计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(6):388−393 [doi: 10.15888/j.cnki.csa.008530]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

关注全局真实度的文本到图像生成①

胡　成, 胡莹晖, 刘兴云

(湖北师范大学 物理与电子科学学院, 黄石 435002)
通信作者: 胡　成, E-mail: 1446429273@qq.com

摘　要: 针对文本和图像模态在高维空间中相互映射的困难问题, 提出以全局句子向量为输入, 以堆叠式结构为基础的生成对抗网络 (GAN), 应用于文本生成图像任务. 该网络融入双重注意力机制, 在空间和通道两大维度上寻求特征融合的更大化, 同时增加真实度损失判别器作为约束. 所提方法在加利福尼亚理工学院的 CUB 鸟类数据集上实验验证, 用 Inception Score 和 SSIM 作为评估指标. 结果表明, 生成图像具有更真实的细节纹理, 视觉效果更加接近于真实图像.

关键词: 文本生成图像; 堆叠式生成对抗网络; 双重注意力机制; 真实度损失; 文本检测

引用格式: 胡成,胡莹晖,刘兴云.关注全局真实度的文本到图像生成.计算机系统应用,2022,31(6):388–393. http://www.c-s-a.org.cn/1003-3254/8530.html

Text-to-image Generation Focusing on Global Fidelity

HU Cheng, HU Ying-Hui, LIU Xing-Yun

(School of Physics and Electronic Science, Hubei Normal University, Huangshi 435002, China)

Abstract: Considering the difficulty in mutual mapping between text and image modalities in high-dimensional space, this study proposes a generative adversarial network (GAN) based on a stacked structure with global sentence vectors as input for the application of text-to-image generation tasks. The network incorporates a dual attention mechanism for greater integration of features in the two dimensions of space and channel. At the same time, we add the discriminator for fidelity loss as a constraint. The proposed method is experimentally verified on the Caltech-UCSD Birds (CUB) dataset, with Inception Score and SSIM as the evaluation indexes. The results show that the generated image has more realistic detail textures, and the visual effect is closer to the real image.

Key words: text-to-image generation; stacked GAN; dual attention mechanism; fidelity loss; text detection

从文本生成图像是计算机视觉领域十分重要的一大方向, 即通过给定输入文本语句, 生成相对应内容的图像, 具有广泛的应用. 例如小说配插图、图像编辑、图像的检索等等. 生成对抗网络 (GAN)[1] 被应用在文本生成图像上取得了一定的可观效果. Reed 等[2] 最先将 GAN 应用到文本生成图像中, 生成了肉眼可接受的64×64分辨率的图像, 验证 GAN 在文本生成图像的可行性. Zhang 等[3] 提出堆叠式的结构 (StackGAN), 将任务阶段化, 逐步细化生成的图片, 生成图像达到256×256分辨率. 后来, Zhang 等人改进了 StackGAN, 提出端到端树状结构的 StackGAN++[4], 通过多尺度的判别器和生成器, 提高了生成图像的质量和清晰度, 但是图像整体亮度偏暗淡, 与数据集样本存在偏差, 同时缺少生成图像真实度的判定.

注意力机制在图像和自然语言处理方面有着广泛的应用. Zhang 等[5] 提出的 SAGAN 首次将自我注意力机制与 GAN 结合, 减少参数计算量的同时, 也聚焦了更多的全局信息. Fu 等[6] 提出双重注意力机制, 在空间和通道两个维度进行特征融合, 用于语义分割. Tang 等[7] 结合双重注意力机制, 应用于语义图像合成.

① 收稿时间: 2021-09-03; 修改时间: 2021-09-26; 采用时间: 2021-10-19; csa 在线出版时间: 2022-05-26

受到以上实验的启发, 针对出现的问题, 我们提出结合双重注意力机制的端到端模型, 该模型基于 StackGAN++基本结构, 以双重注意力机制去最大化融合文本和图像的特征, 树状结构生成低到高分辨率 (128×128) 的图像; 添加 VGG19[8] 预训练模型, 提取生成图像和真实图像特征, 计算相似度损失.

该模型旨在提高生成图像的全局真实度. 全局真实度指图像内容的完整度, 颜色的鲜明度, 场景的对比度和亮度符合人眼视觉感知的程度.

我们在 CUB[9] 鸟类数据集上验证了该方法, 并使用 IS 和 SSIM 指标判定生成图像的多样性、质量和全局真实度. 实验结果分析证明, 与原有技术相比, 我们模型生成的图像一定程度上呈现了更多的鸟类特征, 并提升了整体的亮度和颜色鲜明度, 使生成图像感知上更加接近于真实图像.

1 模型及方法

1.1 模型结构

本文网络结构整体框图如图 1 所示. 结构主要由文本编码器、2 个生成器、3 个判别器和 VGG19 网络构成. 文本编码器使用文献 [10] 中提供的字符级编码器 (char-CNN-RNN), 生成器采用前后级联的方式, 第一个生成器包含 1 个全连接层和 4 个上采样层, 第二个生成器包含连接层, 空间和通道注意力模块, 2 个残差网络[11] 和 1 个上采样层. VGG19 网络作为额外约束, 判别生成图像和真实图像的相似度.

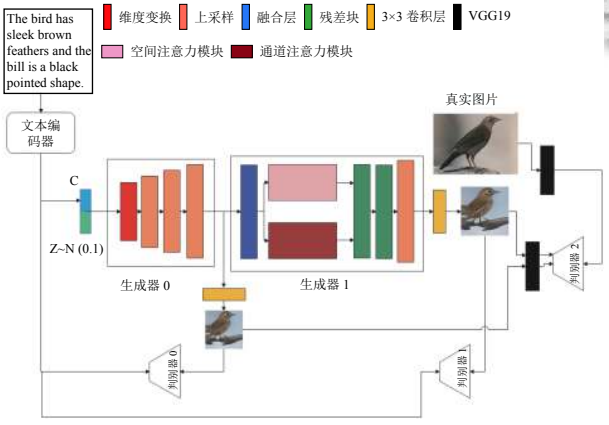

图 1 模型结构

网络大致分为两个阶段, 每个阶段都包含多个输入, 如式 (1) 所示:

$$\begin{cases} c_i = \varphi(t_i) \\ C_i = F_{ca}(c_i) \\ f_0 = F_0(Z) \\ f_i = F_i(f_{i-1}, C_i) \\ I_i = G_j(f_i),\ j \in \{1,2\} \end{cases} \tag{1}$$

其中, φ表示文本编码器, c_i表示全局句子向量, F_{ca}表示条件增强模块, F_i表示全连接层, G_j表示生成器, I_i表示生成器输出.

1.1.1 双重注意力机制

由于图像像素区域和文本存在对应关系, 不同通道存在依赖关系, 我们引入空间和通道注意力机制, 输入为文本向量和低分辨率特征的融合矩阵, 引导生成器更多关注整体特征的关联性和匹配度. 由于高分辨率图像是在低分辨率图像的基础上进行细化, 所以低分辨率图像的好坏决定着最终输出的好坏. 虽然低分辨率图像更加的模糊, 缺少细节, 但是却保留着更多的全局特征. 所以我们将机制放置在 G1 的连接层后, 即残差模块前, 引导生成器在低分辨率维度上关注更多的全局特征. 注意力机制模块如图 2, 图 3 所示.

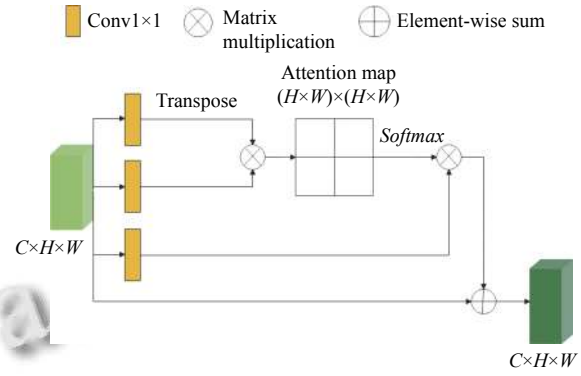

图 2 空间注意力模块 (SAM) 结构

对于通道注意力模块而言, 输入是文本向量和上阶段图像矩阵连接后卷积得到的特征图($h \in R^{C\times H\times W}$). 其流程对应公式如式 (2):

$$\begin{cases} \alpha = \omega_q \omega_k^{\mathrm{T}},\ \alpha \in R^{C\times C} \\ \beta = Softmax(\alpha) \\ \gamma = \beta\omega_v,\ \beta \in R^{C\times H\times W} \\ H = \sigma\gamma + h \end{cases} \tag{2}$$

其中, $\omega_q \in R^{C\times H\times W}$、$\omega_k \in R^{C\times H\times W}$、$\omega_v \in R^{C\times H\times W}$ 分别代表特征图经过三个通道的 1×1 卷积后得到的特征矩阵. 对ω_q和ω_k转置应用一次矩阵乘法, 随后经过*Softmax*层得到位置注意力映射图, 再与特征矩阵ω_v进行一次

矩阵乘法运算, 最后乘上权重因子σ和输入特征图 (h) 逐元素相加得到输出, 以此来增强通道特征图之间的语义依赖性. 权重因子初始化为 0, 并逐步学习变化.

空间注意力机制忽略了通道间的语义关联性, 关注像素间的特征信息, 运算与通道注意力机制类似. 两个模块输出最后从通道维度进行拼接, 得到最终的结果.

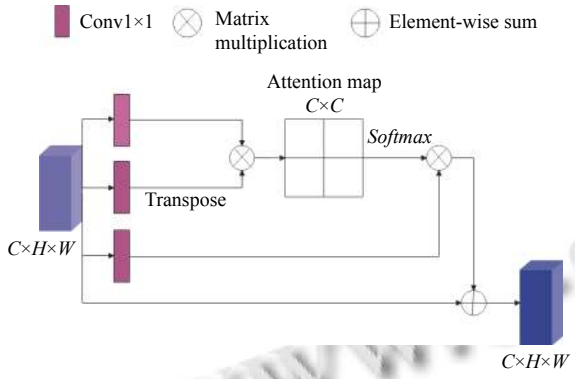

图 3 通道注意力模块 (CAM) 结构

1.1.2 VGG19

增强型超分辨率生成对抗网络 (ESRGAN)[12] 中指出, 使用 VGG19 的第 5 个 maxpool 层前的最后一层卷积层去提取图像特征, 使得生成图像特征在亮度和颜色感知上更接近于真实图像. 受其启发, 我们引入 VGG19 的前 35 层网络层进行预训练处理, 用来提取生成图像和真实图像的特征, 求取两者的 $L1$ 损失, 作为生成图像真实度的判别约束.

1.2 时间复杂度

空间注意力模块输入$C\times H\times W$矩阵, 计算相似特征图的时间复杂度为$\mathrm{O}(CN^2)(N=H\times W)$, *Softmax* 的时间复杂度为$\mathrm{O}(N^2)$, 加权求和的时间复杂度为$\mathrm{O}(CN^2)$, 所以空间注意力模块的时间复杂度为$\mathrm{O}(CN^2)$. 以此类推, 通道注意力模块的时间复杂度为$\mathrm{O}(C^2N)$. 而该模型生成器的最后一层卷积层的时间复杂度为$\mathrm{O}(N4kC^2)$ ($k=3$, 表示卷积核大小) 由于$N=64\times64$, $C=64$, 所以$\mathrm{O}(CN^2)>\mathrm{O}(N4kC^2)$, 即双重注意力模块在本实验中, 虽然取得良好的效果, 但增加了算法的时间复杂度, 在训练时间上并不占优势.

1.3 损失函数

1.3.1 生成器损失

生成器损失包含非条件损失和条件损失两部分. 非条件损失用来判别图像是真实的或是虚假的; 条件损失用来判别图像和文本是否匹配.

$$L_G=\underbrace{-E_{I_i\sim p_{G_i}}\left[\log D_i(G_j(f_i))\right]}_{\text{非条件损失}}\underbrace{-E_{I_i\sim p_{G_i}}\left[\log D_i(G_j(f_i),C)\right]}_{\text{条件损失}} \tag{3}$$

其中, $G_j(f_i)$表示生成器的输出. $j=0,1$, 代表两个生成器. I_i表示生成的第i个图像, 来自于生成图像分布p_{G_i}

两个生成器对应两个尺度的图像分布生成, 各自后面接一个判别器. 不同尺度生成图像送入判别器中, 计算交叉熵损失, 返回真假概率和图像文本匹配概率. 生成器G_j和判别器D_i两者交替优化, 以致收敛. L_G值越小, 代表优化效果越好.

1.3.2 判别器损失

判别器损失包含非条件损失、条件损失和真实度损失 3 部分.

$$\begin{aligned}L_D=&\underbrace{-E_{R_i\sim p_{\text{data}_i}}[\log D_i(R_i)]-E_{I_i\sim p_{G_i}}[\log(1-D_i(G_j(f_i)))]}_{\text{非条件损失}}\\&\underbrace{-E_{R_i\sim p_{\text{data}_i}}[\log D_i(R_i,C)]-E_{I_i\sim p_{G_i}}[\log(1-D_i(G_j(f_i),C))]}_{\text{条件损失}}\\&-\underbrace{\mu\mathrm{L1}}_{\text{真实度损失}}\end{aligned} \tag{4}$$

$$L1=E_{f_i}\|G(f_i)-R_i\|_1 \tag{5}$$

其中, $L1$ 表示真实度损失. 由 VGG19 提取真实图像和不同尺度图像的特征空间, 送入判别器计算 $L1$ 范数距离损失, 通过最小化损失, 达到优化效果.

R_i: 第i个真实图像, 来自于真实图像分布p_{data_i}.

I_i: 生成的第i个图像, 来自于生成图像分布p_{G_i}.

μ: 损失系数, 设其值为 0.001.

非条件损失分别计算真实图像、各个尺度生成图像的交叉熵损失, 优化判别器判别真假的能力. 条件损失采用正负对比计算, 正计算包括真实图像和对应标签, 生成图像和对应标签两个组合, 负计算指真实图像和不对应标签. 通过正负对比学习, 优化判别器判别图像文本匹配能力.

2 实验结果和分析

2.1 实验环境

本文实验基于搭载 GTX1070i 显卡的 CentOS 7 操作系统, 使用 Python 2.7 编程语言, PyTorch 框架.

实验设置训练过程中生成器和判别器学习率为 0.000 1, batch_size 为 8, 迭代次数为 160 次.

2.2 实验数据集及评估指标

2.2.1 数据集

本文实验方法在 CUB200-2011 数据集上进行验证. CUB200-2011 数据集由加州理工学院提出, 共包含 11 788 张鸟类图像, 200 种鸟类, 每张图像对应 10 个文本描述语句. 除类别标签外, 每个图像都会用 1 个边界框、15 个零件关键点和 312 个属性进行进一步注释. 其中, 训练集 8 855 张图像, 测试集 2 933 张图像, 如表 1.

表 1 实验数据集

数据集	CUB
训练集	8 855
测试集	2 933
类别数	20

2.2.2 评估指标

本文采用 Inception Score (*IS*) 和 *SSIM* 作为评估标准. *IS* 基于预先在 ImageNet 数据集[13] 上训练好的 Inception V3 网络. 其计算公式如下:

$$IS(G) = \exp(E_{x\sim p_g} D_{\mathrm{KL}}(p(y|x) \,\|\, p(y))) \tag{6}$$

其中, $x \sim p_G$表示生成的图片, y表示 Inception V3 预测的标签, D_{KL}表示 KL 散度

公式表明, *IS* 评估生成图像的多样性和质量, 好的模型应该生成清晰且多样的图像, 所以边际分布$p(y|x)$和条件分布$p(y)$的 KL 散度越大越好, 即 *IS* 值越大越好. 但是 *IS* 存在不足之处, 它不能判定生成图像的真实度, 所以我们引入 *SSIM* 指标.

SSIM (structural similarity), 结构相似性度量指标, 已被证明更符合人眼的视觉感知特性. 我们用其评估生成图像的真实度. *SSIM* 包含亮度、对比度、结构 3 个度量模块. 其计算公式如下:

亮度对比函数:

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{7}$$

对比度对比函数:

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{8}$$

结构对比函数:

$$s(x,y) = \frac{\sigma_{xy} + c_a}{\sigma_x\sigma_y + c_a} \tag{9}$$

最后把 3 个函数组合起来得到 *SSIM* 指数函数:

$$SSIM(x,y) = [l(x,y)]^{\alpha}[c(x,y)]^{\beta}[s(x,y)]^{\gamma} \tag{10}$$

2.3 实验结果及比较

我们将模型在 CUB 数据集的训练集上进行训练, 并在测试集上进行了验证实验. 下图展示训练过程中收敛的判别器损失和生成器损失, 以及 *IS* 值.

结合图 4、图 5 我们看出, 判别器损失逐步收敛到 (2, 3) 区间, 保持平稳震荡; 生成器损失逐步上升到 (25, 30) 区间, 基本保持缓慢上升的趋势. 模型判别器和生成器形成对抗趋势, 逐步保持平衡状态.

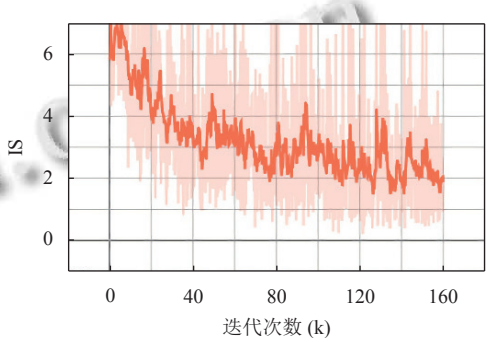

图 4 判别器损失

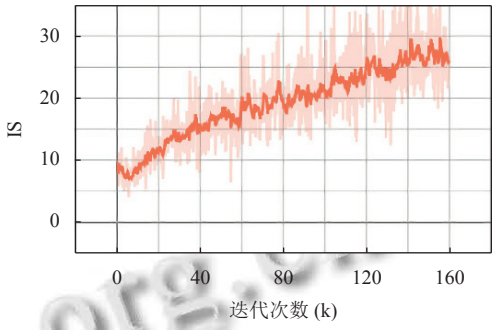

图 5 生成器损失

由图 6 看出, 我们的模型 *IS* 值最高可达到 5.6 左右.

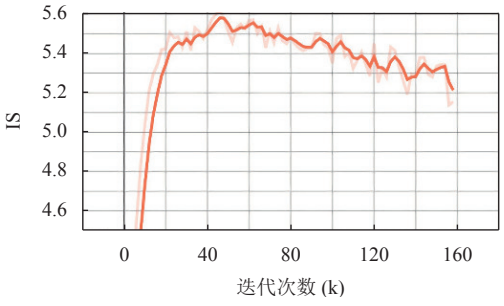

图 6 Inception Score

为了节省内存占用率, 我们将 StackGAN++缩减为两个阶段, 生成128×128分辨率的图像, 在 CUB 数据集上进行训练和测试. 并和我们的方法的测试结果进

行了对比, 实验结果如图 7 所示.

由图 7 可以很明显观察到, StackGAN++模型生成的128×128分辨率的图像亮度偏暗, 与真实图像存在差异. 我们的方法生成的图像颜色更加的鲜艳, 图像整体更加的明亮, 在背景颜色、鸟类形状和整体感知上, 更加地接近真实图像. 同时, 鸟类的羽毛纹理更加的丰富, 例如图 7(b)–图 7(d).

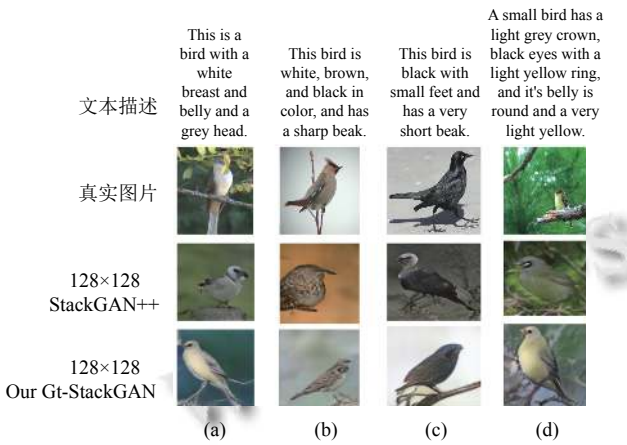

图 7 测试结果示例

我们列举以往不同模型在 CUB 数据集上的 *IS* 值, 进行一个对比, 见表 2. 我们所提方法评估的 *IS* 值能够达到 5.4, 高于所比较的以往模型.

表 2 各模型在 CUB 上的 IS 值

模型	CUB
64×64 GAN-INT-CLS	2.88±0.4
128×128 GAWWN	3.62±0.7
256×256 StackGAN	3.70±0.4
256×256 StackGAN++	3.84±0.6
128×128 Our Gt-StackGAN	5.40±0.5

为了定量地评估我们模型对真实度提升的贡献, 我们用 *SSIM* 指标在生成图像和真实图像做相似性评估, 在 StackGAN++模型和我们模型做了对比实验, 见表 3.

表 3 模型在 CUB 上的 *SSIM* 值

模型	*SSIM*
64×64 StackGAN++	0.15
128×128 StackGAN++	0.20
64×64 Our Gt-StackGAN	0.17
128×128 Our Gt-StackGAN	0.25

由表 3 看出, 相同模型下, 更高分辨率的生成图像具有更高的 *SSIM* 值, 符合图像质量提升导致真实度提升的逻辑. 以此为前提, 对比不同模型在相同分辨率的 *SSIM* 值, 我们的模型值更高, 则图像真实度相比更高. 结合实验结果图来看, 我们模型生成的图像人眼感知与真实图像样本也更加相似.

3 结论

本文提出一种以堆叠式结构为基础, 着重关注图像全局特征真实度的生成对抗网络, 应用于文本生成图像任务. 实验结果证明, 同以往的模型对比, 结果图像更加专注于全局特征, 颜色的鲜明度和整体视觉效果更加具有真实感, 更接近于真实图片. 这是因为我们引入双重注意力机制引导图像学习对应文本的更多特征; 使用真实感损失约束, 提高生成图像的真实感. 在文本单词向量级别, 增添图像子区域的细节, 提升文本和图像的语义一致性, 应用于更加复杂的数据集, 会是接下来研究的一个方向.

参考文献

1 Goodfellow IJ, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Quebec: NIPS, 2014. 2672–2680.

2 Reed S, Akata Z, Yan XC, *et al*. Generative adversarial text to image synthesis. Proceedings of the 33rd International Conference on Machine Learning. New York: JMLR.org, 2016. 1060–1069.

3 Zhang H, Xu T, Li HS, *et al*. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017. 5908–5916.

4 Zhang H, Xu T, Li HS, *et al*. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1947 –1962. [doi: 10.1109/TPAMI.2018.2856256]

5 Zhang H, Goodfellow I, Metaxas D, *et al*. Self-attention generative adversarial networks. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 7354–7363.

6 Fu J, Liu J, Tian HJ, *et al*. Dual attention network for scene segmentation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 3146–3154.

7 Tang H, Bai S, Sebe N. Dual attention GANs for semantic image synthesis. Proceedings of the 28th ACM International

Conference on Multimedia. Seattle: ACM, 2020. 1994–2002.
8 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations. San Diego: ICLR, 2015.
9 Wah C, Branson S, Welinder P, *et al*. The caltech-UCSD birds-200-2011 dataset. California: California Institute of Technology, 2011.
10 Reed S, Akata Z, Lee H, *et al*. Learning deep representations of fine-grained visual descriptions. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 49–58.
11 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 770–778.
12 Wang XT, Yu K, Wu SX, *et al*. ESRGAN: Enhanced super-resolution generative adversarial networks. Leal-Taixé L, Roth S. Computer Vision—ECCV 2018 Workshops. Cham: Springer, 2018. 63–79.
13 Russakovsky O, Deng J, Su H, *et al*. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(3): 211–252. [doi: 10.1007/s11263-015-0816-y]