大连理工大学本科毕业设计（论文）

基于深度学习的虚拟驾驶环境生成

Virtual Driving Environment Generation Based on Deep Learning

学 院（系）：汽车工程学院

专　　业：车辆工程

学 生 姓 名：叶乾

学　　号：201564010

指 导 教 师：祝雪峰

评 阅 教 师：徐金亭

完 成 日 期：2019 年 6 月 17

大连理工大学

Dalian University of Technology

原创性声明

本人郑重声明：本人所呈交的毕业设计（论文），是在指导老师的指导下独立进行研究所取得的成果。毕业设计（论文）中凡引用他人已经发表或未发表的成果、数据、观点等，均已明确注明出处。除文中已经注明引用的内容外，不包含任何其他个人或集体已经发表或撰写过的科研成果。对本文的研究成果做出重要贡献的个人和集体，均已在文中以明确方式标明。

本声明的法律责任由本人承担。

作者签名：　　日　期：

关于使用授权的声明

本人在指导老师指导下所完成的毕业设计（论文）及相关的资料（包括图纸、试验记录、原始数据、实物照片、图片、录音带、设计手稿等），知识产权归属大连理工大学。本人完全了解大连理工大学有关保存、使用毕业设计（论文）的规定，本人授权大连理工大学可以将本毕业设计（论文）的全部或部分内容编入有关数据库进行检索，可以采用任何复制手段保存和汇编本毕业设计（论文）。如果发表相关成果，一定征得指导教师同意，且第一署名单位为大连理工大学。本人离校后使用毕业毕业设计（论文）或与该论文直接相关的学术论文或成果时，第一署名单位仍然为大连理工大学。

论文作者签名：　　日　期：

指导老师签名：　　日　期：

摘　　要

近年来，自动驾驶成为汽车行业研究热点之一。研究指出完全自主的车辆必须行驶数亿英里以证明其在安全性的可靠性，另外强化学习自动驾驶模型需要足够的路面行驶来迭代训练。于法规和时间上的问题，目前很多自动驾驶相关研究机构和厂商将自动驾驶代理模型置于计算机模拟的虚拟驾驶环境中进行训练或测试。

现有的数据驱动法、神经网络生成法能够生成较为逼真的驾驶环境图像，然而无法控制驾驶环境模态（如大气条件、光照条件）。为解决这一问题，本文基于多模态图像生成框架，根据语义布局生成驾驶环境图像并可以构建各种光照下的真实图片。具体工作如下：

1.本文引入部分共享隐空间的概念实现驾驶环境模态控制。图像表示可以分解为关于域不变的内容隐码和捕代表了特定属性的样式隐码。随机样式隐码允许一对多的映射，学习目标域中学习相应图像的条件分布。

2. 本文实现了有语义布局生成街景图像并测试了固定样式编码视频帧连续性的贡献。数值算例表明，本文的方法生成驾驶环境具有高仿真度，固定样式码能提高视频时间连续性。

3.基于多模态方法，实现行车图像光照控制。本文将 Comma2K19 行车数据集分为日间和夜晚两个域，训练了该网络。日间行车图像可转换为夜间图像，也可反向转换，并且能够在一个域中控制光照条件。

综上，本文研究了高仿真度虚拟驾驶环境模态控方法，并以光照条件控制为例进行实验。这对于快速构建自动驾驶虚拟仿真平台具有重要意义，具有重要的研究意义和应用价值。

关键词：虚拟驾驶环境；深度学习；多模态

Virtual Driving Environment Generation Based on Deep Learning

Abstract

In recent years, automatic driving has become one of the hot topics in the automotive industry. It is pointed out that fully autonomous vehicles must travel hundreds of millions of miles to prove their reliability in safety. In addition, reinforcement learning of autopilot models requires sufficient road driving for iterative training. At present, many research institutes and manufacturer of automatic driving put the model of automatic driving agent into the virtual driving environment of computer simulation for training or testing.

Existing data-driven and neural network generation methods can generate more realistic driving environment images, but they can not control driving environment modes (such as atmospheric conditions, lighting conditions). To solve this problem, based on the multi-modal image generation framework, this paper generates driving environment images according to the semantic layout and can construct real images under various illuminations. The specific work is as follows:

1. This paper introduces the concept of partially shared hidden space to realize the modal control of driving environment. Image representation can be decomposed into content-invariant implicit codes with respect to domain and stylistic implicit codes with specific attributes represented by capture. Random Style Hidden Code allows one-to-many mapping, learning the conditional distribution of the corresponding image in the target domain.

2. This paper realizes the generation of street scene images with semantic layout and tests the contribution of frame continuity of fixed-style coded video. The numerical examples show that the method in this paper can generate driving environment with high simulation degree, and the fixed style code can improve the video time continuity.

3. Based on the multi-modal method, the illumination control of driving image is realized. In this paper, Comma2K19 traffic data set is divided into two domains, day and night, and the network is trained. The daytime driving image can be converted into night image or reverse conversion, and the illumination condition can be controlled in a field.

To sum up, this paper studies the modal control method of high simulation virtual driving environment, and takes illumination condition control as an example to carry out experiments.

This is of great significance to the rapid construction of the virtual simulation platform for automatic driving, and has important research significance and application value.

Key Words：Virtual Driving Environment；Deep Learning；Multimodal

目　　录

1 文献综述

1.1 课题背景及意义

近年来，自动驾驶汽车由于其在机动性和安全方面的潜在的巨大社会效益受到了研究人员、风险投资家以及公众的广泛关注。然而，开发高标准的自动驾驶车辆需要大量的行驶测试。一方面有研究指出完全自主的车辆必须行驶数亿英里以证明其在安全性的可靠性。如果需要在自动车辆投放市场前充分证明其性能，那么所需的测试里程将花费研发团队数十年甚至数百年的时间来上路行驶[1]。另一方面自动驾驶汽车的环境感知与控制采用了大量的机器学习方法，如强化学习。一些机器学习模型需要足够的路面行驶来迭代训练。而真实车辆在路面上从时间角度和安全角度都无法完成如此多的试错。所以目前主流解决方案是使用虚拟驾驶模拟器替代路测来进行代理（Agent）的仿真自动驾驶研究。

虚拟驾驶环境需满足两个要求。首先，从环境感知、导航和控制方面测试和验证自动驾驶车辆的的能力。第二是生成大量标记的训练数据，以训练机器学习方法尤其是计算机视觉方面，例如，深层神经网络。

1.2 研究现状

1.2.1 人工建模法

人工建模法是生成虚拟驾驶环境最常见的方法。人工建模法是使用计算机图形、基于物理的建模和机器人运动规划技术的组合来创建一个合成环境，在该环境中可以对移动车辆进行动画和渲染。近来开发出许多模拟器，如英特尔的CARLA[2]、微软的Airsim[3]、谷歌的Carcraft[4]等。CARLA是一种开放式的虚拟城市驾驶模拟器。其虚拟城市环境是由一个专门的数字艺术家团队从零开始人工创建的，包括城市布局、车辆模型、建筑物、行人、街道标志等。除视觉信号外，仿真平台还提供了其他可用于训练驾驶策略的信号，例如GPS坐标、速度、加速度以及有关碰撞和其他相关详细数据。虚拟环境条件可以被指定，包括天气和一天中的时间。Airsim同时提供了的物理和视觉模拟。其能为车辆暴露在物理场中的模拟提供更丰富的支持，包括重力、空气密度、气压和磁场。

虽然所有这些模拟器都运用了了极先进的合成渲染成果，但这些方法仍不足以训练部署于现实世界中的自动驾驶车辆。一个主要障碍是缺少高保真的环境模型。创建与实物高度相似的CG模型成本是非常高的。因此，这些模拟器的合成图像具有独特的CG渲染外观和感觉，即游戏或虚拟现实系统的观感。此外，这些模拟器只有限种类的汽车、行人、建筑物、植被等模型，与现实驾驶环境相差甚远（如图1.1）。

图1.1　CARLA环境截图

1.2.2　数据驱动方法

为解决上述仿真度不高的问题，百度提出了通过摄像头、激光雷达等传感器扫描大量的现实街景实现数据驱动的虚拟驾驶环境生成系统AADS[5]。AADS通过模拟交通流增强真实图像，以创建类似真实照片的渲染场景。该研究使用激光雷达和摄像头扫描街景并将输入数据分解为背景、场景语义和前景对象。

AADS 在数据集采集时使用了一套硬件系统，该系统由两个 Riegl 激光扫描仪、一个实时行扫描激光雷达（Velodyne 64 行）、一个 VMX-CS6 立体摄像机系统和一个高精度 IMU/GNSS 组成。其中 Riegl 扫描仪，使得该系统可以获得比广泛使用的普通激光雷达扫描仪更高密度的点云。而 VMX-CS6 摄像系统配备了高分辨率的宽基线立体摄像机（3384×2710）。数据集中运动物体的形状和位置是利用测速激光雷达得到的。在扫描街道场景前,硬件首先被校准、同步，然后安装在一辆中型 SUV 的顶部。在扫描一个场景的过程中，该 SUV 以平均每小时 30 公里的速度在目标场景周围巡航，每米拍摄一次 RGB 图像。

在给原始数据标注的过程中，与完全手工标注所有二维/RGB 和三维/点云数据的传统方法不同，AADS 开法了一个新的标注流程大幅提高了标注的效率和准确度。因为二维标签在时间和劳动力方面的消耗都很够，所以 AADS 整合了三维标签和二维标签两个阶段的工作。通过使用易于标记的三维标签，AADS 通过三维-二维投影自动生成所有图像帧中静态背景以及对象的高质量二维标注。

对于每个图像帧，在两个 3D 点云数据中标注由五个组覆盖的 25 个不同类。除了汽车、摩托车、交通标识标等标准注释类外，AADS 增加了一个新的“三轮车”类——东亚国家特有的流行交通方式。该数据集还注释了当前在开放数据集中所没有的 35 个不同

车道标记。这些车道标志是根据颜色（如白色和黄色）、类型（如实线和虚线）和用途（如分隔、引导和停车）来分类定义的。

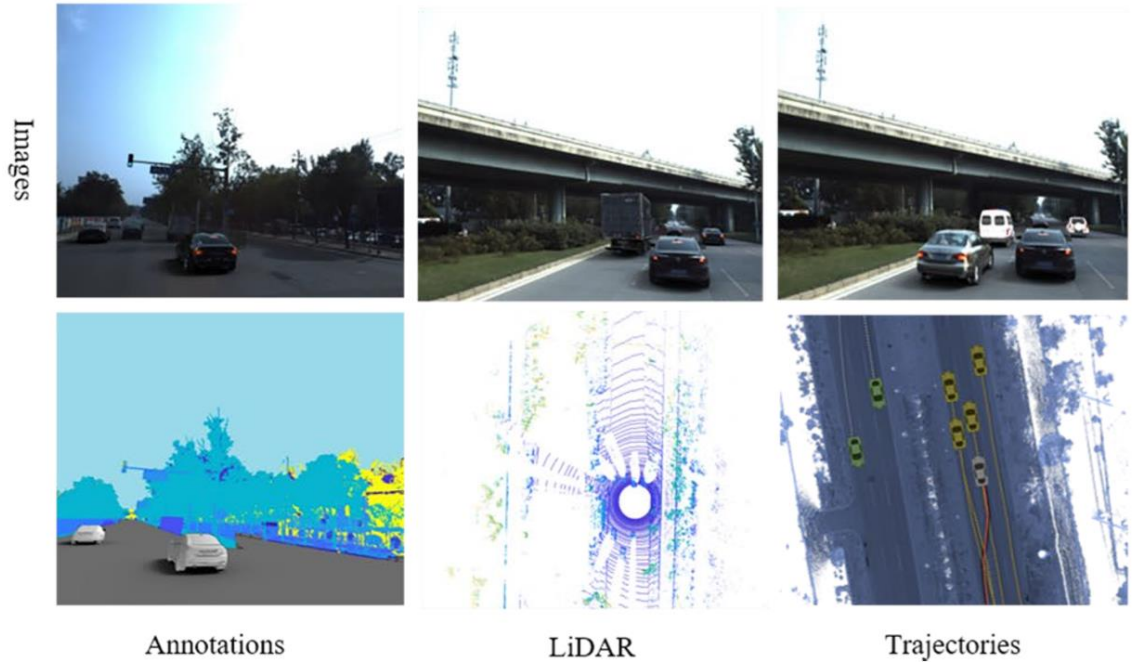


图1.2　AADS环境示意图

AADS使用一种新的视点合成技术，可以在静态背景下改变视点，从而实现环境背景的3D重建。通过精确的室外场景语义解析，可以重新定位3D车辆模型、计算机生成的行人和其他可移动对象，并将其渲染回背景图像中，以创建逼真的街道视图图像。此外，模拟交通流是基于捕捉到的真实世界车辆轨迹，能够还原真实世界场景的复杂性和多样性（如图1.2）。然而，这类方法存对大气条件和光照条件缺乏灵活性。上述条件的改变需要采集目标状况下的数据来完全重新构建目标状况下的虚拟驾驶环境。现实中大气与光照条件可以有多种组合，并且在一些极端气候下数据采集车辆无法正常工作，等待现实自然条件改变而去采集数据从时间上和可操作性上是不可行的。

1.2.3　神经网络合成法

解决CG仿真度不高的另一方式是通过神经网络合成逼真的图像。目前最突出的图像合成方法是基于生成对抗网络（GAN）[6]。GAN于2014年被提出作为替代端到端有监督生成模型的框架，用来规避有些情况下近似复杂概率的计算的困难。在Goodfellow等人的原创工作中[6]，GAN用于合成MNIST数字和32个图像，旨在重现CIFAR-10数据集中不同类别的外观。 WU的工作探讨了vae-gan[8]在生成三维体素模型中的应用，而Wang的工作提出了一个级联GAN[9]，通过结构和样式生成自然图像。Pan研究了将CG模拟器图像翻译为真实图像的方法[10]。由于这两种类型的图像都表达了驾驶场景，该研究通过场景语义分

析配合现实翻译网络来转换它们。现实翻译网络由两个基于CGAN[11]的图像翻译网络组成，第一个图像翻译网络将虚拟图像转化为图像的语义分割。第二个图像翻译网络将分割后图像转化为现实世界中的对应图像。即首先，将模拟器（环境）渲染的虚拟图像分割成场景解析的表现形式，然后通过图像翻译网络（VISRI）[10]将其翻译为合成的真实图像。有研究采用了更为传统的方式，开发了由像素语义布局生成对应图像的合成模型[12]。该模型通过单个级联卷积神经网络的直接端到端监督训练生成图像，训练过程比对抗网络稳定。多样化的损失被使用以一次性生成多样化的图像集合。但是这些图像生成方法仍然存在大气与光照条件不可控的问题，无法充分覆盖自动驾驶车辆的测试环境条件。

1.2.4 图像样式转换

近来发展的图像样式转换技术为实现虚拟驾驶环境大气与光照条件的可控提供了思路。

（1） 图像到图像翻译

图像到图像转换的思想至少可以追溯到Hertzmann等人的图像类比研究[13]。该研究在一个输入输出训练图像对上使用非参数纹理模型[14]。进来有方法使用一组输入输出示例来学习使用CNN的参数化翻译函数[15]。还有的方法建立在Isola等人的Pix2Pix框架之上,使用条件生成对抗网络学习从输入到输出图像的映射[11,16,17]。这类算法可被应用于各种任务，例如从草图或属性和语义布局生成照片[16,18]。

（2） 非成对图像到图像翻译

不成对的图像到图像转换还有其他几种方法也解决了无成对标签的文图，目标是将两个数据域关联起来：x和y。罗塞尔等提出了一个贝叶斯框架，该框架包括一个先验框架，该先验框架基于从源图像计算的基于补丁的马尔可夫随机场和从多样式图像获得的似然项[20]。最近，CoGAN和跨模式场景网络使用权重共享策略来学习跨域的通用表示[21]。Liu等人[22]将上述框架扩展为变分自动编码器[23]和生成对抗网络[6]的组合。另一类并行工作鼓励输入和输出共享特定的“内容”特性，即使它们在“样式”上可能有所不同[24,25,26]。这些方法还使用对抗性网络，附加条件去强制输出接近预定义度量空间中的输入，例如类标签空间[26]、图像像素空间[24]和图像特征空间[25]。L. Karacan对视觉和图形任务的通用解决方案进行了研究[19]。该研究提出的函数不依赖于输入和输出之间任何特定于任务的预定义相似函数，也不假定输入和输出必须位于相同的低维嵌入空间。

（3） 循环一致性

使用可传递性作为规范结构化数据的方法的思想由来已久。在视觉跟踪中，执行简单的前向后向一致性几十年来一直是一个标准技巧[27，28]。在语言领域，通过“反向翻译和协调”来验证和改进翻译是人类译者以及机器共同使用的一种技术[29]。最近，高阶循环一致性被用于运动[30]、三维形状匹配[31]、密集语义对齐[32,33，34]和深度估计[35]的网络架

构。其中，Zhou等人[35]和Godard等人[34]与使用周期一致性损失作为使用传递性监督CNN培训的一种方式。Yi等人[36]在机器翻译的双重学习的启发下，独立地使用类似的目标进行不成对的图像到图像的翻译[37]。

（4） 神经风格转换

神经风格转换是执行图像到图像翻译的另一种方式，它通过匹配预先训练的深层特征的Gram矩阵统计，将一个图像的内容与另一个图像（通常是一幅绘画）的风格结合，合成一个新图像[38,39,40]。最近的研究通过尝试捕获高级外观结构之间的对应关系来学习两个图像集合之间的映射，而不是两个特定图像之间的映射[19]。因此，可以应用于单样本翻译方法不能很好地应用的其他任务，如绘画→照片，物体变形等。

1.2.5 小结

目前虚拟驾驶环境的搭建方法主要有三类：人工建模法、数据驱动法、神经网络合成法。人工建模法基于计算机图形学、物理规律和机器人运动规划技术的人工对驾驶环境建模。该方法可自由对光照和各物理场进行调控，但是存在图像仿真度不高、物体样式有限的问题。数据驱动法使用各类传感器，包括摄像机、激光雷达，对实景进行扫描从而自动构建虚拟驾驶环境。由于环境背景布局、图像直接还原自实景，该方法图像仿真度极高，但是也因此存在灵活度不足的问题，无法改变光照和大气条件。另外该方法所需的取景步骤也带来极大的工作量。神经网络合成法是在已有场景语义布局的基础上将其转换为逼真图像，该方法仿真度高，但同样存在灵活度的问题。近来发展的深度学习图像样式转换技术为实现虚拟驾驶环境大气与光照条件的可控提供了思路。

1.3 本文主要研究内容

本文首先对虚拟驾驶环境的需求做出了分析，介绍现有虚拟驾驶环境的一些生成方法。接着对本文深度学习模型所用到的理论基础进行叙述，包括生成对抗网络、残差网络、实例标准化、自适应实例标准化。本文以从语义布局生成逼真图像为例，研究了基于深度学习的虚拟驾驶环境图像生成方法；以日间行车图像到不同时刻（不同光照条件）下夜晚行车图像转换为例研究了虚拟驾驶环境图像的模态控制方法。提出了部分共享的隐空间概念，将隐空间分为内容空间与样式空间。仅需改变样式隐码即可在不改变图像内容的情况下改变其模态。另外本文还研究了图像重建损失、内容隐码重建损失和样式隐码重建损失对于生成图像质量与多样性的影响。

2 理论基础

2.1 生成对抗网络

生成对抗网络最初是基于博弈论的思想为了实现无监督的图像生成而设计的[6]。深度学习的意义在于挖掘出多样的层次模型，这些模型反映了各种数据概率分布间的复杂非线性映射，如不同语言、图像间的映射。模式分类模型是深度学习最早发展出的典型应用。模式分类通常是那些将高维、不同来源的信息映射到离散的类别标签上。这些模型的分段线性单元具有良好可求导性，所以能使用 BP 算法对网络权重优化，随机失活算法避免过拟合。然而深度生成模型一方面难以近似各种大量相互关联的概率映射；另一方面在生成环境中难以利用分段线性单元的优势。GAN 的提出能够克服这些困难。

在生成对抗网络中，生成模型与敌对模型相抗衡。敌对模型一种能够学习并确定样本是来自模型分布还是数据分布的辨别模型。生成模型的目标是学习照目标数据的分布，制造符合分布的假数据，并且不为辨别模型所检测出。而辨别模型的目标则是分辨目标数据与假数据。在训练过程中，两个模型都在不断改进自身，直到假数据与目标数据无法区分。

生成对抗框架最易应用于由多层感知器构成的的模型。使用多层感知器实现的生成对抗网络能够由学习样本的概率分布并完成随机多维噪声分布到样本的映射。另外此时可以只使用已经非常成熟的 BP 和随机失活算法来优化这两个模型的权重。在使用模型来生成样本时，也只需进行正向传播而不需要马尔可夫链或近似推理。

设有一个先验输入噪声变量 $p_z(z)$，$p_z(z)$ 到目标数据的映射表示为 $G(z;\theta_g)$，其中 G 是一个可微函数，具体形式是参数为 θ_g 的多层感知器，即生成器。生成器输出数据的概率分布为 p_g。生成对抗模型还包括输出为独热向量的辨别器 $D(x;\theta_d)$，其同样由多层感知器构成。$D(x)$ 表示辨别器认为的 x 是来自目标数据的概率。需要优化 D 以最小化 $D(x)$ 与真实标签的误差，即使其能分辨来自目标数据与来自 p_g 的 x。同时优化 G 以最小化 $\log(1-D(G(z)))$：

D 和 G 在进行可有下式表示的为了价值函数 $V(G,D)$ 的双玩家博弈：

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}\left[\log D(x)\right] + E_{z\sim p_z(z)}\left[\log(1-D(G(z)))\right] \tag{2.1}$$

其中 p_{data} 表示目标数据的分布。

由于 D 带来的梯度，在更新 G 的全之后，$G(z)$ 将有更高的概率被 D 识别为来自目标数据。当生成对抗网络接近收敛时，p_g 分布接近与 p_{data} 分布。D 也会成为精确的分类器，收敛于 $D^*(x)=\dfrac{p_{data}(x)}{p_{data}(x)+p_g(x)}$。

2.2 残差网络

2.2.1 残差学习

考虑 $H(x)$ 作为一个底层映射，由几个堆叠的层(不一定是整个网络)构成，x 表示这些层的第一个输入。如果假设多个非线性层可以渐进逼近复杂函数，则等价于假设它们可以渐进逼近残差函数，即 $H(x)-x$ (假设输入和输出的维数相同)。因此，不期望堆叠层近似于 $H(x)$，而是显式地让这些层近似于残差函数 $F(x)=H(x)-x$，因此原始函数变为 $F(x)+x$, 这便是残差学习[41]。虽然这两种形式都应该能够渐进地逼近所需的函数，但残差学习会有更低的学习的难易程度。

这种重新制定的学习方式的动机是关于退化问题的反直觉现象。如果可以将添加的层构造为标识映射，那么较深的模型应该不会比较浅的模型有更大的训练误差。退化问题表明，求解器可能难以用多个非线性层逼近标识映射。在残差学习模型中，如果标识映射已经是最优的，求解器可以简单地使多个非线性层的权值趋近于零来逼近标识映射。残差学习的这种特性使得深度神经网络易于训练并且能够获得更高的精度。

2.2.2 通过捷径进行标识映射

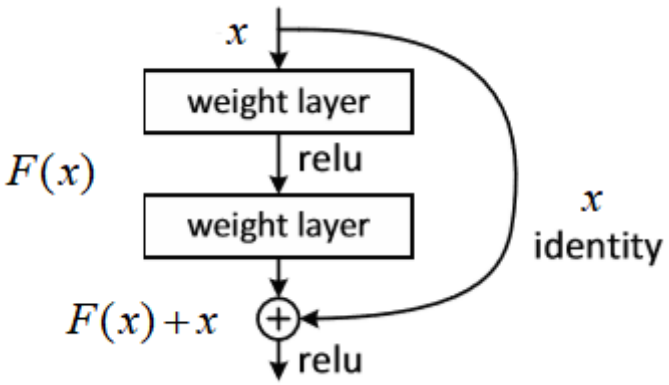


图2.1 残差学习的架构单元

一个有着较少堆叠层的残差学习的架构单元如图所示。一个残差学习单元的数学定义如下：

$$y=F(x,\{W_i\})+x \tag{2.2}$$

这里 x 和 y 分别是架构单元的输入和输出向量。函数 $F(x,\{W_i\})$ 表示要学习的残差映射。对于图 2-1 中具有两个层的示例，$F=W_2\sigma(xW_i)$，其中 σ 表示 relu，为了简化表示式中省略了偏置。操作 $F+x$ 通过捷径连接和元素相加来实现。相加后再经过 relu 激活。

方程式（1）引入的捷径连接并没有带来额外的参数和计算复杂度。在方程（1）中，x和f的维度必须相等。如果不是这样（例如，更改输入/输出通道时），可以通过如下式的线性投影W_s以使两者尺寸匹配：

$$y = F(x,\{W_i\}) + W_s x \tag{2.3}$$

在式(2.3)中也可以使用矩阵W_s。但残差标识映射能够高效的解决退化问题，因此W_s只用于尺寸匹配。

残差函数F的形式是灵活的，它可以由多层感知器构成。但是当F只有一层感知器时，式x类似于线性层: $y = W_1 x + x$，这种形式的残差函数并没有优势。F中的感知器层既可以是全连接层，也可以是卷积层。当$F(x,\{W_i\})$表示多层卷积层时，两矩阵相加就是两特征图通道对通道相加。

2.3 实例标准化

2.3.1 提出背景

实例标准化是为了克服图像风格转换问题中批标准化的缺陷而提出的。对于图像风格转换任务，从过多的样本内容图像x_0中学习模型，如说超过 16 个样本，生成图像的量化结果相较使用较少的此类样本更差。尤其是生成图像的边界上会出现明显错误，可能是由于 Padding 的使用和生成网络中的边界效应。

图像风格转换化的结果一般不应该依赖于内容图像的对比度。而是应该匹配与于内容图像的纹理，即特征图的对比度。图像生成网络应该丢弃内容图像x_0中的对比度信息。但是通过使用标准卷积神经网络模块来学习丢弃对比度信息会带来不必要的困难。在架构中添加实例标准化层是简单有效的方法。实例标准化能够使风格转换网络更易于训练，并且最终获得更低的误差[42]。

2.3.2 数学定义

设$x \in R^{N\times C\times W\times H}$是包含 N 张图片的输入张量，$x_{n,i,j,k}$表示此张量的$nijk$位置的元素，其中$k$和$j$表示图片尺寸空间维度，$i$表示特征通道，$n$是图片在一个批次内的序号。则实例标准化可由下式定义：

$$y_{\mathrm{nijk}} = \frac{x_{\mathrm{nijk}} - \mu_{ni}}{\sqrt{\sigma_{ni}^2 + \varepsilon}} \tag{2.4}$$

$$\mu_{ni} = \frac{1}{HW}\sum_{l=1}^{W}\sum_{m=1}^{H} x_{nilm} \tag{2.5}$$

$$\sigma_{ni}^2 = \frac{1}{HW}\sum_{l=1}^{W}\sum_{m=1}^{H}(x_{nilm} - \mu_{ni})^2 \tag{2.6}$$

实例标准化与一些关于样式转换的文献[43]中采用的批标准化最为关键的不同点在于，后者的标准化是基于一整批图像的，二实例标准化是对标准化分别运用于每张图像上。因为实例标准化能够标准化每张图片的内容x_0，对于设计用途为图像风格转换的神经网络，方程式 12 更为实用。实例标准化通常应用于整个架构中，而不仅仅是网络的输入和输出层。

与批标准化相似之处在于每个实例标准化层后都跟随者一个缩放加偏置操作。然而，当神经网络完成训练投入使用时，并不需要移除或改变实例标准化层。

2.4 自适应实例标准化

自适应实例标准化(AdaIN)是对 IN 的简单扩。IN 将输入规范化为由仿射参数指定的单一样式，自适应实例标准化则通过使用自适应仿射转换使其适用于任意给定的样式。$AdaIN$ 接收一个内容输入x和一个样式输入y，并简单地将x的信道均值和方差与y匹配[44]。与 BN、IN 或 CIN 不同，AdaIN 没有可学习的仿射参数。相反，它自适应地从样式输入计算仿射参数:

$$AdaIN(x,y) = \sigma(y)\left(\frac{x-\mu(x)}{\sigma(x)}\right) + \mu(y) \tag{2.7}$$

在式中，$\sigma(y)$缩放了标准化内容输入，接着将其与$\mu(y)$相加。与实例标准化相似，式中的统计参数是在空间尺度上计算的。

直观地说，考虑一个特征通道，它负责检测某种风格。具有这种风格图像将对该特征通道产生较高的平均激活。AdaIN 生成的输出对该特征具有相同的高平均激活，同时保留了内容图像的空间结构。利用前馈解码器，可以将该风格特征倒转到图像空间。该特征通道的方差可以编码更细微的风格信息，并将其传输到 AdaIN 输出和最终的输出图像中。

简而言之，AdaIN 通过传递特征统计量，特别是信道均值和方差，在特征空间中进行风格转移。AdaIN 层的作用类似于样式交换层[45]。样式交换操作非常耗时和消耗内存，而 AdaIN 层和 IN 层同样简单，只增加极少的计算成本。

3 虚拟驾驶环境多模态转换

本文引入部分共享隐空间的概念实现驾驶环境模态控制。图像表示可以分解为关于域不变的内容隐码和捕代表了特定属性的样式隐码。训练网络时随机样式隐码允许一对多的映射，学习目标域中学习相应图像的条件分布。生成虚拟驾驶环境时通过调整样式隐码便可控制图像模态。

3.1 部分共享的隐空间

假设 $x_1 \in \chi_1$ 和 $x_2 \in \chi_2$ 是来自两个不同图像域的图像。在无监督的图像到图像转换中，样本是从两个边缘分布 $\mathrm{p}(x_1)$ 和 $\mathrm{p}(x_2)$ 中提取的，而不是提取自联合分布 $p(x_1, x_2)$。本文目标是通过完成训练的图像到图像转换模型 $p(x_{1\to 2} \mid x_1)$ 以及 $p(x_{2\to 1} \mid x_2)$ 预测两个条件概率分布 $p(x_2 \mid x_1)$ 和 $p(x_1 \mid x_2)$。其中 $x_{1\to 2}$ 是将 x_1 翻译至 χ_2 产生的样本，$x_{2\to 1}$ 产生方式同上。一般来说，$p(x_2 \mid x_1)$ 和 $p(x_1 \mid x_2)$ 是复杂的多模态分布，确定性翻译模型不能很好地适用于在这种情况。

为解决上述问题，本文提出部分共享的隐空间假设。具体地说，我们假设每个图像 $x_i \in \chi_i$ 是由两个域共享的内容隐码 $c \in C$ 生成的，而样式隐码 $s_i \in S_i$ 是各个与所独有的。换句话说，由符合联合分布的一对相应的图像 (x_1, x_2) 是由 $x_1 \in G_1^*(c, s_1)$ 和 $x_2 \in G_2^*(c, s_2)$ 生成的，其中 c，s_1，s_2 取自先验分布，G_1^*，G_2^* 是底层生成器。进一步假设 G_1^* 和 G_2^* 是确定性函数便可得对应的反编码器 $E_1^* = (G_1^*)^{-1}$ 和 $E_2^* = (G_2^*)^{-1}$。尽管编码器和解码器是确定性的，由于对 s_2 的依赖性，$p(x_2 \mid x_1)$ 仍是连续分布。本文目标是通过神经网络学习底层的生成器和编码器功能。

本文假设与 UNIT[46]中提出的共享隐在空间假设十分相似。UNIT 提出了完全共享的隐空间，而在本文的假设中，只有部分的隐空间（内容空间）可以跨域共享，而另一部分（风格空间）是特定于域的。当跨域映射是多对多时这是一个更合理的假设。

3.2 模型

图 2 展示了本文的大致模型及其学习过程。与 Liu 等人〔15〕的工作相似，我们的翻译模型由每个域 χ_i $(i = 1,2)$ 的编码器 E_i 和解码器 G_i 组成。如图 3.1（a）所示，每个自动编码器的隐码被分解为内容码 c_i 和样式码 s_i，其中 $(c_i, s_i) = (E_i^c(x_i), E_i^s(x_i)) = E_i(x_i)$。如图 3.1（b）所示，图像到图像的转换是通过交换编码器-解码器对来执行的。图像到图像转换模型由两个自动编码器（分别用红色和蓝色箭头表示）组成，每个域各有一个。每个自动编码器的隐码由一个内容隐码 c 和一个样式隐码 s 组成。模型使用用对抗目标训练，确保翻译后的图像与目标域中的真实图像不可区分。同时，模型也使用双向重建

目标训练，以重建图像和隐码。例如，为了将图像 $x_1 \in \chi_1$ 转换为 χ_2，首先提取其内容隐码 $c_1 = E_1^c(x_1)$，接着从先验分布 $q(s_2) \sim N(0,\mathrm{I})$ 中随机抽取样式隐码 s_2。然后使用 G_2 生成最终输出图像 $x_{1\to 2} = G_2(c_1, s_2)$。虽然先验分布是单模太的，但由于解码器的非线性，输出图像分布可以是多模太的。

损失函数包括一个双向重建损失以及一个对抗性损失。双向重建损失确保编码器和解码器功能是完全反向的；对抗性损失确保翻译图像的分布与目标域中的图像分布相同。

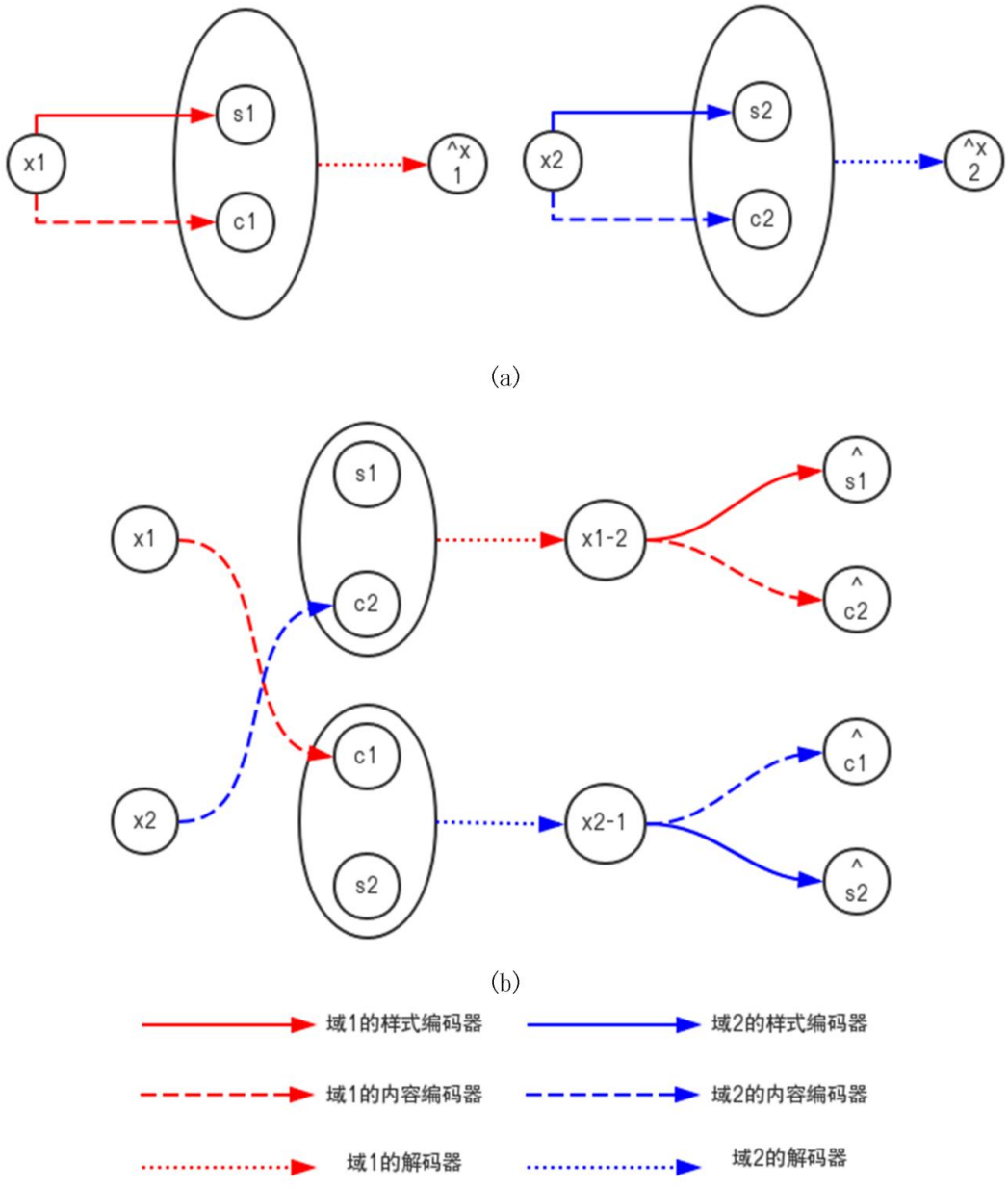


图3.1　模型示意图(a)图像重建示意图(b)图像到图像的转换示意图

3.2.1 双向重建损失

为了学习功能互为相反的编码器和解码器对，我们使用目标函数来同时激励图像→隐码→图像和隐码→图像→隐码方向的重建：

（1） 图像重建

对于从数据分布中采样的图像，在编码和解码后应该能够实现它的重建。

$$L_{recon}^{x_1} = \mathrm{E}_{x_1 \sim p(x_1)}\left[\| G_1(E_1^c(x_1), E_1^s(x_1)) - x_1 \|_1\right] \tag{3.1}$$

（2） 隐码重建

在翻译时对于从隐空间分布中抽取的隐码，我们应该能够在解码和编码后实现它的重建。

$$L_{recon}^{c_1} = \mathrm{E}_{c_1 \sim p(c_1), s_2 \sim q(s_2)}\left[\| E_2^c(G_2(c_1, s_2)) - c_1 \|_1\right] \tag{3.2}$$

$$L_{recon}^{s_2} = \mathrm{E}_{c_1 \sim p(c_1), s_2 \sim q(s_2)}\left[\| E_2^s(G_2(c_1, s_2)) - s_s \|_1\right] \tag{3.3}$$

其中 $q(s_2)$ 是先验分布 $N(0,\mathrm{I})$，$p(c_1)$ 由 $c_1 = E_1^c(x_1)$ 带入 $x_1 \sim p(x_1)$ 得出。

其他损失函数 $L_{recon}^{x_2}$，$L_{recon}^{c_2}$ 和 $L_{recon}^{s_2}$ 的定义方式与 $L_{recon}^{x_1}$ 式相似。本文使用图像重建重建损失 $L_{recon}^{x_2}$，因为它能提高输出图像质量。

风格重建损失 $L_{recon}^{s_i}$ 在不同的样式隐码下，它可以奖励不同的输出。内容重构损失 $L_{recon}^{c_i}$ 奖励翻译后的图像保留输入图像的语义内容。

3.2.2 对抗损失

GAN 被用于翻译图像，使之分布与目标域数据分布相匹配。换句话说，本文模型生成的图像应该与目标域中的真实图像不可区分。

$$L_{GAN}^{x_2} = \mathrm{E}_{c_1 \sim p(c_1), s_2 \sim q(s_2)}\left[\log(1 - D_2(G_2(c_1, s_2)))\right] + \mathrm{E}_{x_2 \sim p(x_2)}\left[\log D_2(x_2)\right] \tag{3.4}$$

其中 D_2 是用于分辨翻译图像与 χ_2 域中真实图像的辨别器。辨别器 D_1 和损失 $L_{GAN}^{x_1}$ 的定义与上文类似。

3.2.3 总损失

联合编码器、译码器和鉴别器共同训练，以优化最终目标，即对抗损失和双向重建损失项的加权和。

$$\begin{aligned}&\min_{E_1,E_2,G_1,G_2}\max_{D_1,D_2} L(E_1,E_2,G_1,G_2,D_1,D_2) = L_{GAN}^{x_1} + L_{GAN}^{x_2} + \\ &\lambda_x(L_{recoon}^{x_1} + L_{recoon}^{x_2}) + \lambda_c(L_{recoon}^{c_1} + L_{recoon}^{c_2}) + \lambda_s(L_{recoon}^{s_1} + L_{recoon}^{s_2})\end{aligned} \tag{3.5}$$

其中，λ_x，λ_c，λ_s是控制重建损失项作用的权重。

4 理论分析

本节建立本文框架的一些理论属性。具体来说，本节证明最小化所提出的损失函数将导致 1）对应于编码和生成过程的隐码分布的匹配；2）由本文框架得到的两个联合图像分布的匹配；3）执行弱形式的循环一致性约束。

首先，当转换数据的分布与目标域数据分布匹配并且编码器、解码器功能完全相反时，式 x 中的总损失最小化。

4.1 隐空间分布匹配

假设已有域 1 编码器 E_1^*，域 2 编码器 E_2^*，域 1 解码器 G_1^*，域 2 解码器 G_2^* 使得：1）$E_1^*=(G_1^*)^{-1}$ 且 $E_2^*=(G_2^*)^{-1}$；2） $p(x_{1\to2})=p(x_2)$ 且 $p(x_{2\to1})=p(x_1)$。那么 E_1^*， E_2^*， G_1^*， G_2^* 将最小化 $L(E_1,E_2,G_1,G_2)=\max_{D_1,D_2}L(E_1,E_2,G_1,G_2,D_1,D_2)$。

隐空间分布匹配：对于图像生成，现有的组合自动编码器和 GAN 的工作需要将编码的隐码分布与解码器在生成时接收的隐码布相匹配，在隐空间使用 KLD 损失或对抗性损失。 如果解码器在生成期间接收到极为不同的隐码分布，则自动编码器训练将无助于 GAN 训练。虽然我们的损失函数不包含明确奖励隐码分布匹配的项，但它具有将隐码分布隐式匹配的效果。

4.2 联合分布匹配

达到优化到最佳状态时，如下等式成立：

$$p(c_1)=p(c_2) \tag{4.1}$$

$$p(s_1)=q(s_1) \tag{4.2}$$

$$p(s_2)=q(s_2) \tag{4.3}$$

上述命题表明，在达到最优时，样式隐码分布的与它们的高斯先验分布相匹配。此外，内容隐码分布的与生成时的分布匹配，即使内容隐码是来自另一域的隐码分布。这表明内容空间在不同域中不变。

联合分布匹配：我们的模型学习了两个条件分布 $p(x_{1\to2}|x_1)$ 和 $p(x_{2\to1}|x_2)$，与数据分布一起定义了两个联合分布 $p(x_1,x_{1\to2})$ 和 $p(x_{2\to1},x_2)$。由于它们为接近相同的基础联合分布 $p(x_1,x_2)$ 而设计，因此它们的期望是彼此一致的，即 $p(x_1,x_{1\to2})=\ p(x_{2\to1},x_2)$。

联合分布匹配为无监督的图像到图像转换提供了重要的约束，并且是近来许多方法成功的基础。本文的模型匹配基础联合分布在达到最优化时。

4.3 样式强化的循环一致性

当模型达到最优时，可以得到 $p(x_1, x_{1\to 2}) = p(x_{2\to 1}, x_2)$ 。

样式强化的循环一致性：如果翻译模型事确定性并且边缘分布是匹配的，联合分布匹配可以通过循环一致性约束实现。但是，这种约束对于多模态图像翻译来说太强了。事实上，如果强制执行循环一致性，翻译模型将退化为确定性函数。下面的命题展示了本文采用一种称为样式化循环一致性的弱形式循环一致性，处于图像样式联合空间之间。这种一致性更适合于多模态图像转换。

设 $h_1 = (x_1, s_2) \in H_1$ 和 $h_2 = (x_2, s_1) \in H_2$ 。 h_1 ， h_2 是图像和风格的联合空间中的点。通过 $F_{1\to 2}(h_1) = F_{1\to 2}(x_1, s_2) \overset{\Delta}{=} (G_2(E_1^c(x_1), s_2), E_1^s(x_1))$ ，本文的模型定义了从 H_1 到 H_2 的确定性映射 $F_{1\to 2}$ （反之亦然）。当模型达到最优时，有 $F_{1\to 2} = F_{2\to 1}^{-1}$ 。

直观地说，样式增强循环一致性意味着如果将图像转换为目标域并使用原始样式将其转换回来，我们应该获得原始图像。样式增强的循环一致性是通过本文的双向重建损失实现的：

$$L_{cc}^{x_1} = E_{x_1 \sim p(x_1), s_2 \sim q(s_2)} \left[\| G_1(E_2^c(G_2(E_1^c(x_1), s_2)), E_1^s(x_1)) - x_1 \|_1 \right] \tag{4.4}$$

5 实验及分析

5.1 模型实例细节

图 5.1 显示了我们的自动编码器的架构。 它由内容编码器，样式编码器和联合解码器组成。

5.1.1 内容编码器

内容编码器由若干个用于对输入进行下采样的跨步卷积层，和若干个用以进一步处理卷积层信息残余块[41]组成。所有卷积层的输出都经过实例标准化（IN）[42]。

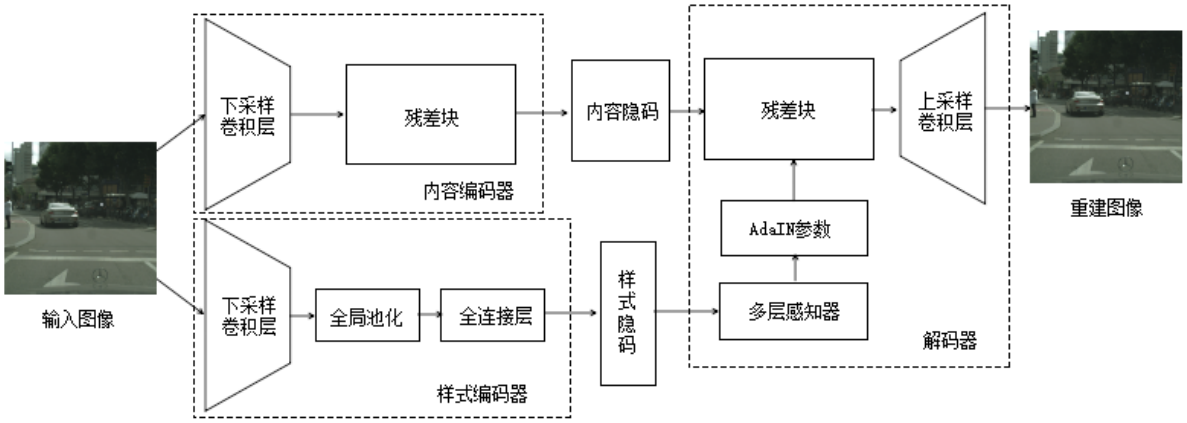


图5.1 自动编码器架构

内容编码器由若干个跨步卷积层和残余块组成。样式编码器包含若干个跨步卷积层，以及后跟的全局平均池化层和全连接层。解码器使用 MLP 从样式隐码生成一组 AdaIN[44]参数。接着内容隐码由具有 AdaIN 层的残余块处理，最后通过上采样和卷积层解码到图像空间。

5.1.2 样式编码器

样式编码器包括若干个跨步卷积层，后面是全局平均池层和全连接层。我们不在样式编码器中使用 IN 层，因为 IN 删除了表示重要样式信息的原始特征均值和方差。

解码器。 我们的解码器根据其内容和样式隐码重建输入图像。 它通过一组残余块处理内容隐码，并最终通过若干个上采样和卷积层产生重建图像。 参考最近在标准化层中使用仿射变换参数来表示样式[44]的工作，残差块配备了自适应实例标准化（AdaIN）[44]层，其参数由多层感知器从样式隐码动态生成（MLP）。

$$AdaIN(z,\gamma,\beta)=\gamma\left(\frac{z-\mu(z)}{\sigma(z)}\right)+\beta \qquad (5.1)$$

其中 z 是先前卷积层的激活，μ 和 σ 是通道平均值和标准偏差，γ 和 β 是由 MLP 生成的参数。 仿射参数是由学习网络产生的，而非如 Huang 等人的工作根据预训练网络的统计数据计算出来的[44]。

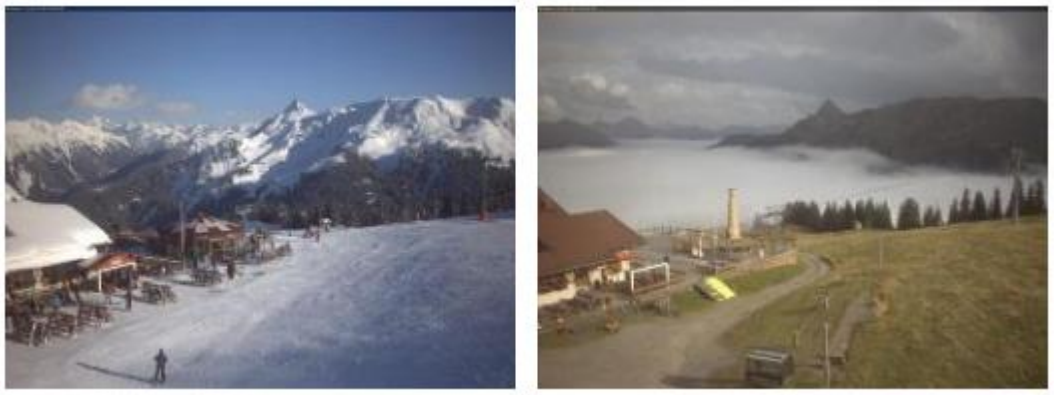
5.1.3 辨别器

为了使得生成器输出即具有现实细节和又具有符合逻辑的全局结构，本文采用多尺度辨别器，目标函数则与 LAGAN 相同[23][47]。

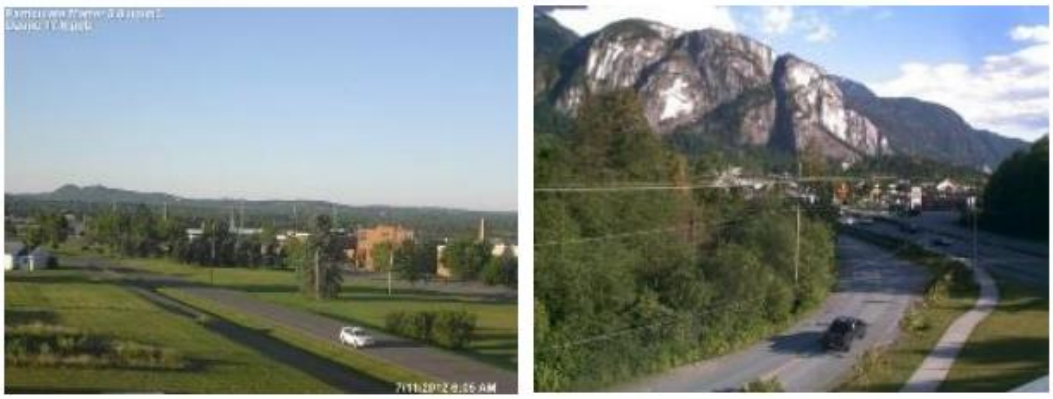
5.1.4 域不变的感知损失

感知损失通常定义为输出和参考图像之间的 VGG [48]特征空间中的距离。感知损失已被证明对有监督的图像到图像的转换任务有益。但是本文使用无监督方法，目标域中没有参考图像。因此本文采用了域不变特性更为显著的感知损失的修改版本，以便使用输入图像作为参考。具体来说，在计算距离之前，我们对输入 VGG 的图像提前执行实例标准化（没有仿射变换），以便删除原始特征均值和方差，这其中包含许多特定于域的信息。域不变的感知损失能加速对高分辨率数据集的训练，因此在本文这些数据集上使用它。

本文进行实验验证了在计算特征距离之前应用 IN 是否确实可以使感知距离更加具有域不变性。实验在[50]中使用的数据集上进行。随机抽样两组图像对：1）来自同一场景但不同域（夏季或冬季）的图像，2）来自同一域但不同场景的图像。图 5.2 展示了来自本文抽取的两组图像对的示例。不应用 IN,计算每组图像对之间的 VGG 特征距离（感知距离）；接着，对所有图像先使用 IN 处理，再按照与上述相同的方式计算 VGG 距离。在图 5.3 中，展示了来自同一场景但不同域的图像与来自同一域但不同场景的图像各自使用或不使用 IN 计算的距离的感知距离-图像数量直方图。在计算距离之前不应用 IN，两组图相对特征距离的分布相似。在使用 IN 的情况下，来自同一场景的图像对具有明显更小的感知距离，即使它们来自不同的域。而不同场景图相对具有较大的感知距离，即使它们来自相同域。因此在计算距离之前应用 IN 使得特征距离更具有域不变性。

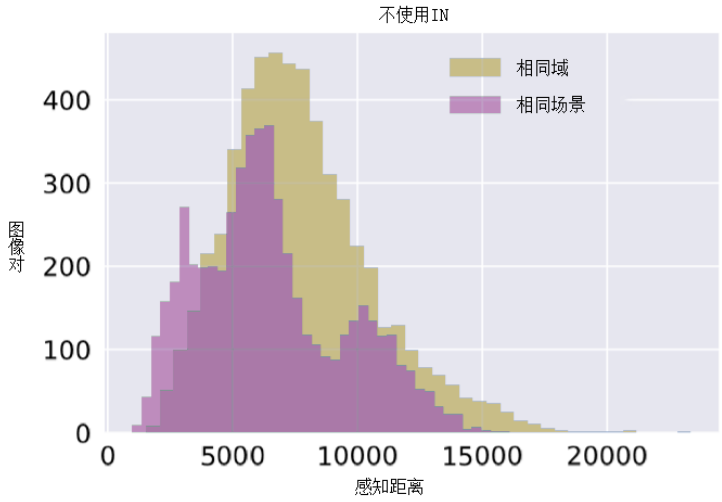


（a）

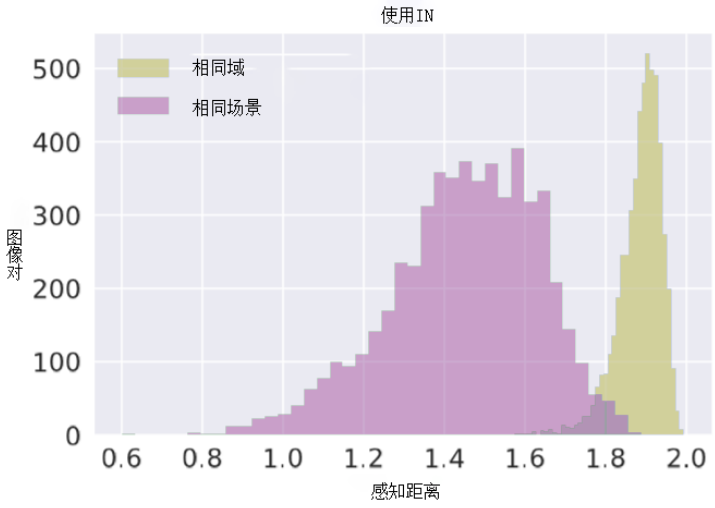


（b）

图5.2　域不变感知损失实验图片对(a)同一场景图像对(b)同一域图像对



(a)



(b)

图 5.3　感知距离直方图(a)不使用 IN(b)使用 IN

5.1.5 网络架构

令 c7s1-k 表示具有 k 个滤波器，卷积核大小为 7×7，步幅为 1 的层；dk 表示具有 k 个滤波器，卷积核大小为 4×4，步幅为 2 的层；Rk 表示包含两层核心大小为 3×3 卷积层的残差块； uk 表示放大倍数为 2 的最近邻上采样层，其后是具有 k 个滤波器，卷积核大小为 5×5，步幅为 1 的层；GAP 表示全局平均池化层；fck 表示具有 k 个滤波器的全连接层。实例标准化（IN）[42]应用于内容编码器，自适应实例标准化（AdaIN）[44]则应用于解码器。在生成器中使用了 relu 激活函数。在辨别器中使用了 Leaky relu 激活函数，函数自变量小于零部分斜率为 0.2。

(1) 生成器架构

内容编码器: c7s1-64, d128, d256, R256, R256, R256, R256

样式编码器: c7s1-64, d128, d256, d256, d256, GAP, fc8

解码器: R256, R256, R256, R256, u128, u64, c7s1-3

(2) 辨别器架构

d64, d128, d256, d512

5.2 训练细节

在所有实验中，使用 Adam 优化器[51]，超参数设置如下：$\beta_1 = 0.5$，$\beta_2 = 0.999$。初始学习步长为 0.0001,学习步长每经过 100,000 次迭代优化减小一半。受限于硬件内存，batch 大小设置 1，每次迭代仅输入一对数据。损失权重设置为$\lambda_x = 10$，$\lambda_x = 1$，$\lambda_x = 1$。样式隐码的维度设置为 8。在训练期间对所有数据应用随机镜像。接下来说明针对不同实验的特殊设定，在模态控制实验中使用权重为 1 的域不变感知损失，网络输入输出图像分辨率都为 256×256。在图像合成实验中不使用域不变感知损失，网络输入输出图像分辨率都为 224×292。将图像变换到目标分辨率的方法是双线性插值。

5.3 数据集

5.3.1 Cityscapes 数据集

Cityscapes 是一个大规模城市街景数据集，其中包含了从 50 个不同城市的街道场景中录制的一组不同的立体视频,除了 20000 粗糙注释帧（如图 5.4 所示）外，还有 5000 帧的高质量像素级注释（如图 5.4 所示）。Cityscapes 旨在用于训练语义视觉算法并评估其在城市场景理解任务中的性能。Cityscapes 共有八组，30 个类别的标签（如表 5.1 所示）。本文在图像合成实验中使用高质量像素级注释的数据集和无标签的视频数据集，图片像素重新插值为 256×256。

图5.4　Cityscapes粗糙注释帧

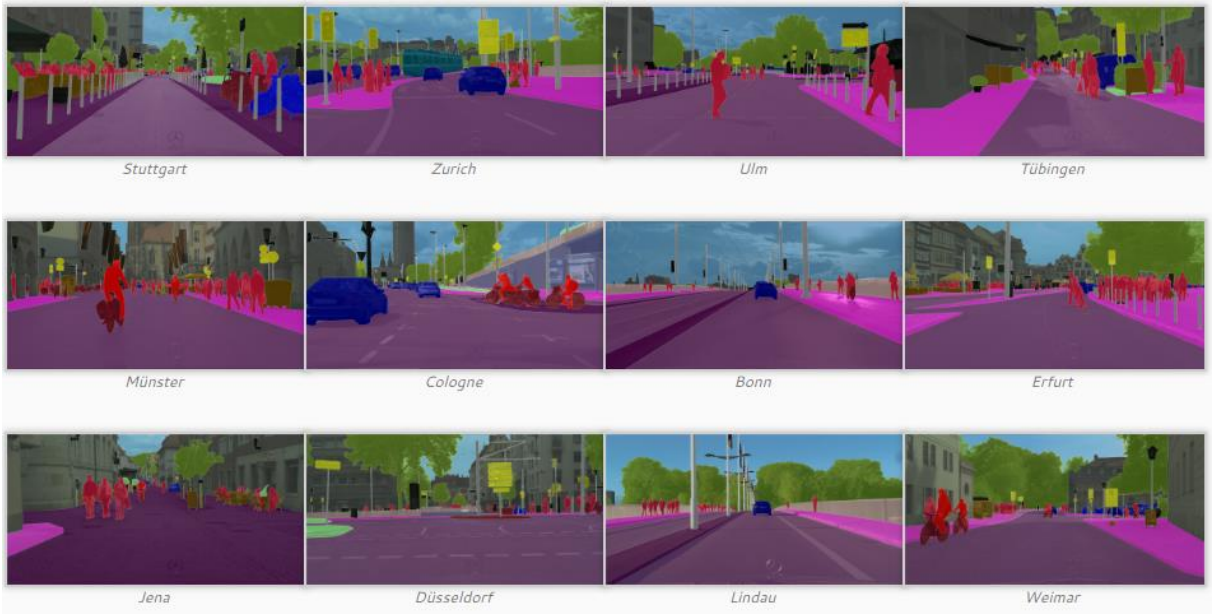


图5.5　Cityscapes高质量注释帧

表5.1　Cityscapes类别定义

组	类别
Flat	Road;sidewalk;parking;rail track
Human	Person;rider
Vehicle	car;truck;bus;on rails;motorcycle;bicycle;caravan;trailer
Construction	Building;wall;fence;guard rail;bridge;tunnel
Object	Pole;pole group;traffic sign;traffic light
Nature	Vegetation;terrain
Sky	sky
Void	Ground;dynamic;static

5.3.2 Comma2K19 数据集

Comma2K19是由Comma AI提供的自动驾驶数据集。该数据集是在美国加利福尼亚280高速公路的加利福尼亚圣若泽与旧金山之间的20公里路段上采集的，累计拍摄时长33小时，即有共2019段视频，每段时长1分钟。Comma2K19是一个完全可复制和可扩展的数据集（视频如图5.6所示）。拍摄的时间段覆盖了一天24小时，即小时内至少有一段视频。这些数据是使用具有与现代智能手机类似传感器（包括面向道路的摄像头、手机GPS、温度计和9轴加速度传感器）的Comma ECONs收集的。此外，eon还捕获原始GNSS测量值并通过Comma grey panda 发送的所有CAN数据。视频分辨率为1164×874，本文将其缩小为292×224用于模态控制实验。

图5.6 Comma2K19数据集视频帧

5.4 评估标准

5.4.1 主观评价

由于自动驾驶车辆最终要投放到真实世界去使用，虚拟驾驶环境的图像不仅需要在细节风格上真实，内容也需要在逻辑上正常，符合现实。因此为了评价本文模型输出的真实性，我们进行了主观评价。具体操作方法是将一个输入图像和网络翻译后的某张生成图像同时展现给评价者，然后给予它们有限的时间来选择哪张图是真实图像。我们为每个评价者随机生成 15 个问题，共计 100 位评价者参与。

5.4.2 LPIPS 距离

为了量化评价图像转换的多样性，我们计算了随机采样的有着相同输入的转换输出之间的平均 LPIPS 距离。LPIPS 由图像深度特征之间的加权欧式距离给出。相关研究已经证明其与人类感知具有很高的相似性[52]。本文使用了 100 个输入图像并且对于每个抽取 10 个输出对作为样本，总共有 1000 个输出样本。本文使用由 ImageNet 数据集预训练的 AlexNet [53]作为深度特征提取器。

5.4.3 图像质量量化评价

在模态控制实验中，为评价多模态图像的质量，本文对于每个输入图像抽取了10个输出作为样本，共取100张输入图像。此实验还需要评价在执行光照条件控制任务时样式隐码重建损失、内容隐码重建损失、图像重建损失对于生成图像质量的影响。因此样本总量巨大，限于时间与精力采用5.4.1中的主观评价方法全面覆盖所有样本。本文采用GAN辨别器作为图像质量量化评价标准。其中辨别器取自在Comma数据集上训练后的本文模型。对于白天到夜晚转换，使用夜晚域的辨别器；对于夜晚到白天转换，使用白天域的辨别器。评价标准为辨别器判断为真实图像的百分比。

5.5 图像合成实验

本实验目的为合成自动驾驶环境图像。网络需要由输入的街景图像语义布局（如图）生成真实的街景图像。这一步是将自动驾驶代理视角下的 CG 模型环境渲染为更为真实驾驶环境的关键一步，语义布局可由语义分割网络标注已经贴图的 CG 模型得到，也可由已经注明类别标签 CG 模型自动生成。本实验使用 Cityscapes 数据集，将街景图像与其语义标签作为两个域供网络训练。

表5.2 本模型与CG建模法主观评价结果

	本模型	CARLA	Airsim	Carcraft	GTA5
Cityscapes	39.73%	0%	0%	0%	0%

表5.3 本模型与其他神经网络合成法主观评价结果

	本模型	CRN	GAN_SemSeg	Isola	Encoder-decoder	Full-resolution network
Cityscapes	39.73%	39.87%	39.40%	38.33%	31.27%	27.00%

首先按照 5.2.1 的方法进行了主观评价。从 Cityscapes 验证集中随机抽取 100 张语义分割标签，将样式码固定在[0,0,0,0,0,0,0,0],令本网络对于每个标签只生成一张街景图像以方便和其他方法的比较。表 5.2 展示了本模型与 CG 建模法主观评价的结果。

英特尔的 CARLA[2]、微软的 Airsim[3]、谷歌的 Carcraft[4]以及 GTA5 游戏是用于自动驾驶代理训练的主流虚拟环境。本文在上述人工建模环境中进行驾驶模拟，截取引擎盖视角的图像用于比较。由表 5.2 可以看出，虽然本方法生成图像相较于来自 Cityscapes 的真实车载图像，被认为更真实的比例仅有 39.76%，但是人工建模法的图像完全没有被认为更真实的。相较于人工建魔法，本文的方法已经很大程度的改善了生成图像的真实性。表 5.3 为本模型与其他神经网络合成法主观评价结果。其他神经网络使用了与本模型同样的语义分割标签作为输入。由表 5.3 可得，本模型真实性排名第二，与效果最好的网络仅相差 0.12%。

图5.7　本文模型生成图像示例

本文模型生成的自动驾驶环境图像如图 5.7 所示。第一第二列图像为输入模型的语义布局，第三第四列为对应的合成图像。所生成的图像在训练集中并未出现过，但内容合理、符合现实逻辑。天空云层、树木、引擎盖镜面反光都渲染真实；在丁字路口，车辆前方路面合成出了停车线，在三叉路分叉处也合成出了引导线，并且这些路面标识如同现实中经过常年使用的路面，具有真实的磨损质感；图中车辆不同于现有的任意款车型，但是前后风挡、车轮、尾灯等所处位置都符合逻辑；路面材质并不单一，处沥青外，第二列第二幅图合成了砖石路面。本文模型生成图片样式具有多样性同时内容又符合逻

辑的性质十分适用于自动驾驶模型训练、测试。

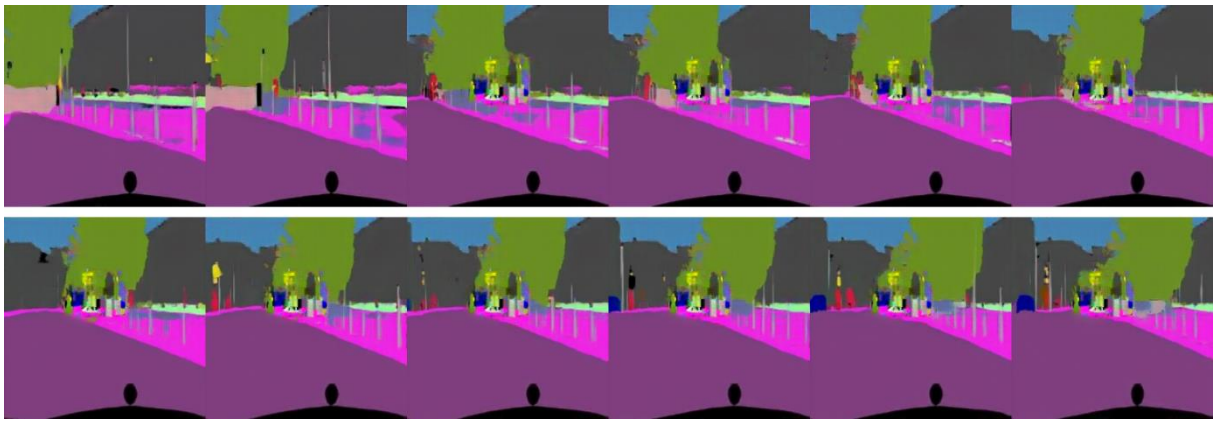


图5.8　连续帧语义分割图像

图5.9　连续帧生成图像

自动驾驶车辆的一些算法是依赖于时间序列信息的，如光流法，因此虚拟驾驶环境图像在时间序列上也需要具有连续性，不应在内容或样式上产生突变。本文针对这一需求进行了基于视频实验，视频取自 cityscapes 数据集。由于视频没有语义分割标签，本文使用在 cityscapse 上训练过的语义分割网络获得语义分割图像，如图 5.8 所示为对 12 张连续帧分割的结果。将语义分割图输入模型，样式码固定，对于所有帧使用相同的样式码，得到的合成图像如图 5.9 所示。由图可见，光照条件、树木、云彩、建筑在相邻帧上变化不大，但是当帧数间隔较大时，本该相同的建筑结构发生较大变化，如第六帧道路左侧房屋墙面大片污渍在第十二帧变为上下两个窗口。其次，在播放网络输出帧编码的视频时发现，路面标识在连续几帧中的常处于相同位置，这带来了车辆静止的视觉假象。样式码对于图像在时间序列上的连续性由一定作用，但仍有缺陷。

5.6　模态控制实验

本实验目的为验证本文模型具有对图像模态进行控制的能力。虚拟驾驶环境的光照

条件是一个重要属性，对自动驾驶算法会产生很大的影响。因此选取光照条件控制作为模态控制的典型示例进行试验。本实验使用 commma2k19 行车视频数据集，每 25 帧采样一次作为训练数据。拍摄于早九点到晚六点之间的视频与其他时间段的视频被分开作为两个域。完成优化的网络应当能将驾驶坏境的光照在白天与夜晚之间转换，并且能可控的渲染出白天黑夜不同时间段的光照。

表5.4　光照条件转换图像量化分析结果

	白天到夜晚		夜晚到白天	
	质量	多样性	质量	多样性
无图像重建损失	23.4%	0.291	7.5%	0.382
无内容隐码重建损失	15.3%	0.176	13.4%	0.198
无样式隐码重建损失	34.2%	0.92	23.7%	0.127
完整网络	56.9%	0.235	41.8%	0.189
真实数据	N/A	0.311	N/A	0.287

首先，我们定量的分析了本文模型及其三个变体——分别去除图像重建损失 L^{x}_{recon}，内容隐码重建损失 L^{c}_{recon} 与样式隐码重建损失 L^{s}_{recon}。样式隐码从符合八维正太分布的噪声中采样，对于每张输入图像，采样 100 个样式隐码生成对应的 100 张图像。驾驶环境日夜转换后的质量与多样性如表 5.4 所示。图像质量受 L^{x}_{recon} 和 L^{c}_{recon} 影响最大，缺少其中之一，转换的图像的质量会大幅降低。 在没有 L^{s}_{recon} 的情况下，模型输出的多样性降低。而完整模型生成的图像多样性对染相较于无图像重建损失略低，但图像质量得到大幅提升，达到了较好的平衡。

图5.10　入夜光照条件引导图像

图5.11　卤素路灯光照条件引导图像

图5.12　深夜光照条件引导图像

接着进行模态控制实验，目标为将日间行车图像转换为夜间行车图像，并且控制其光照，定向生成入夜时、卤素路灯照明、深夜三种光照条件下的图像。我们分别搜索了入夜时（图 5.10）、卤素路灯照明（图 5.11）和深夜（图 5.12）这三种条件下的图像作为引导图像。通过夜晚域的编码器抽取引导图像的样式隐码，由此获得了入夜时、卤素路灯照明和深夜光照条件的样式隐码。

(1) 入夜光照条件样式隐码：[0.59932866, -0.63002128, 0.07587199, -0.07933834, -1.2467873, 0.54258754, -0.82982448, -0.66341979]

(2) 卤素路灯光照条件样式隐码：[-0.7843677,　-1.73333418, -0.59294801, 0.3364004,　-0.50422754, -0.55436861, -0.74849017,　1.44244231]

(3) 深夜光照条件样式隐码：[-0.00501374, 2.08136842, -0.03810636, -0.19362248, -1.03384617, -0.51903289, 0.53973022, 0.25999474]

将日间图像和样式隐码同时输入网络，便可得到同时符合日间图像布局及所需光照条件的夜间行车图像。

如图 5.10-图 5.15 所示，本文模型成功地将日间行车图像转换为夜晚行车图像。给定日间行车输入图像，通过输入不同的样式隐码，能控制转换来的夜晚图像的光照条件。样式 1 对应入夜光照、样式二对应卤素路灯光照、样式三对应深夜光照。如图 5.10，输出的第一种样式与刚入夜的光照相似，远方天空微亮，由远及近亮度逐渐降低；前方车量尾灯亮起；路面出现车辆大灯照射效果。第二种样式与有卤素路灯照明路面的光照相似，整体色调偏暖。样式三与深夜无路灯道路的光照条件相似，在车灯照射范围外的景物漆黑一片。虽然图片中的光照条件经历了大幅变化，但是车道、车辆、树木、天空的位置、形状与布局都保持不变。

图5.10　日间转夜间行车图像示例1

图5.11　日间转夜间行车图像示例2

图5.12　日间转夜间行车图像示例3

图5.13　日间转夜间行车图像示例4

图5.14　日间转夜间行车图像示例5

结　论

本文提出了一种基于深度学习的虚拟驾驶环境图像的生成方法。该模型是属于无监督方法，实现了由语义布局合成全新的逼真驾驶环境图像，并且在不影响图像内容的基础上控制图像模态。隐码在模型中被分为内容码与样式码，通过固定样式码，对在时间序列相近的图像的连续性有了改善；通过改变样式码，可以控制图像中一些属性的改变，如光照条件。本文主要创新点如下：

（1）提出面向自动驾驶的多模态图像对抗生成算法。我们依据语义分割图像生成多种样式的街景真实图像。

（2）本文算法可实现日夜行车图像转换，并且通过改变样式码控制驾驶环境光照条件。

（3）本文方法所生成视频帧连续性与较好，说明本文算法健壮性。

此外，本文的方法在合成图像逼真度的主观测试中，优于 CG 建模法，在深度学习方法中处于领先位置。在多模态图像转换的量化分析中，本文采用的多种损失函数大幅提升了图像质量及多样性。未来的工作可以研究将本方法与长短期记忆网络结合使得生成视频相隔较远的帧也有很好的连续性。

修改记录

第一次修改记录：

第 11 页图 3.1 图名，**修改前：**模型示意图

修改后：模型示意图(a)图像重建示意图(b)图像到图像的转换示意图

第 18 页图 5.2 图名，**修改前：**域不变感知损失实验图片对

修改后：域不变感知损失实验图片对(a)同一场景图像对(b)同一域图像对

第 18 页图 5.3 图名，**修改前：**感知距离直方图

修改后：感知距离直方图(a)不使用 IN(b)使用 IN

第二次修改记录：

第 31 页结论，**修改前：**

本文提出了一种基于深度学习的虚拟驾驶环境图像的生成方法。该模型是属于无监督方法，实现了由语义布局合成全新的逼真驾驶环境图像，并且在不影响图像内容的基础上控制图像模态。隐码在模型中被分为内容码与样式码，通过固定样式码，对在时间序列相近的图像的连续性有了改善；通过改变样式码，可以控制图像中一些属性的改变，如光照条件。本文的方法在合成图像逼真度的主观测试中，优于CG建模法，在深度学习方法中处于领先位置。在多模态图像转换的量化分析中，本文采用的多种损失函数大幅提升了图像质量及多样性。未来的工作可以研究将本方法与长短期记忆网络结合使得生成视频相隔较远的帧也有很好的连续性。

修改后：

本文提出了一种基于深度学习的虚拟驾驶环境图像的生成方法。该模型是属于无监督方法，实现了由语义布局合成全新的逼真驾驶环境图像，并且在不影响图像内容的基础上控制图像模态。隐码在模型中被分为内容码与样式码，通过固定样式码，对在时间序列相近的图像的连续性有了改善；通过改变样式码，可以控制图像中一些属性的改变，如光照条件。本文主要创新点如下：

（1）提出面向自动驾驶的多模态图像对抗生成算法。我们依据语义分割图像生成多种样式的街景真实图像。

（2）本文算法可实现日夜行车图像转换，并且通过改变样式码控制驾驶环境光照条件。

（3）本文方法所生成视频帧连续性与较好，说明本文算法健壮性。

此外，本文的方法在合成图像逼真度的主观测试中，优于 CG 建模法，在深度学习方法中处于领先位置。在多模态图像转换的量化分析中，本文采用的多种损失函数大幅

提升了图像质量及多样性。未来的工作可以研究将本方法与长短期记忆网络结合使得生成视频相隔较远的帧也有很好的连续性。

第三次修改记录：

第 19 页 5.2，**修改前：**网络输入输出图像分辨率都为 256*256

修改后：网络输入输出图像分辨率都为 256×256

记录人（签字）：

指导教师（签字）：

致　　谢

四年的求学生涯，在老师，同学的全力支持下，我走得辛苦却也收获颇丰。在值此论文即将付梓之际，我思绪万千，心情久久不能平静。

感谢我的导师祝雪峰老师，从大二的大创项目到现在的毕业论文都给我悉心指导。他严肃的科学态度激励着我，我的论文写作规范性得到不断提高。 同时，祝雪峰老师对我的科研工作都会提出更高的要求，在他的要求下我尝试改进工作获得了效果，让我懂得了要不断努力精益求精。在此，我谨向祝雪峰老师致以十二分诚挚的谢意。

感谢郑仁成老师，郑老师对我的科研工作给与了大量的指导建议，他对实事求是、一丝不苟，值得我学习。

感谢韩小强老师，韩老师为我在车队的工作提供了极大的帮助，也传授了许多一线工艺上的经验。

感谢刘笑晨师兄，我们差不多在同一时间开始学习神经网络，一起交流过理论、共同调试过计算机集群。

最后还要感谢在百忙之中为我评审论文的各位专家和参加答辩的各位老师，感谢所有给予过我关心和帮助的人。

参考文献

[1] N. Kalra, S. M. Paddock, Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?[J]. Transportation Research Part A: Policy and Practice. 2016, 94: 182-193.

[2] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, V. Koltun. CARLA: An Open Urban Driving Simulator[C].1st Annual Conference on Robot Learning, the US, 2017, 1-16.

[3] S. Shah, D. Dey, C. Lovett, A. Kapoor, AirSim: High-fidelity visual and physical simulation for autonomous vehicles[J]. Field and Service Robotics. 2018, 5: 621–635.

[4] A. C. Madrigal. Inside Waymo's Secret World for Training Self-Driving Cars[J/OL]. The Altantis.2017, 8, 23. https://www.theatlantic.com/technology/archive/2017/08/inside-waymos-secret-testing-and-simulation-facilities/537648/.

[5] Li W , Pan C , Zhang R , et al. AADS: Augmented Autonomous Driving Simulation using Data-driven Algorithms[J]. Science Robotics. 2019, 4(28):1-24.

[6] Goodfellow I J , Pouget-Abadie J , Mirza M , et al. Generative Adversarial Nets[C]. International Conference on Neural Information Processing Systems, Montréal. 2014.

[7] Wu J , Zhang C , Xue T , et al. Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling[C]. International Conference on Neural Information Processing Systems, Barcelona. 2016.

[8] Larsen A B L , Sønderby, Søren Kaae, et al. Autoencoding beyond pixels using a learned similarity metric[C]. International Conference on International Conference on Machine Learning, Lille. 2015.

[9] Wang X , Gupta A . Generative Image Modeling using Style and Structure Adversarial Networks[C]. European Conference on Computer Vision, Amsterdam. 2016.

[10] Pan X , You Y , Wang Z , et al. Virtual to Real Reinforcement Learning for Autonomous Driving[J/OL]. Ithaca, NY: arXiv.org, 2017(2017,9,26)[2019/6/15]. https://arxiv.org/abs/1704.03952.

[11] Mirza M , Osindero S . Conditional Generative Adversarial Nets[J/OL]. Ithaca, NY: arXiv.org, 2014(2014,10,6)[2019/6/15]. https://arxiv.org/abs/1411.1784.

[12] Chen, Q., Koltun, V. Photographic image synthesis with cascaded refinement networks[C]. IEEE International Conference on Computer Vision, Venice, 2017.

[13] Hertzmann A , Jacobs C E , Oliver N , et al. Image analogies[C]. Annual conference on Computer graphics and interactive techniques, New York. 2001: 327-340.

[14] A. A. Efros and T. K. Leung. Texture synthesis by non-parametric sampling[C]. IEEE International Conference on Computer Vision, Corfu, 1999.

[15] Efros A A , Leung T K . Texture Synthesis by Non-Parametric Sampling[C]. IEEE International Conference on Computer Vision, Corfu. 1999.

[16] J. Long, E. Shelhamer, T. Darrell. Fully convolutional networks for semantic segmentation[C]. IEEE Conference on Computer Vision and Pattern Recognition, Boston. 2015.

[17] P. Sangkloy, J. Lu, C. Fang, et al. Scribbler: Controlling deep image synthesis with sketch and color[C]. IEEE Conference on Computer Vision and Pattern Recognition, Honolulu. 2017.
[18] P. Isola, J.-Y. Zhu, T. Zhou, et al. Imageto-image translation with conditional adversarial networks[C]. IEEE Conference on Computer Vision and Pattern Recognition, Honolulu. 2017.
[19] L. Karacan, Z. Akata, A. Erdem, et al. Learning to generate images of outdoor scenes from attributes and semantic layouts[J/OL]. Ithaca, NY: arXiv.org, 2016(2016,12,1)[2019/6/15]. https://arxiv.org/abs/1612.00215.
[20] Jun-Yan Zhu, Taesung Park, Phillip Isola, et al. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks[J/OL]. Ithaca, NY: arXiv.org, 2017(2017. 5, 30)[2019/6/15].
https://arxiv.org/abs/1703.10593.
[21] R. Rosales, K. Achan, B. J. Frey. Unsupervised image translation[C]. ICCV, Nice, 2003.
[22] M.-Y. Liu, O. Tuzel. Coupled generative adversarial networks[C]. Advances in Neural Information Processing Systems, Barcelona, 2016.
[23] M.-Y. Liu, T. Breuel, J. Kautz. Unsupervised image-to-image translation networks[C]. Advances in Neural Information Processing Systems, Long Beach, 2017.
[24] D. P. Kingma ,M. Welling. Auto-encoding variational bayes[C]. International Conference on Learning Representations, Banff, 2014.
[25] A. Shrivastava, T. Pfister, O. Tuzel, et al. Learning from simulated and unsupervised images through adversarial training[C]. IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, 2017.
[26] Y. Taigman, A. Polyak, L. Wolf. Unsupervised cross-domain image generation[C]. International Conference on Learning Representations, Toulon, 2017.
[27] K. Bousmalis, N. Silberman, D. Dohan, et al. Unsupervised pixel-level domain adaptation with generative adversarial networks[C]. IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, 2017.
[28] Z. Kalal, K. Mikolajczyk, J. Matas. Forwardbackward error: Automatic detection of tracking failures[C]. International Conference on Pattern Recognition,Istanbul, 2010.
[29] N. Sundaram, T. Brox, K. Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow[C]. European Conference on Computer Vision, Crete, 2010.
[30] R. W. Brislin. Back-translation for cross-cultural research[J]. Journal of cross-cultural psychology. 1970, 1(3): 185–216.
[31] C. Zach, M. Klopschitz, M. Pollefeys. Disambiguating visual relations using loop constraints[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, 2010.
[32] Q.-X. Huang ,L. Guibas. Consistent shape maps via semidefinite programming[J]. Symposium on Geometry Processing. 2013, 32(5): 177-186.
[33] F. Wang, Q. Huang, L. J. Guibas. Image cosegmentation via consistent functional maps[C]. IEEE International Conference on Computer Vision, Sydney, 2013.
[34] T. Zhou, Y. J. Lee, S. Yu, and A. A. Efros. Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences[C]. IEEE Conference on Computer Vision and Pattern Recognition, BOSTON, 2015.

[35] T. Zhou, P. Krahenbuhl, M. Aubry, et al. Learning dense correspondence via 3dguided cycle consistency[C]. IEEE Conference on Computer Vision and Pattern Recognition, LAS VEGAS, 2016.

[36] C. Godard, O. Mac Aodha, ,G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency[C]. IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, 2017.

[37] Z. Yi, H. Zhang, T. Gong, et al. Dualgan: Unsupervised dual learning for image-to-image translation[C]. IEEE International Conference on Computer Vision, Venice, 2017.

[38] D. He, Y. Xia, T. Qin,et al. Dual learning for machine translation[C]. Advances in Neural Information Processing Systems, Barcelona, 2016.

[39] L. A. Gatys, A. S. Ecker, M. Bethge. Image style transfer using convolutional neural networks[C]. IEEE Conference on Computer Vision and Pattern Recognition, LAS VEGAS, 2016.

[40] J. Johnson, A. Alahi, L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution[C]. European Conference on Computer Vision, Amsterdam, 2016.

[41] D. Ulyanov, V. Lebedev, A. Vedaldi, et al. Texture networks: Feed-forward synthesis of textures and stylized images[C]. International Conference on Machine Learning, New York City, 2016.

[42] He, K., Zhang, X., Ren, S., et al.Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition, LAS VEGAS,2016.

[43] Ulyanov D , Vedaldi A , Lempitsky V . Improved Texture Networks: Maximizing Quality and Diversity in Feed-forward Stylization and Texture Synthesis[C]. IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii,2017.

[44] D. Ulyanov, V. Lebedev, A. Vedaldi, et al. Texture networks: Feed-forward synthesis of textures and stylized images[C]. International Conference on Machine Learning, New York City, 2016.

[45] Huang, X., Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization[C]. IEEE International Conference on Computer Vision, Venice, 2017.

[46] T. Q. Chen ,M. Schmidt. Fast Patch-based Style Transfer of Arbitrary Style[J/OL]. Ithaca, NY: arXiv.org, (2016,12,13)[2019,6,15]. https://arxiv.org/abs/1612.04337.

[47] Wang, T.C., Liu, et al. Highresolution image synthesis and semantic manipulation with conditional gans[C]. IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake, 2018.

[48] Mao, X., Li, et al. Least squares generative adversarial networks[C]. IEEE International Conference on Computer Vision, Venice, 2017.

[49] Simonyan, K., Zisserman, et al. Very deep convolutional networks for large-scale image recognition[C]. International Conference on Learning Representations, San Diego, 2015.

[50] Isola, P., Zhu, J.Y., Zhou, et al. Image-to-image translation with conditional adversarial networks[C]. IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, 2017.

[51] Kingma, D.P., Ba, et al. Adam: A method for stochastic optimization[C]. International Conference on Learning Representations, San Diego, 2015.

[52] Zhang R , Isola P , Efros A A , et al. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric[C]. IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake, 2018.

[53] Krizhevsky A , Sutskever I , Hinton G . ImageNet Classification with Deep Convolutional Neural Networks[C]. Advances in Neural Information Processing Systems, Lake Tahoe, 2012.

[54] Salimans T., Goodfellow I., Zaremba W., et al. Improved techniques for training gans[C]. Advances in Neural Information Processing Systems, Barcelona, 2016.

[55] Szegedy C., Vanhoucke V., Ioffe S., et al. Rethinking the inception architecture for computer vision[C]. IEEE Conference on Computer Vision and Pattern Recognition, LAS VEGAS, 2016.