利用 Pytorch 构建与分析 Pix2Pix 模型的优化与改进技术报告

作者：梨花先雪

摘要

我们研究了生成对抗网络的基本原理，并探索这些网络如何实现图像到图像的转换任务。这类网络不仅能够学习从输入图像到输出图像的映射关系，还能够自动学习适合该映射的损失函数，从而优化模型训练。Pix2Pix 是深度生成网络中的一个典型模型，它基于条件 GAN 实现图像生成。在本研究中，我们重点解析了 Pix2Pix 深度生成模型的工作原理，完成了该模型的实现，并利用它进行图像生成任务。此外，我们通过对比实验和消融实验等方法，分析了不同参数对 Pix2Pix 模型训练效率及生成效果的影响。最后，我们对模型的归一化方法进行了改进，引入了注意力机制，从而显著提升了实验结果。

关键词— Pix2Pix, GAN, 图像生成, 损失函数, U-Net, PatchGAN

1. 引言

计算机视觉的迅猛发展为图像生成与翻译任务带来了全新的机遇与挑战。图像翻译技术的价值日益显现，其核心目标是将图像从一个领域映射到另一个领域，为图像处理与合成提供了高效的解决方案。在该领域中，生成对抗网络（GANs）的引入为图像生成任务注入了强大的动力，而条件生成对抗网络（CGAN）的创新应用则进一步推动了图像翻译技术的发展。特别是 Pix2Pix 模型，凭借其独特的结构设计与卓越的性能，成为该研究方向中的经典代表之一。

1.1 GAN

生成对抗网络（GAN）通常由生成器（生成网络）和判别器（判别网络）两部分组成。生成器的主要功能是通过学习训练数据的特征，在判别器的反馈指导下，将随机噪声分布逐步调整为接近训练数据的真实分布，从而生成具有训练数据特征的相似样本。而判别器的任务则是判断输入数据的真实性，即区分数据是来源于真实分布还是生成器生成的伪数据，并将判断结果作为反馈提供给生成器。两者通过交替训练，共同提升能力，直至生成器能够生成足以以假乱真的数据，与判别器的判别能力达到一种平衡状态。

由于 GAN 模型具有结构简洁、原理易懂的特点，同时还能生成训练集中未出现过的图像，因此在很长一段时间内成为生成模型研究的核心对象。为了应对不同的应用需求，GAN 还衍生出多种变体，用以解决各类下游问题。

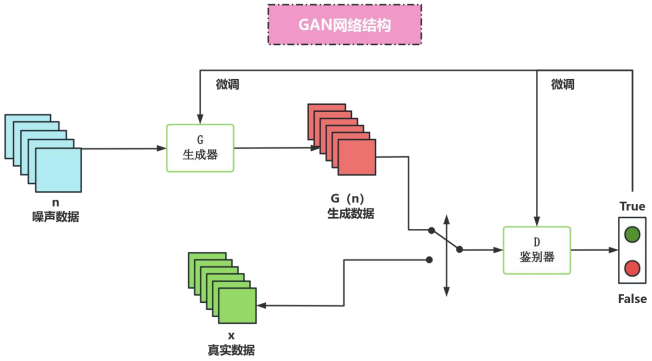


图 1 GAN 工作原理的示意图

1.2 cGAN

GAN 的输入是一个 n 维随机向量，输出则为某一类别的图像。尽管这些图像属于同一类别并呈现相似风格，但我们无法直接控制生成数据的具体外观。然而，如果希望生成的图像可控，则需要额外引入一个输入标签作为指导条件。这类模型通常被称为条件模型（conditional model），基于此理念扩展而来的生成对抗网络被称为条件生成对抗网络（conditional GAN，简称 cGAN）。

简单来说，在 cGAN 中，指导条件（通常表示为 y）会被编码为向量，并与随机向量 z 拼接（通过 concatenate 操作），然后输入到生成器中，生成具有特定条件的图像 G(z,y)。在判别阶段，条件 y 同样作为额外信息，通过多层映射后与真实数据 x 或生成数据 G(z,y)融合形成新的向量，并输入判别器进行真假判定。

1.3 pix2pix

Pix2Pix 是一种基于条件生成对抗网络（cGAN）的深度学习图像转换模型，由 Phillip Isola 等人在 2017 年的 CVPR 会议上提出。该模型能够实现多种图像转换任务，包括语义图或标签到真实图像、灰度图到彩色图、航空图到地图、白天到黑夜、线条草图到实际图像的转

化。作为将 cGAN 应用于有监督图像到图像翻译的经典模型，Pix2Pix 包括两个主要组件：生成器和判别器。

与传统方法不同，尽管这些任务的目标都是通过像素间的预测来完成图像转换，往往需要为每个任务设计不同的专用模型，而 Pix2Pix 则提供了一个统一框架，使用相同的架构和目标，在不同的数据集上进行训练，从而得到令人满意的转换效果。正因如此，许多研究者和艺术家已经采用该网络，并通过其创作了各类艺术作品。

尽管 Pix2Pix 取得了显著的成果，仍然面临一些挑战和待解决的问题。其中之一是如何在复杂场景中保持图像的细节，尤其是在高分辨率图像转换的过程中。此外，模型的训练和收敛也存在一定的技术难题。为了解决这些问题，研究人员正在不断提出新的方法和技术。

在本节的最后一段中，我将对本技术报告的主要贡献总结如下：
1) 实现 Pix2Pix 深度生成模型：完成模型的搭建与训练，并生成目标图像。
2) 参数优化与关键因素探究：对模型的超参数进行调节，分析和验证不同参数对图像生成质量的影响。
3) 消融实验：通过移除或替换模型中的某些组件，评估其对整体性能的贡献，深入理解模型的关键组成部分。
4) 模型改进：对归一化方法进行优化，引入注意力机制，以提升模型的图像生成质量与表现。

2. 研究方法

2.1 pix2pix 方法

（1）条件生成网络（cGAN）

GAN 是一种生成模型，旨在学习从随机噪声向量 z 到输出图像 y 的映射关系，即 G:z→y。与之相比，条件 GAN 则扩展了输入范围，学习从观测图像 x 和随机噪声向量 z 到目标图像 y 的映射，即 G:{x,z}→y。通过对抗训练，判别器 D 被训练以尽可能准确地区分生成器 G 产生的图像与真实图像。同时，生成器 G 的目标是生成足够逼真的输出，使判别器 D 无法区分真假。整个训练过程如图所示。

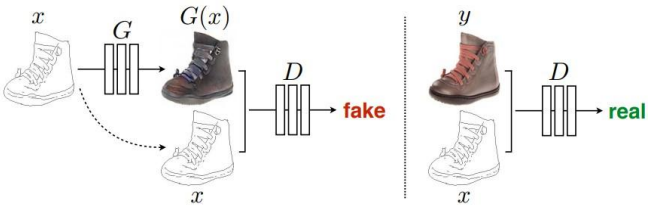


图 2 GAN 训练过程如图

图 2 展示了训练条件 GAN 来实现边缘到照片的映射过程。判别器 D 被训练用来区分由生成器 G 合成的“假”图像与真实的 {边缘，照片} 对。在此过程中，生成器 G 的目标是生成足够逼真的图像，以欺骗判别器 D。与无条件 GAN 不同，条件 GAN 中的生成器和判别器都会同时接收到输入的边缘图作为条件信息，以此指导生成和判别过程。

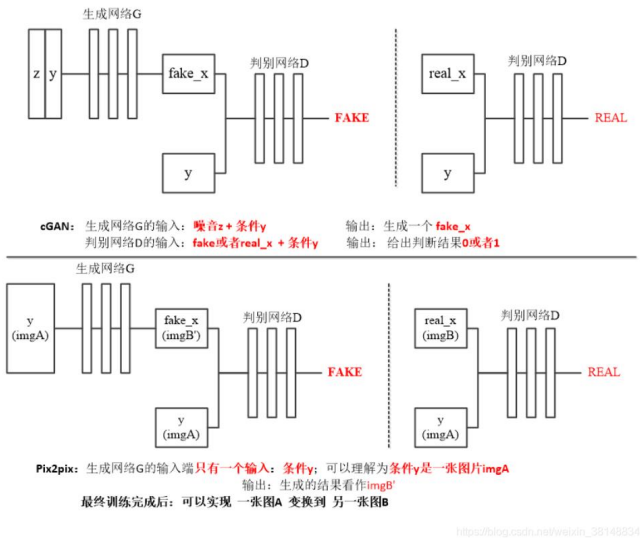


图 3 普通 CGAN 的结构（上），Pix2Pix 的结构（下）

（2）Loss 函数

x 是输入图像，y 是真实图像，z 是噪声；判别器 G 想让 Loss 最大化，而生成器 D 则想让 Loss 最小化。引入 Conditional GAN 的代价函数：

$$L_{cGAN}(G,D) = E_{x,y}[\log D(x,y)] + E_{x,z}[\log(1 - D(x,G(x,z)))]$$

作为对比，以下列一个普通 GAN 的 loss 函数：

$$L_{GAN}(G,D) = E_{y}[\log D(y)] + E_{x,z}[\log(1 - D(G(x,z)))]$$

受到前人工作启发，在这基础上又加入了人 L1 代价函数：

$$L_{L1}(G) = E_{x,y,z}[||y - G(x,z)||_1]$$

所以最终的目标函数就是：

$$G^* = \arg\min_G \max_D L_{cGAN}(G,D) + \lambda L_{L1}(G)$$

2.2 pix2pix 网络结构

与传统 GAN 使用多层感知机（MLP）作为模型结构不同，Pix2Pix 采用了卷积神经网络（CNN）中常见的 卷积 + 批归一化（BN）+ ReLU 的结构设计。该结构在 Pix2Pix 中的具体实现和细节如下所述。

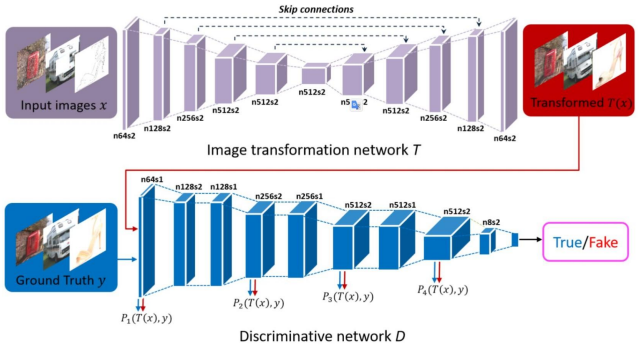


图 4 pix2pix 网络结构

（1）带跳跃的生成器（U-NET）

图像到图像转换的一个显著特征是将高分辨率的输入网格映射到同样高分辨率的输出网格。此外，对于我们研究的问题，尽管输入与输出的表面外观不同，但二者往往源于相同的基础结构，因此输入和输出之间的结构通常大致对齐。基于这一特性，我们设计了适合该场景的生成器网络架构。

在以往的解决方案中，编码器-解码器网络是常见的选择。这类网络通过逐层降采样将输入映射至一个瓶颈层，再通过逐层上采样恢复至原始分辨率。然而，这种架构要求所有信息必须通过瓶颈层流动。对于许多图像翻译任务，由于输入与输出共享大量底层信息，直接穿过网络传递这些信息更为高效。例如，在图像着色任务中，输入与输出共享显著边缘的位置。

为了让生成器能够跳过瓶颈层传递类似的信息，我们借鉴了 U-Net 的结构设计，在网络中添加了跳跃连接。具体而言，我们在第 i 层与第 n−i 层之间添加了跳跃连接，其中 n 是网络的总层数。每个跳跃连接会直接将第 i 层的所有通道与第 n−i 层的通道相连，从而实现信息的高效传递。

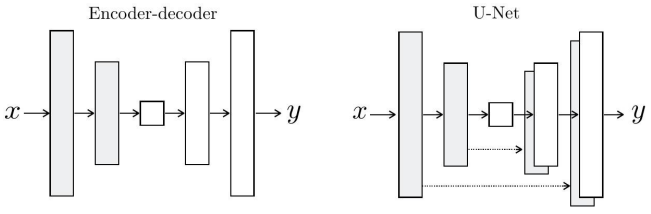


图 5 generator 有两种结构选择

U-Net "是一个编码器-解码器，在编码器和解码器堆栈的镜像层之间有跳过的连接。

- Pix2Pix 的网络结构：采用了基于 U-Net 的设计，通过在压缩路径和扩张路径之间添加跳跃连接来实现特征的高效传递。
- 输入图像尺寸：模型处理的输入图像尺寸为 256×256。
- 降采样策略：整个网络在压缩路径中进行了三次降采样，每次降采样后通道数均增加一倍，初始通道数为 64。

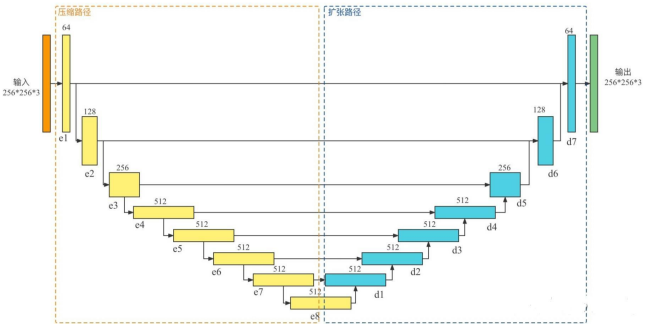


图 6 U-Net 网络结构图

- 压缩路径的操作：每个降采样操作由大小为 4×4 的卷积核、批归一化（BN）和 ReLU 激活函数组成，卷积步长由是否降采样决定。
- 扩张路径的操作：扩张路径通过反卷积操作进行上采样，以恢复空间分辨率。
- 特征融合方式：压缩路径与扩张路径之间通过拼接操作进行特征融合，以充分利用多层次特征信息。

（2）马尔可夫判别器(PatchGAN)

传统 GAN 的一个难点在于生成的图像通常较为模糊，其重要原因之一是判别器将整张图像作为输入进行判断。而 Pix2Pix 模型的判别器采用了 PatchGAN，这种设计对输入图像的每个局部区域（patch）输出一个预测概率值，从而将传统的“整图判断”转变为“局部区域判断”。

例如，当判别器的输入尺寸为 1×3×256×256（批次大小为 1，3 个通道，图像分辨率为 256×256），若设置 N=8，则判别器的输出为 1×1×32×32。其中，输出中的每个值表示对应输入图像中 8×8 区域的真实性概率。

Pix2Pix 利用了具备马尔科夫特性的 PatchGAN 判别器，通过将低频成分的重建与高频成分的生成相结合，提升了图像质量。PatchGAN 的核心思想是仅用于构建高频信息，因此无需整图输入判别器，而是对图像的每个 N×N 区域（patch）单独进行真假判断，并假设不同的 patch 相互独立。最终，将整张图像中所有 patch 的判断结果取平均，作为判别器的最终输出。

在具体实现中，PatchGAN 使用一个全卷积小网络，其输入为任意大小的 N×N 区域。网络的最后一层通过 sigmoid 激活函数输出每个 patch 的真实概率，并使用二元交叉熵损失（BCEloss）计算最终的损失值。这种方式显著降低了输入维度，减少了模型参数量，提升了计算效率，同时适用于任意大小的图像输入。

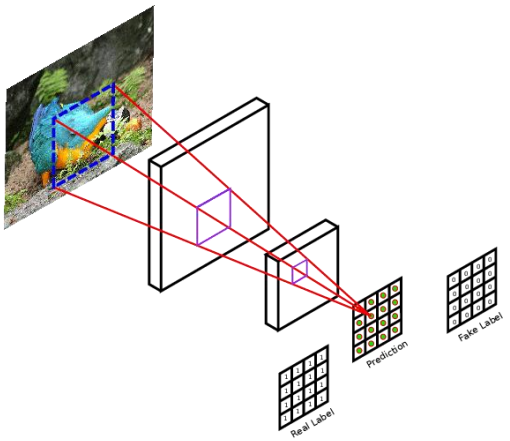


图 7 PatchGAN

3. 实验与结果分析

实验数据集：facades
实验平台：windows
基本的实验参数如下：

Epoch_num	100，200
Batch_size	1-16
Learning_rate	0.0002
Betas	（0.5，0.999）

（1）迭代次数对实验的影响（Batch_size=1, Learning_rate =0.0002）

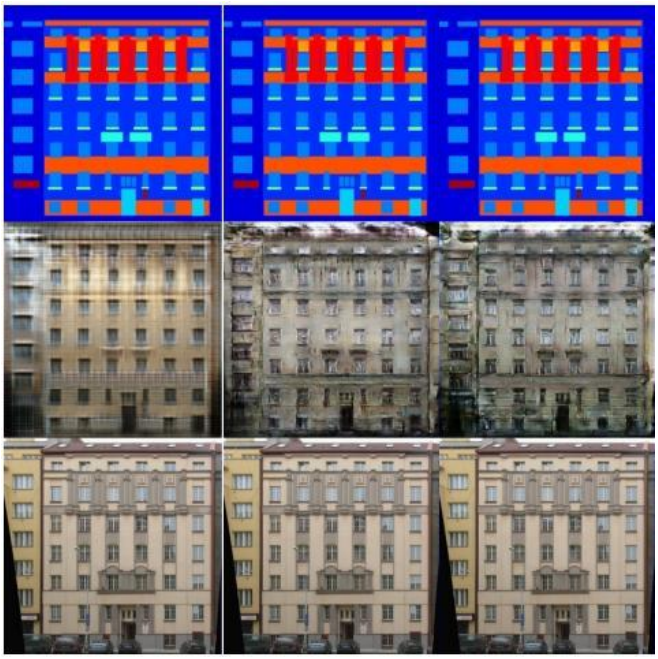


图 8

上面一排是输入图像，中间一排从左到右分别为 pix2pix 训练 10 轮、100 轮、200 轮生成的图像，下面一排是真实图像。

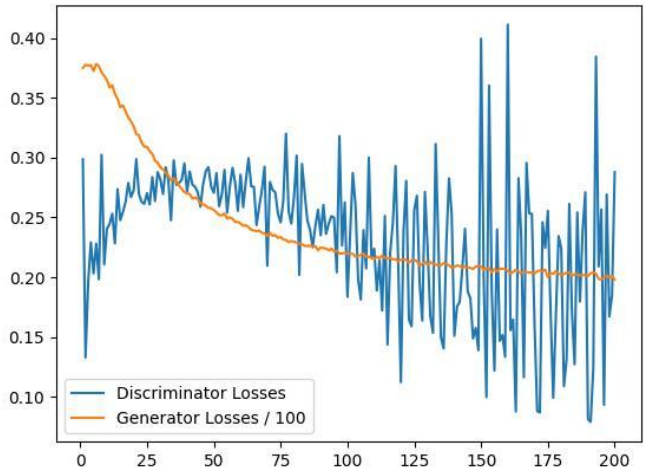


图 9 随着迭代次数的变化，D_loss 和 G_loss 的变化图

通过图 8 和图 9 可以看出，随着 Pix2Pix 模型迭代次数的增加，生成器的损失（G_loss）逐渐减小并趋于收敛，而判别器的损失（D_loss）则在 0.2 附近波动。观察生成的图像质量，迭代 10 次时效果较为模糊，而迭代 100 次和 200 次时图像质量有明显提升。

这一现象主要是由于 Pix2Pix 模型在训练初期尚未完全收敛，随着训练轮数的增加，模型对数据特征的学习逐步深入，从而生成的图像效果不断改善。

（2）Batch_size 对实验的影响（Epoch_num=200, Learning_rate =0.0002）

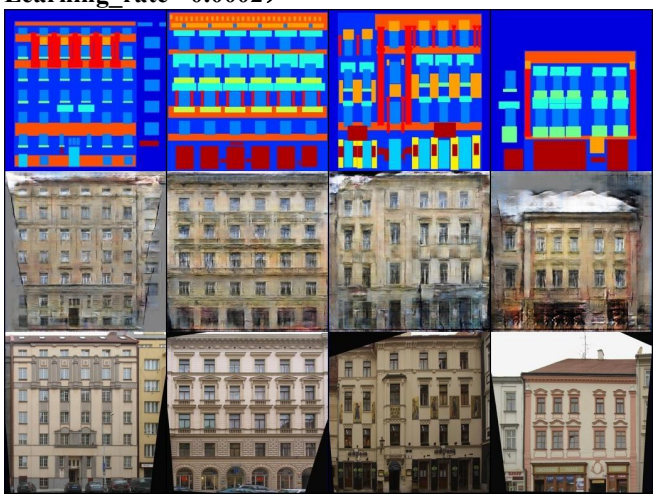


图 10 左边第一列 batch_size=1，右边四列为 batch_size=4

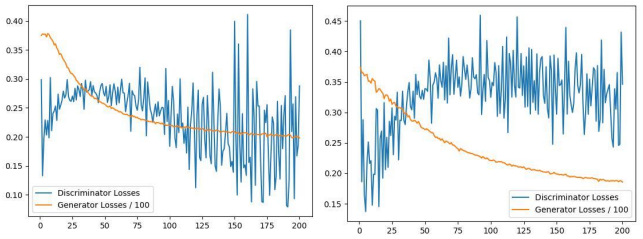


图 11 batch_size=1 batch_size=4

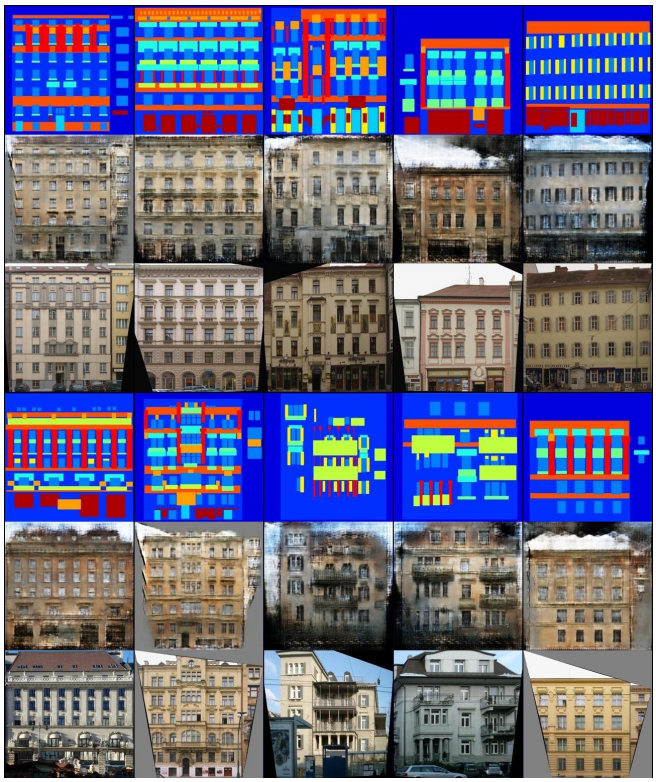


图 12 batch_size=16

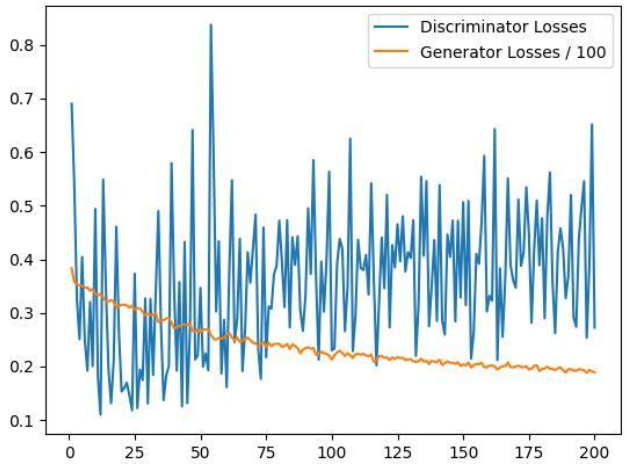


图 13 batch_size=16

这是 Pix2Pix 模型在不同批量大小（batch_size = 1、4、16）下训练的结果，展示的是第 200 轮迭代生成的图像。从结果可以看出，随着 batch_size 的增大，相同轮数训练所需的时间有所减少。然而，当 batch_size 较大时，生成的图像相对模糊；而在 batch_size 较小时，生成的图像更为清晰。

此外，生成器和判别器的损失（Loss）在 batch_size 较小时下降幅度更显著，模型的收敛效果也更优。

（3）PatchGAN 大小对实验的影响（Batch_size=1, Epoch_num=100, Learning_rate =0.0002）

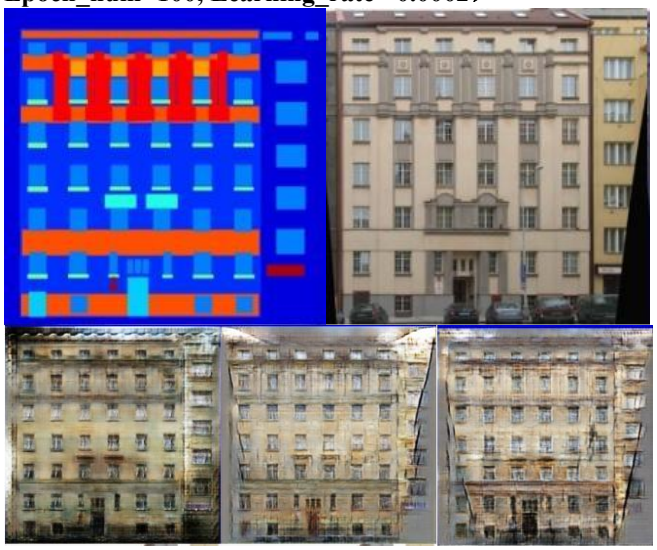
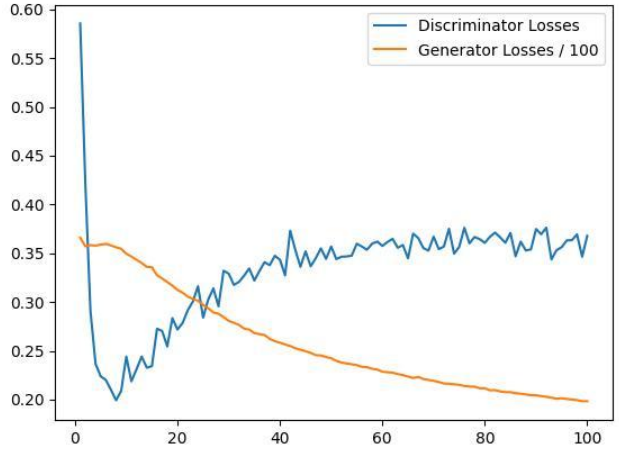
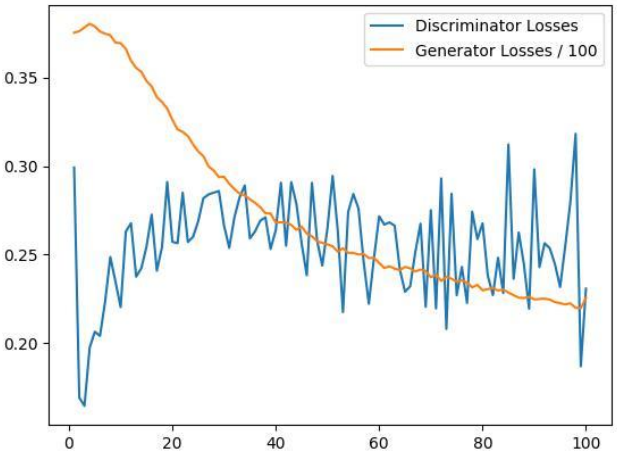


图 14

下面一排生成图像从左到右分别对应着 PatchGAN 大小为 16x16,70x70,286x286。





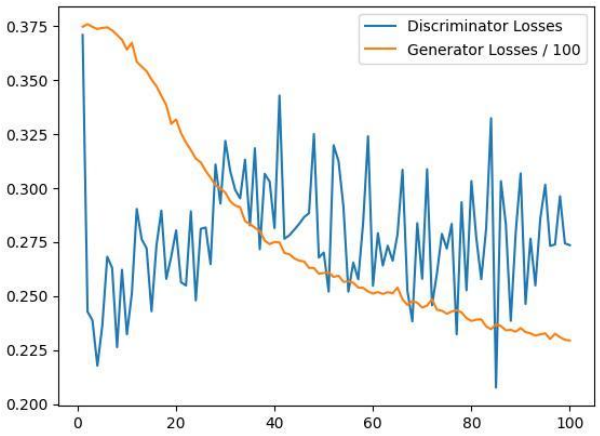


图 15

上中下分别对应 PatchGAN 大小为 16x16,70x70,286x286。

这是 Pix2Pix 模型在不同 PatchGAN 感受野大小（16x16、70x70、286x286）下的实验结果。尽管三张生成图像差异不大，肉眼难以区分（需要在 FCN 模型下评估准确率表现），这里主要通过生成器和判别器的损失（Loss）来分析感受野大小的选择。

从实验结果来看，三种感受野下，生成器的损失（G_loss）均表现良好，逐步下降。在判别器的损失（D_loss）方面，感受野大小为 70x70 时，D_loss 在 0.24 附近波动；感受野为 16x16 时，D_loss 甚至有增大的趋势；而感受野为 286x286 时，D_loss 约为 0.275。

因此，综合考虑各项指标，选择 70x70 作为 PatchGAN 的感受野大小是最优的。

（4）损失函数的消融实验（Batch_size=1, Epoch_num=100, Learning_rate =0.0002）

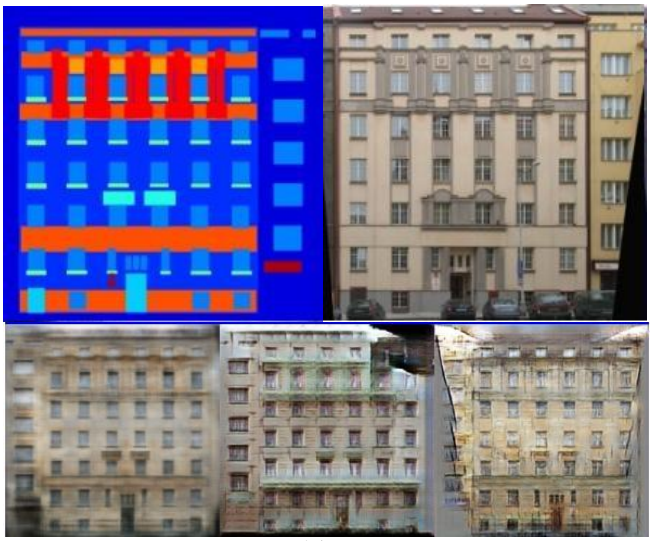


图 16

下面一排生成图像从左到右分别对应着 L1 损失, cGAN 损失,L1+cGAN 损失。

这是 Pix2Pix 模型在不同损失函数配置下（仅使用 L1 损失，仅使用 cGAN 损失，以及同时使用 L1 和 cGAN 损失）进行的消融实验。实验结果表明，单独使用 L1 损失时，生成的图像非常模糊，但大致的色块分布和结构信息已经被学习到。而仅使用 cGAN 损失时，图像中出现了大量高频信号（即颜色突变明显），导致图像整体锐化过度。

因此，结合 L1 损失和 cGAN 损失（L1 + cGAN）的方法相比单独使用 L1 / cGAN 损失，能够得到更为平衡和高质量的图像生成效果。

（5）归一化方法（改进）（BATCH_SIZE=4, EPOCH_NUM=200, LEARNING_RATE =0.0002）

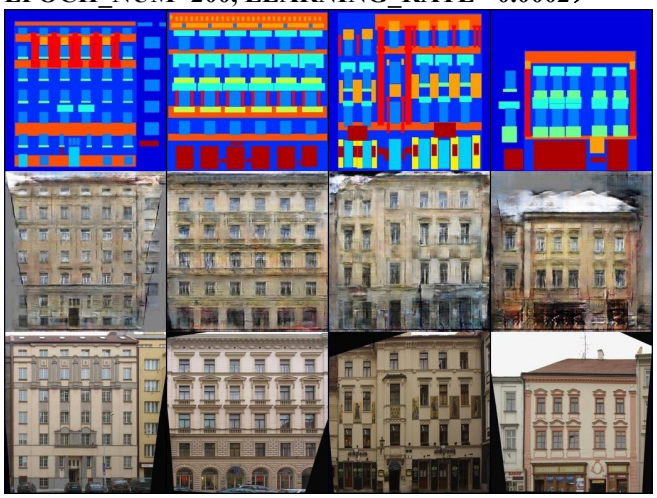


图 17 使用 BN 归一化

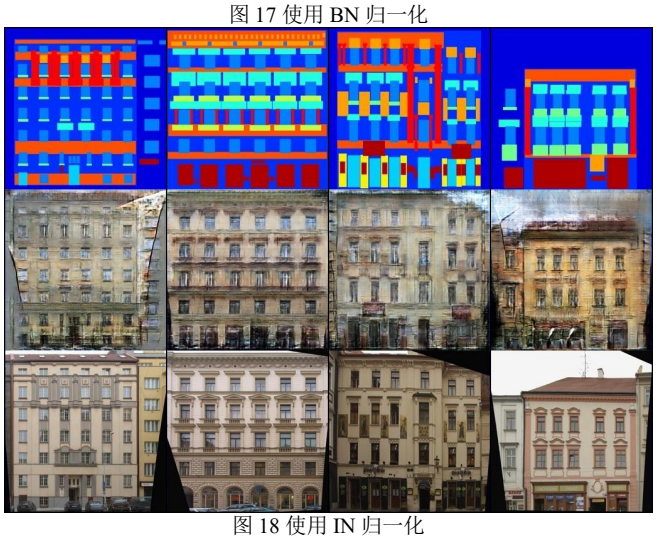


图 18 使用 IN 归一化

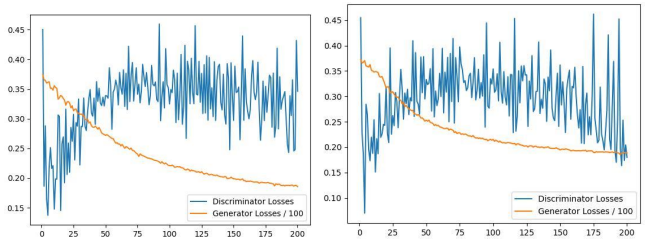


图 19 BN 归一化　　IN 归一化

由于在 batch_size=1 时，BN 归一化和 IN 归一化效果相似，因此我们在 batch_size=4 的条件下进行了实验。原论文中的 Pix2Pix 模型使用的是 BN 归一化，在此基础上，我们将 BN 归一化替换为 IN 归一化进行改进。经过 100 轮迭代，实验结果表明，使用 IN 归一化的 Pix2Pix 模型在收敛速度上明显优于使用 BN 归一化的模型。从生成器和判别器的损失曲线来看，IN 归一化的表现也更为优越。

在图像风格转换任务中，生成的风格结果主要依赖于特定的图像实例，因此对整个 batch_size 数据进行归一化并不理想。而 IN 归一化只对 HW 维度进行操作，这种方式更加合理。使用 IN 归一化不仅加速了 Pix2Pix 模型的收敛，而且能够保持每个图像实例之间的独立性，进而提高了模型的效果。

（6）加入注意力机制（改进）（BATCH_SIZE=4, EPOCH_NUM=100, LEARNING_RATE =0.0002）

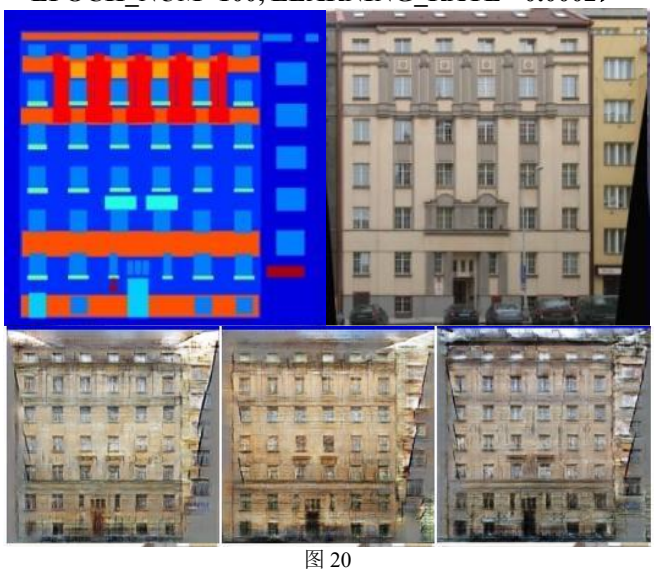
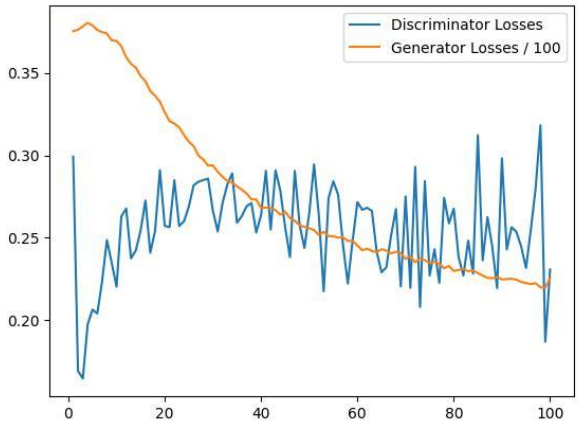


图 20

下面一排从左到右分别对应着没有加入注意力机制、下采样过程加入注意力机制、上采样过程加入注意力机制。



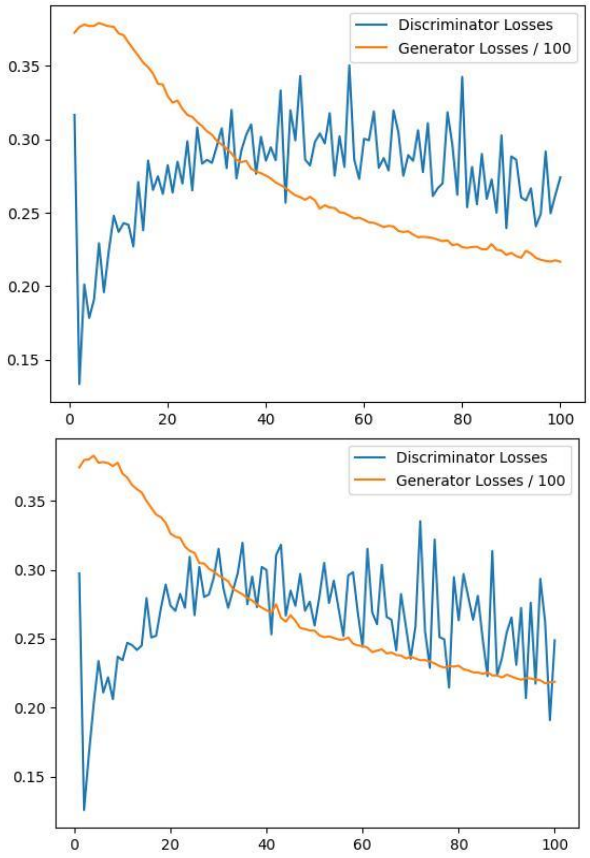


图 21

上中下分别对应着没有加入注意力机制、下采样过程加入注意力机制、上采样过程加入注意力机制。

这是 Pix2Pix 模型在三种不同设置下的实验结果：没有加入注意力机制、在下采样过程中加入注意力机制、以及在上采样过程中加入注意力机制。从理论上分析，下采样过程通常用于捕捉输入图像的全局特征，而引入空间注意力机制能够使生成器更加关注输入图像中的重要部分，从而获得更丰富的信息，有助于提升生成图像的质量和细节。

在上采样过程中加入注意力机制的优势在于，生成器能够更好地结合下采样过程中捕捉到的特征，并有选择性地强调对生成目标有贡献的部分。理论上，注意力机制应用于下采样和上采样过程都应具有正面效果。然而，从生成器和判别器的损失函数来看，加入注意力机制的下采样过程反而表现较差，而在上采样过程中加入注意力机制略微有所提升，尽管这种提升并不显著。

4. 结论

本次实验主要基于原论文实现了 Pix2Pix 深度生成模型，并在此基础上进行了调参和改进。通过一系列对

比实验和消融实验，验证了不同参数对模型性能的影响，并得出了若干结论。

首先，通过对比不同的训练轮数、batch_size 和 PatchGAN 感受野大小等超参数，实验结果表明，Pix2Pix 模型在训练轮数较多、batch_size 较小时生成的图像效果最好。对于输入大小为 256x256 的图像，选择 PatchGAN 感受野大小为 70x70 时效果最佳。此外，通过对损失函数进行消融实验，发现 L1+cGAN 损失函数比单独使用 L1 或 cGAN 损失的效果更好。

在归一化方法方面，我们对 Pix2Pix 模型进行了改进，采用了 IN 归一化替代 BN 归一化。实验结果显示，IN 归一化在图像转换任务中优于 BN 归一化，能够有效加速模型的收敛并提升性能。另外，我们还加入了注意力机制，实验表明，在上采样过程中加入注意力机制能对模型性能略有提升，虽然提升幅度不大。

然而，实验中仍存在一些不足。首先，模型生成图像的质量评估不够准确，传统的 per-pixel mean-squared error 无法有效评估结构性损失，导致难以准确评估生成图像的质量。尽管论文中提到了使用 AMT 平台和 FCN 模型进行语义分割评测，但由于各种限制，未能采用这些方法。此外，模型的性能评估主要依赖生成器和判别器的 Loss 指标，这也使得对图像质量的评估较为粗略。

总体而言，Pix2Pix 模型为图像翻译领域作出了重要贡献，巧妙地利用了 GAN 框架，为“Image-to-Image translation”类问题提供了一个通用框架。其核心技术包括基于条件 GAN 的损失函数、基于 U-Net 的生成器和基于 PatchGAN 的判别器。尽管 Pix2Pix 在许多图像翻译任务上取得了令人瞩目的效果，但模型需要大量的成对图像数据进行训练，并且当输入图像与训练集的偏差较大时，生成结果的质量可能会大打折扣。因此，如何提高模型在不同数据集和任务上的泛化能力，仍然是未来研究的一个重要方向。

5. 参考文献

[1] Isola, Phillip, et al. “Image-to-Image Translation with Conditional Adversarial Networks.” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017): 5967-5976.
[2] Ioffe, Sergey, and Christian Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.” ArXiv abs/1502.03167 (2015): n. pag.