计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20190479
57(7): 1522-1530, 2020

# 基于多模态输入的对抗式视频生成方法

于海涛[1] 杨小汕[2] 徐常胜[1,2]
[1](合肥工业大学计算机与信息学院 合肥 230031)
[2](模式识别国家重点实验室(中国科学院自动化研究所) 北京 100190)
(yuht@mail.hfut.edu.cn)

# Antagonistic Video Generation Method Based on Multimodal Input

Yu Haitao[1], Yang Xiaoshan[2], and Xu Changsheng[1,2]
[1](*School of Computer and Information, Hefei University of Technology, Hefei* 230031)
[2](*National Laboratory of Pattern Recognition*(*Institute of Automation, Chinese Academy of Sciences*), *Bejing* 100190)

**Abstract** Video generation is an important and challenging task in the field of computer vision and multimedia. The existing video generation methods based on generative adversarial networks (GANs) usually lack an effective scheme to control the coherence of video. The realization of artificial intelligence algorithms that can automatically generate real video is an important indicator of more complete visual appearance information and motion information understanding. A new multi-modal conditional video generation model is proposed in this paper. The model uses pictures and text as input, gets the motion information of video through text feature coding network and motion feature decoding network, and generates video with coherence motion by combining the input images. In addition, the method predicts video frames by affine transformation of input images, which makes the generated model more controllable and the generated results more robust. The experimental results on SBMG (single-digit bouncing MNIST gifs), TBMG (two-digit bouncing MNIST gifs) and KTH (kungliga tekniska högskolan human actions) datasets show that the proposed method performs better on both the target clarity and the video coherence than existing methods. In addition, qualitative evaluation and quantitative evaluation of SSIM(structural similarity index) and PSNR(peak signal to noise ratio) metrics demonstrate that the proposed multi-modal video frame generation network plays a key role in the generation process.

**Key words** deep learning; video generation; video prediction; convolutional neural network; generative adversarial network (GAN)

**摘 要** 视频生成是计算机视觉和多媒体领域一个重要而又具有挑战性的任务.现有的基于对抗生成网络的视频生成方法通常缺乏一种有效可控的连贯视频生成方式.提出一种新的多模态条件式视频生成模型.该模型使用图片和文本作为输入,通过文本特征编码网络和运动特征解码网络得到视频的运动信息,并结合输入图片生成连贯的运动视频序列.此外,该方法通过对输入图片进行仿射变换来预测视频

**收稿日期**:2019-07-09;**修回日期**:2019-11-13
**基金项目**:国家重点研发计划基金项目(2018AAA0100604);国家自然科学基金项目(61702511,61720106006,61728210,61751211,U1836220,U1705262,61872424);模式识别国家重点实验室自主课题(Z-2018007)
This work was supported by the National Key Research and Development Program of China (2018AAA0100604), the National Natural Science Foundation of China (61702511, 61720106006, 61728210, 61751211, U1836220, U1705262, 61872424), and the Research Program of National Laboratory of Pattern Recognition (Z-2018007).
**通信作者**:杨小汕(xiaoshan.yang@nlpr.ia.ac.cn)

帧,使得生成模型更加可控、生成结果更加鲁棒.在 SBMG(single-digit bouncing MNIST gifs),TBMG(two-digit bouncing MNIST gifs)和 KTH(kungliga tekniska högskolan human actions)数据集上的实验结果表明:相较于现有的视频生成方法,生成结果在目标清晰度和视频连贯性方面都具有更好的效果.另外定性评估和定量评估(SSIM(structural similarity index)与 PSNR(peak signal to noise ratio)指标)表明提出的多模态视频帧生成网络在视频生成中起到了关键作用.

**关键词** 深度学习;视频生成;视频预测;卷积神经网络;生成对抗网络

**中图法分类号** TP391

视频的自动生成技术具有广泛的应用前景,例如视频编辑、增强现实、电影和游戏制作等.早期针对图像/视频生成技术的研究主要集中在计算机图形学领域展开[1-3].最先进的计算机图形学算法能够合成逼真的照片和视频,但这些技术需要依赖于专用的设计软件和大量专家的手工劳动,而且通常被限制在特定的人物、物体或者场景.近年来,随着深度学习技术在物体检测、行为识别等领域取得突破性进展[4-6],视频生成这一更具有挑战性的问题逐渐走入了计算机视觉和多媒体等领域研究人员的视野.实现能够自动生成真实视频的人工智能算法是更完备的视觉表观信息和运动信息理解的一个重要标志.

传统的深度神经网络分类模型需要监督式地在大规模标注样本上进行训练,而生成对抗网络(generative adversarial network, GAN)[7]通过对抗式地训练生成式网络和判别器网络来无监督地学习样本的特征分布,进而可以根据随机种子,生成真实样本.基于这一思想,GAN 在图片风格化和图像生成领域取得了优异性能[8-10],同时也被应用于视频生成并成为目前的主流方法[11-14].相比图像生成算法,视频生成是一项具有更多挑战的任务.尽管视频只比图像数据多了一个时间维度,但因此带来的运动信息的动态变化以及视觉内容的多样性都使得可能生成的结果空间变得十分巨大.此外视频是对执行各种动作的对象的视觉信息进行时空记录,生成模型除了要学习对象的外观模型外,还需要学习对象的物理结构.这些都是视频生成的困难所在.Vondrick 等人[15]把视频表示为潜在隐空间中的特征点,可以训练生成网络来表示从隐空间到视频片段的映射.Tulyakov 等人[12]把视频的潜在特征空间分解为运动子空间与内容子空间,大大减小了模型的复杂度.但由于这些方法是基于随机噪声生成视频,生成的视频存在视觉外观模糊不清、运动信息规律性不强的问题.

针对上述问题,大量基于条件式生成对抗网络的方法被提出.Li 等人[14]提出用自然语言作为输入条件来指导视频生成.虽然自然语言对描述视频中的关键内容和主要运动信息有很大的帮助,但是仅用语言作为条件,最终生成的视频难以准确表达物体/背景的细节信息以及长期的动态变化.在一些条件视频生成[16-18]中,运动轨迹、人脸 AUs(action units)值和语义图等信息分别被作为输入条件来指导视频内容生成.虽然这些方法在特定领域的视频上得到了较好结果,但这些输入条件的标注仍然需要较为专业的技术人员才能提供.

本文我们希望建立更为简单、有效的输入条件来得到更加鲁棒、可控的视频生成模型.为了提供充足的视觉外观信息,我们采用图片作为输入来表达视频中包含的主要物体和场景信息.考虑到自然语言是人类用于描述事物或者表达意图的最有效的工具,因此在运动信息方面,我们采用自然语言作为引导.基于以上讨论,我们提出基于图片和文本输入的多模态对抗式视频生成模型.一方面,我们将输入的文本信息通过循环神经网络进行编码来提取语义特征.这些语义特征将被解码为运动特征来辅助视频中的视觉信息生成和运动信息生成.另一方面,考虑到视频片段中的物体或者场景在较短时间内通常比较相似,我们学习输入图片到视频帧的仿射变换来得到更为准确和连贯的视频序列.由于缺乏运动信息的监督标签,我们采用了生成对抗网络捕捉帧与帧之间的运动信息,为特征提取网络提供反馈,使其能够生成连续有意义的运动特征.

本文的主要贡献是提出了一个新的多模态对抗式视频生成模型,将文本信息和图片信息同时引入视频生成,使得生成模型更加可控、生成结果更加鲁棒.

## 1 相关工作

我们简要地将相关工作分为两大类:图像生成

和视频生成,下面将分别围绕这 2 个方面详细介绍相关工作.

随着深度学习在图像分类和物体检测领域取得突破性进展,如何生成真实的图像在人工智能领域也得到了广泛的研究和分析.最早在 2014 年 Goodfellow 等人[7]提出了 GAN 网络的理论框架,利用 GAN 以无监督的方式生成图像.虽然早期的 GAN 为图像生成提供了一个独特而有前景的方向,但是生成结果存在模糊不清、细节丢失等问题.为了得到高质量的图像,Denton 等人[19]进一步将拉普拉斯金字塔引入 GAN.最近,Reed 等人[20]利用 GAN 基于给定的文本描述进行图像生成,实现了从字符级到像素级的翻译.Zhang 等人[21]将 2 个生成网络叠加在一起,逐步渲染出逼真的图像.CoupledGAN[8]构建了在不同域中生成图像的模型,可以无监督地将一个域中的图像转换为另一个域中的图像.InfoGAN[22]学习了一种更具解释性的隐特征来表示图像.Arjovsky 等人[23]提出了一种更稳定的对抗网络算法框架 Wasserstein GAN.

视频生成在计算机视觉领域并不是一个全新的问题.由于计算、数据和建模工具的限制,早期的视频生成工作侧重于生成动态纹理模式[1-3].近年来,随着 GPU(graphics processing unit)、网络视频和深度神经网络的出现,越来越多的基于深度学习的视频生成方法被提出.但要将对抗式图片生成模型扩展到视频,需要对空间和时间的复杂变化进行描述,这使得问题具有更多的挑战.最早基于 GAN 的视频生成模型是 Vondrick 等人[15]提出的,该算法采用时空 3D 反卷积分别生成前景与背景.最近,基于 GAN 的 3D 反卷积[13]被进一步分解为 1D 反卷积层和 2D 反卷积层来生成视频.同时大量基于条件式生成对抗网络的视频生成方法被提出.Li 等人[14]提出了用自然语言编码来指导视频生成.Marwah 等人[24]采用了循环的 VAE(variational autoencoder)和分层的注意机制来通过文本生成图像序列.Pumarola 等人[17]以人脸 AUs 值和图片作为输入,通过无监督的方式训练,并借助连续变化的 AUs 值生成动态的表情视频.Pan 等人[18]提出基于语义图的视频预测,通过语义图实现多样化的图片生成,同时使用 VAE 对视频帧中的运动信息进行编码,最终生成真实的街景视频.

根据上述分析,本文的工作是提出了以图片和文本作为输入条件的对抗式视频生成模型.与已有的基于条件对抗网络的视频生成方法相比,我们提出的多模态视频生成方法的输入条件更简洁、有效.

## 2 多模态对抗视频生成方法

图 1 显示了本文所使用的基于多模态输入的条件视频生成模型的框架图.整个网络结构由 5 个子网络组成,包括文本特征编码网络 $R_T$、运动特征解码网络 $D_V$、图片生成网络 $G_I$、图片判别网络 $D_I$、视频判别网络 $D_V$.整个网络基于 GAN 框架进行训练.

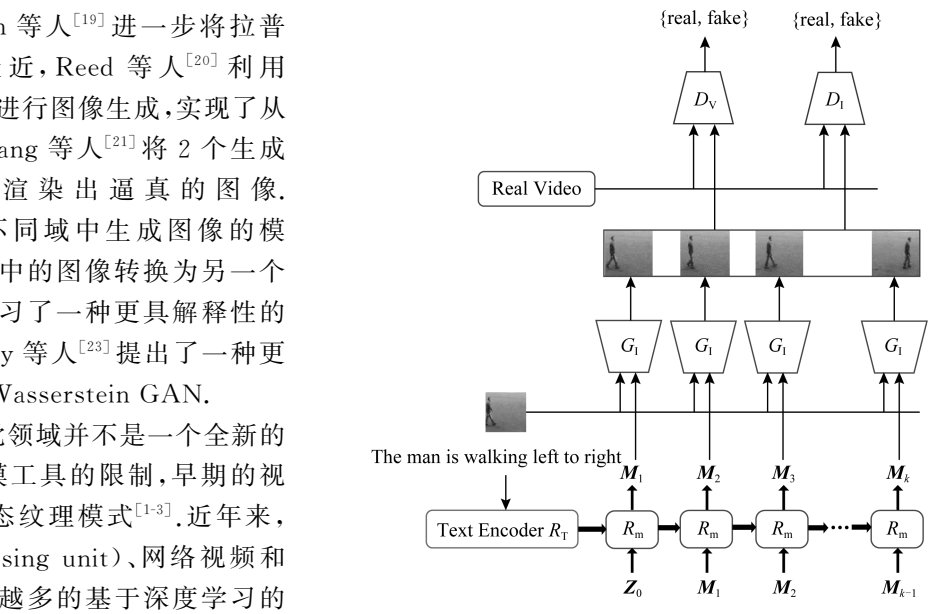



Fig. 1 Framework of antagonistic video generation method based on multimodal input

图 1 基于多模态输入的对抗式视频生成方法框架

文本特征编码网络 $D_V$ 用于提取输入文本的语义特征,运动特征解码网络 $R_m$ 根据文本的语义特征进一步生成运动特征来表达目标的运动信息.图片生成网络 $G_I$ 能够根据输入图片和对应的运动特征生成最终的视频帧.在对抗式训练中,视频判别网络 $D_V$ 用于捕捉帧与帧之间的运动信息,从而为图片生成网络 $G_I$,文本特征编码网络 $R_T$ 和运动特征解码网络 $R_m$ 提供反馈.而图片判别网络 $D_I$ 则专注于单帧图片的视觉内容判别,为输出更清晰的图片增加更多的细节约束.每个模块的实现细节将在后面章节进行详细介绍.

### 2.1 文本特征编码网络 $R_T$ 和运动特征解码网络 $R_m$

长短期记忆网络(long short term memory, LSTM)是一种针对序列型数据而设计的前馈神经网络,主要用来处理序列有关数.其通过在相邻时刻的隐藏层神经元之间加入连接形成循环结构,LSTM 可以重复利用之前时刻的历史信息,为了提

取文本的语义特征以及图像序列的运动特征,本文采用 LSTM 搭建了文本特征编码网络 $R_T$ 和运动特征解码网络 $R_m$.

本文提出的模型中,我们通过图片输入得到待生成目标的视觉内容信息.但要想生成视觉上连续变化的视频序列,则还需要为模型引入运动信息.我们使用文本描述来提供运动信息.由于 LSTM 处理序列数据时的优势,在自然语言处理(natural language processing, NLP)领域常被用于机器翻译[25]和句子语义特征提取.类似地,我们采用 LSTM 搭建了一个编码-解码结构.为了处理变长的文本输入信息,我们用一个 LSTM 网络来对文本信息进行编码,将输入的文本信息编码为一个定长向量.同时为了产生前后关联的运动编码信息,使用另一个 LSTM 对定长向量进行解码,得到一系列的运动编码信息.具体如下.

首先,描述语句分词后被表示为词向量并依次输入文本特征提取网络 $R_T$,网络 $R_T$ 是初始状态为 $\boldsymbol{h}_0$ 的 LSTM.最终文本描述语句($\boldsymbol{W}^1,\boldsymbol{W}^2,\cdots,\boldsymbol{W}^K$)被表示为最后一个单元对应的隐藏层特征(记为 $\boldsymbol{M}_0$).其中网络 $R_T$ 的初始状态 $\boldsymbol{h}_0$ 用全 0 向量表示,$\boldsymbol{W}^i$ 为第 $i$ 个词向量,$K$ 为句子长度.文本特征提取网络的输出 $\boldsymbol{M}_0$ 将用于运动特征解码网络 $R_m$ 的输入.运动特征解码网络 $R_m$ 由另一个 LSTM 构成.以文本特征 $\boldsymbol{M}_0$ 作为初始状态,全 0 向量作为初始输入,以后每一次的输入为前一层的输出,$R_m$ 网络将 $\boldsymbol{M}_0$ 解码为生成每帧图像所需的运动特征($\boldsymbol{M}_1,\boldsymbol{M}_2,\cdots,\boldsymbol{M}_k$),$\boldsymbol{M}_i$ 表示第 $i$ 帧的运动特征,$k$ 表示生成视频的长度,在实验中我们固定 $k=16$.

### 2.2 基于图片和运动特征的视频帧生成网络 $G_1$

根据 Hao 等人[16]的研究表明:视频生成任务需要预测的输出帧中大部分像素都可以直接从第 1 帧复制,这些像素只在位置上发生了一定的偏移.而第 1 帧中的少数像素区域由于被遮挡和剧烈运动等因素,需要用算法重新生成.基于以上分析,本文的视频生成网络 $G_1$ 采取与 Hao 等人类似的分治方法,把要预测的视频帧分解为变换图和新生成图.其中变换图由输入图片根据光流特征扭曲变换得到,而新生成图由图片和文本特征直接解码得到,最后通过合并变换图和新生成图得到最终的输出结果.视频生成网络 $G_1$ 的结构如图 2 所示.具体实现过程将在下面详细展开介绍.

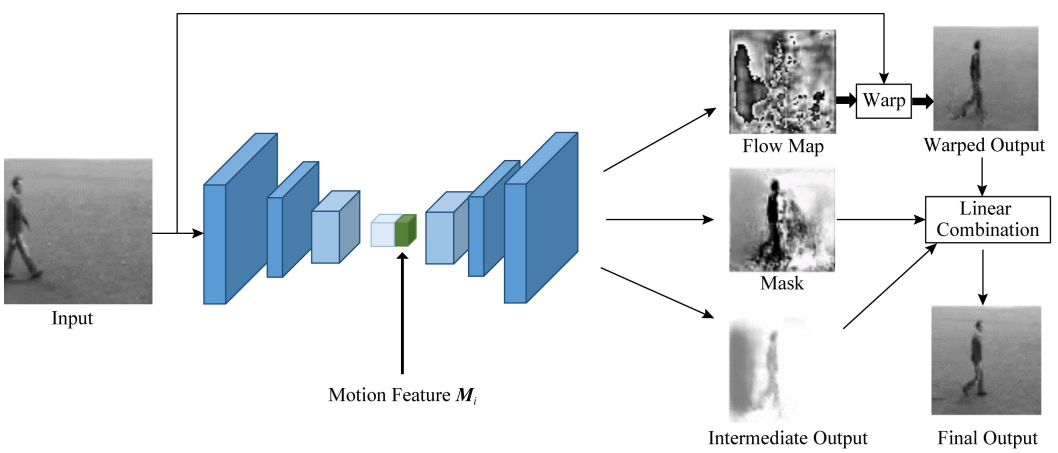



Fig. 2 Video generation network based on pictures and motion features

图 2 基于图片和运动特征的视频帧生成网络 $G_1$

给定图片 $I_0\in\mathbb{R}^{W\times H\times N}$ 和运动特征 $\boldsymbol{M}_i\in\mathbb{R}^{D_M}$,$i=1,2,\cdots,k$,视频生成网络 $G_1$ 将输出当前时刻的视频帧 $O\in\mathbb{R}^{W\times H\times N}$.对于图片 $I_0$,我们首先用一个具有 9 个 3×3 卷积层和 3 个池化层结构的卷积网络得到 $(W/8)\times(H/8)$ 尺度的视觉特征.$\boldsymbol{M}_i$ 是一个 $D_M$ 维向量,其在输入后将被扩展为 $\boldsymbol{M}_i\in\mathbb{R}^{\frac{W}{8}\times\frac{H}{8}\times D_M}$,$i=1,2,\cdots,k$,每个大小为 $(W/8)\times(H/8)$ 的特征通道上的值都相同,以适应图片的卷积特征的尺度.$W$ 和 $H$ 分别表示图片的宽和高,$N$ 为输入图片的通道数.最终扩展后的运动特征将与图片卷积特征按通道进行合并.

基于合并后的图片视觉特征和运动特征,我们首先用一个具有 9 个 3×3 卷积层和 3 个反卷积层结构的卷积网络进行上采样.最后通过 3 个不同的卷积层生成 3 个子图,包括稠密光流图 $D$、掩模图 $M$ 和新生成图 $O^h$.其中稠密光流图 $D$ 用于描述原始输入图片中每个像素的位移情况,掩模图 $M$ 用于描述输入图片中哪些区域因为遮挡或者目标快速移

动需要新生成像素,新生成图 $O^{\mathrm{h}}$ 则表示新生成的图片像素信息.

在得到稠密光流图 $D$ 之后,我们采用一个可微分的扭曲变换将输入图片 $I_0$ 做仿射变换得到变换图 $O^{\mathrm{f}}$.具体来说,变换图 $O^{\mathrm{f}}$ 的$(x,y)$位置的像素值是从原始图片 $I_0$ 的$(x_0,y_0)=(x+\Delta x,y+\Delta y)$位置变换得来,其中 $D_{x,y}=(\Delta x,\Delta y)$.由于 $D_{x,y}$ 生成结果是实数,因此采用双线性插值来计算变换图 $O^{\mathrm{f}}$ 中的每个像素值:

$$O^{\mathrm{f}}_{x,y}=\sum_{i,j}(1-|x_0-i|)(1-|y_0-j|)I_0^{i,j},\quad(1)$$

其中,$(i,j)$是$(x_0,y_0)$的四邻域.

最终输出图片可以通过合并变换图 $O^{\mathrm{f}}$ 和新生成图 $O^{\mathrm{h}}$ 来得到:

$$O_{x,y}=M_{x,y}\times O^{\mathrm{f}}_{x,y}+(1-M_{x,y})\times O^{\mathrm{h}}_{x,y}.\quad(2)$$

### 2.3 视频判别器 $D_{\mathrm{V}}$ 和图片判别器 $D_{\mathrm{I}}$

传统的 GAN 网络由 2 个网络组成:生成网络和判别网络.生成网络的目的是生成尽可能真实的图像,而判别网络的目的是尽可能区分真实图像和模型生成图像.这 2 个网络在一个最大最小的博弈游戏中不断优化,共同提升.在实际应用中,生成网络和判别网络都被实际化为卷积神经网络.其目标函数为

$$\min_G\max_D V(D,G)=E_{x\sim pdata(x)}[\ln(D(x))]+E_{Z\sim Pz(z)}[\ln(1-D(G(z)))].\quad(3)$$

在本文模型中我们在生成图像的同时还要保证视频序列之间的动态连贯性.因此我们的网络框架包含了 2 个判别器:视频判别器 $D_{\mathrm{V}}$ 和图片判别器 $D_{\mathrm{I}}$.这 2 个判别器与生成网络协同训练,使得模型能够提供更为高质量的生成结果.同时我们采用改进的 WGAN[23] 来稳定模型的训练.本文模型的优化目标可以表示为

$$\mathcal{L}_{\mathrm{WGAN}}(G_{\mathrm{I}},R_{\mathrm{M}},R_{\mathrm{T}},D_{\mathrm{I}},D_{\mathrm{V}})=E_{\tilde{v}\sim p_{\tilde{v}}}[D_{\mathrm{I}}(G_{\mathrm{I}}(I_0,\boldsymbol{M}_i))]-E_{v\sim p_v}[D_{\mathrm{I}}(v)]+E_{\tilde{v}\sim p_{\tilde{v}}}[D_{\mathrm{V}}(G_{\mathrm{I}}(I_0,\boldsymbol{M}_i))]-E_{v\sim p_v}[D_{\mathrm{V}}(v)],\quad(4)$$

其中,$\tilde{v}$ 是由生成网络 $G_{\mathrm{I}}$ 得到的视频序列,$v$ 是训练集中的真实视频.

图片判别网络 $D_{\mathrm{I}}$ 采用常规的包含 4 个 3×3 卷积层(每层都包含一个池化操作)和 1 个全连接层的 2D 卷积网络来实现.训练时,当输入真实图片时使其输出 1,当输入由网络生成的图片时使其输出 0.而视频判别网络 $D_{\mathrm{V}}$ 则采用包含 4 个 3×3 卷积层(每层都包含一个池化操作)和 1 个全连接层的 3D 卷积网络来实现.这是因为 $D_{\mathrm{V}}$ 的输入是连续的 $k$ 帧图片,而 3D 卷积能够提取其中的时域信息.同样,当输入真实视频时使其输出 1,当输入由模型生成的视频时使其输出 0.视频判别网络可以捕捉帧与帧之间的动态变化信息,为文本特征编码网络 $R_{\mathrm{T}}$ 和运动特征解码网络 $R_{\mathrm{m}}$ 提供反馈.

我们将总损失函数定义为

$$\min_{G_{\mathrm{I}},R_{\mathrm{m}},R_{\mathrm{T}}}\max_{D_{\mathrm{I}},D_{\mathrm{V}}}\mathcal{L}=\mathcal{L}_{\mathrm{WGAN}}(G_{\mathrm{I}},R_{\mathrm{m}},R_{\mathrm{T}},D_{\mathrm{I}},D_{\mathrm{V}})+\lambda_{\mathrm{I}}\mathcal{L}_{\mathrm{I}}(G_{\mathrm{I}},R_{\mathrm{m}},R_{\mathrm{T}}),\quad(5)$$

其中:

$$\mathcal{L}_{\mathrm{I}}(G_{\mathrm{I}},R_{\mathrm{m}},R_{\mathrm{T}})=\sum_{i=0}^{T}(G_{\mathrm{I}}(I_0,\boldsymbol{M}_i)-I_i)^2,$$

$I_i$ 为真实的第 $i$ 帧图片,该项使得生成图片与真实图片更接近,加快模型收敛速度.

### 2.4 训练方式

与传统的 GAN 网络框架相同,我们先训练图片判别网络 $D_{\mathrm{I}}$、视频判别网络 $D_{\mathrm{V}}$,再对抗式地训练文本特征编码网络 $R_{\mathrm{T}}$、运动特征解码网络 $R_{\mathrm{m}}$、图片生成网络 $G_{\mathrm{I}}$.我们使用 Adam 优化器进行训练,batch size 为 16,学习率为 0.000 5,$\beta_1=0.9$,$\beta_2=0.99$.

## 3 验证与实验结果

本文所用数据集示例如图 3 所示:

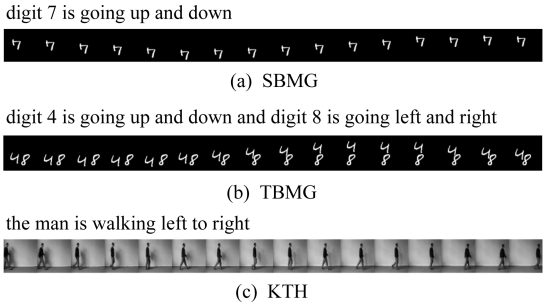



Fig. 3 Dataset examples

图 3 数据集示例

### 3.1 实验使用数据集介绍

SBMG(single-digit bouncing MNIST gifs)数据集为了验证模型的有效性,我们采用了和 Mittal 等人[26]一样的方法合成动态的手写数字视频样本.SBMG 是包含单个数字运动的视频,每个视频样本是由手写数字数据集 MNIST[27] 中随机采样的图像生成.对于给定大小为 64×64 的数字图像,根据指定运动描述语句(例如数字 7 上下移动)移动手写数字对应的白色像素点来模拟生成数字的运动视频.我们生成了 60 000 个视频样本,每个视频样本都对

应着一个描述语句.图 3(a)显示了由数字 7 构造生成的视频以及对应的描述语句.实验中,我们随机选取 50 000 个视频用于训练,10 000 个用于测试.

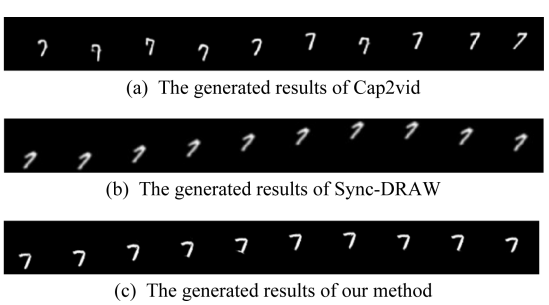
TBMG(two-digit bouncing MNIST gifs)数据集也是由 MNIST 数据集中的数字图片生成.区别在于 TBMG 是包含 2 个数字同时运动的视频.采用 MNIST 数据集中的 2 张图片,按照描述语句移动 2 张图片的白色像素区域并叠加得到视频样本.TBMG 数据集包含 30 000 个视频样本.图 3(b)显示了由数字 4 和 8 构造生成的视频以及对应的描述语句.实验中,我们随机选取 25 000 个视频用于训练,5 000 个用于测试.

KTH (kungliga tekniska högskolan human actions)数据集为了在更真实的数据集上评估本文模型的性能,使用了 KTH 数据集[28].这个数据集包含超过 2 000 个视频序列,是通过拍摄 25 个人执行 6 种不同的动作得到.我们选取人物步行的视频序列来进行实验.通过把人物步行的视频进行切分和人工标记,我们得到了 200 个包含"从右向左走"和"从左向右走"2 种行为的视频.每个视频有 16 帧,视频帧大小为 64×64.图 3(c)显示了人物"从左向右走"的视频和其对应的描述语句.实验中,我们随机选取 175 个视频用于训练,25 个用于测试.

### 3.2 与现有对比模型评估

为了评估我们的模型,我们与现有的模型 Cap2vid[24] 和 Sync-DRAW[26] 进行了比较.Cap2vid 通过学习文本与视频帧之间的长期和短期依赖关系,通过 LSTM 建模以增量的方式生成视频.Sync-DRAW 使用一个循环的变分自编码器(R-VAE)和一个分层的注意机制来创建一个随时间逐渐变化的视频帧.

图 4 显示了不同方法在 SBMG 数据集上的数字视频生成结果.



(a) The generated results of Cap2vid

(b) The generated results of Sync-DRAW

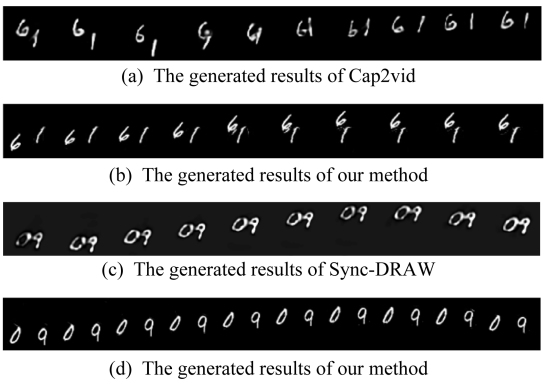
(c) The generated results of our method

Fig. 4 Comparison of effects of different models on SBMG

图 4 不同模型在 SBMG 上的效果对比

由图 4 可以看到 Cap2vid 生成的视频中数字外观上前后有一定差异.而 Sync-DRAW 和本文的方法基本保持了原有的内容.

图 5 展示了不同方法在 TBMG 数据集上的数字视频生成结果.



(a) The generated results of Cap2vid

(b) The generated results of our method

(c) The generated results of Sync-DRAW

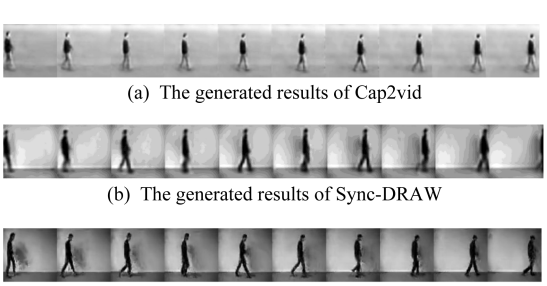
(d) The generated results of our method

Fig. 5 Comparison of the effects of different models on TBMG datasets

图 5 不同模型在 TBMG 数据集上的效果对比

由图 5 可以看到,Cap2vid 仅通过文本生成视频,生成结果虽然能保持数字的运动,但是数字在前后帧的变化过大.而我们的方法生成的视频具有更好的清晰度和连贯性.Sync-DRAW 方法生成的结果同样不够清晰,而我们通过对输入图片进行变换得到视频帧,能够为生成任务提供更多的细节,同时保证了视频内容在前后帧的连贯性.

图 6 显示了不同方法在 KTH 数据集上的结果对比.



(a) The generated results of Cap2vid

(b) The generated results of Sync-DRAW

(c) The generated results of our method

Fig. 6 Qualitative comparison of models on KTH

图 6 模型在 KTH 上的定性对比

由图 6 可以看到 Cap2vid 生成的人物姿态过于单一,未能模拟人物行走的完整动作,而 Sync-DRAW 的生成结果同样不够清晰.我们的方法生成的视频结果在人物的清晰度以及动作的完整性上都有更好的表现.

由于 Cap2vid 与 Sync-DRAW 为无监督方法，无法计算定量指标，我们基于模型生成的结果进行了定性比较，同时我们将本文提出的方法与 Hao 等人[16]提出的方法进行了定量比较.从表 1 来看，PSNR，SSIM(2 个指标值越大越好)都有所提升，其中 BMG 为 SBMG 和 TBMG 的合并.

**Table 1 Model Contrast Analysis**

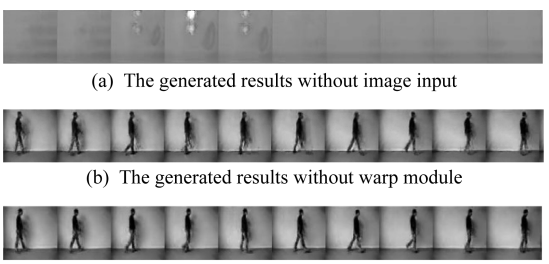
**表 1 模型性能量化分析**

| Dataset | Ref[16] | | Our Method | |
|---|---|---|---|---|
| | PSNR | SSIM | PSNR | SSIM |
| BMG | 15.69 | 0.74 | 15.87 | 0.76 |
| KTH | 27.04 | 0.78 | 27.85 | 0.81 |

### 3.3 模型变体的定性与定量评估

为了验证本文模型各个组件的有效性，我们设置了模型变体进行对比实验：图 7(a)基于文本输入(将图片输入置 0)的视频生成方法；图 7(b)基于图片和文本特征直接解码生成视频帧，不采用变换图与新生成图合并的方式；图 7(c)本文提出的完整生成模型.

如图 7 所示，通过 3 个方法的实验对比我们可以看到，仅输入文本的模型图 7(a)能生成一部分运动信息，但由于缺乏目标的视觉信息，生成视频过于模糊.模型图 7(b)虽然能够生成可辨识的视频结果，但由于生成结果是直接通过解码图片和文本特征得到，因此细节信息不如完整模型图 7(c)通过合并变换图和新生成图得到的结果.



(a) The generated results without image input

(b) The generated results without warp module

(c) The generated results of full model

Fig. 7 Comparison of model variants on KTH datasets

图 7 模型变体在 KTH 数据集上的结果对比

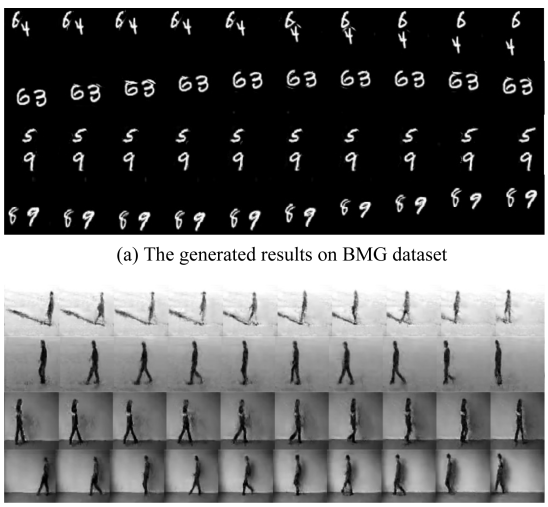
下面我们通过计算预测视频帧和真实视频帧之间的 PSNR 和 SSIM[29]指标进一步评估不同模型性能.如表 2 所示，通过比较 SSIM 和 PSNR，可以看到完整模型图 7(c)在 BMG 数据集和 KTH 数据集上的结果都优于模型变体图 7(a)和图 7(b)的结果.图 6 和表 2 中的实验对比都表明我们提出的多模态输入更有益于得到鲁棒可控的视频结果.此外，通过把视频生成结果分解为由输入图片变换得到的内容与新生成内容，可以得到更清晰、连贯的视频序列.

**Table 2 Quantitative Analysis of Model Variation and Complete Model**

**表 2 模型变种与完整模型定量分析**

| Dataset | Variant Model | | | | Full Model | |
|---|---|---|---|---|---|---|
| | Fig.7(a) | | Fig.7(b) | | Fig.7(c) | |
| | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| BMG | 14.68 | 0.59 | 15.49 | 0.74 | 15.87 | 0.76 |
| KTH | 16.19 | 0.56 | 26.54 | 0.77 | 27.85 | 0.81 |

图 8 显示了我们模型更多的生成结果.由图 8 可以看出我们的模型在清晰度以及动作的连贯性上都有不错的表现.



(a) The generated results on BMG dataset

(b) The generated results on KTH dataset

Fig. 8 Model generation results

图 8 模型生成结果

## 4 总结与展望

基于生成对抗网络的视频生成算法近年来得到了研究人员的广泛关注.本文提出一种新的基于多模态输入的条件式视频生成模型.一方面基于图片信息输入为生成视频提供更多细节，并基于仿射变换来预测视频帧；另一方面使用文本特征编码网络和运动特征解码网络得到运动信息，进而辅助生成网络输出连贯的视频序列.在 SBMG，TBMG，KTH 数据集上的实验结果表明，我们的模型在运动连贯性和内容前后一致性上都优于现有的模型.我们的方法使得生成模型更加可控、生成结果更加鲁棒.但本文提出的视频生成算法仍然有很大的改进空间，在未来我们将继续探索更有效的模型，以适应更为复杂环境下的视频生成需求.

## 参 考 文 献


[1] Szummer M, Picard R W. Temporal texture modeling[C]// Proc of IEEE Int Conf on Image Processing. Piscataway, NJ: IEEE, 1996: 823-826

[2] Wei Liyi, Levoy M. Fast texture synthesis using tree-structured vector quantization[C]//Proc of the 27th Annual Conf on Computer Graphics and Interactive Techniques. New York: ACM, 2000: 479-488

[3] Doretto G, Chiuso A, Wu Yingnian, et al. Dynamic textures [J]. International Journal of Computer Vision, 2003, 51(2): 91-109

[4] Wang Limin, Xiong Yuanjun, Wang Zhe, et al. Temporal segment networks: Towards good practices for deep action recognition[C]//Proc of European Conf on Computer Vision. Berlin: Springer, 2016: 20-36

[5] Liu Wei, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector[C]//Proc of European Conf on Computer Vision. Berlin: Springer, 2016: 21-37

[6] He Kaiming, Georgia G, Piotr D, et al. Mask R-CNN[C]// Proc of Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 2961-2969

[7] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Proc of Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 2672-2680

[8] Liu Mingyu, Tuzel O. Coupled generative adversarial networks[C]//Proc of Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2016: 469-477

[9] Wang Tingchun, Liu Mingyu, Zhu Junyan, et al. High-resolution image synthesis and semantic manipulation with conditional GANs[C]//Proc of Int Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 8798-8807

[10] Balakrishnan G, Zhao A, Dalca A V, et al. Synthesizing images of humans in unseen poses[C]//Proc of Int Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 8340-8348

[11] Mathieu M, Couprie C, LeCun Y. Deep multi-scale video prediction beyond mean square error[OL]. (2015-11-17)[2019-04-08]. https://arxiv.or g/abs/1511.05440

[12] Tulyakov S, Liu Mingyu, Yang Xiaodong, et al. MoCoGAN: Decomposing motion and content for video generation[C]//Proc of Int Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 1526-1535

[13] Saito M, Matsumoto E, Saito S. Temporal generative adversarial nets with singular value clipping[C]//Proc of Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 2830-2839

[14] Li Yitong, Min M R, Shen Dinghan, et al. Video generation from text[OL]. (2017-10-01)[2019-04-05]. https://arxiv.org/abs/1710.00421

[15] Vondrick C, Pirsiavash H, Torralba A. Generating videos with scene dynamics[C]//Proc of Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2016: 613-621

[16] Hao Zekun, Huang Xun, Belongie S. Controllable video generation with sparse trajectories[C]//Proc of Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2018: 7854-7863

[17] Pumarola A, Agudo A, Martinez A M, et al. GANimation: Anatomically-aware facial animation from a single image[C]//Proc of European Conf on Computer Vision. Berlin: Springer, 2018: 818-833

[18] Pan Junting, Wang Chengyu, Jia Xu, et al. Video generation from single semantic label map[C]//Proc of Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 3733-3742

[19] Denton E, Chintala S, Szlam A, et al. Deep generative image models using a Laplacian pyramid of adversarial networks[C]//Proc of Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2015: 1486-1494

[20] Reed S, Akata Z, Yan Xinchen, et al. Generative adversarial text to image synthesis[C]//Proc of Int Conf on Machine Learning. New York: ACM, 2016: 1060-1069

[21] Zhang Han, Xu Tao, Li Hongsheng, et al. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks[C]//Proc of Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 5907-5915

[22] Xi Chen, Yan Duan, Rein H, et al. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets[C]//Proc of Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2016: 2172-2180

[23] Arjovsky M, Chintala S, Bottou L. Wasserstein GAN[OL]. (2017-12-06)[2019-04-13]. https://arxiv. org/abs/1701.07875

[24] Marwah T, Mittal G, Balasubramanian V N. Attentive semantic video generation using captions[C]//Proc of Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 1426-1434

[25] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Proc of Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 3104-3112

[26] Mittal G, Marwah T, Balasubramanian V. Sync-DRAW: Automatic gif generation using deep recurrent attentive architectures[C]//Proc of ACM Int conf on Multimedia. New York: ACM, 2017: 1096-1104

[27] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J.] Proceedings of the IEEE, 1998, 86(11): 2278-2324

[28] Schuldt C, Laptev I, Caputo B. Recognizing human actions: A local SVM approach [C] //Proc of Int Conf on Pattern Recognition. Piscataway, NJ: IEEE, 2017: 32-36

[29] Wang Zhou, Bovik A C, Sheikh H R, et al. Image quality assessment: From error visibility to structural similarity [J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612

**Yu Haitao**, born in 1996. Master candidate. His main research interests include computer vision and multimedia.

**Yang Xiaoshan**, born in 1989. PhD, associate professor. Member of CCF. His main research interests include recognition/ranking of image and video, deep learning.

**Xu Changsheng**, born in 1969. PhD, professor. Distinguished member of CCF. His main research interests include multimedia content analysis/indexing/retrieval, pattern recognition and computer vision.