2022

基于空间自适应归一化的街景语义图像合成

汇报人：田瑶

时间：2022.11.20

目录

CONTENTS

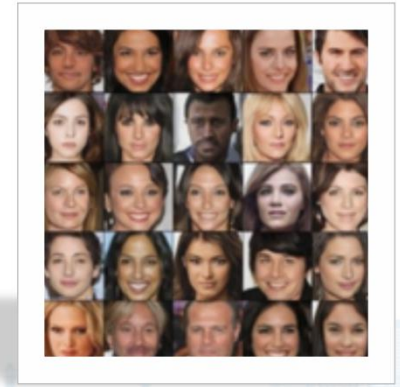
PART ONE

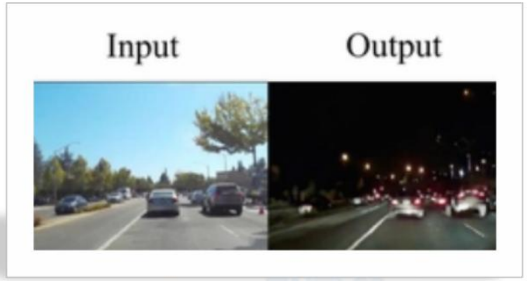
引言

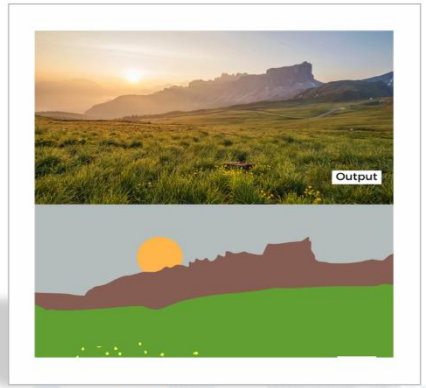
1 引言

研究背景

图像到图像的翻译是将图像从一个域转换到另一个域的过程，其目标是学习输入图像和输出图像之间的映射。包括从轮廓草图生成真实图像以及合成人像、图片风格转换等。







1 引言

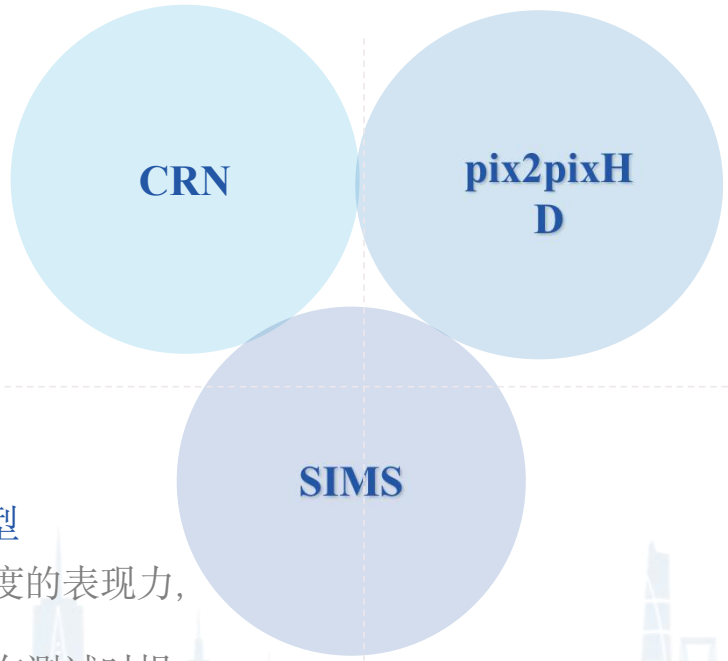
研究背景

本文研究的目标是从街景语义图像合成真实图像，这对训练自动驾驶汽车模型非常有用

语义图像合成

1 引言

相关研究



CRN—级联提炼网络

●提出了一种级联提炼网络在输入语义布局的条件下生成图片

●用带回归损失的损失函数来训练卷积神经网络

pix2pixHD

●生成器从U-Net升级为多级生成器。

●判别器升级为多尺度判别器。

●优化目标上增加了基于判别器的特征匹配损失。

●增加实例级别的信息。

SIMS—半参数模型

●参数模型的优点是具有高度的表现力，可进行端到端的训练。

●非参数模型的优点是可以在测试时提取大型的真实图片数据集里的素材。

●结合这两种方法的优势，提出了SIMS。

相关研究

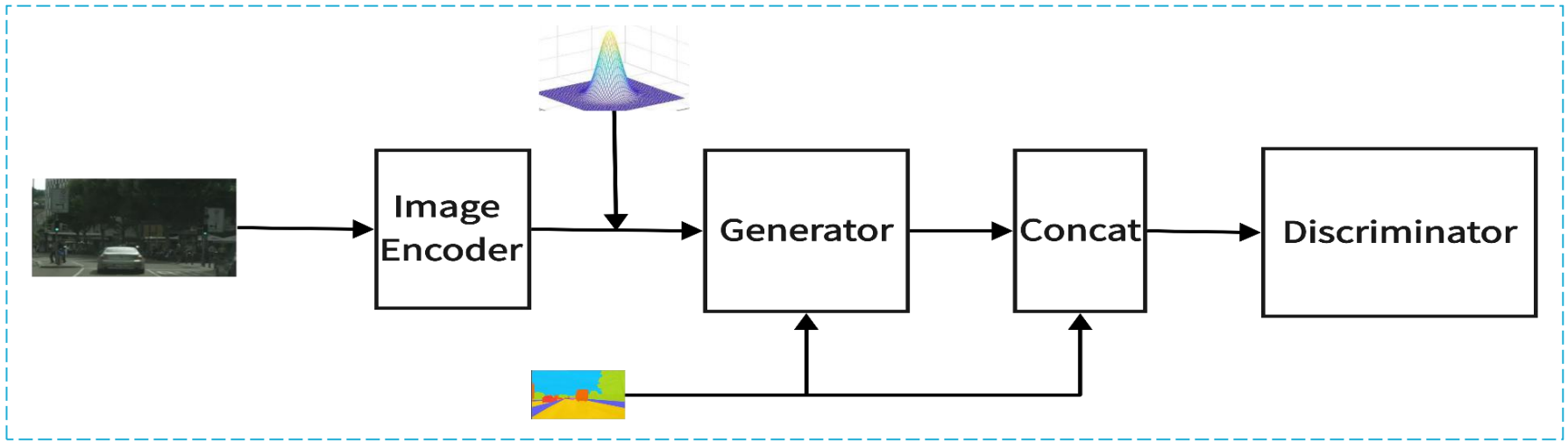
- 直接将语义布局加噪声作为网络输入
- 网络中的归一化层倾向于“洗去”语义信息

PART TWO

模型与方法

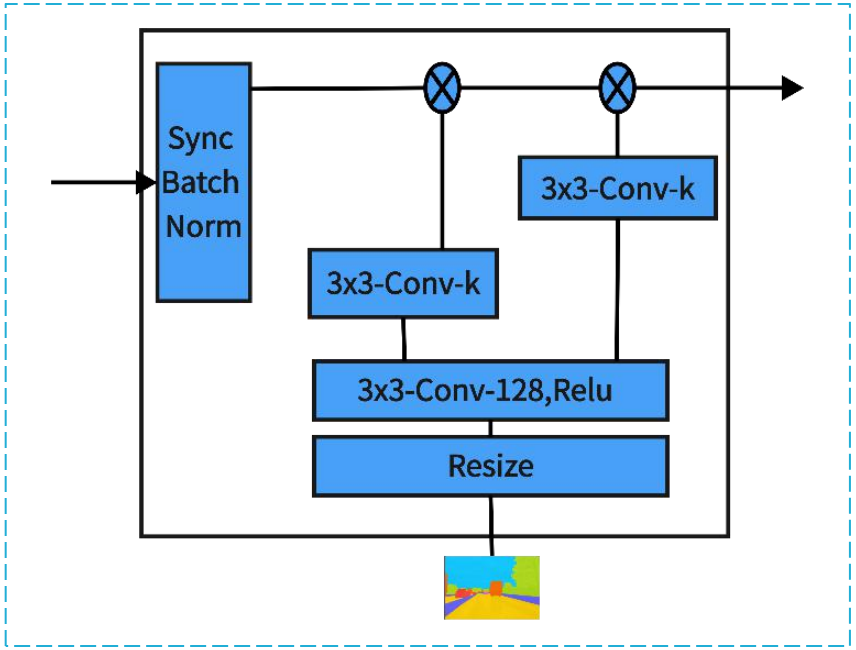
2 模型与方法

总体网络结构



整体的网络结构是以生成对抗网络（GAN）为基础框架

SPADE模块



由于语义图与特征图大小不一致，因此先将 语义图“resize”成与特征图大小一致。

经过一层卷积，relu激活，到一个中间层。

再有两个不同的卷积层，得到对应的 γ 和 β ，再分别作为缩放和偏置系数，作用到图片上。

SPADE模块的计算公式

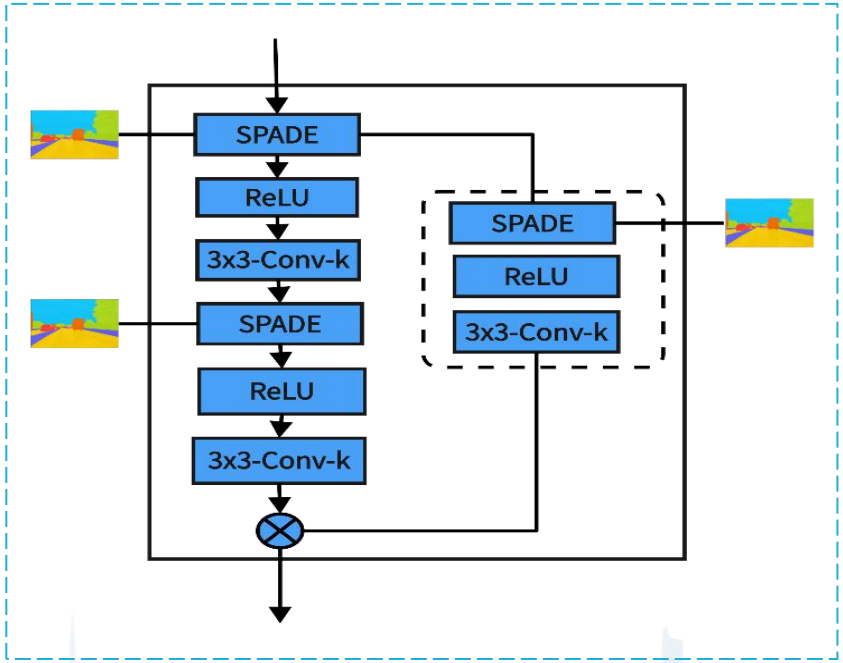
$$\gamma^{i}_{c,y,x}(m)\frac{h^{i}_{n,c,y,x}-\mu^{i}_{c}}{\sigma^{i}_{c}}+\beta^{i}_{c,y,x}(m)$$

SPADE模块是在BN的基础上做了修改，修改内容就在于 γ 和 β 计算的不同。在BN中 γ 和 β 的计算是通过网络训练得到的，而SPADE中 γ 和 β 是通过语义图像计算得到的，计算公式如下：

在BN中， γ 和 β 是一维的，其中每个值对应输入特征图的每个通道，而在SPADE当中， γ 和 β 是三维的，除了通道维度外，还有宽和高两个维度，因此公式中 γ 和 β 下标包含c,y,x三个符号

2 模型与方法

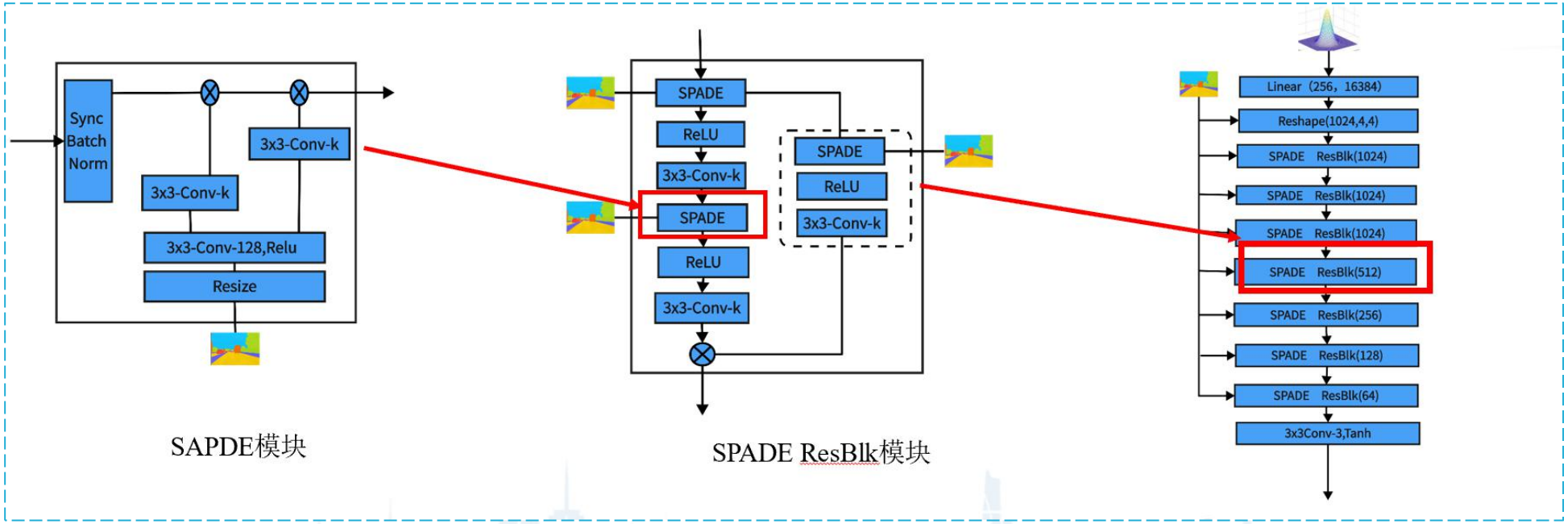
SPADE ResBlk模块



SPADE ResBlk模块中，将常用的“卷积→激活→归一化”模块替换成了“SPADE→激活→卷积”，将顺序颠倒之外，使得SPADE模块可以利用语义图信息来指导归一化

2 模型与方法

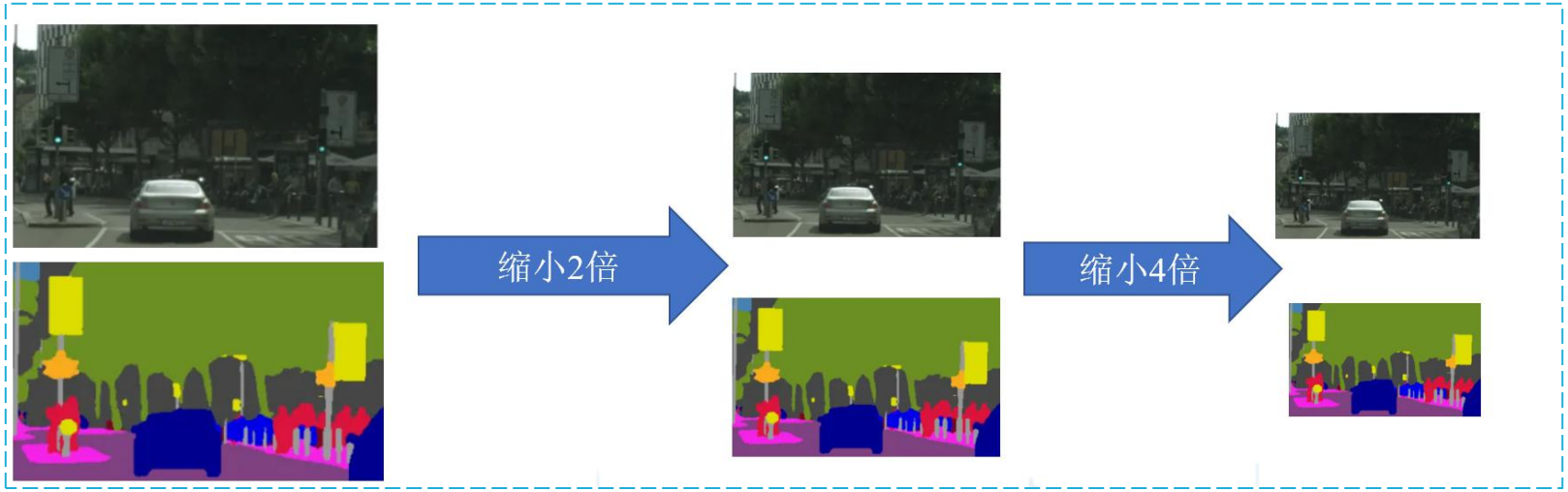
生成器



生成器采用堆叠多个SPADE ResBlk实现

2 模型与方法

判别器

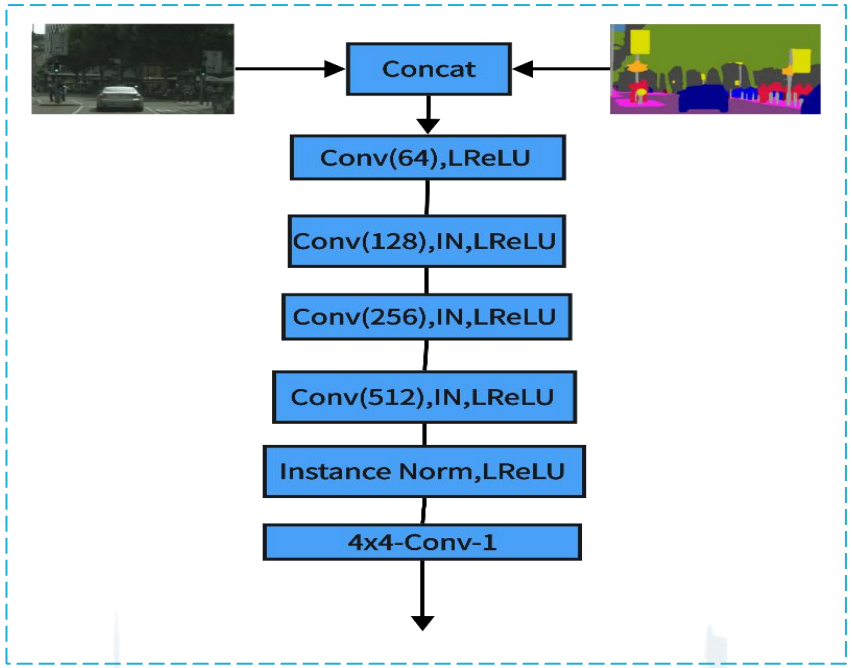


判别器采用多尺度判别器:

- 先基于真实图像和合成图像构建三层图像，每层的尺寸大小不一样（像一个金字塔的形状）；
- 针对每层图像分别训练一个判别器进行判别，三个判别器的结构是一样的，只不过处理的输入尺寸不一样;
- 这一部分的设计是希望判别器促进图像整体和细节方面的合成。

2 模型与方法

判别器



三个判别器的结构都是一样的，都是一个卷积网络，对图像的一个图像块进行卷积，卷积得出的结果就是该图像块属于真实图像的概率，然后将整张图像的卷积结果做平均，得出最终的判断。

网络训练优化目标

生成器和判别器的博弈目标式：

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

- 判别器D1，D2，D3的优化目标是最大化辨别出图片是合成图片的概率；
- 生成器G的目标是最小化合成图片和真实图片之间的差距；
- 引入特征匹配损失L_{FM}为正则项，增强判别器的判别效果，使得生成图像和真实图像在不同的网络层都具有类似的特征，优化图像合成效果；

PART THREE

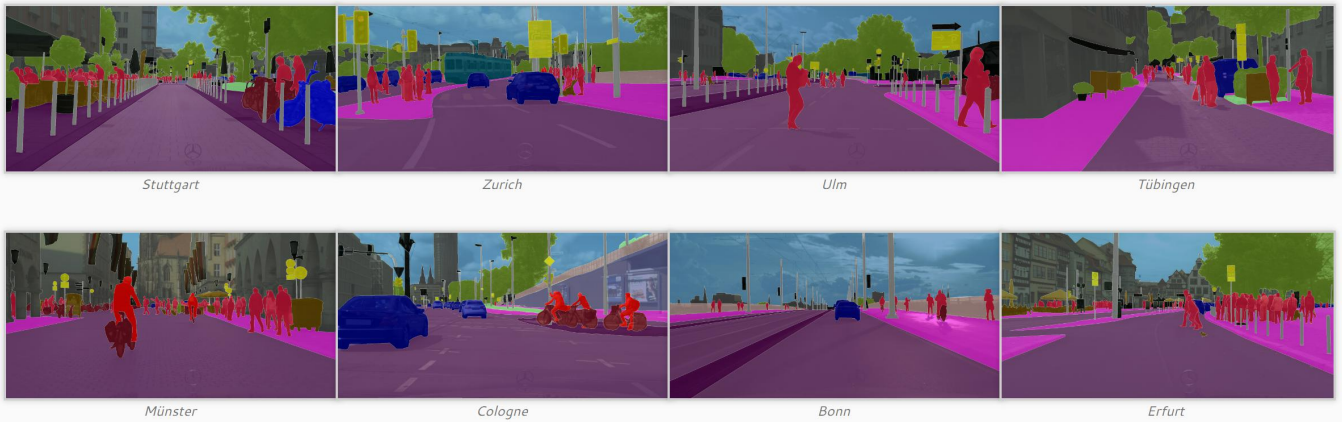
实验结果与分析

3 实验结果与分析

数据集介绍

cityScapes数据集：

Below are examples of our high quality dense pixel annotations that we provide for a volume of 5 000 images. Overlayed colors encode semantic classes (see class definitions). Note that single instances of traffic participants are annotated individually.



以城市街道场景的语义图片为基础的数据集，研究使用5000张驾驶场景的高质量像素级注释图像，其中的2975张图用于训练模型，500张用作验证集（验证集的作用就是调整模型的超参数且初步评估模型的能力），剩下的1525张图片用作测试集，用来评估最终模型的泛化能力。

3 实验结果与分析

实验结果

通过在cityScapes数据集上实验，得到的部分实验结果展示

3 实验结果与分析

实验结果量化分析

使用均交并比和像素准确率来衡量本文网络模型的性能，结果如下表所示

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}}$$

方法	平均交并比
CRN	52.4
SIMS	47.2
pix2pixHD	58.3
本文方法	**62.3**

方法	像素准确率
CRN	77.1
SIMS	75.5
pix2pixHD	81.4
本文方法	**81.9**

PART FOUR

结论

4 结论

研究主要基于现有的语义图像合成方法，提出了一种针对街景语义图像合成的改进空间自适应归一化方法。

实验结果表明，研究获得的街景语义合成图像均交并比为62.3，像素准确率为81.9，均优于其他常见的方法，表明本文研究的网络模型，能够在像素级别和感知效果上都更接近真实的图片。

4 展望

在数据集方面，使用的是cityScapes数据集，比较单一，未来将实现在更多的图片集上进行实验并分析实验结果改进模型，提高模型的泛化能力。

在语义图像合成方面，本文实现的是一对一的转换，未来的研究将致力于进行多模块的转换，应用到单模块无法解决的场景，增强模型的适用性，拓宽其使用的范围。

感谢观看

汇报人：田瑶

时间：2022.11.20