Journal of Computer Applications
计算机应用, 2020, 40(6): 1621 - 1626
ISSN 1001-9081
CODEN JYIIDU
2020-06-10
http://www.joca.cn

文章编号:1001-9081(2020)06-1621-06
DOI:10.11772/j.issn.1001-9081.2019101802

基于生成对抗双网络的虚拟到真实驾驶场景的视频翻译模型

刘士豪,胡学敏*,姜博厚,张若晗,孔　力
(湖北大学 计算机与信息工程学院,武汉 430062)
(*通信作者电子邮箱huxuemin2012@hubu.edu.cn)

摘　要:针对虚拟到真实驾驶场景翻译中成对的数据样本缺乏以及前后帧不一致等问题,提出一种基于生成对抗网络的视频翻译模型。为解决数据样本缺乏问题,模型采取"双网络"架构,将语义分割场景作为中间过渡分别构建前、后端网络。在前端网络中,采用卷积和反卷积框架,并利用光流网络提取前后帧的动态信息,实现从虚拟场景到语义分割场景的连续的视频翻译;在后端网络中,采用条件生成对抗网络框架,设计生成器、图像判别器和视频判别器,并结合光流网络,实现从语义分割场景到真实场景的连续的视频翻译。实验利用从自动驾驶模拟器采集的数据与公开数据集进行训练和测试,在多种驾驶场景中能够实现虚拟到真实场景的翻译,翻译效果明显好于对比算法。结果表明,所提模型能够有效解决前后帧不连续和动态目标模糊的问题,使翻译的视频更为流畅,并且能适应多种复杂的驾驶场景。

关键词:虚拟到真实;视频翻译;生成对抗网络;光流网络;驾驶场景
中图分类号:TP391.4　　**文献标志码**:A

Video translation model from virtual to real driving scenes based on generative adversarial dual networks

LIU Shihao, HU Xuemin*, JIANG Bohou, ZHANG Ruohan, KONG Li
(*School of Computer Science and Information Engineering, Hubei University, Wuhan Hubei* 430062, *China*)

Abstract: To handle the issues of lacking paired training samples and inconsistency between frames in translation from virtual to real driving scenes, a video translation model based on Generative Adversarial Networks was proposed in this paper. In order to solve the problem of lacking data samples, the model adopted a "dual networks" architecture, where the semantic segmentation scene was used as an intermediate transition to build front-part and back-part networks, respectively. In the front-part network, a convolution network and a deconvolution network were adopted, and the optical flow network was also used to extract the dynamic information between frames to implement continuous video translation from virtual to semantic segmentation scenes. In the back-part network, a conditional generative adversarial network was used in which a generator, an image discriminator and a video discriminator were designed and combined with the optical flow network to implement continuous video translation from semantic segmentation to real scenes. Data collected from an autonomous driving simulator and a public data set were used for training and testing. Virtual to real scene translation can be achieved in a variety of driving scenarios, and the translation effect is significantly better than the comparative algorithms. Experimental results show that the proposed model can handle the problems of the discontinuity between frames and the ambiguity for moving obstacles to obtain more continuous videos when applying in various driving scenarios.

Key words: virtual to real; video translation; Generative Adversarial Networks (GAN); optical flow network; driving scene

0　引言

自动驾驶是人工智能的主要研究领域之一。目前,由于技术的不成熟和稳定性不够等原因,自动驾驶的最新方法往往只能在虚拟场景中训练和测试[1],只有少数成熟的技术才能在实际场景中测试。然而,由于虚拟场景不同于真实场景,在色彩、纹理等方面和真实场景存在一定的差别,在虚拟场景中训练的模型难以较好地适应真实场景,泛化能力较差,而在真实场景中训练自动驾驶模型,成本较高且具有一定的风险。图像翻译是将图像的内容从一种展现形式转换成另一种展现形式,在图像美化、风格迁移、场景设计和视频特效等方面取得了较好的研究成果[2]。如果能将图像翻译技术应用于自动驾驶中虚拟场景到真实场景的视频翻译,不仅能解决自动驾驶模型训练泛化能力差的问题,而且能有效减少训练的成本

收稿日期:2019-10-24;**修回日期**:2019-12-11;**录用日期**:2019-12-20。
基金项目:国家自然科学基金资助项目(61806076);湖北省自然科学基金资助项目(2018CFB158);湖北省大学生创新创业训练计划项目(S201910512026);湖北大学楚才学院大学生科学研究项目(20182211006)。
作者简介:刘士豪(1999—),男,湖北天门人,主要研究方向:计算机视觉;　胡学敏(1985—),男,湖南岳阳人,副教授,博士,主要研究方向:计算机视觉、机器学习;　姜博厚(1999—),男,湖北武汉人,主要研究方向:计算机视觉;　张若晗(1997—),女,湖北襄阳人,硕士研究生,主要研究方向:机器学习;　孔力(1995—),男,湖北咸宁人,硕士研究生,主要研究方向:深度学习。

和风险。因此，研究自动驾驶从虚拟场景到真实场景的视频翻译模型具有重要的学术意义和商业价值。

国内外的研究人员在图像翻译方面做了许多研究，并取得了一定的成果。传统的图像翻译方法主要是基于模型框架的构建和细节纹理的合成。Efros 等[3]对输入样本的纹理采集，并进行纹理拼接，然后合成新的风格的纹理；Hertzmann 等[4]采取多图像训练的方法，最终得到细节纹理生成模型，将模型应用于新的目标图像，合成一个相似的图像。这些方法能够有效合成新的风格的图像，但是这类非参数的图像翻译只能提取图像的底层特征，并非高层抽象特征。当用于自动驾驶场景这类具有复杂信息的图像时，如人、车道线等关键信息会有较多模糊和扭曲，难以满足实际的需求。

近年来，随着深度学习的兴起，许多极具创新性的方法应用在图像翻译领域，并取得了突破性的进展，这些方法主要分为两大类：第一大类是基于卷积神经网络（Convolutional Neural Network，CNN）的方法，这类方法通常是采取卷积神经网络对图像进行内容特征提取和风格特征提取，然后再通过内容重塑和风格重塑，实现图像的风格翻译。Gatys 等[5]利用深度卷积神经网络进行特征的提取，然后通过迭代优化的方式生成具有新风格的图像；Luan 等[6]在 Gatys 的基础上改进损失函数，使其只改变图像中目标物体的颜色，并不改变其他纹理特征。第二类是基于生成对抗网络（Generative Adversarial Network，GAN）的方法，这类方法通常是基于 GAN 及其衍生模型，旨在进一步提高控制图像风格的能力和生成图片的分辨率以及质量。Isola 等[7]提出的基于条件生成对抗网络的方法 pix2pix，可以实现任意风格图片之间的翻译；Wang 等[8]在 GAN 基础上提出的感知损失对抗网络，将普通对抗损失和感知损失结合，生成更加真实的图片。由于在训练时必须使用成对的图片来进行训练，而虚拟到真实场景的成对的图像数据集难以采集，所以无法用于驾驶场景的转换。Zhu 等[9]提出的 Cycle-GAN、Yi 等[10]提出的 DualGAN、Kim 等[11]提出的 DiscoGAN 都实现了任意风格之间的相互转换，且训练时并不需要成对的风格不同的图片，但由于模型没有关联前后帧的信息，使得在动态生成视频时，存在前后帧的关键信息不一致的现象，因此难以应用于自动驾驶中虚拟到真实场景的视频翻译。

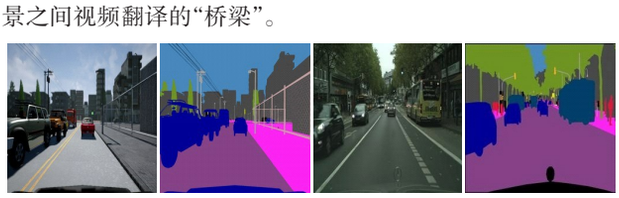
综上所述，图像场景翻译方法应用到自动驾驶场景翻译中存在的问题主要有两个：一是成对的虚拟场景和真实场景的图像样本缺乏；二是现有方法都是基于图像翻译的方法，没有考虑视频中前后帧的连贯性问题。针对这两种问题，本文提出一种基于生成对抗双网络（Generative Adversarial Dual Networks，GADN）的从虚拟到真实驾驶场景的视频翻译模型。该方法中，基于深度生成对抗网络，提出一个“双网络”的结构，把语义分割图像作为“桥梁”，将虚拟场景和真实场景连接起来，先将虚拟场景视频翻译成语义分割视频，然后再将语义分割视频翻译成真实场景视频，从而解决虚拟场景和真实场景成对的图像数据集缺乏的问题。此外，本文在双网络中采用光流法，并在生成对抗网络中设计图像判别器和视频判别器两个判别器，关联前后帧信息，解决视频前后帧的连贯性问题。本文方法既解决了自动驾驶模型的训练难题，又为视频翻译提供了新的解决方法。

1 虚拟到真实的驾驶场景视频翻译模型

为实现虚拟场景到真实驾驶场景的翻译，本文首先将语义分割图像作为中间桥梁，提出一种“双网络”的视频翻译框架。其中，前端网络实现从虚拟场景到语义分割场景的视频翻译，后端网络实现从语义分割场景到真实场景的视频翻译。然后基于该框架，利用深度生成对抗网络，设计前端网络和后端网络的结构。

1.1 “双网络”的视频翻译框架

图像的语义分割是对分割后的图像加上语义标签，一般是用不同的颜色代表不同类别的物体。图 1(a)是从 Carla 自动驾驶模拟器[12]采集的驾驶场景的虚拟帧和其对应的语义分割图，图 1(b)是从 Cityscapes 数据集[13]中获取的驾驶场景的真实帧和其对应的语义分割图，并且对两者用统一标准进行语义分割。从图 1 中可以看出，虽然驾驶的虚拟场景和真实场景之间在光线、色彩、纹理等方面存在着较大差异，但是如果对两者进行语义分割处理，则处理之后的图片具有一致的风格。并且，语义分割图像中不包含光线、色彩和纹理等细节，但是又涵盖了自动驾驶所需的道路、车辆、建筑、行人等目标内容，因此本文采用语义分割图像作为虚拟场景和真实场景之间视频翻译的“桥梁”。



(a) 虚拟场景及其语义分割　　(b) 真实场景及其语义分割

图 1　虚拟场景和真实场景的语义分割

Fig. 1　Semantic segmentations of virtual and real scenes

基于“虚拟场景-语义分割场景-真实场景”的思想，本文设计一种“双网络”的视频翻译模型，如图 2 所示。该网络分为前端网络和后端网络两部分，前端网络负责将虚拟场景翻译为语义分割场景，后端网络负责将语义分割场景翻译成真实场景。通过“双网络”模型，首先将虚拟场景视频翻译成语义分割场景视频，然后将语义分割场景视频翻译成真实场景视频，从而实现虚拟场景到真实场景的视频翻译。

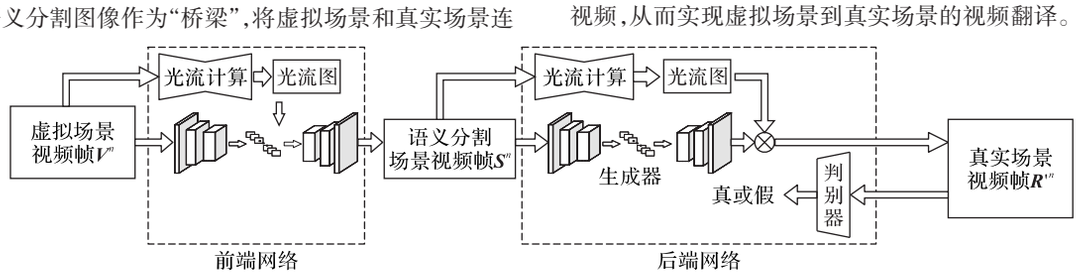


图 2　虚拟到真实的视频翻译模型

Fig. 2　Video translation model from virtual to real scenarios

在前端网络中，向网络输入虚拟驾驶场景的视频帧序列 $\boldsymbol{V}^n=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_n\}$，通过卷积网络提取每一帧的特征。为保障语义分割的前后帧之间的一致性，计算每相邻两帧之间的光流。生成当前帧时，前一帧的特征中融入前一帧到当前帧

的光流,得到当前帧的特征,再通过反卷积网络得到当前帧的语义分割图。通过多帧的计算,生成语义分割场景视频帧 $\boldsymbol{S}^n=\{\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_n\}$。在后端网络中,输入生成的语义分割场景视频帧 $\boldsymbol{S}^n=\{\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_n\}$,通过卷积网络提取每一帧的特征。为解决前后帧不一致问题,本文采用光流网络计算相邻两帧之间的光流。生成当前帧时,将计算得到的光流与反卷积得到的帧融合,最后生成语义分割图对应的真实场景图。同样通过多帧的计算,生成真实场景视频帧 $\boldsymbol{R'}^n=\{\boldsymbol{r}'_1,\boldsymbol{r}'_2,\cdots,\boldsymbol{r}'_n\}$。

1.2　从虚拟到语义分割场景的前端网络

前端网络的作用是将虚拟场景视频翻译成语义分割场景视频。现有的针对视频的语义分割往往是对视频的每一帧图像单独进行语义分割,然而这种对每一帧都进行语义分割的方法计算量较大,同时也无法关联前后帧之间的信息。由于本文提出的双网络框架是基于连续帧的视频而非单帧图像,故设计一种动态语义分割前端网络,以实现快速高效的、前后帧一致的虚拟场景视频语义分割。

利用光流计算前后帧之间的关系,是视频分析的一种常用方法[14]。受到文献[14]启发,本文设计前端网络的思想为:在视频序列帧中选取一部分关键帧,在关键帧上采取直接通过常用的单帧图像语义分割方法进行语义分割得到关键帧的语义分割图;在当前帧上通过光流网络计算当前帧和关键帧之间的光流,通过关键帧的特征图和光流网络来预测当前帧的特征图,再反卷积生成当前帧的语义分割图。

假设一段虚拟驾驶场景序列有 N 帧(设 N 为 j 的倍数),将每连续的 j 帧为一段,则有 N/j 段连续帧。设 $m=N/j$,将每段中的第一帧作为关键帧,则一共有 m 帧关键帧。故关键帧的集合可以表示为 $\boldsymbol{V}_k^m=\{\boldsymbol{v}_{k1},\boldsymbol{v}_{k2},\cdots,\boldsymbol{v}_{km}\}$。将关键帧设为 $\boldsymbol{v}_{kn}$,其中 $n\in[1,N/j]$,当前帧设为 $\boldsymbol{v}_{kn+i}$,其中 $i\in[1,j]$。如图3所示,通过关键帧和当前帧的虚拟图像生成对应的语义分割图。

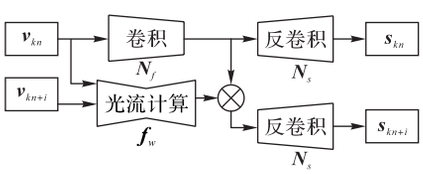


图3　视频翻译前端网络

Fig. 3　Front-part network for video translation

在生成关键帧的语义分割图时,通过卷积网络 $\boldsymbol{N}_f$ 对关键帧 $\boldsymbol{v}_{kn}$ 提取特征,再直接通过反卷积重建网络 $\boldsymbol{N}_s$ 得到关键帧的语义分割图 $\boldsymbol{s}_{kn}$。在生成当前帧的语义分割图时,通过光流网络 $\boldsymbol{f}_w$ 计算关键帧 $\boldsymbol{v}_{kn}$ 到当前帧 $\boldsymbol{v}_{kn+i}$ 之间的光流,并利用卷积网络 $\boldsymbol{N}_f$ 提取关键帧 $\boldsymbol{v}_{kn}$ 特征,再融合两者得到当前帧的特征,最后通过反卷积重建网络 $\boldsymbol{N}_s$ 得到当前帧的语义分割图 $\boldsymbol{s}_{kn+i}$。整个过程用公式可以表达为式(1)、(2):

$$\boldsymbol{s}_{kn}=\boldsymbol{N}_s(\boldsymbol{N}_f(\boldsymbol{v}_{kn})) \tag{1}$$

$$\boldsymbol{s}_{kn+i}=\boldsymbol{N}_s(W(\boldsymbol{N}_f(\boldsymbol{v}_{kn+i}),\boldsymbol{f}_w(\boldsymbol{v}_{kn},\boldsymbol{v}_{kn+i}))) \tag{2}$$

其中,$i\in[1,j]$,$n\in[1,m]$。W 是将关键帧的特征与关键帧和当前帧之间的光流融合的函数,本文采用的是双线性插值的融合方法。由于卷积、反卷积和光流计算已有成熟的网络模型,故本文中的特征提取卷积网络采用文献[15]中提出的网络,反卷积生成语义分割图的网络采用文献[16]的网络,光流网络采用文献[17]中提出的网络。为便于采用卷积网络,本文将处理的视频帧尺寸统一缩放为 256 × 256 像素。

1.3　从语义分割到真实场景的后端网络

本文提出的"双网络"中的后端网络的目的是实现语义分割场景视频为输入,真实场景视频为输出,如式(3)所示:

$$\boldsymbol{R}^n=\boldsymbol{G}(\boldsymbol{S}^n) \tag{3}$$

为了更好地达到这个目的,本文采用文献[18]中提到的条件生成对抗网络的方法作为基础模型。在原始GAN中,生成器 $\boldsymbol{G}$ 和判别器 $\boldsymbol{D}$ 不断博弈,生成器 $\boldsymbol{G}$ 学习到逼近真实样本数据的分布。条件GAN是对原始GAN的一个扩展,生成器和判别器都增加额外信息 y 为条件,y 可以是任意信息。条件GAN的目标函数是带有条件概率的极大极小值博弈。如式(4)所示:

$$\min_{\boldsymbol{G}}\max_{\boldsymbol{D}}\boldsymbol{V}(\boldsymbol{D},\boldsymbol{G})=E_{x\sim Pdata(x)}\left[\log\boldsymbol{D}\left(x|y\right)\right]+E_{z\sim P_z(z)}\left[\log\left(1-\boldsymbol{D}\left(\boldsymbol{G}(z)|y\right)\right)\right] \tag{4}$$

本文中的条件 y 为语义分割序列图 $\boldsymbol{S}^n$。由式(3)、(4)可以推导出后端网络训练的目标函数如式(5):

$$\min_{\boldsymbol{G}}\max_{\boldsymbol{D}}\boldsymbol{V}(\boldsymbol{D},\boldsymbol{G})=E_{(\boldsymbol{R}^n,\boldsymbol{S}^n)}\left[\log\boldsymbol{D}\left(\boldsymbol{R}^n|\boldsymbol{S}^n\right)\right]+E_{\boldsymbol{S}^n}\left[\log\left(1-\boldsymbol{D}\left(\boldsymbol{G}\left(\boldsymbol{S}^n\right)|\boldsymbol{S}^n\right)\right)\right] \tag{5}$$

其中:$\boldsymbol{G}$ 是生成器;$\boldsymbol{D}$ 是判别器;$\boldsymbol{R}^n$ 是真实图序列;$\boldsymbol{S}^n$ 是语义分割图序列。后端网络整体框架如图4所示。为关联前后帧之间的信息,采取将前后两帧输入生成器 $\boldsymbol{G}$,生成当前帧的真实图,通过多帧的计算,生成真实场景序列帧。判别器方面,采用两个判别器,图像判别器 $\boldsymbol{D}_{\mathrm{I}}$ 和视频判别器 $\boldsymbol{D}_{\mathrm{V}}$。图像判别器用来判别单帧生成图的真假,视频判别器用来判别生成真实场景序列的一致性。

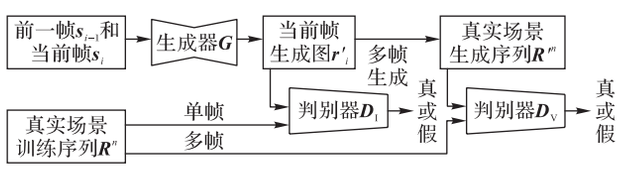


图4　视频翻译后端网络

Fig. 4　Back-part network for video translation

1)生成器设计。

在视频预测中,获得前一帧到当前帧的转换函数,则可以基于前一帧来预测当前帧。受到文献[19]的启发,转化函数可以由前一帧到当前帧的光流矢量来进行描述。因此,在总时间为 N 帧的序列中,时刻 i 的生成真实图像可以由三个因素得到:第 i 时刻的语义分割图 $\boldsymbol{s}_i$;过去 k 个时刻的语义分割图 $\boldsymbol{s}_{i-k+1},\boldsymbol{s}_{i-k+2},\cdots,\boldsymbol{s}_{i-1}$;过去 k 个时间由语义分割图生成的真实图 $\boldsymbol{r}'_{i-k+1},\boldsymbol{r}'_{i-k+2},\cdots,\boldsymbol{r}'_{i-1}$。

依据以上分析,本文设计的生成网络如图5所示。为使训练更加稳定,同时减少GPU显存的消耗,本文将 k 设为1,则第 i 帧生成的真实图取决于 $\boldsymbol{s}_{i-1}$、$\boldsymbol{s}_i$、$\boldsymbol{r}'_{i-1}$ 这3幅图像。在整段总时间为 N 的序列中,第 i 时刻输入第 i 帧的语义分割图 $\boldsymbol{s}_i$,通过单帧翻译网络,直接生成初步真实场景图 $\boldsymbol{i}_i$。然后通过光流网络计算第 $i-1$ 时刻和第 i 时刻的语义分割图的光流矢量 $\boldsymbol{f}_{i-1\to i}$,再输入第 $i-1$ 时刻生成的真实场景图 $\boldsymbol{r}'_{i-1}$。在 $\boldsymbol{r}'_{i-1}$ 基础上,通过光流矢量 $\boldsymbol{f}_{i-1\to i}$,计算得到扭曲图 $\boldsymbol{w}_i$,即利用光流场计算像素点在空间位置的偏移,得到经过像素点移位后的图像[17]。由于扭曲图 $\boldsymbol{w}_i$ 是在真实场景图 $\boldsymbol{r}'_{i-1}$ 基础上扭曲生成,会产生一定的模糊[20],这将使得最后效果不理想。此时,本文计算扭曲图 $\boldsymbol{w}_i$ 的模糊程度掩模作为权重 $\boldsymbol{\delta}$,加权平均初步真实场景图 $\boldsymbol{i}_i$ 和扭曲图 $\boldsymbol{w}_i$,得到最后生成的当前帧的真实场景图 $\boldsymbol{r}'_i$。整个步骤可以用式(6)来表达:

$$\boldsymbol{r}'_i=(\boldsymbol{1}-\boldsymbol{\delta})\odot\boldsymbol{w}_i+\boldsymbol{\delta}\odot\boldsymbol{i}_i \tag{6}$$

其中⊙为逐像素相乘运算符。为保证生成结果的真实程度,最后的生成结果由模糊程度掩模的值对扭曲图和直接生成图

加权平均得出。当 $\boldsymbol{w}_i$ 越模糊时，δ 越大，则 $\boldsymbol{w}_i$ 所占的比重越小，$\boldsymbol{i}_i$ 所占的比重越大，最后的生成图更倾向于 $\boldsymbol{i}_i$；反之，则更倾向于 $\boldsymbol{w}_i$。由于文献[7]提出的单帧图像翻译能够有效地将一类图像翻译成另外一类图像，因此本文采用文献[7]中改进的U-net结构作为单帧翻译网络的框架，掩模计算网络采用文献[15]中提出的网络。光流计算网络与前端网络一样，采用文献[17]中提出的FlowNet2。

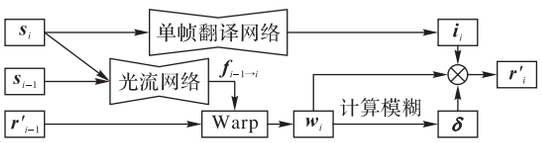


图5 后端网络的生成器模型

Fig. 5 Generator model in back-part network

2)判别器设计。

普通的条件GAN模型中只有单一的图像判别器，但是单一的图像判别器只能对单帧图像进行判别，将其用到视频翻译中时，由于没有考虑到前后帧之间的联系，会出现前后帧不一致的现象。因此，为了保证图像的真实性以及前后帧的一致性，本文将采用两个判别器：图像判别器 $\boldsymbol{D}_\mathrm{I}$ 和视频判别器 $\boldsymbol{D}_\mathrm{V}$，两个判别器都将采用条件判别器。

图像判别器 $\boldsymbol{D}_\mathrm{I}$：设置 $\boldsymbol{D}_\mathrm{I}$ 旨在使得图4中生成的真实场景图 $\boldsymbol{r}'_i$ 更加真实。采用文献[7]中的PatchGAN结构，将生成的初步结果图分割为很多小图像块Patch，然后判别图像中 $N\times N$ 个图像块是否为真。这个步骤通过卷积层来实现，逐次叠加卷积层的最终输出 $N\times N$ 的矩阵，最后通过平均所有的响应来得到最终输出。由于是基于生成的真实场景图来进行判别，故本文采用条件判别器，在每次判别时，从序列中随机抽取语义分割图和其对应的生成真实场景图和真实场景训练图。因此，图片损失函数如式(7)所示：

$$\boldsymbol{L}_\mathrm{I}=E_{(\boldsymbol{s}_i,\boldsymbol{r}'_i)}\left[\log\left(1-\boldsymbol{D}_\mathrm{I}(\boldsymbol{s}_i,\boldsymbol{r}'_i)\right)\right]+E_{(\boldsymbol{s}_i,\boldsymbol{r}_i)}\left[\log\left(\boldsymbol{D}_\mathrm{I}(\boldsymbol{s}_i,\boldsymbol{r}_i)\right)\right] \tag{7}$$

其中：$\boldsymbol{D}_\mathrm{I}$ 为文献[7]中的图像判别器；$\boldsymbol{r}'_i$ 表示生成的真实场景图；$\boldsymbol{r}_i$ 为实际的真实场景图；$\boldsymbol{s}_i$ 为对应的语义分割图。

视频判别器 $\boldsymbol{D}_\mathrm{V}$：设置 $\boldsymbol{D}_\mathrm{V}$ 旨在保证生成的连续帧之间的一致性。视频判别器基于相同的光流矢量，判别原本真实场景图和生成的真实场景图。在判别时，从序列中随机抽取出 t 帧的真实场景图以及 t 帧真实场景图中每相邻两帧之间的光流矢量图(共计是 $t-1$ 帧光流矢量图)，还有这 t 帧对应的语义分割图。因此视频损失函数表达式为(8)：

$$\begin{aligned}\boldsymbol{L}_\mathrm{V}=&E_{(\boldsymbol{r}'^{\,i-1}_{i-t},\boldsymbol{s}^{i-1}_{i-t},\boldsymbol{f}^{i-2}_{i-t})}\left[\log\left(1-\boldsymbol{D}_\mathrm{V}(\boldsymbol{r}'^{\,i-1}_{i-t},\boldsymbol{f}^{i-2}_{i-t})\right)\right]+\\&E_{(\boldsymbol{r}^{i-1}_{i-t},\boldsymbol{s}^{i-1}_{i-t},\boldsymbol{f}^{i-2}_{i-t})}\left[\log\boldsymbol{D}_\mathrm{V}(\boldsymbol{r}^{i-1}_{i-t},\boldsymbol{f}^{i-2}_{i-t})\right]\end{aligned} \tag{8}$$

其中：$\boldsymbol{r}^{i-1}_{i-t}$ 表示 t 帧的真实场景图；$\boldsymbol{s}^{i-1}_{i-t}$ 表示这 t 帧对应的语义分割图；$\boldsymbol{f}^{i-1}_{i-t}$ 表示 t 帧真实场景图中每相邻两帧之间的光流矢量图；$\boldsymbol{r}'^{\,i-1}_{i-t}$ 表示 t 帧的生成的真实场景图。$\boldsymbol{D}_\mathrm{V}$ 视频判别器采用文献[20]中提出的视频序列一致判别器。

3)目标函数。

在普通的GAN模型中，目标函数由判别器判别真实数据损失和判别由生成器生成的伪真实数据的损失组成。其中：$\boldsymbol{G}$ 需要最小化，使得判别器难以判别生成的数据，$\boldsymbol{D}$ 需要最大化，能够鉴别出数据的真假。在本文中，存在一个生成器和两个判别器，需要最小化 $\boldsymbol{G}$，最大化 $\boldsymbol{D}_\mathrm{I}$ 和 $\boldsymbol{D}_\mathrm{V}$。因此，本文提出的生成模型的优化目标函数最终设计为：

$$\min_{\boldsymbol{G}}\max_{\boldsymbol{D}_\mathrm{I}\&\boldsymbol{D}_\mathrm{V}}\boldsymbol{V}(\boldsymbol{D},\boldsymbol{G})=\min_{\boldsymbol{G}}\left(\max_{\boldsymbol{D}_\mathrm{I}}\boldsymbol{L}_\mathrm{I}(\boldsymbol{G},\boldsymbol{D}_\mathrm{I})+\max_{\boldsymbol{D}_\mathrm{V}}\boldsymbol{L}_\mathrm{V}(\boldsymbol{G},\boldsymbol{D}_\mathrm{V})\right) \tag{9}$$

其中：$\boldsymbol{G}$ 是生成模型中的生成器，用来生成连续帧的真实场景图；$\boldsymbol{L}_\mathrm{I}$ 是图像判别器的损失；$\boldsymbol{L}_\mathrm{V}$ 是视频判别器的损失。为使得训练更加稳定，采用最小二乘损失函数。在优化算法中，采用自适应矩估计(ADAptive Moment estimation，ADAM)法。

2 实验结果与分析

本文进行实验的硬件环境：CPU为Core i7-8700K(四核4.7 GHz)、GPU为NVIDA GTX 1080ti(显存11 GB)、内存为32 GB。软件环境：操作系统为Ubuntu、训练测试平台为PyTorch，编程语言为Python。

在数据集方面，本文使用从Carla自动驾驶模拟器平台[12]采集的数据集以及Cityscapes数据集[13]。其中Carla自动驾驶模拟器平台是Intel公司开发的开源模拟器，模拟器场景中有各种街道、树木、行人、车辆等；Cityscapes数据集是一个大规模的公开数据集，记录在50个不同城市的驾驶时街道场景。所选取的数据集全部都是连续段的视频帧序列，在Carla中采集虚拟图和其对应的语义分割图，在Cityscapes数据集中获取连续的真实场景视频帧和其对应的语义分割图。从Carla中采集的视频帧尺寸为1 024 × 512像素；Cityscapes数据集中视频帧尺寸为2 048 × 1 024像素。训练时，分别采集2 975段连续序列用于单独训练前、后端网络，同时将所有数据集尺寸调整到256 × 256。其中，前端网络训练集包含虚拟场景和对应的语义分割图，后端网络训练集包含真实场景和对应的语义分割图；测试时，额外采集300段连续序列(全部为虚拟场景)用于测试。训练和测试的每个片段包含30个视频帧。

为验证本文方法的有效性，本文对比两种经典的基于GAN的图像翻译模型：pix2pix[7]与Cycle-GAN[9]。Cycle-GAN可以实现任意两类图像的翻译，因此采用翻译视频中每一帧图像的方法进行对比；由于pix2pix方法只能实现语义图像到真实图像的翻译，因此本文采取“pix2pix+前端网络”(以下简称为“pix2pix”)的方式进行虚拟到真实场景的翻译。在评判标准方面，由于目前在图像翻译领域，若无两种不同风格成对的数据集，则往往从人眼观察的角度来进行效果对比[7-9,20]，因此本文将从图像定性方面对比实验结果。图6~8选取的分别从不同场景、连续性和动态目标三个方面来对比的实验结果。

从图6~8中可以得出以下结论：

1)本文方法可以有效地将多种不同虚拟场景翻译成真实场景，并且比现有方法更加清晰和准确。由于Cycle-GAN方法没有使用中间过渡而是直接翻译，生成器无法针对细节进行准确处理。而pix2pix方法与GADN方法都采用了“双网络”框架，使场景中每一个事物对应的语义得到更好的翻译，因此这两种方法都能准确地翻译每一个场景中的对应物体。如图6所示，在直行和T字路口场景中，方框标出的区域是天空，在Cycle-GAN方法的翻译效果中却是一片树林；在L型路口场景中，方框标出的是一片树林，在Cycle-GAN方法的翻译效果中却是一片树林和房屋的混合体；在有行人的场景中，方框标出的为行人，在Cycle-GAN方法的翻译效果中行人十分模糊；在有小车的场景中，方框标出的是小车，在Cycle-GAN方法的翻译效果中小车并没有被准确翻译出。

2)本文方法能实现前后帧图像的连续翻译，解决了传统图像翻译方法在视频翻译时前后帧不一致的问题。由于本文采用了关联前后帧的方法，在视频序列帧生成时加入光流信息，所以对比现有方法，在视频翻译中具有明显的优势。图7为一段连续视频帧的翻译对比结果，方框选取的目标为建筑房屋，在Cycle-GAN的翻译效果中并不准确，翻译成了树林；同时随着时间的推移，目标物体不停地在变换，模糊不清，出

现前后帧不一致现象。在 pix2pix 与 GADN 方法中房屋建筑的形状和大小则翻译地较为准确;但是由于 pix2pix 方法中也没考虑前后帧信息,因此随着时间的推移,房屋的颜色和形状在变换,前后帧并不连续;而只有在本文提出的 GADN 方法中则表现最好,房屋建筑的颜色和形状在前后帧之间的一致性较高。

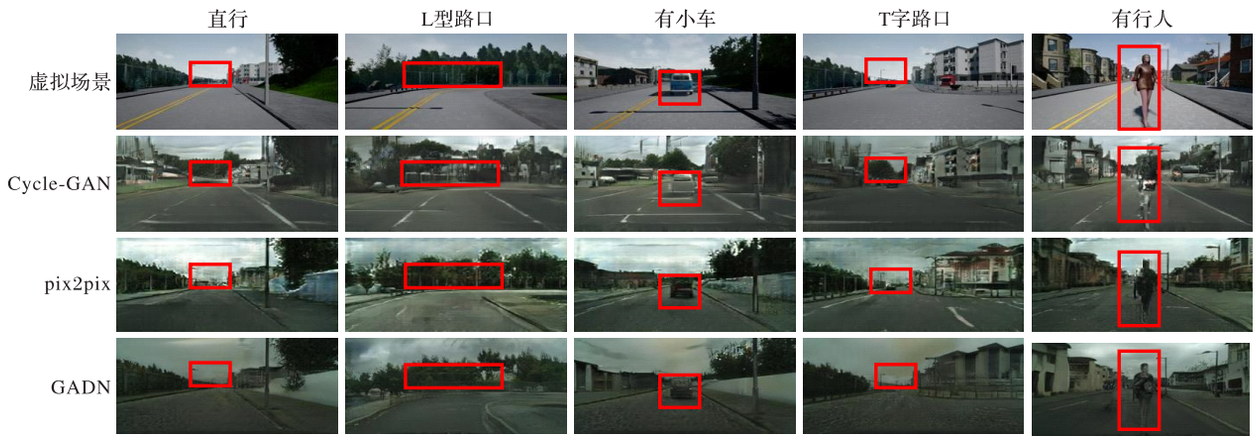


图 6　不同静态场景对比

Fig. 6　Comparative results in different static scenes

图 7　连续帧场景对比

Fig. 7　Comparative results in continuous frames

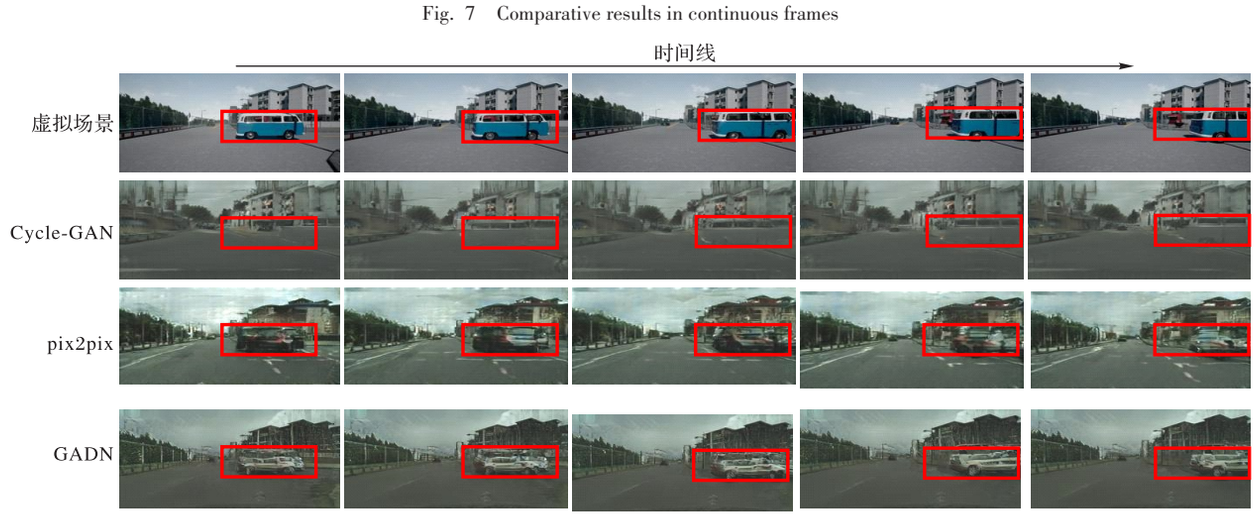


图 8　动态目标场景对比

Fig. 8　Comparative results in scenes with dynamic objects

3)本文方法能够有效解决视频翻译中动态目标不完整以及突变的问题。由于本文方法采用语义分割作为中间过渡,同时在视频序列帧生成时融入光流,所以本文的方法在有动态目标的场景中翻译效果更好。如图 8 所示,本车和目标车辆(方框标出)分别通过右转和左转进入 T 字路口车道。在 Cycle-GAN 方法的翻译效果中却并没有出现车辆;在 pix2pix 方法中,车辆目标虽然被翻译出来,但是其形态并不完整,且在前后帧出现突变的情况,导致翻译有误。只有本文提出

GADN 方法翻译的真实场景中动态车辆障碍物相对比较完整，且没有出现前后帧突变的情况，效果明显好于前两种方法。

3 结语

本文提出了一种基于生成对抗网络的“双网络”模型，并利用该模型实现从虚拟驾驶场景到真实场景的动态视频翻译。在该方法中，在输入端输入虚拟的驾驶场景序列帧，采用语义分割场景作为中间过渡，同时计算前后帧之间的光流来关联前后帧信息，实现在输出端输出真实驾驶场景视频帧。实验结果表明，本文提出的模型能有效地解决直接对视频每一帧图像翻译中存在的翻译图像不准确、前后帧不一致以及动态目标不完整等问题，在虚拟到真实驾驶场景的动态转换中具有较好的表现，适用于自动驾驶算法的训练和测试。同时，本文方法也存在一定的不足。由于前后端网络都需要用到光流网络等已训练好的网络，所以对训练要求较高。此外，本文方法主要是为自动驾驶强化学习中训练和测试服务，因此未来的工作将会致力于将本文方法用于自动驾驶中，实现本文方法的最大利用。

参考文献（References）

[1] 白丽赟，胡学敏，宋昇，等. 基于深度级联神经网络的自动驾驶运动规划模型[J]. 计算机应用，2019，39(10)：78-84.（BAI L Y，HU X M，SONG S，et al. Motion planning model based on deep cascaded neural networks for autonomous driving [J]. Journal of Computer Applications，2019，39(10)：78-84.）

[2] 陈淑環，韦玉科，徐乐，等. 基于深度学习的图像风格迁移研究综述[J]. 计算机应用研究，2019，36(8)：2250-2255.（CHEN S H，WEI Y K，XU L，et al. Survey of image style transfer based on deep learning [J]. Application Research of Computers，2019，36(8)：2250-2255.）

[3] EFROS A A，FREEMAN W T. Image quilting for texture synthesis and transfer [C]// Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. New York：ACM，2001：341-346.

[4] HERTZMANN A，JACOBS C E，OLIVER N，et al. Image analogies [C]// Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. New York：ACM，2001：327-340.

[5] GATYS L A，ECKER A S，BETHGE M. Texture synthesis using convolutional neural networks [EB/OL]. (2015-11-09) [2019-05-11]. https://arxiv.org/pdf/1505.07376.pdf.

[6] LUAN F J，PARIS S，SHECHTMAN E，et al. Deep photo style transfer [C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Washington，DC：IEEE Computer Society，2017：6997-7005.

[7] ISOLA P，ZHU J Y，ZHOU T，et al. Image-to-image translation with conditional adversarial networks [C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Washington，DC：IEEE Computer Society，2017：1125-1134.

[8] WANG C，XU C，WANG C，et al. Perceptual adversarial networks for image-to-image transformation [J]. IEEE Transactions on Image Processing，2018，27(8)：4066-4079.

[9] ZHU J Y，PARK T，ISOLA P，et al. Unpaired image-to-image translation using cycle-consistent adversarial networks [C]// Proceedings of the 2017 IEEE International Conference on Computer Vision. Washington DC：IEEE Computer Society，2017：2223-2232.

[10] YI Z，ZHANG H，TAN P，et al. DualGAN：unsupervised dual learning for image-to-image translation [C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Washington DC：IEEE Computer Society，2017：2849-2857.

[11] KIM T，CHA M，KIM H，et al. Learning to discover cross-domain relations with generative adversarial networks [C]// Proceedings of the 34th International Conference on Machine Learning. New York：JMLR. org，2017，70：1857-1865.

[12] DOSOVITSKIY A，ROS G，CODEVILLA F，et al. CARLA：an open urban driving simulator [EB/OL]. (2017-11-10) [2019-06-08]. https://arxiv.org/pdf/1711.03938.pdf.

[13] CORDTS M，OMRAN M，RAMOS S，et al. The cityscapes dataset for semantic urban scene understanding [C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Washington，DC：IEEE Computer Society，2016：3213-3223.

[14] ZHU X，XIONG Y，DAI J，et al. Deep feature flow for video recognition [C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Washington，DC：IEEE Computer Society，2017：2349-2358.

[15] HE K，ZHANG X，REN S，et al. Identity mappings in deep residual networks [C]// Proceedings of the 2016 European Conference on Computer Vision，LNCS 9908. Cham：Springer，2016：630-645.

[16] CHEN L C，PAPANDREOU G，KOKKINOS I，et al. DeepLAB：Semantic image segmentation with deep convolutional nets，atrous convolution，and fully connected CRFs [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2017，40(4)：834-848.

[17] ILG E，MAYER N，SAIKIA T，et al. FlowNet 2.0：Evolution of optical flow estimation with deep networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington，DC：IEEE Computer Society，2017：2462-2470.

[18] 陈文兵，管正雄，陈允杰. 基于条件生成式对抗网络的数据增强方法[J]. 计算机应用，2018，38(11)：3305-3311.（CHEN W B，GUAN Z X，CHEN Y J. Data augmentation method based on conditional generative adversarial net model [J]. Journal of Computer Applications，2018，38(11)：3305-3311.）

[19] OHNISHI K，YAMAMOTO S，USHIKU Y，et al. Hierarchical video generation from orthogonal information：optical flow and texture [C]// Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. Menlo Park：AAAI Press，2018：2387-2394.

[20] WANG T C，LIU M Y，ZHU J Y，et al. Video-to-video synthesis [C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems. Lake Tahoe，Nevada：Curran Associates Inc.，2018：1152-1164.

This work is partially supported by the National Natural Science Foundation of China (61806076)，the Natural Science Foundation of Hubei Province (2018CFB158)，the Undergraduate Innovation and Entrepreneurship Training Plan of Hubei Province (S201910512026)，the Student Science Research Project of Chucai Honors College of Hubei University (20182211006).

LIU Shihao，born in 1999. His research interests include computer version.

HU Xuemin，born in 1985，Ph. D.，associate professor. His research interests include computer vision，machine learning.

JIANG Bohou，born in 1999. His research interests include computer version.

ZHANG Ruohan，born in 1997，M. S. candidate. Her research interests include machine leaning.

KONG Li，born in 1995，M. S. candidate. His research interests include deep learning.