DOI：10.7652/xjtuxb201810020

应用残差生成对抗网络的路况视频帧预测模型

袁帅，秦贵和，晏婕
（吉林大学计算机科学与技术学院，130000，长春）

摘要：在路况视频帧的预测领域中，针对当前大部分模型所存在的预测图像分辨率低、图像模糊和局部细节缺失等问题，提出了一种应用残差生成对抗网络的路况视频帧预测模型(RB-GAN)。该模型用于在给定一段路况视频流的情况下更好地预测未来的一帧路况图像，应用多个级联的残差模块初步提取输入视频流的图像特征；利用感知网络强化对视频流中物体运动特征的提取；通过使用双重判别器提高生成对抗网络生成的图像的质量；用 Adam 方法来优化网络权值的深度学习过程。基于生成对抗网络这种半监督的学习框架，训练后的模型可以预测出一段路况视频流下一时刻的同输入视频流具有时空一致性的帧图像。应用车辆检测领域常用的 KITTI 数据集对生成对抗网络模型进行训练和测试，实验结果表明：与主要依赖于像素均值的方法相比，RB-GAN 模型预测图像的分辨率提高了 2～4 倍，达到 256 像素×512 像素，在图像锐度标准上提高了 1～2 个数量级，同时图像也更加符合人眼视觉的主观感受，所预测得到的路况视频帧图像质量更高，更具有实用性价值，可以更好地为诸如检测算法等其他下游算法提供有效的特征信息。

关键词：生成对抗网络；深度学习；自动驾驶；路况视频帧预测
中图分类号：TP38 **文献标志码**：A **文章编号**：0253-987X(2018)10-0146-07

A Prediction Model for Video Frames of Road Conditions Using Residual Blocks Generating Adversarial Networks

YUAN Shuai, QIN Guihe, YAN Jie
(College of Computer Science and Technology, Jilin University, Changchun 130000, China)

Abstract: A novel model to predict video frames of road conditions is proposed to solve the problem that in the prediction area of road condition video frames, most of predicting models are of such as low resolution, image blurring and lack of local details, and the model uses residual blocks to generate adversarial networks. Multi-cascade residual units are used to extract feature maps preliminarily and two discriminators are adopted to raise the quality of images generated by the GANs. The model power to extract feature maps is improved by using perceptual loss network, and the Adam optimizer is used to update network weights. The proposed model produces future frames in the next moment which have great temporal coherency with the input frames through using semi-supervised deep learning framework. The model is trained and tested using datasets KITTI, and the results show that the generated frames by the proposed model have the resolution of 256×512, which are 2 to 4 times higher than the results from pixel mean based functions and past foreign works, the sharpness accuracy increases by 1 to 2 magnitudes,

收稿日期：2018-03-28。 作者简介：袁帅(1992－)，男，硕士生；秦贵和(通信作者)，男，教授，博士生导师。 基金项目：吉林省重点科技攻关资助项目(20150204034GX)。
网络出版时间：2018-08-16 网络出版地址：http://kns.cnki.net/kcms/detail/61.1069.T.20180816.1705.008.html

and the generated frames are more responsive to human visual perception. The frame images predicted by the model are of higher quality and more practical value, and provide more effective features for other downstream models such as detection algorithms.

Keywords: generative adversarial networks; deep learning; auto driving; road condition frames prediction

深度学习的理论和技术极大地促进了诸多计算机视觉相关领域的发展，在机器人、自动驾驶、物体检测分类等方面都提供了新的思路方法，但是这类模型通常需要大量的人工处理费用高昂的标签化数据集去进行训练，缺乏标准化数据在某种程度上是深度学习领域研究的瓶颈，然而视频图像的预测研究则不受标签化数据集不足的限制，视频数据无需过多人工的预处理，数据体量趋近于无限。在自动驾驶等领域所使用的物体检测、分类的方法均是基于对静态图片的学习，如 YOLO、Faster RCNN 等模型对视频内容的学习本质上仍然是不断对单张图像进行物体检测，如果可以通过一段连续的视频帧去预测未来下一帧的图像将有助于提升视频分类等方法的效果。2015 年 Srivastava 等验证了预测得到视频中提取的特征图可以提高在两个行为识别数据库上分类的准确率[1]。

Srivastava 等应用以编码解码器(EDF)为基础结构、以长短期记忆网络(LSTM)为组成单元生成图像[2]。Bhattacharjee 等通过使用由 2 维卷积核构造的 2 级生成器来生成帧图像，在每一级生成器中通过图像反池化技术来提高图像分辨率[3]，使用了生成对抗模型的框架以及变种的归一化互相关算法、对比差异函数来共同控制网络权值的变化，但是其损失函数中对抗部分的损失权值设置的极小，生成对抗的框架思想在整体模型中起到的影响有限，主要还是依赖传统的像素级别的对比方法，如生成图像与真实图像像素的范式距离。Mathieu 等在视频帧的预测上同样使用了多级生成器的思想[4]，通过预测同一张图像不同尺度上的采样结果来逐步提高最终预测结果的分辨率。Liang 等通过构造一对互成对偶的生成对抗网络来分别对视频流中未来的帧和光流进行预测[5]，并通过将帧生成器直接生成的帧图像与由光流生成器生成的光流所包裹的视频流中的最后一帧图像进行融合，获取尺寸为 256 像素×256 像素的预测帧图像，并且在对抗训练的过程中帧生成器和光流生成器不仅会使用自身对应的判别器还会利用对方的判别器来获得梯度反馈，但是其实光流本身就是一种蕴含于原始图像数据中的可以由人工直接从两帧图像计算得出的信息，直接使用多个判别器对单一的帧图像生成器进行评价以提取更多有效特征。Xiong 等更为创造性地尝试通过一帧图像去直接预测未来连续的多帧图像[6]，同样也利用了对偶对抗网络的思想，但是不同于 Liang 等提出的网络模型直接通过两个生成器去生成和细化最终需要预测出的帧图像，而不是通过图像的后期融合。

上述方法大多应用在一定的假设条件下：如视频中主体对象运动变化幅度较小、没有突发事件、背景为静态、背景不参与图像质量评价，且仅在预测较小尺度图像时是可行的，但是当分辨率的要求提高、图像尺度增大时上述方法会产生模糊的结果。本文方法不基于以上假设，模型的训练与测试均使用 KITTI 训练集，数据集中路况多变，视频流中的物体对象变化剧烈，在评价最终预测结果时对预测生成的帧图像包括前景与背景在内的全部像素进行计算以衡量生成的帧图像的整体效果，一方面由于背景像素参与评价，同尺度图像参与峰值信噪比与结构一致性度量等指标计算的像素点数大幅增加，评价指标更容易受图像品质影响。另一方面，抛开相邻帧图像之间的运动特征提取的问题，仅图像生成本身就是一个复杂的过程，Ledig 等通过使用残差模块构成的生成器网络提高了可生成图像的分辨率[7]，然而在这个过程中图像信息的损失仍然不可避免。2017 年 Chen 等使用多级生成模型串连[8]的方式去解决这个问题，在每一级网络中重新输入与本级模型特征图分辨率相对应的图像数据来强化图像特征，然而用这种方法直接应用在视频帧的预测中却不足以提取出相邻帧之间的运动对象的运动变化特征。

本文提出一种基于生成对抗框架的端到端的网络模型(RB-GAN)，利用级联排布的残差模块构造生成器[9]，应用双判别器增强生成对抗网络捕捉图像细节纹理的能力，同时应用感知网络来进一步提取视频流中图像之间的运动特征，在保证生成内容与视频流具有时空一致性的前提下，能够生成更大分辨率的预测图像。

1 生成对抗网络模型

近年来 Goodfellow 等提出的生成对抗网络(GANs)成为图像检测领域研究的热点[10]。Mirza 等提出的条件对抗网络可以更加便捷地控制网络的生成结果[11]。Isola 等探索了利用生成对抗网络的框架[12],将 U-net 作为网络的生成器[13],从简单的语义图像生成复杂的现实世界图像,但是 GANs 局限于小尺度图像的生成,难以呈现复杂场景中的细节。对于生成对抗网络来说生成大尺度的图像仍然是一个难点,网络训练的不稳定性仍然困扰着研究者。生成对抗网络主要由生成器 G 和判别器 D 两部分构成,传统的生成对抗网络通过将一个噪音向量映射为输出图像,判别器则用来区分输入的数据是来自判别器还是来自生成器。在这里给出生成对抗网络的损失函数

$$\min_G \max_D \mathscr{L}_{\mathrm{GAN}}(A,B) \tag{1}$$

其中 $\mathscr{L}_{\mathrm{GAN}}(A,B)$的定义如下

$$\mathscr{L}_{\mathrm{GAN}}(A,B) = E(A,B)[\lg D(A,B)] + E(A)[1-\lg D(A,G(A))] \tag{2}$$

式中:A、B 为计算期望的因子。与 VAE 等深度网络相比,在生成图像的问题上原始的生成对抗网络能够生成人脸、室内等给多样化情景下的具有更高锐度的图像数据,其交替训练的生成器与判别器可以捕捉到更全面细致的训练数据的特征分布而不必依赖人工定义的特征,但是在生成更高分辨率的图像时,对生成器和判别器捕捉数据分布的能力提出了更高的要求:生成的图像不仅包含准确的全局特征(如本文中涉及到的街道、汽车、房屋等场景)也包含细致的局部特征(如本文中车辆表面的纹理、图案)。另一方面预测帧图像也不仅仅是单纯的生成一幅图像,更重要的是要捕捉输入的视频流中的运动特征以进行合理的预测,过去的研究在一定程度上证明了更深的网络模型和将单纯的几个卷积操作制作成有跨层连接的单元模块(如本文所用的残差模块)有利于捕捉图像特征[14],能够有效避免深度网络中梯度消失的问题,本文算法通过使用多个级联的残差模块实现这种深度生成器以生成具有复杂信息的图像,同时多尺度判别器的联合训练可以对全局和局部特征进行全面的捕捉[15],感知网络在图像风格转换领域上所取得的成功也启发了我们将其运用在运动特征的捕捉上[16],本文提出的模型通过引入一个深度网络作为感知网络使用,利用其中间层特征差异进一步加强梯度反馈以预测生成具有时空一致性的高分辨率视频帧图像。

1.1 生成器模型

定义 $A=\{A_0,A_1,A_2,\cdots,A_{N-1}\}$是一段视频流中的帧图像的集合,下标 $0\sim N-1$ 代表帧的时间序列,A^i 代表经过路况视频流中每一帧都经过 i 次降采样之后的数据版本,A_j^i 代表视频流 A 经过 i 次降采样的第 j 帧,类似的可以定义生成器网络模型预测出的帧图像集合 $\hat{B}=\{\hat{B}_0,\hat{B}_1,\hat{B}_2,\cdots,\hat{B}_{M-1}\}$,$B=\{B_0,B_1,B_2,\cdots,B_{M-1}\}$为生成器生成的帧图像数据所对应的真值,由此给出 $\hat{B}$ 的生成公式 $\hat{B}=G(A^0,A^1,A^2,\cdots,A^K)$。本实验中仅预测未来 1 帧,取 M 为 1。这里 K 的取值并不固定,通常取降采样 K 次后图像仍能保留充分的全局特征的 K 值。本实验中定义生成器为 G,G 分为两个部分:第 1 部分,如图 1 所示,将降采样次数最多版本的 A^{i+1}(图像尺度最小的版本)输入网络,经过 N 个残差卷积模块进行处理来提取特征,将生成的中间过程特征图传递给生成器网络模型第 2 部分的上采样单元,这里上采样的方式没有使用传统的反卷积操作,而是使用了双线性插值的方法,这样有助于避免在图像的生成过程中出现伪影,提高图像的生成质量。第 2 部分,如图 2 所示,数据经过 K 次上采样图像分辨率得到提升后再进行卷积操作提取特征,重复上采样和卷积的过程直至分辨率满足生成图像分辨率的需求,最后再利用一层映射卷积将提取的特征图映射为预测结果图。

1.2 多重判别器模型

传统的生成对抗网络采用一个单一的判别器,反馈给生成器网络的信息有限,尤其是生成包含光影纹理等现实世界中复杂的细节信息的图像。通过

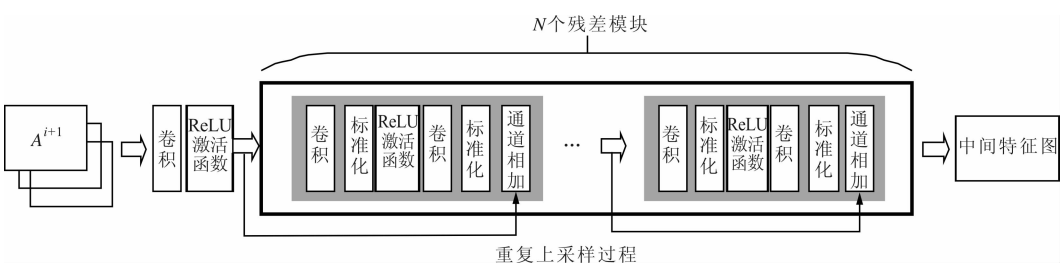

图 1 生成器网络模型的第 1 部分

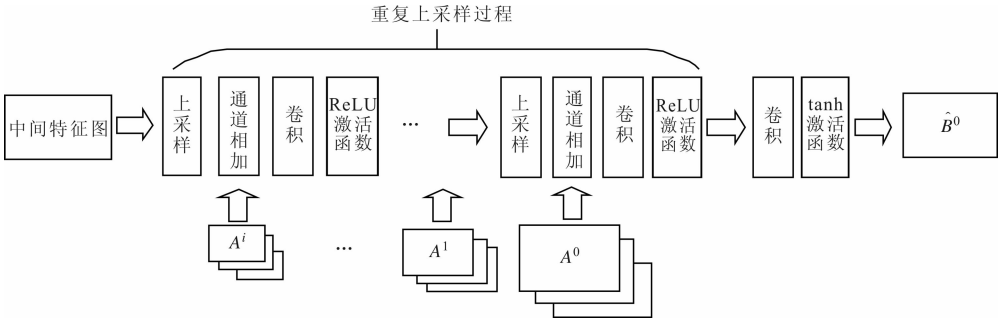

图 2　生成器网络模型的第 2 部分

采用了两个判别器，分别在不同尺度上对最终生成的结果进行评价，判别器网络结构如图 3 所示。本文模型不仅将判别器的最终输出结果用于损失函数，两个判别器网络的中间层的特征图也被提取出来用于比较不同输入的差异变化，这样一次网络的计算过程可以更加充分发挥判别器的作用。定义 D_1、D_2 分别为两个判别器，D_1 应用在 A、B 和 $\hat{B}$；D_2 应用在 A^1、B^1 和 $\hat{B}^1$（A、B 和 $\hat{B}$ 降采样了一次的版本）。首先给出本文中所使用的对抗网络的损失函数

$$\min_G \max_{D_1,D_2} \sum_{p=1,2} \mathscr{L}_{\mathrm{GAN}}(G,D_p) \tag{3}$$

判别器中间层的损失函数为

$$\mathscr{L}_{\mathrm{MF}}(A,B) = E(A,B)\sum_{p=1}^{2}\sum_{q=1}^{R}\frac{1}{W_q}\cdot \left[\left\| F_p^q(A,B) - F_p^q(A,B)\right)\right\|_\tau] \tag{4}$$

式中：W_q 为判别器网络的第 q 个 ReLU 激活层的元素数（设判别器网络共存在 R 个 ReLU 激活层）；p 为判别器的序号；F_p^q 为判别器网络 D_p 中第 q 个 ReLU 激活层所产生的特征图；τ 为两幅特征图所使用的距离度量方法，本文取 L1 方法。

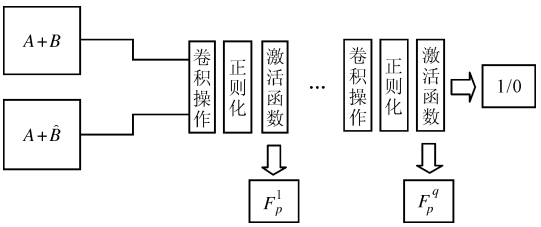

图 3　判别器网络结构

1.3　感知网络模型

Bhattacharjee 等提出 TCDM-GAN 模型用于预测一段分辨率为 128 像素×256 像素的视频流未来连续的多帧，但是该方法在高分辨率的情况下即使仅预测一帧图像误差也较大，网络收敛较慢。TCDM-GAN 模型进一步通过采用变种的归一化互相关（NCC）算法来提高预测图像的准确率，但这种方法假定图像中主体对象运动特征微小，并且需要人为设定 NCC 模板大小等关键参数。当 TCDM-GAN 模型应用在大尺度的图像上的时候，超参数的设定将会变得难以确定，如果过小生成器难以模拟出图像中物体运动的过程，如果过大图像容易模糊，并且在不同场景下无法自适应，如车辆在高速路和在行人较多的城市交通路行驶时，速度是不同的，反映到图像上物体运动变化的幅度也有很大差异，NCC 算法带来了很大的计算量，使训练变得缓慢。本文提出一个自适应的方法，不仅可以进一步提高图像的清晰度，还能提取出图像帧之间这种运动变化的特征。受 Johnson 等工作的启发，本文算法引入一个预训练的 VGG-19 模型作为整个模型的感知网络[17]，将预测的帧与对应的真值图像分别与输入序列的最后两帧拼接并分别输入 VGG-19 模型，模型分别提取出两类输入在 VGG-19 模型中的中间特征，并计算两类中间特征的 L1 距离作为网络损失函数的一部分，感知网络的损失函数为

$$\mathscr{L}_{\mathrm{PER}}(A_{N-1},A_{N-2},\hat{B},B) = \sum_{z=1}^{N}\frac{1}{M_z}[\left\| V_Z(A,B) - V_Z(A,G(A))\right\|_\theta] \tag{5}$$

式中：N 为 VGG 网络中 ReLU 激活层数；M_z 为第 z 个激活层所产生的特征图数；$V_Z(A,B)$ 代表第 z 个激活层在输入 A 与 B 的帧图像后所产生的 VGG 网络特征图。本文中 θ 取 L1 距离。

1.4　组合的损失函数

本文模型将生成器网络、判别器网络以及感知网络的损失函数加权后线性组合在一起，构成一个组合的损失函数并以此控制生成对抗网络梯度下降的过程。组合的损失函数定义如下

$$\mathscr{L}_{\mathrm{combin}}(A,B) = \lambda_{\mathrm{GAN}}\min_G\max_{D1,D2}\sum_{h=1,2}\mathscr{L}_{\mathrm{GAN}}(G,D_h) + \lambda_{\mathrm{PER}}\mathscr{L}_{\mathrm{PER}}(A_{N-1},A_{N-2},B,B) + \lambda_{\mathrm{MF}}\mathscr{L}_{\mathrm{MF}}(A,B) \tag{6}$$

2　实验数据集

本实验采用 KITTI 数据集作为训练集和测试

集[18]，此数据集包含城市、居民区、道路、学校等多个不同场景下的具有较高分辨率的视频流，并且视频已经被预处理为独立的帧图像。本实验取 KITTI 原始数据集中城市和居民区这两个类别中的数据集，使用其中的灰度图像，并将所有图像预处理为统一大小 1 024 像素×512 像素。这两类数据共 49 个视频流，共计 36 881 个图像，本实验中将其中 27 809 张图像作为训练集，余下 9 072 张图像作为测试集，检验模型训练结束后的准确率。

3 残差对抗网络实验

3.1 生成器结构设置

实验模型的输入为 4 张单通道 256 像素×512 像素的灰度图像，预测结果同样为 256 像素×512 像素的灰度图像。模型通过使用 3×3 的卷积核提取输入图像特征，将其映射到 1 024 通道的特征图，再经过 ReLU 激活函数处理。本实验取 7 个残差网络模块作为初始输入图像的特征提取。每个残差网络模块包含两个卷积层，并且在每个卷积层之后接一个规范化层，本模型不使用批规范化，而是实例规范化，实例规范化在生成清晰图像时有更好的效果[19]。并且第一个卷积层的实例规范化层之后还会额外接一个 ReLU 激活函数来增加整体网络的非线性程度。每个残差模块生成的特征图的通道数都是 1 024。每个残差模块的输入都采用映射方式填充 1 圈。残差模块最后生成的特征图经过上采样后的结果与对应尺度的输入图像在通道维度相加（残差网络输出 1024 通道的特征图，输入的为 4 张单通道的灰度图像，相加后为 1 028 通道的输入数据），结果作为下一层卷积层的输入数据，重复这两步操作至分辨率满足要求。每个卷积层仍然使用 3×3 的卷积核，搭配 ReLU 激活函数，第一层卷积结果为 64 像素×128 像素×512 像素，经过上采样与卷积操作后，特征图分辨率逐渐提升为 128 像素×256 像素×256 像素、256 像素×512 像素×128 像素，最后采用一个卷积核为 1×1 的映射层将特征图映射为 256 像素×512 像素的灰度图像。所有卷积步长均为 1。

3.2 判别器结构设置

本文所提出的变种的生成对抗模型使用两个判别器，除了输入数据的尺度不同外，网络结构完全相同：判别器具有 5 个卷积层，均采用 4×4 的卷积核，前 3 层不采用池化技术而用卷积核移动步长为 2 的操作代替以达到更好的保留特征和减小中间过程特征图尺寸的目的。第 4、5 层卷积操作的移动步长更改为 1 以进行充分的提取特征。生成的特征图的通道数分别为 64、128、256、512、1。判别器每层输入均使用常量 0 填充 1 圈。

3.3 对抗训练设置

实验中生成器、判别器交替训练。原始的生成对抗网络的训练不稳定，训练过程容易崩溃，在实验中使用了 Wasserstein GAN[20] 模型作为对抗训练的基础，Wasserstein GAN 模型具有更好的稳定性和更加合理的对抗结构。本网络模型不使用随机梯度下降优化器（SGD）而是选择 Adam 作为优化器[21]，初始学习律设置为 0.00 2，全部数据训练完 100 次前保持不变，接下来的 100 次学习率线性衰减到 0。实验中感知器网络采用 VGG19 网络。损失函数权值分别取 $\lambda_{GAN}=1$、$\lambda_{PER}=10$、$\lambda_{MF}=10$。

3.4 模型评价方法

本文使用评价模型常用的峰值信噪比（PSNR，设为 P）和结构相似性度量（SSIM，设为 S）的方法来评价结果，在以往的工作中，往往只关注网络输出结果在运动区域的指标，将背景视为静态的或者忽略背景较小的运动变化，这是不合适的，因为从实用的角度出发，运动特征的捕捉虽然重要，但是生成的图像整体效果（包括背景和前景）是否逼真也影响图像最终的实用性。所以在实验中对完整的图像进行评估比较。峰值信噪比的计算方式如下

$$P = 10\lg\left(\frac{W^2}{R}\right) \tag{7}$$

式中：R 是预测的图像与对应真值的均方误差，W 为图像强度最大值点，本文 R 取值 255.0。

通过采用结构相似性指标 S 来进一步衡量模型生成的图像与真实图像在结构上的相似程度，实验中所采用的结构相似性度量的计算方式如下

$$S(A,B) = \frac{2(\mu_A\mu_B + C_1)(2\sigma_{AB} + C_2)}{(\mu_A^2 + \mu_B^2 + C_1)(\sigma_A^2\sigma_B^2 + C_2)} \tag{8}$$

式中：μ_A、μ_B 分别为图像 A、B 的像素值均值，σ_A、σ_B 分别为图像 A、B 的像素值标准差，取超参数 $C_1=(Q_1S)^2$，$C_2=(Q_2S)^2$，取 Q_1 为 0.01，Q_2 为 0.03，S 为 255.0（即像素强度的最大值）。上述的传统评测方法计算上均过于依赖于图像数据的均值，无法真实表现出图像的清晰度与锐度，与人眼主观感受差距较大，本文通过采用 Mathieu 提出的 Sharpness 方法来评价图像锐度，设为 Y，通过计算图像相邻像素数值的差异给出了一种更加符合人眼主观视觉感受的图像评价指标，锐度的计算方式如下

$$Y(A,\hat{A})=\frac{\sum_i\sum_j|A_{i,j}-A_{i-1,j}|+|A_{i,j}-A_{i,j-1}|}{\sum_i\sum_j|\hat{A}_{i,j}-\hat{A}_{i-1,j}|+|\hat{A}_{i,j}-\hat{A}_{i,j-1}|} \quad (9)$$

式中：A 为生成图像真值，$\hat{A}$ 为生成图像，生成图像越接近真值，锐度越高，Y 越接近于 1。

3.5　与传统方法的对比

在 TensorFlow1.4 平台下使用一块英伟达 1080ti 显卡进行实验。实验以连续 3 帧单通道灰度图像作为输入（$T=1,2,3$），通过使用不同的模型方法训练 200 轮后预测出第 4 帧图像作为最终结果（$T=4$）。实验输入输出的效果如图 4 所示，预测的准确率结果见表 1。从图 4 中可以明显看出，RB-GAN 模型生成了最清晰的结果。从表 1 中可以看出，本文提出的 RB-GAN 模型在预测未来一帧时生成的图像虽然在 P 指标上低于其他方法，尤其是低于以计算像素数值距离为主要策略的 L1 和 L2 方法，但在图像锐度上则明显高于其他方法，锐度指标趋近于 1，较其他方法高出了 1～2 个数量级，更好地还原图像的细节纹理部分（图 4 中方框区域被局部放大，便于观察图像细节）。

应当注意到 P 取得最好效果的 L1 方法其本身就是完全基于两幅图像像素数值的绝对距离计算梯度，方法的本质就是令生成的图像与真值图像在整体像素值上尽可能接近而不关心生成图像像素彼此之间的相对关系（所以这类方法都倾向于生成模糊的图像，两幅图像每一对像素点之间都存在一定差异，但是整体的像素值差异很小），因此在使用 L1 和 L2 方法时即使生成的图像模糊，这些图像的数值指标也极高，尤其是图像尺度变大的时候，这种现象就更加明显，这也是峰值信噪比这种基于计算两幅图像像素均方误差的衡量指标的弊端，而结构相似性度量指标采用了更加丰富的构成方式，在计算的过程中采用图像均值、标准差及其他超参数，从表 1 中可以看出，本文方法在 S 这一指标上与其他方法是较为接近的。而在不依赖于均值、标准差等反映平均特性的 Y 指标上我们取得了数量级上的优势，同时基于这项评价指标所生成的图像也更加符合人眼主观感受。

表 1　实验预测帧的各项指标对比

方法	P	S	$\lvert Y-1\rvert$
L1	23.203	0.971	0.858
L2	20.352	0.936	1.773
TCDM-GAN	20.931	0.950	1.331
RB-GAN	18.929	0.926	0.045

4　结　论

本文提出了一种基于生成对抗网络框架利用残差模块构造的网络模型，可以有效预测一段路况视频在时序上未来的一帧图像，避免了传统深度网络模型过多的难以手动设置的超参数。以往大部分研

(a) $T=1$ 的输入图像　(b) $T=2$ 的输入图像　(c) $T=3$ 的输入图像　(d) $T=4$ 的预测图像真值

(e) $T=4$ 的 L1 方法预测结果　(f) $T=4$ 的 L2 方法预测结果　(g) $T=4$ 的 TCDM-GAN 模型预测结果　(h) $T=4$ 的 RB-GAN 模型预测结果

(i) $T=4$ 的 L1 方法局部图　(j) $T=4$ 的 L2 方法局部图　(k) $T=4$ 的 TCDM-GAN 模型局部图　(l) $T=4$ 的 RB-GAN 模型局部图

图 4　4 种方法预测结果的对比实验

究工作均以生成 256 像素×256 像素的图像为基本目标，而本文模型生成的图像的分辨率达到了 256 像素×512 像素，比过去相关研究所预测的图像的分辨率高 2～4 倍，输出的帧图像与输入的一系列帧图像在时空上具有良好一致性。在更加严格的 Sharpness 测评标准下本文提出的模型的准确率提高了一个数量级，达到了 0.045，图像的光影纹理等细节呈现的更加丰富准确，符合人眼主观感受。

参考文献：

[1] SRIVASTAVA N, MANSIMOV E, SALAKHUDINOV R. Unsupervised learning of video representations using LSTMS [C]// Proceedings of the 32nd International Conference on Machine Learning. New York, USA: ACM, 2015: 843-852.

[2] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[3] BHATTACHARJEE P, DAS S, BHATTACHARJEE P, et al. Temporal coherency based criteria for predicting video frames using deep multi-stage generative adversarial networks [C]// Proceedings of the Advances in Neural Information Processing Systems. New York, USA: NIPS, 2017: 4271-4280.

[4] MATHIEU M, COUPRIE C, LECUN Y. Deep multi-scale video prediction beyond mean square error [EB/OL]. (2015-11-17) [2018-02-28]. https://arxiv.org/abs/1511.05440.

[5] LIANG X, LEE L, DAI W, et al. Dual motion GAN for future-flow embedded video prediction [C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ, USA: IEEE, 2017: 1762-1770.

[6] XIONG W, LUO W, MA L, et al. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks [EB/OL]. (2017-09-22) [2018-02-28]. https://arxiv.org/abs/1709.07592.

[7] LEDIG C, THEIS L, HUSZÁR F, et al. Photo-realistic single image super-resolution using a generative adversarial network [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2017: 105-114.

[8] CHEN Q, KOLTUN V. Photographic image synthesis with cascaded refinement networks [C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ, USA: IEEE, 2017: 1520-1529.

[9] YANG C, LU X, LIN Z, et al. High-resolution image inpainting using multi-scale neural patch synthesis [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2017: 4076-4084.

[10] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks [C]// Proceedings of the Advances in Neural Information Processing Systems. New York, USA: NIPS, 2014: 2672-2680.

[11] MIRZA M, OSINDERO S. Conditional generative adversarial nets [EB/OL]. (2014-11-06) [2018-02-28]. https://arxiv.org/abs/1411.1784.

[12] ISOLA P, ZHU J Y, ZHOU T, et al. Image-to-image translation with conditional adversarial networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2017: 5967-5976.

[13] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation [C]// Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Berlin, Germany: Springer, 2015: 234-241.

[14] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognitio. Piscataway, NJ, USA: IEEE, 2016: 770-778.

[15] WANG T C, LIU M Y, ZHU J Y, et al. High-resolution image synthesis and semantic manipulation with conditional GANs [EB/OL]. (2017-11-30) [2018-02-25]. https://arxiv.org/abs/1711.11585.

[16] JOHNSON J, ALAHI A, LI F F. Perceptual losses for real-time style transfer and super-resolution [C]// Proceedings of the European Conference on Computer Vision. Berlin, Germany: Springer, 2016: 694-711.

[17] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2014-09-04) [2018-02-24]. https://arxiv.org/abs/1409.1556.

[18] GEIGER A, LENZ P, STILLER C, et al. Vision meets robotics: the KITTI dataset [J]. Journal of Robotics Research, 2013, 32(11): 1231-1237.

（下转第 166 页）