文章编号：2095-2163(2021)10-0033-05 中图分类号：TP391；TP183 文献标志码：A

基于图卷积的手势骨架生成

曾 瑞，张海翔，马汉杰，蒋明峰，冯 杰
（浙江理工大学 信息学院，杭州 310018）

摘 要：目前手势生成的工作多用于从语音或文本中产生协同的手势以及实现手势数据增强。前者作为非语言信号辅助交流，却难以单独表达语义。对于后者，大多数都是将骨骼关节点当作图像的一个像素，整体当作图像处理，而没有考虑到关节点间丰富的人体结构信息，从而可能导致生成的结果是扭曲的、不自然的。本文提出了基于图卷积的生成式模型，以有效地编码结构信息到手势生成中。研究中将本文的方法与基于全连接神经网络以及基于卷积神经网络的方法进行了对比，实验结果表明，本文生成的手势在定量和定性结果上有了明显的改善。图卷积在手势骨架生成上的成功应用，可以进一步指导手势骨架到真实手势的生成工作，因而对生成自然、真实的手势有重要意义。

关键词：手势骨架生成；生成式对抗网络；图卷积神经网络

Skeleton-based gesture generation based on Graph Convolutional Networks

ZENG Rui, ZHANG Haixiang, MA Hanjie, JIANG Mingfeng, FENG Jie
(School of Informatics Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

【Abstract】 Currently, the work of gesture generation is mostly to generate coordinated gestures from speech or text and realize the data augmentation for gesture. The former is used as nonverbal signals to conduce communication, but it is difficult to express semantics alone. The latter is that in most cases, the skeleton joints are regarded as pixels of the image and a frame of the gesture as an image. However, those do not take the rich structure information among joints into consideration, so it may cause the generated result to be distorted and unnatural. The paper proposes a generative model based on Graph Convolutional Networks to efficiently encode structural information into gesture generation. The research compares the proposed method with methods based on Fully Connected Neural Network and Convolutional Neural Network. The results show that the gestures generated by the proposed method have been significantly improved in quantitative and qualitative results. The successful application of Graph Convolutional Networks in the generation of skeleton-based gestures can further guide the work of generating real gestures from skeleton-based gestures, and it is of great significance for generating natural and real gestures.

【Key words】 skeleton-based gesture generation; Generative Adversarial Network; Graph Convolutional Neural Networks

0 引 言

近来，手势生成的工作多用于从语音或文本中产生协同的手势[1-3]，以及实现手势数据的增强[4-5]。生成式对抗网络因其在生成上的优异表现，在手势生成上也得到了广泛的应用[1-4]。但是，对于给定话语生成对应手势的主要问题是，手势作为非语言信号辅助语言，使得交流更加顺畅，却难以单独表达语义。这是由于语音到手势的高度非确定性映射，即使是同一个人说相同的短语，也可能在每次重复时伴随不同的手势动作，并且生成的结果会特定于个人手势风格。另外，手势数据增强工作主要针对的是真实的手势，并且大多数都是将人体的骨骼关节点当做图像的一个像素，将动作的一帧当做一个图像，而没有考虑到骨骼关节点间丰富的人体结构信息，从而可能导致生成的结果是扭曲的、不自然的。研究发现，图卷积神经网络能够处理非欧式空间的数据，而不同于传统的网络模型如 CNN、LSTM 等只能处理欧式空间的网格结构的数据。因此，为了能更好地利用手部的结构信息，本文采用了基于图卷积的生成式对抗网络模型来直接生成手势骨架。实验结果表明，文中的方法对手势骨架的生成的确有了更自然更高质量的结果。本文工作的主要贡献概括为 2 个方面：

第一，提出了基于图卷积神经网络的手势骨架生成方法，可以有效地将手部的结构信息编码到手

作者简介：曾 瑞(1996-)，女，硕士研究生，主要研究方向：计算机视觉；张海翔(1973-)，男，博士，副教授，主要研究方向：计算机视频图像处理技术、计算机视觉技术、深度几何学习方法；马汉杰(1982-)，男，博士，副教授，主要研究方向：视频图像处理；蒋明峰(1977-)，男，博士，教授，主要研究方向：计算机医学图像处理、生物医学信号处理；冯 杰(1980-)，男，博士，讲师，主要研究方向：视频处理。
通讯作者：张海翔 Email：zhhx@zstu.edu.cn

收稿日期：2021-07-25

势建模中。

第二,在手势骨架生成任务上,通过有效地利用手部结构信息,文中的方法比基于全连接神经网络以及基于卷积神经网络的生成方法在定性和定量结果上都取得了更好的结果。

1 相关工作

1.1 生成式对抗网络

生成式对抗网络[6-8](Generative Adversarial Network,GAN)是一种优秀的生成式模型,能够学习已有样本的分布并生成与之相似的样本,已然成为学界研究热点。生成器 G 与判别器 D 是 GAN 模型的重要组成部分,这两者之间的相互对抗使双方都得到增强,最终使生成模型尽可能生成逼真的样本,示意图如图 1 所示。

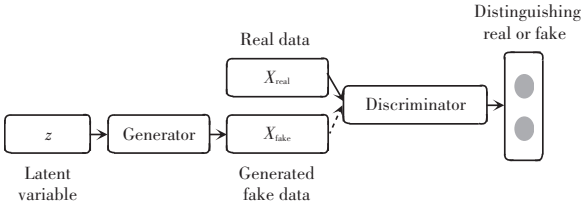


图 1 生成式对抗网络

Fig. 1 Generative Adversarial Network

图 1 中,z 为表示随机噪声的隐变量,可通过生成器生成假样本。判别器则对输入数据进行判别区分。训练时,生成器和判别器交替训练,不断往复。优化的目标函数[9]如下:

$$\min_G \max_D V(D,G) = E_{X\sim P_{data}(x)}[\log D(x)] + E_{z\sim P_z(z)}[\log(1 - D(G(z)))] \quad (1)$$

其中,E 为分布函数的期望值;x 为真实数据;z 为噪声。式(1)其实就是一个最大最小优化问题,生成器与判别器都进行优化,在交替训练中双方都逐步得到增强。

GAN 提出之后,各种 GAN 的衍生模型相继提出,在结构改进、应用等方面进行创新,用于诸如图像生成、图像转换、图像修复等多个领域。在结构改进上,如 2017 年提出的 wgan[10]、began[11]等改进了目标函数,使得训练更加稳定。在应用方面,如 CycleGAN[12] 和 Pix2Pix[13] 实现了风格迁移,TPGAN[14]能根据半边人脸生成整张人脸的前向图。

1.2 图卷积神经网络

卷积神经网络通过局部化的卷积核来学习局部的稳定结构,然后通过层级堆叠将其变为层次化的多个尺度的结构模式,其强大的建模能力使得在图像处理、对象检测、自然语言处理等任务上都取得了不错的效果。但是,平移不变性却使其只能处理欧式数据,而处理不了如交通网络这样非欧结构的数据。

2013 年,图的基于谱域和基于空间的卷积神经网络[15]被首次提出。谱方法[15-18]和空间方法[19-20]是目前图卷积[21-22]的 2 种主要方法。前者把图的信号变换到谱域,在谱域进行卷积后再变换到空间域,以此完成图卷积。后者则直接在空间域定义节点相关性。其应用主要集中于计算机视觉、交通预测、推荐系统、生物化学、自然语言处理等领域。比如在计算机视觉中,Marino 等人[23]将知识图谱引入到图片分类中,使用图卷积神经网络更好地利用知识图谱中的先验知识,在 COCO 数据集的多标签分类任务上取得了提升。

2 基于图卷积的手势骨架生成

2.1 模型介绍

本文提出了一种基于图卷积的手势骨架生成方法。模型框架采用的是通用的生成式对抗网络[24],由一个生成器和一个判别器组成。

生成器结构见图 2,图 2(a)中的 Graph conv layer 具体结构在图 2(b)中说明。在图 2 中,*noise* 为服从标准正态分布的随机噪声,*label* 为手势种类的标签,图卷积模块具体见图 2(b)。图 2(b)中的 H 即为图 2(a)中图卷积层的输入,$\hat{\boldsymbol{A}}$ 为邻接矩阵经归一化处理后的结果,见公式(2):

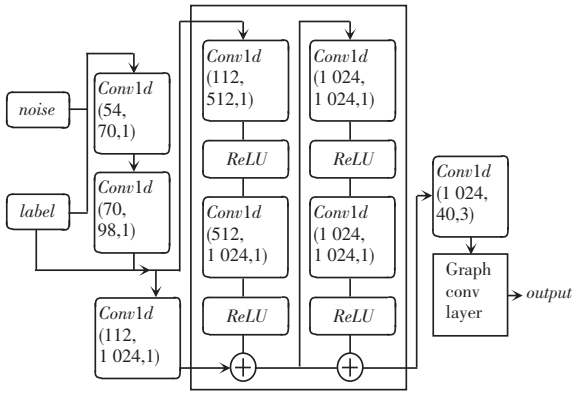
$$\hat{\boldsymbol{A}} = \boldsymbol{D}^{-1}(\boldsymbol{A} + \boldsymbol{I}) \quad (2)$$

其中,邻接矩阵 $\boldsymbol{A}$ 为表示手势各关节点之间相邻关系的矩阵;$\boldsymbol{I}$ 为单位矩阵;$\boldsymbol{D}$ 为对应的度矩阵。邻接矩阵 $\boldsymbol{A}$ 加上一个单位矩阵 $\boldsymbol{I}$,是希望在进行信息传播的时候关节点自身的特征信息也得到保留,那么 $\boldsymbol{A} + \boldsymbol{I}$ 就聚合了各关节点本身以及相邻关节点的特征信息。而进行归一化操作 $\boldsymbol{D}^{-1}(\boldsymbol{A} + \boldsymbol{I})$ 则是为了信息传递的过程中保持原有分布,防止一些相邻关节点多的节点和相邻关节点少的节点在特征分布上产生较大的差异。

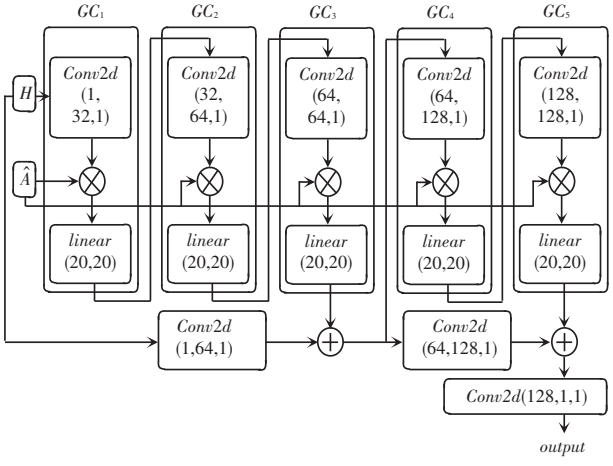
图卷积模块中共有 5 个图卷积层,即图 2(b)中的 GC_1、GC_2、GC_3、GC_4 以及 GC_5。第一个图卷积层以 $\boldsymbol{H}$ 和 $\hat{\boldsymbol{A}}$ 为输入,输出为 $\boldsymbol{H}^{(1)}$,见公式(3):

$$\boldsymbol{H}^{(1)} = \hat{\boldsymbol{A}}\boldsymbol{H}\boldsymbol{W}^{(1)} \quad (3)$$

第二个图卷积层以 $\boldsymbol{H}^{(1)}$ 和 $\hat{\boldsymbol{A}}$ 为输入,输出为 $\boldsymbol{H}^{(2)}$。这样,经过 5 次图卷积之后,得到生成器的输出,也就是手势的各关节点的坐标。



（a）总体结构图



（b）图卷积模块结构图

图 2　生成器结构图

Fig. 2　Generator structure chart

判别器的结构如图 3 所示。判别器以生成器生成的或数据集中的手势各关节点为输入，经过多个卷积层和激活层，最终得到 2 个输出：$output_1$ 结果在 0 到 1 之间，用来判别输入为真或假；$output_2$ 结果为输入手势的类别。

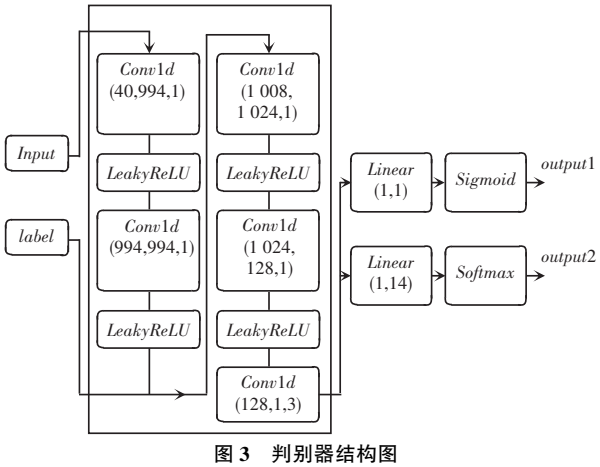


图 3　判别器结构图

Fig. 3　Discriminator structure chart

2.2　训练细节

在训练过程中，判别器和生成器交替训练，通过相互对抗让这两个模型同时得到增强。两者都使用 *Adam* 作为优化器，学习率为0.000 2。目标函数见上文公式(1)，并采用了交叉熵损失函数。另外，为了增加网络的抗干扰能力，使用了单侧标签平滑，用标签 0.9 代替 1 表示真实的数据。

2.3　数据集

本文数据集采用的是由数据堂提供的静态手势识别数据。本文实验采用了其中的数字 1、数字 2、比心、点赞、握拳等 14 种单手手势，共14 000条数据。每条数据含一张手势图像以及一个标注文件，标注文件中写明了手势的 21 个关节点及手势类别等信息。本文实验按 8∶2 将所有数据分为训练集和测试集，采用了手势的 21 个关节点坐标信息。手势关节点的标注情况见图 4。

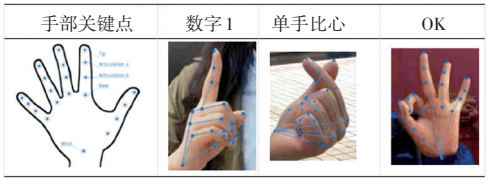


图 4　手势关节点标注示意图

Fig. 4　Schematic diagram of gesture joint labeling

2.4　评价指标

Maximum Mean Discrepancy（MMD，最大平均差异）[25] 以样本 $x \sim P(X)$ 和 $y \sim Q(Y)$ 来度量 2 个分布 $P(X)$ 和 $Q(Y)$ 之间的相似性。其本质上是 2 个分布的数据经过映射函数变化后的期望之差的上确值，但由于直接计算期望十分困难，可以采取计算期望的无偏估计（unbiased estimate）-均值。通过两者间的差值来判别 2 个分布的相似程度。值越小，那么这 2 个数据分布越相似。*MMD* 的具体计算见公式(4)：

$$MMD_u^2[F,X,Y] = \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m}k(x_i,x_j) + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}k(y_i,y_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}k(x_i,y_i) \tag{4}$$

其中，$X:=\{x_1,\dots,x_m\}$ 和 $Y:=\{y_1,\dots,y_n\}$ 分别为服从 P 与 Q 分布的样本。x_i 与 x_j 为服从 P 分布的独立的随机数据，y_i 与 y_j 同理。$k(\cdot,\cdot)$ 为高斯核函数，具体计算见公式(5)：

$$k(x,x') = e^{-\frac{\|x-x'\|^2}{2\sigma^2}} \tag{5}$$

另外，文献[26]比较了 Inception Score，Mode Score，MMD 等 6 种 GAN 具有代表性的基于样本的评估度量，结果表明 MMD 能够区分真实图像和生

成图像,可以在一定程度上衡量模型生成图像的优劣性,是最合适的评估指标之一。在深度生成模型[27]和贝叶斯采样[28]中,则被用于衡量生成样本相较于真实数据的质量。该指标还被用于评估生成的动作与真实动作之间的相似性[24,29]。

3 实验分析

本次研究进行了实验以评估所提出的方法在静态手势骨架生成上的有效性,并采用 MMD 来衡量生成手势的质量。仿真时在同一数据集上进行了 3 次实验:首先是本文的方法、即基于图卷积的生成方法,其次是 wgan_gp 方法、即基于全连接的生成方法,最后是消融实验、即基于卷积神经网络的生成方法。结果显示,本文的方法更好。下面进行了详细的阐释与分析。

由于目前没有生成静态手势骨架的工作,文中的对比实验选择了 wgan_gp[31]方法。wgan_gp 以 wgan[10]为基础,将梯度截断替换为梯度惩罚,以解决梯度消失、梯度爆炸的问题。优化的目标函数如下:

$$L = \underset{\tilde{x}\sim P_g}{E}[D(\tilde{x})] - \underset{x\sim P_r}{E}[D(x)] + \lambda \underset{\hat{x}\sim P_{\hat{x}}}{E}[(\| \nabla_{\hat{x}} D(\hat{x}) \|_2 - 1)^2] \tag{6}$$

其中,$\tilde{x}$ 为生成器生成的样本,$\tilde{x} \leftarrow G_\theta(z)$,$\hat{x}$ 为生成数据和真实数据之间的插值,$\hat{x} \leftarrow \epsilon x + (1-\epsilon)\tilde{x}$,$\epsilon \sim U[0,1]$。式(6)的最后一项即为添加的梯度惩罚项。这样既满足了 1-L 条件,也可以保证权重变化不那么剧烈,模型不容易坏掉。本次实验中,wgan_gp的生成器与判别器采用的皆是全连接网络架构。本文方法与该方法的比较结果见表 1,本文在 MMD 上的结果比 wgan_gp 方法的要小,也就是本文提出的方法更优。

表 1 对比实验结果

Tab. 1 Comparative experiment results

	wgan_gp	The proposed
MMD	0.265 2	0.169 7

此外,还在本文的方法上进行了消融实验以验证图卷积模块的有效性,实验结果见表 2。具体来说,使用 5 个 CNN 层替换了生成器中的 5 个 GCN 层,并使用相同的隐藏维数和核大小,其他部分保持完全相同。

表 2 消融实验结果

Tab. 2 Ablation results

	The proposed without GCN	The proposed
MMD	0.196 5	0.169 7

各方法生成的手势骨架图如图 5 所示。图 5 中列出了数字 1、数字 2、数字 4、数字 6、单手比心、OK、握拳、LOVE 共 8 种手势。从图 5 中可以看出,本文的方法能够生成更加合理、自然的手势,而 wgan_gp 与消融实验生成的手势较为扭曲。从定量结果来看,本次研究提出的方法在 MMD 上也更加优越。

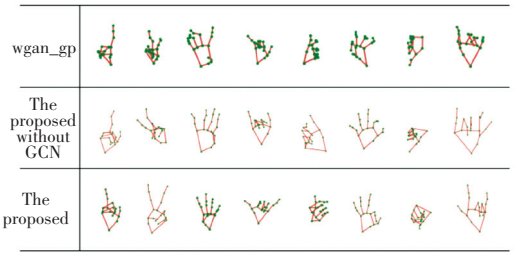


图 5 生成的手势骨架对比图

Fig. 5 Comparison of the generated gesture skeleton

4 结束语

在本文中,研究提出了基于图卷积的生成式对抗网络模型,以有效地编码手部结构信息到基于骨架的人体手势生成。从对比实验和消融实验的结果可以观察到,文中的生成结果在定量和定性评估上有了明显的改善,证明了图卷积用于手势骨架的有效性。在后续的工作中,将进一步研究基于图卷积的序列手势生成。

参考文献

[1] REBOL M, GÜTI C, PIETROSZEK K. Passing a non-verbal turing test: Evaluatina gesture animations generated from speech [C]// 2021 IEEE Virtual Reality and 3D User Interfaces (VR). Lisbon, Portugal: IEEE, 2021: 573-581.

[2] TUYEN N T V, ELIBOL A, CHONG N Y. Learning from Humans to Generate Communicative Gestures for Social Robots [C]// 2020 17th International Conference on Ubiquitous Robots (UR). Kyoto, Japan: The Ritsumeikan University, 2020: 284-289.

[3] FERSTL Y, NEFF M, MCDONNELL R. Adversarial gesture generation with realistic gesture phasing [J]. Computers & Graphics, 2020, 89: 117-130.

[4] 刘田丰, 马力. 一种基于 GAN 的手势图像生成方法[J]. 计算机与数字工程, 2020, 48(8): 2014-2017, 2023.

[5] LINDGREN K, KALAVAKONDA N, CABALLERO D E, et al. Learned hand gesture classification through synthetically generated training samples [C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid, Spain: IEEE, 2018: 3937-3942.

[6] 罗佳, 黄晋英. 生成式对抗网络研究综述[J]. 仪器仪表学报, 2019, 40(3): 74-84.

[7] HONG Yongjun, HWANG U, YOO J, et al. How generative adversarial networks and their variants work: An overview [J].

arXiv preprint arXiv 1711.05914,2019.
[8] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems. Montreal, Canada:NIPS,2014: 2672-2680.
[9] 邹秀芳,朱定局. 生成对抗网络研究综述[J]. 计算机系统应用,2019,28(11):1-9.
[10] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein GAN[J]. arXiv preprint arXiv:1701.07875, 2017.
[11] BERTHELOT D, SCHUMM T, METZ L. BEGAN: Boundary equilibrium generative adversarial networks[J]. arXiv preprint arXiv:1703.10717, 2017.
[12] ZHU Junyan, PARK T, ISOLA P. Unpaired image-to-image translation using cycle-consistent adversarial networks[J]. 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy:IEEE, 2017:2242-2251.
[13] 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 & Pattern Recognition. Honolulu:IEEE, 2017:1125-1134.
[14] HUANG Rui, ZHANG Shu, LI Tianyu, et al. Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis[C]// 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy:IEEE Computer Society, 2017:2458-2467.
[15] BRUNA J, ZAREMBA W, SZLAM A, et al. Spectral networks and locally connected networks on graphs[C]// ICLR. Banff, Canada:dblp, 2014:1-14.
[16] DEFFERRARD M, BRESSON X, VANDERGHEYNST P. Convolutional neural networks on graphs with fast localized spectral filtering[C]//Advances in Neural Information Processing Systems. Barcelona:NIPS,2016:3844-3852.
[17] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv: 1609.02907,2016.
[18] ZHANG Xiaotong, LIU Han, LI Qimai, et al. Attributed graph clustering via adaptive graph convolution[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao China:AAAI,2019:4327-4333.
[19] ATWOOD J, TOWSLEY D. Diffusion-convolutional neural networks[C]// Advances in Neural Information Processing Systems. Barcelona:NIPS, 2016:1993-2001.
[20] ZHUANG Chenyi, MA Qiang. Dual graph convolutional networks for graph based semi-supervised classification[C]//Proceedings of the World Wide Web Conference on World Wide Web. LYON, France: International World Wide Web Conferences Steering Committee, 2018:499-508.
[21] 徐冰冰, 岑科廷, 黄俊杰,等. 图卷积神经网络综述[J]. 计算机学报, 2020,43(5):755-780.
[22] 王健宗, 孔令炜, 黄章成,等. 图神经网络综述[J]. 计算机工程, 2021,47(4):1-12.
[23] MARINO K, SALAKHUTDINOV R, GUPTA A. The more you know: Using knowledge graphs for image classification[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA:IEEE,2017:20-28.
[24] YU Ping, ZHAO Yang, LI Chuanyuan, et al. Structure-aware human-action generation[M]//VEDALDI A, BISCHOF H, BROX T, et al. Computer Vision-ECCV 2020. ECCV 2020. Lecture Notes in Computer Science. Cham: Springer , 2020, 12375:18-34.
[25] GRETTON A, BORGWARDT K M , RASCH M J, et al. A kernel two-sample test[J]. Journal of Machine Learning Research, 2012,13:723-773.
[26] XU Qiantong, HUANG Gao, YUAN Yang, et al. An empirical study on evaluation metrics of generative adversarial networks[J]. arXiv preprint arXiv:1806.07755, 2018.
[27] ZHAO Yang, ZHANG Jiangyi, CHEN Changyou. Self-adversarially learned bayesian sampling[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii, United States:AAAI, 2019,33: 5893-5900.
[28] HAN Jun, LIU Qiang. Stein variational gradient descent without gradient[J]. arXiv preprint arXiv:1806.02775,2018.
[29] WALKER J, MARINO K, GUPTA A, et al. The pose knows: Video forecasting by generating pose futures[C]// 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy:IEEE,2017:3352-3361.
[30] WANG Zhenyi, YU Ping, ZHAO Yang, et al. Learning diverse stochastic human-action generators by learning smooth latent transitions[J]. arXiv preprint arXiv:1912.10150,2019.
[31] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of Wasserstein GANs[C]//Advances in Neural Information Processing Systems.Long Beach:NIPS, 2017: 5767-5777.

(上接第 32 页)

[5] TOUNSI Y, KUMAR M, NASSIM A,et al. Speckle denoising by variant nonlocal means methods[J]. Applied Optics, 2019, 58(26):7110-7120.
[6] SANTOS C A N, MARTINS D L N, MASCAREHAS N D A. Ultrasound image despeckling using stochastic distance-based BM3D[J]. IEEE Transactions on Image Processing : A publication of the IEEE Signal Processing Society,2017,26(6): 2632-2643.
[7] 黄小芬. 运动模糊车牌图像的复原算法研究与实现[J]. 山西大同大学学报(自然科学版),2018,34(3):35-38.
[8] 曹莹,段玉波,刘继承. 基于多尺度的形态滤波降噪方法[J]. 化工自动化及仪表,2015,42(11):1202-1205.
[9] 刘琬臻,付忠良. 基于局部方差改进的超声图像各向异性扩散去噪算法[J]. 计算机应用,2013,33(9):2599-2602.
[10] 刘勍,温志贤,杨筱平,等. 现代数字图像噪声滤除技术及其评价[J]. 自动化与仪器仪表,2012(2):146-148.
[11] 陈文青,王佰玲,倪国强. 基于各向异性扩散的单幅图像去雾算法[J]. 光学技术,2017,43(4):354-358.
[12] 郑良仁,代文征,靳宗信,等. 基于 DWT 和奇异值分解的图像增强算法[J]. 现代电子技术,2017,40(15):21-24.
[13] 孟丽茹,赵岩,王世刚,等. 基于 2D 视觉注意模型的全参考图像质量评价方法[J]. 吉林大学学报(信息科学版),2014,32(6):563-568.
[14] 闫乐乐,李辉,邱聚能,等. 基于区域对比度和 SSIM 的图像质量评价方法[J]. 应用光学,2015,36(1):58-63.
[15] 谢小甫,周进,吴钦章. 基于无参考结构清晰度的自适应自动对焦方法[J]. 光电工程,2011,38(2):84-89.
[16] 许春冬,周静,龙清华,等. 基于 coif-5 小波的心音自适应阈值降噪方法[J]. 科学技术与工程,2019,19(2):106-113.