的双重问题。在协同性层面，草图控制易受绘制精度影响(手绘线条易引发轮廓误判，矢量草图需专业技能)；语音控制面临语音转文本语义丢失(方言、情感语气误识别)及模态融合权重失衡问题；布局控制则依赖手动标注边界框，操作复杂度高。从用户体验视角看，现有多模态控制方案的操作成本超出普通用户承载范围，如嘈杂环境下语音识别准确率骤降、缺乏多轮交互修正功能(如“调整生成物体颜色”)，导致技术难以在创意设计、教育课件制作等非专业场景普及。

4.2. 展望

针对当前技术挑战与行业应用需求，未来文本生成图像技术将以提升核心性能与拓展应用价值为目标，在语义控制精度、交互体验优化、技术适配范围等维度展开探索：一方面，将通过强化文本语义解析能力与多模态融合效率，推动生成结果与文本描述的精准匹配，同时简化多模态控制操作流程，降低非专业用户使用门槛；另一方面，将着力优化模型架构与硬件适配方案，在控制生成质量的前提下实现模型轻量化，满足移动端等资源受限场景需求，并结合不同行业特性开发定制化解决方案。未来我们可以研究的题目像实时化扩散模型：探索步数-质量帕累托前沿，结合 DPM-Solver++与量化技术，在移动端实现<10 步的交互式草图编辑；多条件冲突消解：当文本“红色猫”与草图“蓝色轮廓”矛盾时，建立条件可信度评估机制，通过条件 dropout 或加权 cross-attention 动态调和。此外，技术发展过程中还需同步构建完善的伦理安全体系，通过生成内容溯源、AI 图像识别等手段防范潜在风险，确保技术在合规、公平的框架下实现可持续发展，更好地服务于创意设计、教育、医疗等多元领域。

参考文献

[1] Zhang, M., Liu, F., Li, B., Liu, Z., Ma, W. and Ran, C. (2024) CrePoster: Leveraging Multi-Level Features for Cultural Relic Poster Generation via Attention-Based Framework. *Expert Systems with Applications*, **245**, Article ID: 123136. https://doi.org/10.1016/j.eswa.2024.123136

[2] Rathod, V.S., Tiwari, A. and Kakde, O.G. (2024) Folded Ensemble Deep Learning Based Text Generation on the Brain Signal. *Multimedia Tools and Applications*, **83**, 69019-69047. https://doi.org/10.1007/s11042-024-18124-z

[3] Huang, L.G. and Li, H.Y. (2023) Research on Image Generation Based on VAE and CGAN Fusion Model. In

Proceedings of the 2023 *International Conference on Computer Vision and Pattern Recognition*, Jingdezhen Ceramic Institute.

[4] Wang, P. and Yang, W. (2023) Text to Multi-Object Images Synthesis Based on Non-Local Self-Attention. In: *Proceedings of the* 2023 *International Conference on Artificial Intelligence and Pattern Recognition*, Chongqing University of Technology, 340-347. https://doi.org/10.1117/12.2681140

[5] Jiang, H., Kazi, R.H., Dontcheva, M., Zhao, S. and Shi, K. (2021) Automatic Layout for Interactive UI Elements. *ACM Transactions on Graphics* (*TOG*), **40**.

[6] Kang, M., Zhu, J., Zhang, R., Park, J., Shechtman, E., Paris, S., *et al*. (2023) Scaling up GANs for Text-to-Image Synthesis. 2023 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 18-22 June 2023, 1-10. https://doi.org/10.1109/cvpr52729.2023.00976

[7] Li, J., Yang, J., Zhang, J., Liu, C., Wang, C. and Xu, T. (2021) Attribute-Conditioned Layout GAN for Automatic Graphic Design. *IEEE Transactions on Visualization and Computer Graphics*, **27**, 4039-4048. https://doi.org/10.1109/tvcg.2020.2999335

[8] Sun, P., Liu, X., Weng, L. and Liu, Z. (2025) Generative Adversarial Network Based on Self-Attention Mechanism for Automatic Page Layout Generation. *Applied Sciences*, **15**, Article No. 2852. https://doi.org/10.3390/app15052852

[9] Wang, Y., Pu, G., Luo, W., Wang, Y., Xiong, P., Kang, H., *et al*. (2022) Aesthetic Text Logo Synthesis via Content-Aware Layout Inferring. 2022 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 1-10. https://doi.org/10.1109/cvpr52688.2022.00247

[10] Li, J., Yang, J., Hertzmann, A., Zhang, J. and Xu, T. (2021) LayoutGAN: Synthesizing Graphic Layouts with Vector-Wireframe Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**, 2388-2399. https://doi.org/10.1109/tpami.2019.2963663

[11] Xu, Y., Xia, M., Hu, K., Zhou, S. and Weng, L. (2025) Style Transfer Review: Traditional Machine Learning to Deep Learning. *Information*, **16**, 157-168. https://doi.org/10.3390/info16020157

[12] Tan, Y. (2022) Feature Recognition and Style Transfer of Painting Image Using Lightweight Deep Learning. *Computational Intelligence and Neuroscience*, **2022**, Article ID: 1478371. https://doi.org/10.1155/2022/1478371