一种空间控制的快速图像风格化方法

侯泽宇 许 毅
(武汉理工大学计算机科学与技术学院 湖北 武汉 430061)

摘 要 针对现有的图像风格化方法中内容与风格不匹配而导致风格化效果较差的问题,在分析快速图像风格化方法的基础上,提出一种空间控制的快速图像风格化方法(FSTSC)。将图像分为多个小块(patch),使用块匹配方法寻找风格图像和内容图像最匹配的 patch。利用图像语义分割方法对图像进行预处理,根据不同内容类别的引导向量计算语义风格损失,控制内容图像的不同内容区域获得对应的风格化处理,从而提升图像风格化的效果。实验将该方法与现有的快速图像风格化方法在图像风格化效果和时间两方面进行对比,分别提升约 12.5% 和 25.5%,表明该方法是一种合理和有效的方法。

关键词 空间控制 快速图像风格化 块匹配 语义分割 风格转换网络

中图分类号 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2020.10.038

A METHOD OF FAST IMAGE STYLE TRANSFER BASED ON SPATIAL CONTROL

Hou Zeyu Xu Yi
(*School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430061, Hubei, China*)

Abstract To solve the problem of poor image style transfer effect of existing image style transfer methods caused by the mismatch between content and style, based on the analysis of fast image style transfer method, this paper proposes a fast image style transfer method based on spatial control(FSTSC). It divided the image into multiple patches and used the patch matching method to find the best matching patch between style image and content image. Then the semantic segmentation method was used to preprocess the image, calculated the semantic style loss according to the guidance vector of different content categories, and controlled the different content regions of the content image to obtain the corresponding style transfer processing, thus improving the image style transfer effect. In the experiment, this method was compared with the existing fast image style transfer method in terms of image style transfer effect and time, which improved about 12.5% and 25.5% respectively. The results show that this method is reasonable and effective.

Keywords Spatial control Fast image style transfer Patch matching Semantic segmentation Style transfer network

0 引 言

图像风格化技术可以将艺术作品的风格迁移到另一种现实生活的图像中,使得图像在保留原始内容的情况下加入卡通、动漫、黑白、油画、抽象画等极具特色的艺术风格,创作出具有艺术风格的作品。目前,图像风格化已经在摄影和视频方面有了较为广泛的应用,在动漫制作、电影特效和照片滤镜等领域有了重大突破,极大地减少了艺术设计工作者的工作量,同时可以帮助非专业用户方便快速地创作出具有艺术风格的作品,有重要的现实意义和实际应用价值。

近年来,学者们在深度学习的图像风格化方向进行了很多研究,在图像风格化速度方面,可分为慢速图像风格化方法和快速图像风格化方法。慢速图像风格化方法的研究源于 Gatys 等[1]使用深度卷积神经网络

收稿日期:2019-07-19。**侯泽宇**,硕士生,主研领域:计算机视觉,机器学习。**许毅**,副教授。

(Convolutional Neural Network,CNN)进行纹理合成的工作。之后他们又提出了迭代更新的图像风格化方法[2],该方法利用卷积神经网络 VGGNet 学习到的通用特征表示分离图像的内容与风格,将风格转换问题转变成为神经网络的优化问题,但其生成一幅风格化图像需要反复进行前反向计算迭代,耗时较长。Li 等[3]在 Gatys 的基础上使用马尔可夫随机场(Markov Random Field,MRF)的方法代替 Gram 矩阵来确定风格损失,图像风格化的效果优于 Gatys 的方法,但生成风格化图像依然需要反复的迭代,耗时较长。Johnson 等[4]首次提出了快速图像风格化方法,该方法通过训练一个风格转换网络来进行风格的迁移,从而将图像风格化的速度提升了近 1 000 倍,但是作者提出的风格转换网络参数约为 100 万个,过多的网络参数导致了风格转换网络的训练耗时较长。Ulyanov 等[5]在 Johnson 的基础上提出了多尺度的风格转换网络 Texture Nets,网络更加轻巧,图像风格化速度具有一定的提升,但网络的训练仍需要较长的时间。

除了风格化速度方面的研究之外,学者们还提出了一系列针对特定情况下的图像风格化效果进行提升的方法。Gatys 等[6]提出使用颜色控制的方法,在图像风格化的过程中保留内容图像原本的颜色信息,在需要颜色保留的情况下可以提升图像风格化效果,但是对于需要颜色变化的图像风格化处理,该方法并不适用。吴联坤[7]使用前景背景分离的方法,提升了纯色背景图像的风格化效果,但是对于非纯色背景的图像,图像风格化效果较差。

根据上述的分析,针对图像风格化速度和效果两个关键点,本文提出利用更轻量的风格转换网络,并结合图像物体空间语义信息控制的方法,在各种情况下均可以获得更优秀的风格化生成图像。

1 方法设计

1.1 空间控制

现实世界中的图像场景一般比较复杂,在给内容图像进行风格转换的过程中应该尊重内容图像的场景语义信息,在图像的空间信息层级进行一定的控制,约束风格化的空间范围。因此,本文在图像块匹配方法的基础上提出语义风格损失,以控制图像语义内容与风格的匹配程度,空间控制和语义风格损失的计算原理如图 1 所示。

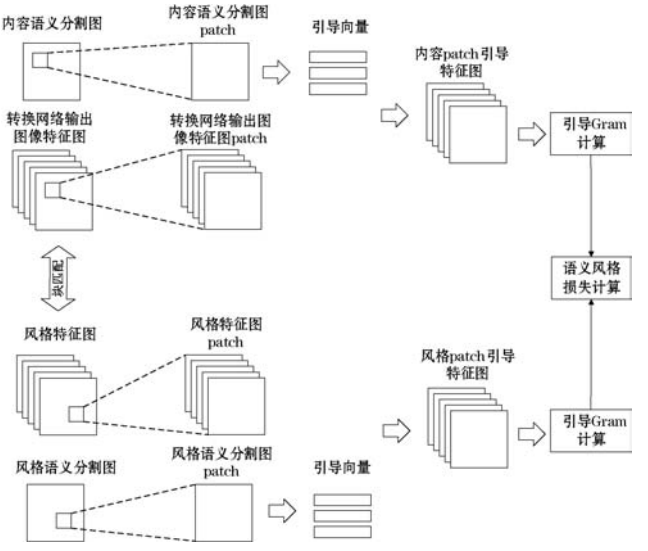

图 1 空间控制和语义风格损失计算原理图

FSTSC 方法使用一个尺寸为 3×3 的窗口,以 1 为步长在内容图像经过转换网络的输出图像的特征图上进行滑动,每次滑动该窗口都作为一个 patch,通过块匹配方法在风格特征图中寻找最匹配的 patch。

块匹配方法使用归一化交叉相关(Normalization Cross Correlation,NCC)函数作为匹配准则[8],通过计算图像之间的互相关性来确定匹配的程度,即在风格特征图 $\boldsymbol{\Phi}_l(s)$ 中搜索与 $\boldsymbol{\Phi}_l(y)$ 中选取的 patch 块 NCC 最高的区域作为对应匹配,计算方法为:

$$\boldsymbol{\Psi}_{NN(i)}(\boldsymbol{\Phi}_l(s)) = \underset{j=1,2,\cdots,P_s}{\operatorname{argmin}} \frac{\boldsymbol{\Psi}_i\boldsymbol{\Phi}_l(y) \cdot \boldsymbol{\Psi}_l\boldsymbol{\Phi}_l(s)}{|\boldsymbol{\Psi}_i\boldsymbol{\Phi}_l(y)| \cdot |\boldsymbol{\Psi}_l\boldsymbol{\Phi}_l(s)|} \tag{1}$$

式中:$\boldsymbol{\Psi}_i\boldsymbol{\Phi}_l(y)$ 为在转换网络输出图像的特征图上选取的第 i 个 patch;$\boldsymbol{\Psi}_j\boldsymbol{\Phi}_l(s)$ 为在风格特征图上选取的第 j 个 patch。

在寻找到匹配的 patch 之后,使用 DeepLab-v2 图像语义分割算法[9]对风格图像和内容图像进行语义分割操作,可以获得具有多个内容类别的语义分割图,生成引导矩阵,该矩阵记录了原图像各个像素对应的内容类别。利用引导矩阵获取 patch 中各个内容类别对应的 patch 引导向量,从而计算出 patch 引导 Gram 矩阵。假设 c 为 patch 块 p 的内容类别之一,首先通过引导矩阵得到类别 c 对应的引导向量,即矢量化后的二进制掩码,之后将 patch 特征图 $\boldsymbol{F}_l(p)$ 与类别 c 的 patch 引导向量 $\boldsymbol{T}_c(p)$ 相乘的结果作为 patch 引导特征图:

$$\boldsymbol{F}_{l,c}(p) = \boldsymbol{F}_l(p)\boldsymbol{T}_c(p) \tag{2}$$

进一步计算出 patch 引导 Gram 矩阵:

$$\boldsymbol{G}_{l,c}(p) = \boldsymbol{F}_{l,c}(p) \cdot \boldsymbol{F}_{l,c}(p)^{\mathrm{T}} \tag{3}$$

l 层 patch 语义风格损失根据 VGG-16 损失网络中第 l 层输出特征图的 patch 引导 Gram 矩阵进行比较,将通过块匹配方法获得的两个最匹配 patch 之间的所有内容类别引导 Gram 矩阵的欧氏距离之和作为 patch

语义风格损失,l 层语义风格损失为所有 patch 语义风格损失之和:

$$L_{\text{style}}^{l} = \sum_{i=1}^{P}\sum_{c=1}^{C} \| \boldsymbol{G}_{l,c}(\boldsymbol{\Psi}_i(\boldsymbol{\Phi}_l(y))) - \boldsymbol{G}_{l,c}(\boldsymbol{\Psi}_{NN(i)}(\boldsymbol{\Phi}_l(s))) \|^2 \quad (4)$$

式中:P 为所有 patch 的总数;C 为每个 patch 包含的内容类别总数。总语义风格损失为 relu1_2、relu2_2、relu3_3 和 relu4_3 四层语义风格损失之和:

$$L_{\text{style}} = \sum_{l=1}^{L} L_{\text{style}}^{l} \quad (5)$$

对于内容损失,使用内容图像经过转换网络的输出图像 y 和内容图像 c 在损失网络的 relu4_2 层输出的特征图之间的欧氏距离作为内容损失:

$$L_{\text{content}} = \| \boldsymbol{\Phi}_l(y) - \boldsymbol{\Phi}_l(c) \|^2 \quad (6)$$

FSTSC 方法使用以上的方式对图像风格化过程中图像风格化的空间范围进行控制,为了使生成图像更加平滑,还加入了全变差正则化损失 L_{TV}。因此,总损失 L_{total} 为内容损失、语义风格损失与全变差正则化损失的加权之和:

$$L_{\text{total}} = \lambda_c L_{\text{content}} + \lambda_s L_{\text{style}} + \lambda_{\text{TV}} L_{\text{TV}} \quad (7)$$

式中:λ_c 为内容损失的权重;λ_s 为语义风格损失的权重;λ_{TV} 为全变差正则化损失的权重。

1.2　总体结构

为了进一步降低风格转换网络的训练时间,FSTSC 方法设计了一个新的风格转换网络,精简了网络架构,减少了网络的参数数量,FSTSC 方法的总体结构如图 2 所示。

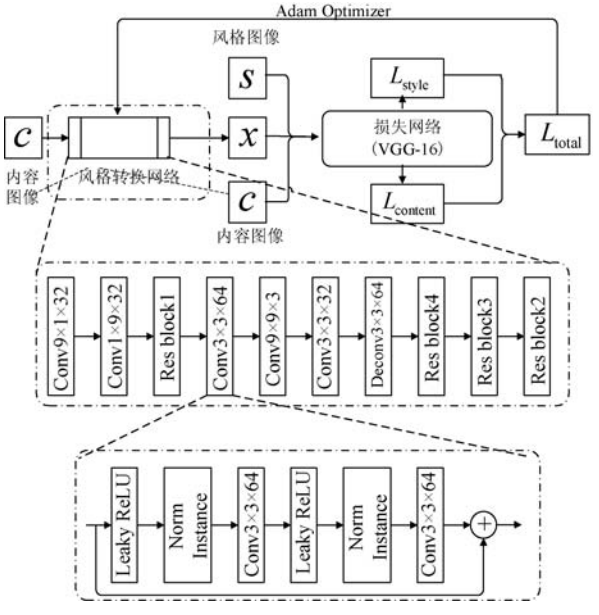

图 2　总体结构图

FSTSC 方法主要包括一个风格转换网络和一个损失网络。风格转换网络以残差网络[10]作为基础,共包含三个部分。第一部分为下采样,根据 GoogLeNet 中非对称卷积的设计思想[11],将一个 9×9 卷积层分解为9×1 和 1×9 的两个卷积层,可以在保证感受野相同的情况下减少网络的参数数量,并采纳了 Radford 等的建议并未使用池化层,而是使用小步长的卷积层进行下采样[12],所以该部分共包含了 9×1 卷积层、1×9 卷积层和 3×3 卷积层。第二部分由 4 个相同的残差块组成,每个残差块均包含两个 3×3 卷积层,在卷积层之间使用实例归一化(Instance Norm)层和 Leaky ReLU 层进行连接,以增强网络的泛化能力并解决部分神经元失效的问题[13]。第三部分为上采样,使用与下采样相同数量的反卷积层将图像进行放大,还原至与输入图像相同的尺寸,最后的卷积层使用 tanh 作为激活函数以保证输出值在[0,255]内[14]。该风格转换网络降低了网络模型的复杂度,能够提升图像风格化速度,提高图像风格化效率。

该模型利用在 ImageNet 数据集上进行预训练的 VGG-16 网络作为损失网络,基于此网络提取图像的高维特征,通过不同层的响应结果构建内容损失 L_{content} 和语义风格损失 L_{style} 分别用于衡量图像之间的内容和风格差异,从而计算出总损失。之后使用 Adam 优化方法逐步减小总损失值来进行风格转换网络权值 W 的更新,完成风格转换网络的训练:

$$W = \arg\min_{w} L_{\text{total}} \quad (8)$$

最终将风格图像的风格保存到风格转换网络模型中,在使用阶段,将内容图像输入到风格转换网络就能够快速地得到生成图像,从而进行快速图像风格化处理。

2　实　验

2.1　实验环境

FSTSC 方法的实验环境采用六核心的 Intel Core i7-8700K CPU,内存为 16 GB,运行 64 位 Ubuntu 18.04 操作系统,使用 NVIDIA GTX 1080Ti GPU,显存为 11 GB。实验主要使用 Python 作为编程语言,通过开源的深度学习框架 TensorFlow 进行实现,并使用 CUDA 和 CuDNN 技术实现 GPU 加速。

2.2　训练细节

风格转换网络的训练使用 MSCOCO 2014 数据集,共采用 80 000 幅图像作为训练集,设置 batch size 为 10,epoch 为 2,采用学习率自适应的优化算法 Adam Optimizer,设置学习率为 0.001。

2.3 实验结果及分析

将 FSTSC 方法与 Gatys 等[2]和 Johnson 等的方法[4]进行对比,实验包括风景、建筑、动植物等场景,实验结果如图 3 所示。可以看出,Gatys 等的方法和 Johnson 等的方法忽略了内容图像与风格图像在内容上的差异,在风格转换的过程中没有考虑语义内容信息,导致生成图像部分内容与风格不匹配,比如在第二行结果中的草地出现了鲜花的风格,第三行结果中的天空出现了大桥的风格,以及第四行结果中的夜空出现了建筑物的风格等。而 FSTSC 方法在进行图像风格化时,考虑到图像子空间区域中语义内容与风格的匹配关系,对风格图像和内容图像分别进行语义分割预处理,使得图像的内容与风格相匹配,图像风格化的结果更加真实。此外,相比现有方法,FSTSC 方法的生成图像边缘更加平滑,效果更加优秀。

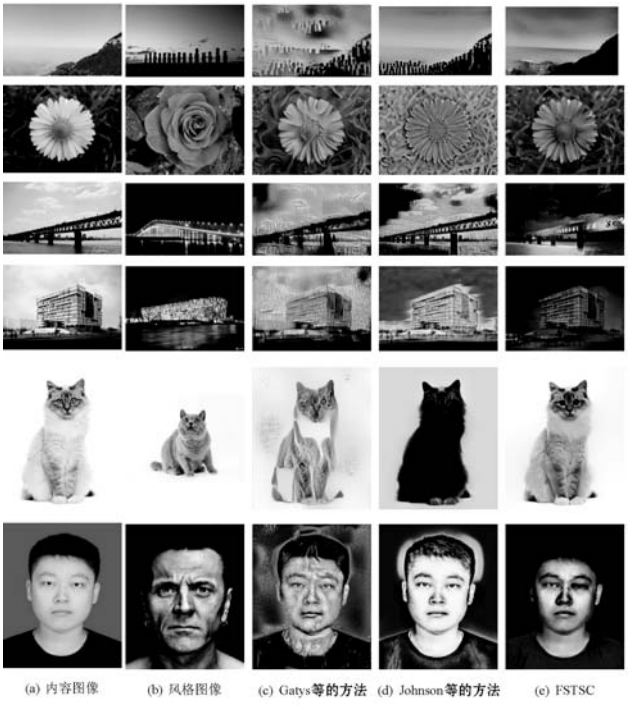

图 3 3 种图像风格化方法效果对比

对实验结果的评价主要包括主观评价和客观评价两方面,其中主观评价主要是对图像风格化的效果进行评价,以人为观察者,对生成图像的优劣作出主观的定性评价。共选取 20 人作为观察者参与实验,向每位观察者展示风格图像及对应的 3 种方法的生成图像,要求观察者在观看图像的同时根据与风格图像风格的相似程度给出相应的评价分值,使用分值 1 ~ 5 表示很差、较差、一般、较好、很好五个等级,实验共进行 20 次,最终将每幅结果图的平均分值作为最后的评价分值[15],结果如图 4 所示。实验结果表明,FSTSC 方法的生成图像的风格与风格图像最相似,图像风格化效果最好。

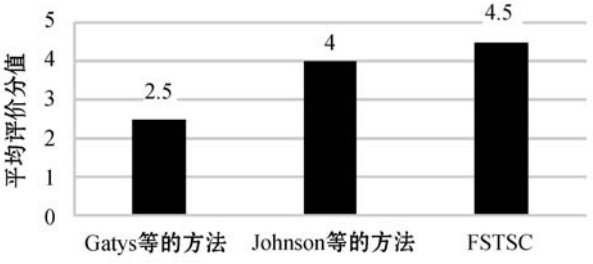

图 4 图像风格化效果主观评价

客观评价主要从风格转换网络的训练时间和图像风格化速度两方面进行比较。对于风格转换网络训练时间的比较,本文在相同实验环境下将 Johnson 等提出的残差转换网络[4]、Ulyanov 等提出的多尺度转换网络[5]与 FSTSC 方法使用的风格转换网络,经过 40 000 次迭代训练所需时间进行对比,结果如表 1 所示。

表 1 各风格转换网络训练时间对比

参数	Johnson 等的方法	Ulyanov 等的方法	FSTSC
训练时间/h	4.0	2.1	1.9

对于图像风格化速度的比较,由于 Gatys 等的风格化方法原理与其他方法不同,因此本文使用 Gatys 方法迭代 500 次的图像风格化时间与 Johnson 等的方法、Ulyanov 等的方法、FSTSC 方法进行对比,结果如表 2 所示。

表 2 各图像风格化方法风格化时间对比 s

图像尺寸	Gatys 等的方法	Johnson 等的方法	Ulyanov 等的方法	FSTSC
256 × 256	18.47	0.017	0.023	0.012
512 × 512	51.65	0.054	0.075	0.041
1 024 × 1 024	208.93	0.226	0.301	0.174

从表 1 和表 2 的结果可以看出,FSTSC 方法设计的风格转换网络结构更加精简,参数数量减少,在降低计算成本和训练时间的同时可提升图像风格化的速度。因此,FSTSC 方法在图像风格化效果和速度上具有一定的优化效果,针对快速图像风格化方法进行有效的提升。

3 结 语

本文提出了一种 FSTSC 方法,使用块匹配方法寻找风格图像和内容图像最匹配的 patch,并利用图像语义分割方法进行预处理,根据不同的内容类别计算语义风格损失,以保证生成图像的局部内容与其风格相

匹配。另外,设计了结构更加精简的风格转换网络,从而可以加快图像风格化速度的同时有效提升图像风格化的效果。从图像风格化效果的主观评价和图像风格化时间的客观评价两方面进行分析,结果表明 FSTSC 方法优于现有的一些方法。

下一步的研究工作是解决在图像风格化过程中当内容图像与风格图像的语义内容差别过大时,由于很难保证内容与风格的完美匹配所导致的在一定程度上影响生成图像视觉观感的问题。

参 考 文 献

[1] Gatys L A, Ecker A S, Bethge M. A neural algorithm of artistic style[EB]. arXiv:1508.06576, 2015.

[2] Gatys L A, Ecker A S, Bethge M. Image style transfer using convolutional neural networks[C]//Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.

[3] Li C, Wand M. Combining markov random fields and convolutional neural networks for image synthesis[C]//Computer Vision and Pattern Recognition. IEEE, 2016.

[4] Johnson J, Alahi A, Li F F. Perceptual losses for real-time style transfer and super-resolution[C]//European Conference on Computer Vision, 2016.

[5] Ulyanov D, Lebedev V, Vedaldi A, et al. Texture networks: feed-forward synthesis of textures and stylized images[C]// International Conference on Machine Learning (ICML), 2016.

[6] Gatys L A, Ecker A S, Bethge M, et al. Controlling perceptual factors in neural style transfer[C]//Computer Vision and Pattern Recognition. IEEE, 2017.

[7] 吴联坤. 基于 TensorFlow 分布式与前景背景分离的实时图像风格化算法[D]. 杭州:浙江大学, 2017.

[8] Briechle K, Hanebeck U D. Template matching using fast normalized cross correlation[J]. Proceeding of SPIE on Optical Pattern Recognition XII, 2001, 4387: 95 – 102.

[9] 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 & Machine Intelligence, 2018, 40(4): 834 – 848.

[10] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Computer Vision and Pattern Recognition. IEEE, 2016.

[11] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Computer Vision and Pattern Recognition. IEEE, 2016.

[12] Radford A, Metz L, Chintala S, et al. Unsupervised representation learning with deep convolutional generative adversarial networks[C]//International Conference on Learning Representations, 2016.

[13] 操江峰. 基于深度学习的图像与视频风格化研究与实现[D]. 北京:中国科学院工程管理与信息技术学院, 2017.

[14] Chikontwe P, Lee H J. Towards robust face sketch synthesis with style transfer algorithms[M]//IT Convergence and Security 2017, 2018: 172 – 179.

[15] 赵梦. 基于主成分分析法的图像质量评价方法研究[D]. 杭州:杭州电子科技大学, 2013.

(上接第 179 页)

[9] Miao Q, Li Y, Ouyang W, et al. Multimodal gesture recognition based on the ResC3D network[C]//2017 IEEE International Conference on Computer Vision Workshops (ICCVW). IEEE, 2018: 3047 – 3055.

[10] Haghighat M, Abdel M M, Alhalabi W. Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition[J]. IEEE Transactions on Information Forensics and Security, 2016, 11(9): 1984 – 1996.

[11] Chaib S, Liu H, Gu Y, et al. Deep feature fusion for VHR remote sensing scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(8): 4775 – 4784.

[12] Hotelling H. Relations between two sets of variates[J]. Biometrika, 1936, 28(3/4): 321 – 377.

[13] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25(2): 1097 – 1105.

[14] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB]. [2019 – 07 – 05]. arXiv:1409.1556, 2014.

[15] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2818 – 2826.

[16] Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1 – 9.

[17] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]//The Thirty-First AAAI Conference on Artificial Intelligence, 2017: 4278 – 4284.

[18] Li F F, Fergus R, Perona P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories[J] Computer vision and Image Understanding, 2007, 106(1): 59 – 70.