**Computer Science and Application** 计算机科学与应用, 2025, 15(2), 220-229
Published Online February 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.152049

Hans汉斯

# 基于条件GANs的高分辨率图像合成模型

**张居正**

江西理工大学信息工程学院，江西 赣州

收稿日期：2025年1月22日；录用日期：2025年2月21日；发布日期：2025年2月28日

## 摘　要

**本研究提出了一种创新的方法，旨在通过条件生成对抗网络(Conditional Generative Adversarial Networks, cGANs)从语义标签图合成高分辨率、照片级逼真的图像。尽管条件性GANs在多个领域展现出广泛的应用潜力，但其生成的图像通常分辨率较低，且与真实图像的相似度存在显著差距。针对这一挑战，本研究引入了一种新颖的对抗性损失函数，并设计了一种多尺度生成器和判别器架构，以提升图像合成的质量和分辨率。具体而言，我们的方法能够产生2048 × 1024像素的高分辨率图像，这些图像在视觉吸引力上取得了显著的提升。通过与现有技术的比较，我们的方法在深度图像合成和编辑的质量及分辨率方面均展现出明显的优越性。本研究的创新之处在于提出了一种新的对抗性学习目标和多尺度架构，有效解决了cGANs在生成高分辨率图像时的不稳定性问题，并显著提高了图像细节和纹理的真实性，为高分辨率图像合成领域提供了新的技术路径。**

## 关键词

**条件生成对抗网络，多尺度生成器和判别器架构，对抗性损失**

# A High-Resolution Image Synthesis Model Based on Conditional GANs

**Juzheng Zhang**

School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou Jiangxi

Received: Jan. 22nd, 2025; accepted: Feb. 21st, 2025; published: Feb. 28th, 2025

## Abstract

**This study presents an innovative approach aimed at synthesizing high-resolution, photo-realistic images from semantically labeled graphs via Conditional Generative Adversarial Networks (cGANs). Although conditional GANs show a wide range of potential applications in several domains, the images they generate are usually of low resolution and have significant gaps in similarity to real images.**

文章引用: 张居正. 基于条件 GANs 的高分辨率图像合成模型[J]. 计算机科学与应用, 2025, 15(2): 220-229.
DOI: 10.12677/csa.2025.152049

**To address this challenge, this study introduces a novel adversarial loss function and designs a multi-scale generator and discriminator architecture to enhance the quality and resolution of image synthesis. Specifically, our method is able to produce high-resolution images of 2048 × 1024 pixels, which achieve remarkable results in terms of visual appeal. By comparing with existing techniques, our method demonstrates significant superiority in both quality and resolution of deep image synthesis and editing. The innovation of this study lies in the proposal of a new adversarial learning target and a multi-scale architecture, which effectively solves the instability problem of cGANs in generating high-resolution images and significantly improves the authenticity of image details and textures, providing a new technical path for the field of high-resolution image synthesis.**

## Keywords

**Conditional GANs, Multi-Scale Generator and Discriminator Architectures, Adversarial Loss**



## 1. 引言

在传统的图形学领域，实现高保真度图像渲染面临着显著的挑战，这要求对物体的几何结构、表面材质以及光线传播进行精确的模拟。虽然现有的图形算法能够处理这些复杂的模拟任务，但它们在构建和编辑虚拟场景时往往需要投入巨大的成本和时间。本研究提出了一种新途径，即通过学习数据驱动的模型来生成照片级真实的图像，从而将复杂的图形渲染任务转化为模型学习与推理的简化问题。本文介绍了一种新颖的方法，该方法能够从语义标签图生成高分辨率图像，并具有多方面的应用潜力。例如，我们可以利用这种方法生成合成的训练数据以训练视觉识别算法，因为相比于生成训练图像，创建所需场景的语义标签要简单得多。通过语义分割技术，我们能够将图像转换为语义标签，对标签中的对象进行编辑，然后再将它们转换回图像。

在从语义标签合成图像的过程中，我们采用了 pix2pix 方法，这是一种基于条件生成对抗网络(cGANs) [1]的图像到图像的翻译框架[2]。然而，Chen 和 Koltun 最近的研究[3]指出，在高分辨率图像生成任务中，传统的对抗性训练可能不够稳定，容易遭遇失败。他们提出了使用修改后的感知损失[4]-[6]来合成高分辨率图像，尽管这些图像在分辨率上达到了要求，但在细节的精细度和纹理的真实感方面仍有不足。在本研究中，我们针对现有先进方法中存在的两个主要问题提出了解决方案：1) 使用 cGANs 生成高分辨率图像时的不稳定性；2) 以往高分辨率图像在细节和纹理真实性上的不足。我们展示了通过引入一个新的、稳健的对抗性学习目标和多尺度生成器及判别器架构，我们能够成功合成 2048 × 1024 分辨率的逼真图像，这些图像不仅分辨率高，而且在细节和纹理上也更加接近真实世界。

## 2. 相关工作

生成式对抗网络(GANs) [1]的目标是通过生成与自然图像无法区分的样本，从而模拟自然图像的分布。GANs 在图像生成[1] [7]、表示学习[8]、图像处理[9]、物体检测[10]以及视频应用[11] [12]等多个领域展现了其强大的应用潜力。为了在无条件的情况下合成更大尺寸的图像(例如 256 × 256)，研究者们已经提出了多种从粗到细的方案[13] [14]。受到这些研究成果的启发，我们设计了一种新的从粗到细的生成器和多尺度判别器架构，以适应更高分辨率条件图像的生成需求。

在图像到图像的翻译领域，许多研究者借助对抗性学习来实现图像从一个域到另一个域的转换，训练数据由输入–输出图像对提供。相较于 L1 损失，后者常导致生成图像模糊[2] [6]，对抗性损失[1]已成为众多图像到图像任务的流行选择[15] [16]。这是因为判别器能够学习到一个可训练的损失函数，并自动适应生成图像与目标域真实图像之间的差异。例如，最近的 pix2pix 框架[2]利用图像条件的 GANs [17]，在不同应用中取得了成功，如将谷歌地图转换为卫星视图，以及从用户草图生成猫的图像。

然而，Chen 和 Koltun [3]指出，由于训练过程中的不稳定性以及优化问题，条件型 GANs 在生成高分辨率图像时面临挑战。为了解决这一难题，他们采用了一种基于感知损失的直接回归目标[4]-[6]，并成功开发出首个能够合成 $2048\times1024$ 图像的模型。尽管该模型生成的图像分辨率很高，但细节的精细度和纹理的真实性仍有待提高。我们的研究正是在他们的成功基础上进一步发展而来的。

## 3. 实例级的图像合成

我们构建了一个条件对抗框架，旨在从语义标签图合成高分辨率且具有照片级真实感的图像。首先，我们回顾了基线模型 pix2pix 的相关内容(详见第 3.1 节)。随后，我们阐述了通过优化目标函数和网络架构来提升生成图像逼真度与分辨率的方法(详见第 3.2 节)。此外，我们还利用额外的实例级对象语义信息来进一步提高图像质量(详见第 3.3 节)。最终，我们引入了一种实例级特征嵌入方案，以更有效地处理图像合成过程中固有的多模态特性(详见第 3.4 节)。

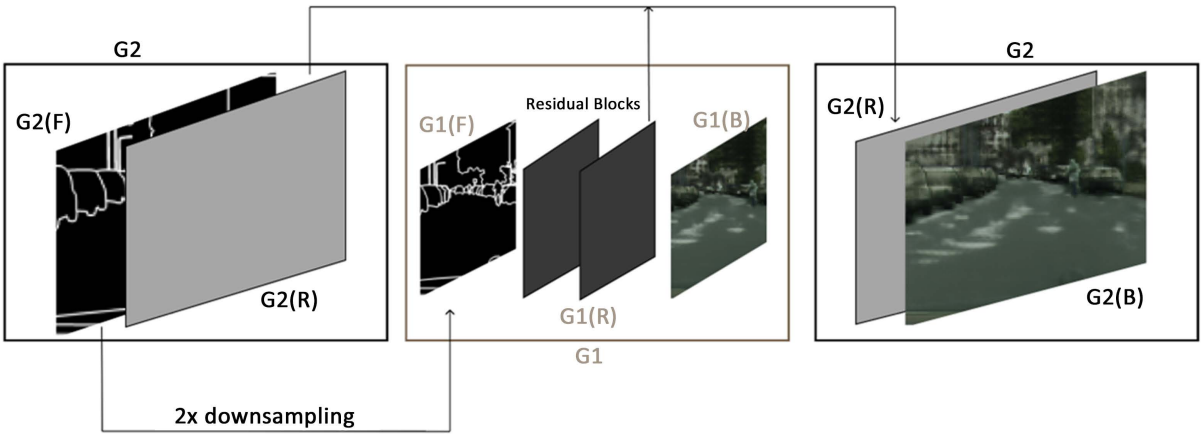


**Figure 1.** Coarse-to-fine generator network architecture
**图 1.** 粗到细生成器网络结构

### 3.1. pix2pix 基线

pix2pix 方法[2]是一种图像到图像翻译任务中使用的条件生成对抗网络(cGAN)框架，它包括一个生成器 $G$ 和一个判别器 $D$。生成器 $G$ 致力于将输入的语义标签图转换为逼真的图像，而判别器 $D$ 的任务是区分生成的图像与真实的图像。该训练数据集是一组对应的 Images$\{(s_i, x_i)\}$，其中 $s_i$ 是一个语义标签图，$x_i$ 是相应的自然照片。条件型 GANs 的目的是通过以下的最小极限博弈对真实图像的条件分布进行建模：$\min_G \max_D L_{GAN}(G, D)$，其中目标函数 $L_{GAN}(G, D)$ 由以下公式给出：

$$L_{GAN}(G,D) = E_{(s,x)}\left[\log D(s,x)\right] + E_{(s)}\left[\log\left(1 - D\left(s, G_{(s)}\right)\right)\right] \tag{1}$$

其中，$\log D(s,x)$ 表示判别器 $D$ 对真实图像对$(s,x)$的对数概率。$E_{(s,x)}\left[\log D(s,x)\right]$表示判别器 $D$ 在真实图像对$(s,x)$上的期望损失。$G_{(s)}$ 是生成器 $G$ 根据语义标签图 $s$ 生成的图像。$\log\left(1 - D\left(s, G_{(s)}\right)\right)$表示判别器

$D$ 对生成图像对 $\left(s,G_{(s)}\right)$ 的对数概率。$E_{(s)}\left[\log\left(1-D\left(s,G_{(s)}\right)\right)\right]$ 表示判别器 $D$ 在生成图像对 $D\left(s,G_{(s)}\right)$ 上的期望损失。判别器 D 的目标是尽可能正确地识别出真实图像对，因此 $D$(s,x)越接近 1 越好，$D\left(s,G_{(s)}\right)$ 越接近 0 越好。在 pix2pix 方法中，生成器采用了 U-Net 架构[18]，而判别器则基于补丁的全卷积网络[19]构建。判别器的输入是将语义标签映射与对应的图像通道进行连接。该方法生成的图像分辨率最高为 256 × 256。然而，在尝试直接利用 pix2pix 框架生成更高分辨率的图像时，我们发现训练过程不稳定，且生成的图像质量未能达到预期效果。

### 3.2. 提高逼真度和分辨率

我们对 pix2pix 框架进行了改进，主要通过引入一个从粗到细的生成器、一个多尺度判别器架构以及一个稳健的对抗性学习目标函数来实现。

**粗到细的生成器** 我们将生成器分解成两个子网络：$G_1$ 和 $G_2$。我们称 $G_1$ 为全局生成器网络，$G_2$ 为局部增强器网络。然后，发生器由元组 $G=\{G_1,G_2\}$ 给出，如图 1 所示。全局生成器网络的分辨率为 1024 × 512，而局部增强器网络输出的图像的分辨率是前者输出尺寸的 4 倍(每个图像维度为 2 倍)。为了以更高的分辨率合成图像，可以利用额外的局部增强器网络。例如，生成器 $G=\{G_1,G_2\}$ 的输出图像分辨率是 2048 × 1024，而输出图像 $G=\{G_1,G_2,G_3\}$ 的输出图像分辨率是 4096 × 2048。我们的全局发生器是建立在 Johnson 等人[6]提出的架构上的，该架构已被证明是成功地在 512 × 512 的图像上的神经风格转移。它由 3 个部分组成：一个卷积前端 $G_1^{(F)}$，一组残余块 $G_1^{(R)}$，和一个转置的卷积后端 $G_1^{(B)}$。一个分辨率为 1024 × 512 的语义标签图依次通过这 3 个组件，输出分辨率为 1024 × 512 的图像。

局部增强器网络也由 3 个部分组成：一个卷积前端 $G_2^{(F)}$，一组残余块 $G_2^{(R)}$，和一个转置的卷积后端 $G_2^{(B)}$。$G_2$ 的输入标签图的分辨率为 2048 × 1024。与全局生成器网络不同，残差块 $G_2^{(R)}$ 的输入是两个特征图的元素之和：$G_2^{(F)}$ 的输出特征图和全局生成器网络后端 $G_1^{(B)}$ 的最后一个特征图。这有助于整合从 $G_1$ 到 $G_2$ 的全局信息。

在训练过程中，我们首先训练全局发生器，然后按照它们的分辨率顺序训练局部增强器，以完成图像合成任务。

**多尺度判别器** 高分辨率图像合成对 GAN 判别器的设计提出了巨大挑战。为了区分高分辨率的真实图像和合成的图像，鉴别器需要有一个大的接受域，这就需要一个更深的网络或更大的卷积核。另外，过拟合将成为一个更令人担忧的问题。

为了解决这个问题，我们使用 3 个判别器，它们具有相同的网络结构，但在不同的图像尺度上运行。我们将这些判别器称为 $D_1$、$D_2$ 和 $D_3$。具体来说，我们将真实的和合成的高分辨率图像按 2 和 4 的系数进行降样，以创建一个 3 个尺度的图像。然后训练鉴别器 $D_1$、$D_2$ 和 $D_3$，以分别区分 3 个不同尺度的真实和合成图像，即判别器 $D_1$ 处理原始分辨率的图像，判别器 $D_2$ 处理降样 2 倍的图像，判别器 $D_3$ 处理降样 4 倍的图像。由于将低分辨率的模型扩展到更高的分辨率只需要在最细的级别增加一个额外的鉴别器，而不是从头开始重新训练，这使得训练从粗到细的生成器更加容易。

有了多个判别器，学习问题就变成了一个多任务学习问题，即：

$$\min_{G}\max_{D1,D2,D3}\sum_{k=1,2,3}L_{GAN}\left(G,D_K\right) \tag{2}$$

在同一图像尺度上使用多个 GAN 判别器已经在无条件的 GANs 中被提出[20]。Iizuka 等人[21]在条件 GANs 中加入了一个全局图像分类器来进行画像。在这里我们将这一设计扩展到不同图像尺度的多个判别器上，以便为高分辨率图像建模。

**改进的对抗性损失** 我们对 GAN 的损失函数进行了改进，通过在公式(1)中引入基于特征匹配的损失项

来实现。这一改进有助于稳定训练过程，因为生成器需要在多个尺度上生成符合自然图像统计特性的数据。具体来说，我们从多层判别器中提取特征，并学习如何在真实图像和合成图像之间匹配这些中间表征。为了便于表述，我们将鉴别器 $D_k$ 的第 $i$ 层特征提取器表示为 $D_k^{(i)}$。那么，特征匹配损失 $\mathrm{L}_{\mathrm{FM}}\ (\mathrm{G}, \mathrm{D}_{\mathrm{k}})$就是：

$$L_{FM}\left(G, D_k\right)=E_{(s,\text{x})} \sum_{i=1}^{T} \frac{1}{N_i}\left[\left\|D_k^{(i)}(s, x)-D_k^{(i)}(s, G(s))\right\|_1\right] \tag{3}$$

其中 $T$ 是总的层数，$N_i$ 表示每层的元素数。$D_k^{(i)}(s, x)$ 表示真实图像在第 $i$ 层的特征，$D_k^{(i)}(s, G(s))$ 则表示生成图像在第 $i$ 层的特征，$\left\|D_k^{(i)}(\mathrm{s}, \mathrm{x})-D_k^{(i)}(\mathrm{s}, G(\mathrm{s}))\right\|_1$ 则为真实图像和生成图像在第 $i$ 层特征图上的 L1 距离，即第 $i$ 层的特征匹配损失。我们的 GAN 判别器特征匹配损失与感知损失相关[4] [6]，而感知损失已被证明在图像超分辨率[16]和风格迁移[6]等任务中非常有效。

### 3.3. 使用实例图

目前的图像合成技术主要依赖于语义标签图[2] [3] [15]，这类图像通过每个像素的值来表示该像素所对应的物体类别，但它们无法区分属于同一类别的不同物体。相比之下，我们认为实例图能够提供至关重要的信息，即对象的边界信息，这是语义标签图所缺失的。例如，在多个同级别的物体紧密相邻的情况下，仅凭语义标签图无法将它们有效区分，而实例图的存在则使得这一区分过程变得更加简单。

为了提取这些关键信息，我们首先计算了实例边界图(如图 2 所示)。在我们的实现中，如果一个像素的对象标识与其四个相邻像素中的任何一个不同，那么该像素在实例边界图中的值为 1，否则为 0。随后，我们将实例边界图与语义标签图的单个向量表示进行串联，并将结果输入到生成器网络中。同样地，判别器的输入是实例边界图、语义标签图和语境图的通道连接。图 3 和图 4 展示了一个示例，证明了利用物体边界信息所带来的改进效果。

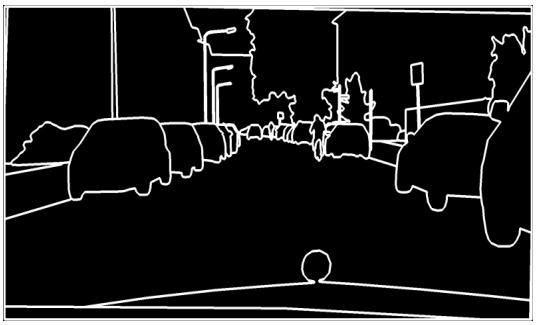

**Figure 2.** Instance boundary map
**图 2.** 实例边界图

**Figure 3.** Before boundary improvement
**图 3.** 边界改进前

**Figure 4.** After boundary improvement
**图 4.** 边界改进后

### 3.4. 学习实例级特征嵌入

从语义标签图进行图像合成是一个一多映射问题。理想的图像合成算法应该能够使用相同的语义标签图生成不同的现实图像。为了生成多样化的图像并允许实例级的控制，我们为图像中的每个实例增加了额外的低维特征通道作为生成器的输入。我们表明，通过操纵这些特征，我们可以对合成过程进行灵活的控制。

我们训练一个编码器 $E$ 来学习一个对应于地面真实图像的特征图。为了使每个实例中的特征保持一致，我们在编码器的输出中加入一个实例平均池层。然后，平均特征被广播到同一实例的所有像素位置。

在得到这个特征图 $E(x)$后，我们通过将标签图和 E(x)串联起来，将公式(4)中的 $G(s)$替换为 $G(s, E(x))$，并与生成器共同进行端到端的训练。这使得编码器能够捕获最具代表性的特征供生成器使用，例如道路的纹理，而无需明确告诉编码器什么是“纹理”。为了在推理时进行交互式编辑，在编码器被训练后，我们首先在训练图像中的所有实例上运行它，并记录获得的特征。然后，我们对每个语义类别的这些特征进行 K-means 聚类。因此，每个聚类都对特定风格的特征进行编码。而在推理时，我们随机挑选一个聚类中心并将其作为编码特征。结合公式(2)和(3)，其中 $\lambda$ 作为超参数用于平衡两种损失的权重，最终我们推出将公式(4)如下：

$$\min_{G}\left(\left(\max_{D1,D2,D3}\sum_{k=1,2,3}L_{GAN}\left(G,D_{K}\right)\right)+\lambda\sum_{k=1,2,3}L_{FM}\left(G,D_{k}\right)\right) \tag{4}$$

## 4. 实验结果

### 4.1. 提高逼真度和分辨率的测试

**实验目的**：测试对比前文中提到的生成的高分辨率图像的质量。

**数据集**：

**训练数据集**主要包括实例图、标签图以及真实的图像，其中我们使用数据集每种类型 8 张图片。

**测试数据集**则由标签图和实例图组成，每种类型 8 张。

**实验过程及结果**：我们先用训练集训练了 1024 × 512 的模型，并在此基础上通过粗到细的生成器来生成 2048 × 1024 的模型。通过提高分辨率的对比测试，可以使生成的图片质量有着明显地提升。通过视觉对比(如图 5 和图 6 所示)，可以发现 2048 × 1024 分辨率的图像在纹理清晰度和细节丰富度上有明显提升。例如，道路的纹理、建筑物的轮廓以及车辆的细节在高分辨率图像中更为逼真。使用拉普拉斯方差对生成的图像进行评估，1024 × 512 模型生成图片的拉普拉斯方差为 92.39，而 2048 × 1024 模型生成图片的拉普拉斯方差提升到 190.2，可见逐步精细的生成策略在提高图像分辨率和清晰度方面是有效的。

**Figure 5.** 1024 × 512 resolution
**图 5.** 1024 × 512 分辨率

**Figure 6.** 2048 × 1024 resolution
**图 6.** 2048 × 1024 分辨率

## 4.2. 是否使用实例图的比较测试

**实验目的：**对比训练时有无实例图数据集对生成图片的细节影响。

**数据集：**

**测试数据集**同 4.1 一样。

**训练数据集**主要分为带有实例图的训练集和不带实例图的训练集。

**实验过程及结果：**与 4.1 中的实验过程类似，分别使用带有实例图和不带实例图的训练集进行模型训练。如图 7 和图 8 所示，相比较于未使用实例图的数据集训练的模型，可以发现使用实例图的模型在物体边界和细节处理上更为清晰和自然。例如，道路的边界在使用实例图的模型中更为明显。

**Figure 7.** Training set with instance map at 2048 × 1024 resolution
**图 7.** 使用实例图作训练集 2048 × 1024

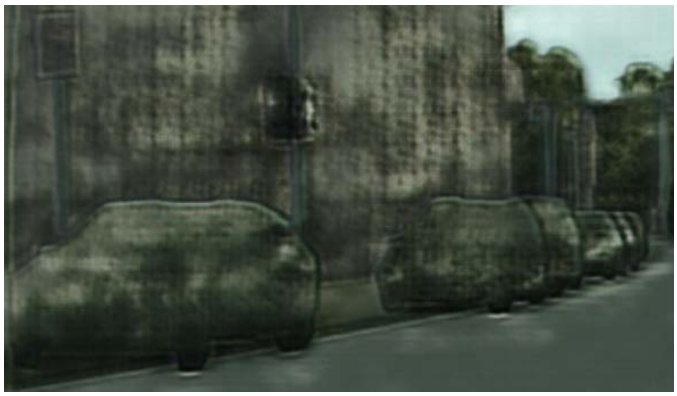

**Figure 8.** Training set without instance map at 2048 × 1024 resolution
**图 8.** 未使用实例图作训练集 2048 × 1024

### 4.3. 实例级特征嵌入的比较测试

**实验目的**：用实例级的特征嵌入做图片细节的比较。

**数据集**：这里的数据集同 4.1 的数据集一样。

**实验过程及结果**：我们先提取训练数据集中的真实图像和实例图的特征编码，存储到对应的标签图的特征矩阵中，再对于标签的特征进行特征聚类。之后，我们用这些特征编码分别代入到了模型的训练中。如图 9 与图 10 所示，经由实例级特征嵌入的对比分析，能够明显地看出图像中汽车细节部分的处理在视觉呈现上更为直观且清晰。

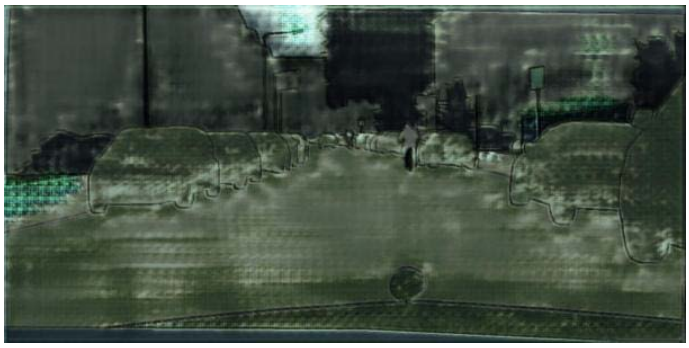

**Figure 9.** Without instance feature embedding at 2048 × 1024 resolution
**图 9.** 未使用实例特征嵌入 2048 × 1024

**Figure 10.** With instance feature embedding at 2048 × 1024 resolution
**图 10.** 使用实例特征嵌入 2048 × 1024

### 4.4. 实验结论与分析

为了进一步验证所提模型在高分辨率图像合成任务中的性能，我们将其与当前主流方法PerceptionGAN [22]和StackGAN [23]进行了对比实验。实验采用第4.1节中相同的数据集进行训练，并以2048 × 1024分辨率的生成图像作为对比对象。定量评估结果显示：PerceptionGAN的拉普拉斯方差为176.3，NIQE值为8.5，BRISQUE值为42.6；StackGAN的拉普拉斯方差为188.6，NIQE值为7.8，BRISQUE值为40.2；而我们的方法拉普拉斯方差达到190.2，NIQE值为7.7，BRISQUE值为38.5。分析表明，PerceptionGAN虽然在生成语义一致的图像方面表现出色，但在高分辨率图像的清晰度和细节表现上存在不足。这主要是由于其感知损失主要关注全局语义一致性，对局部细节的优化能力有限。StackGAN通过堆叠多个生成对抗网络逐步细化图像，能够在一定程度上生成较为清晰的图像，但在处理复杂场景时，尤其是在实例级细节的生成上，仍可能出现模糊或细节丢失的情况。相比之下，我们的方法通过在多尺度判别器中引入特征匹配损失，并为每个实例分配独特的特征向量，使得生成器能够生成更自然的细节，有效避免了细节的重复和不自然感，从而在图像清晰度、细节表现和自然度方面均优于其他两种方法。

在本次实验中，我们通过多个对比测试深入探究了不同因素对图像生成质量的影响。首先，在提高逼真度和分辨率的测试中，我们发现从1024 × 512模型基础上通过粗到细的生成器生成的2048 × 1024模型，其生成的图片质量有明显提升，这说明了逐步精细的生成策略在提高图像分辨率和逼真度方面是有效的。其次，在是否使用实例图的比较测试里，对比结果清晰地表明，带有实例图的训练集所训练出的模型在生成图片的各个细节方面，如汽车、道路等，都优于不带实例图训练集的模型，这凸显了实例图在丰富图像细节、增强模型对具体物体特征把握能力方面的重要性。接着，在实例级特征嵌入的比较测试中，通过提取特征编码并进行特征聚类后再代入模型训练，使得生成图片的汽车细节部分处理更为直观，这证明了实例级特征嵌入能够进一步优化模型对特定物体细节的呈现效果。最后，我们与主流方法进行了比较实验，通过评估表明我们的方法优于主流方法。

## 5. 总结

本文提出了一种基于条件生成对抗网络(cGANs)的高分辨率图像合成方法，旨在通过语义标签图和实例图生成高分辨率、照片级逼真的图像。通过一系列实验和对比分析，我们验证了该方法在图像清晰度、细节表现和自然度方面的优势，并与当前主流方法进行了详细对比，通过评估拉普拉斯方差、NIQE和BRISQUE等指标，我们的方法在图像清晰度、自然度和整体质量方面均优于PerceptionGAN和StackGAN。这表明我们的方法在高分辨率图像合成任务中具有显著优势。在未来的工作中仍有一些方向可以进一步探索和改进，例如在更大规模和更多样化的数据集上验证模型的泛化能力和稳定性，持续探索与Diffusion Models等最新技术[24]-[26]的结合，以实现更高质量的图像合成和更广泛的应用。

## 参考文献

[1] Jia, Y., Yu, W. and Zhao, L. (2024) Generative Adversarial Networks with Texture Recovery and Physical Constraints for Remote Sensing Image Dehazing. *Scientific Reports*, **14**, Article No. 31426. https://doi.org/10.1038/s41598-024-83088-x

[2] Isola, P., Zhu, J., Zhou, T. and Efros, A.A. (2017) Image-to-Image Translation with Conditional Adversarial Networks. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 5967-5976. https://doi.org/10.1109/cvpr.2017.632

[3] Chen, Q. and Koltun, V. (2017) Photographic Image Synthesis with Cascaded Refinement Networks. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 1520-1529. https://doi.org/10.1109/iccv.2017.168

[4] Dosovitskiy, A. and Brox, T. (2016) Generating Images with Perceptual Similarity Metrics Based on Deep Networks.

*Advances in Neural Information Processing Systems*, Barcelona, 5-10 December 2016, 29.

[5] Choi, J., Lee, J., Shin, C., Kim, S., Kim, H. and Yoon, S. (2022). Perception Prioritized Training of Diffusion Models. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 11462-11471. https://doi.org/10.1109/cvpr52688.2022.01118

[6] Johnson, J., Alahi, A. and Fei-Fei, L. (2016) Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In: *Lecture Notes in Computer Science*, Springer, 694-711. https://doi.org/10.1007/978-3-319-46475-6_43

[7] Li, Z., Xia, B., Zhang, J., *et al*. (2022) A Comprehensive Survey on Data-Efficient GANs in Image Generation.

[8] Salimans, T., Goodfellow, I., Zaremba, W., *et al*. (2016) Improved Techniques for Training Gans. *Advances in Neural Information Processing Systems*, Barcelona, 5-10 December 2016, 29.

[9] Zhu, J., Krähenbühl, P., Shechtman, E. and Efros, A.A. (2016) Generative Visual Manipulation on the Natural Image Manifold. In: *Lecture Notes in Computer Science*, Springer, 597-613. https://doi.org/10.1007/978-3-319-46454-1_36

[10] Li, J., Liang, X., Wei, Y., Xu, T., Feng, J. and Yan, S. (2017) Perceptual Generative Adversarial Networks for Small Object Detection. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 1951-1959. https://doi.org/10.1109/cvpr.2017.211

[11] Mathieu, M., Couprie, C. and LeCun, Y. (2015) Deep Multi-Scale Video Prediction beyond Mean Square Error.

[12] Tulyakov, S., Liu, M., Yang, X. and Kautz, J. (2018). Mocogan: Decomposing Motion and Content for Video Generation. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 1526-1535. https://doi.org/10.1109/cvpr.2018.00165

[13] Burt, P.J. And Adelson, E.H. (1987) The Laplacian Pyramid as a Compact Image Code. In: *Readings in Computer Vision*, Elsevier, 671-679. https://doi.org/10.1016/b978-0-08-051581-6.50065-9

[14] Denton, E.L., Chintala, S. and Fergus, R. (2015) Deep Generative Image Models Using a Laplacian Pyramid of Adversarial Networks. 2015 *Advances in Neural Information Processing Systems*, Montreal, 7-12 December 2015, 28.

[15] Karacan, L., Akata, Z., Erdem, A., *et al*. (2016) Learning to Generate Images of Outdoor Scenes from Attributes and Semantic Layouts.

[16] Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., *et al*. (2017) Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 4681-4690.

[17] Mirza, M. (2014) Conditional Generative Adversarial Nets.

[18] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. In: *Lecture Notes in Computer Science*, Springer, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[19] Long, J., Shelhamer, E. and Darrell, T. (2015) Fully Convolutional Networks for Semantic Segmentation. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 3431-3440. https://doi.org/10.1109/cvpr.2015.7298965

[20] Durugkar, I., Gemp, I. and Mahadevan, S. (2016) Generative Multi-Adversarial Networks.

[21] Iizuka, S., Simo-Serra, E. and Ishikawa, H. (2017) Globally and Locally Consistent Image Completion. *ACM Transactions on Graphics*, **36**, 1-14. https://doi.org/10.1145/3072959.3073659

[22] Garg, K., Singh, A.K., Herremans, D. and Lall, B. (2020) Perceptiongan: Real-World Image Construction from Provided Text through Perceptual Understanding. 2020 *Joint 9th International Conference on Informatics, Electronics & Vision* (*ICIEV*) and 2020 4*th International Conference on Imaging, Vision & Pattern Recognition* (*icIVPR*), Kitakyushu, 26-29 August 2020, 1-7. https://doi.org/10.1109/icievicivpr48672.2020.9306618

[23] Rao, A.S., Bhandarkar, P.A., Devanand, P.A., Shankar, P., Shanti, S., *et al*. (2023) Text to Photo-Realistic Image Synthesis Using Generative Adversarial Networks. 2023 2*nd International Conference on Futuristic Technologies* (*INCOFT*), Belagavi, 24-26 November 2023, 1-6. https://doi.org/10.1109/incoft60753.2023.10425482

[24] Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., *et al*. (2023) Diffusion Models: A Comprehensive Survey of Methods and Applications. *ACM Computing Surveys*, **56**, 1-39. https://doi.org/10.1145/3626235

[25] Graikos, A., Yellapragada, S., Le, M., Kapse, S., Prasanna, P., Saltz, J., *et al*. (2024) Learned Representation-Guided Diffusion Models for Large-Image Generation. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 16-22 June 2024, 8532-8542. https://doi.org/10.1109/cvpr52733.2024.00815

[26] Peebles, W. and Xie, S. (2023) Scalable Diffusion Models with Transformers. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Paris, 1-6 October 2023, 4172-4182. https://doi.org/10.1109/iccv51070.2023.00387