DOI:10. 13471/j. cnki. acta. snus. 2019. 10. 11. 2019A079

基于层间互相关感知损失的风格迁移方法*

庄轩权[1],李彩霞[1]，黎培兴[1,2]
（1. 中山大学数学学院,广东 广州 510275;
2. 中山大学广东省计算科学重点实验室,广东 广州 510275）

摘　要：深度学习在风格迁移领域的应用使一系列以图片艺术风格化为核心的产品真正落地，而从像素级损失向基于Gram矩阵的感知损失的转变是其中最关键的跨越。Gram矩阵在艺术风格特征的提取上有良好的效果，但其局限于同等级语义特征间相关性统计的做法并不能作为艺术风格的充分表示。自Gram矩阵被提出以来，一系列研究并未对其进行充分的研究和改进，而是关注于模型结构的设计以提高风格迁移的速度。提出使用层间互相关矩阵作为Gram矩阵的代替或补充进行风格迁移任务的风格损失函数计算。实验表明，在得到相似水平输出结果的情况下，使用层间互相关矩阵方法可以降低20%的计算时间。

关键词：风格迁移；Gram矩阵；卷积神经网络；风格损失函数；感知损失；深度学习
中图分类号：TP183　　**文献标志码：**A　　**文章编号：**0529-6579（2020）06 -0126 -10

Style transfer based on cross-layer correlation perceptual loss

ZHUANG Xuanquan[1], *LI Caixia*[1], *LI Peixing*[1,2]
（1. School of Mathematics, Sun Yat-sen University, Guangzhou 510275, China;
2. Guangdong Province Key Laboratory of Computational Science,
Sun Yat-sen University, Guangzhou 510275, China）

Abstract: Great success in deep-learning-based style transfer is accelerating the development of photo artistic stylization applications. And the change of loss function from per-pixel loss to perceptual loss based on the Gram matrix is the most critical part of this progress. Gram matrix shows good performance in style feature extraction, but it only focuses on correlations among same level features. Therefore, Gram matrix cannot be considered as a complete representation of styles. However, most of the research focus on how to improve transfer speed by designing new model structure instead of analyzing and modifying the Gram matrix. The cross-layer correlation matrix is used to calculate style loss function as a replacement or supplement to the Gram matrix. By experiments, it is shown that this method can reduce 20% of the calculation time in comparison with the Gram matrix method while yielding similar outputs.

Key words: style transfer; Gram matrix; convolutional neural network; style loss function; perceptual loss; deep learning

风格迁移技术指对某个图像进行渲染，使其艺术风格与某个艺术绘画作品相似，且画面的主体内容不变（见图1）。2015年Gatys等[1-2]开创性地将卷积神经网络运用到风格迁移领域，提出了度量特征相关性的Gram矩阵用于风格表示，开创了现代风格迁移时代。Gram矩阵的核心思想是利

* **收稿日期：**2019 - 10 - 11
基金项目：广东省基础与应用基础研究基金（2020B1515310007）；中山大学广东省计算科学重点实验室（2020B1212060032）
作者简介：庄轩权（1995年生），男；**研究方向：**深度学习与图像处理；E-mail：andrezhuang@tencent. com
通信作者：黎培兴（1971年生），男；**研究方向：**机器学习与数据挖掘；E-mail：lnslpx@mail. sysu. edu. cn

用预训练网络强大的特征提取能力得到有意义的特征映射输出，并将特征映射之间的相关性作为风格的度量。此后，Justin 等[3] 基于 Gram 矩阵设计了基于前馈网络的快速风格迁移模型，使得 Prisma 等图像艺术风格化的应用得以流行。

(a) 怀士堂　(b) 风格迁移结果　(c) The Scream

(d) 中大北门　(e) 风格迁移结果　(f) Starry Night

图 1　风格迁移示例

Fig. 1　Examples of style transfer

然而自 Gatys 等提出 Gram 矩阵以来，风格迁移研究领域对于损失函数的构造一直没有足够的探索，本文提出使用层间互相关矩阵作为 Gram 矩阵的代替或补充，在得到良好结果的情况下缩短 20% 以上的计算时间，从而提高训练效率。

1　风格迁移技术

2015 年，Gatys 等[1-2] 提出使用 Gram 矩阵来度量图像的风格，由此开辟了基于深度学习的图像风格迁移领域。对于一张图片，使用预训练好的分类网络，如 VGG-16[4]，将图片输入得到某一层的特征映射，对特征映射的各个通道两两做互相关计算得到对称矩阵，这个矩阵就称为 Gram 矩阵。严格来说，衡量图像之间风格差异的损失函数 $L_S(I_s, X)$ 定义为

$$L_S(I_s, X) = \sum_{i=1}^{n} w_{l_i} L^{l_i}$$

其中，I_s 和 X 分别表示风格图像和待优化图像，l_i 表示特征映射处于预训练网络的层数，L^{l_i} 为第 l_i 层两图像间的风格损失，w_{l_i} 为对应的权重参数。第 l_s 层的风格损失函数定义为

$$L^{l_i} = \frac{1}{4N_{l_i}^2 M_{l_i}^2} \sum_{j,k} (G_{jk}^{l_i} - A_{jk}^{l_i})^2, 1 \leqslant j, k \leqslant N_{l_i}$$

其中，N_{l_i} 为第 l_i 层卷积核的个数，M_{l_i} 为第 l_i 层特征映射的长宽乘积，$G_{jk}^{l_i}$ 和 $A_{jk}^{l_i}$ 表示以待优化图像和风格图像为输入，预训练网络的第 l_i 层的特征映射输出的第 j 个和第 k 个卷积核的互相关函数，即

$$G_{jk}^{l_i} = \sum_{s} F_{js}^{l_i}(X) F_{ks}^{l_i}(X), 1 \leqslant s \leqslant M_{l_i}$$

$F_{js}^{l_i}(X)$ 为第 l_i 层的第 j 个卷积核输出的特征映射的第 s 个像素点的值。而以 $G_{jk}^{l_i}$ 和 $A_{jk}^{l_i}$ 为元素组成的大小为 $N_{l_i} \times N_{l_i}$ 的矩阵 G^{l_i} 和 A^{l_i} 则分别是待优化图像和风格图像在预训练网络第 l_i 层输出的特征映射计算得到的 Gram 矩阵，它们都是对称矩阵。实验结果表明这种矩阵可以很好地度量图像之间风格的差异，以此作为损失函数使用梯度下降法更新图像的像素值就能够得到跟风格图像相似的纹理风格。

风格迁移任务的目标是使用风格图像的纹理特点绘制内容图像的内容信息。因此，除了对风格进行迭代逼近之外，对内容的逼近也同样重要。在 Gatys 等的实验中，直接使用预训练网络的中间几层特征映射之间的像素值差异作为衡量语义信息相似度的标准，并获得了良好的效果。这其中的原因是这些预训练网络都是以图像分类任务为目标进行训练的，训练集包含了大量的物体类别，因此卷积神经网络在训练中降低多类别交叉熵损失函数的过程中，卷积层的卷积核在试图提取各种能描述不同物体差异的信息，这其中就包括了低层到高层的语义信息。以内容图像和待优化图像作为输入，预训练网络在第 l_i 层得到的特征映射的损失严格定义为

$$L_c(I_c, X, l_i) = \frac{1}{2} \sum_{j,k} (F_{jk}^{l_i}(I_c) - F_{jk}^{l_i}(X))^2$$

其中，I_c 表示内容原始图像，X 表示待优化图像，j 表示在第 l_i 层特征映射中第 j 个卷积核的编号，k 表示该卷积核得到的特征映射中第 k 个位置的像素点。可以看出，衡量两张图像内容上差异的损失函数仅仅只是简单使用特征映射之间的平方损失，我们不得不感叹于卷积神经网络强大的特征提取能力。由各层内容损失得到的总体内容损失为

$$L_c(I_c, X) = \sum_{i=1}^{n} w_{l_i} L_c(I_c, X, l_i)$$

其中，w_{l_i}为赋予第l_i层内容损失的权重。由此我们得到了以待优化图像，内容图像，风格图像作为共同输入的三元损失

$$L_t(I_c, I_s, X) = \alpha L_c(I_c, X) + \beta L_s(I_s, X)$$

其中α和β为内容损失和风格损失的权重参数。在原文中，作者选择的预训练网络是VGG-16，VGG-16在VGG系列的卷积神经网络中是应用最为广泛的，因为其具有良好的准确率以及不错的效率。在内容损失方面，作者选取了conv4_2作为计算内容损失的特征映射。在风格损失方面，作者选取了conv1_1、conv2_1、conv3_1、conv4_1、conv5_1作为计算风格损失的特征映射，并赋予各层相等的权重，而α/β选取10^{-3}或10^{-4}。

尽管使用Gram矩阵进行的图像风格迁移取得了良好的效果，但对于Gram矩阵的本质、是否有其他方式度量风格差异等问题，文中并没有给出答案。Li等[5]从迁移学习的角度出发去看图像风格迁移。文章将风格迁移任务看做是一种域适应的问题，并从理论上证明了Gram矩阵实际上与二阶多项式核的最大均值差异等价。从这种等价关系可知：

（i）图像的风格可以本质上表示为卷积神经网络中不同卷积层下的特征分布；

（ii）风格迁移的过程可以看成是从内容图像到风格图像的一种分布调整。

Li等尝试将二阶多项式核函数替换成其他多项式核函数或高斯核函数，实验结果表明，不同的核函数替代Gram矩阵进行风格迁移能得到类似的良好结果，同时又有许多不同的细节上的变化，从最小均值差异出发使用不同的核函数度量风格差异的做法大大丰富了风格迁移结果的多样性。

针对Gatys等[1]通过迭代更新像素值的方式效率较低的问题，以Justin等[3]为代表的一系列研究提出了使用前馈网络直接输出风格迁移结果的快速风格迁移方法，按模型可承载的风格数量可划分为单模型单风格方法[3, 6]、单模型多风格方法[7]、单模型任意风格方法[8-9]。

自2014年Goodfellow等[10]提出生成式对抗网络（generative adversarial networks，GAN）以来，有关生成式对抗网络的研究便一直活跃在众多研究领域当中。深度卷积生成式对抗网络[11]提出使用卷积层和转置卷积层对生成式对抗网络进行改进，为生成式对抗网络在图像领域的发展开拓了更优的思路。

近年来，部分研究将生成式对抗网络运用于两个域之间的图像的相互转换，这可以看作是一种广义上的风格迁移方法。Zhu等[12]提出使用两个对称的生成式对抗网络构造一种循环一致性损失，从而实现将输入的图像向特定分布转换的目的，如将夏季的图像转换为冬季的图像。类似的想法还有DualGAN[13]和DiscoGan[14]等。StyleGAN[15]利用风格迁移领域提出的AdaIN模块[8]对生成式对抗网络进行优化，成功实现了对输入图像的各种细节进行微调的重大突破，成为生成式对抗网络和广义风格迁移领域的里程碑之一。

尽管使用生成式对抗网络系列的方法也能做到将输入的图片进行纹理风格上的转化并保持图片内容上的一致性，但和基于Gram矩阵的一系列风格迁移方法有许多差异：

（i）基于Gram矩阵的风格迁移方法使用Gram矩阵衡量图像之间的风格差异，是一个可以被计算的统计量，可以量化地给出任意一张图片的风格数值；生成式对抗网络系列方法无法显式给出风格的定义，而是通过对抗训练对两个域的分布进行学习，试图让网络自行拟合出一套进行风格转换的参数。

（ii）基于Gram矩阵的风格迁移方法可以学习任意单一图像的纹理风格特征并将其迁移到任意图像之上；而生成式对抗网络由于损失函数和训练的动机限制，只能学习一类图像的风格而无法刻画单一图像的纹理风格特征，比如其通过训练可以学习画家梵高的画作整体风格，但无法很好地学习梵高的《Starry Night》这幅画作的风格。尽管一些工作尝试通过生成式对抗网络学习单一图像的纹理分布，但基本只能输出与原风格图像在内容上极度统一的结果[16]，而基于Gram矩阵的方法通过将对多个图像分别进行学习得到的参数进行简单组合就可以得到融合后的风格输出，因而也能学习一类图像的风格纹理特征。

（iii）由于（ii）中提及的原因，基于Gram矩阵的风格迁移方法得到一个输出结果良好的模型所需的数据获取成本会远低于基于生成式对抗网络的方法；另外，由于期望网络自行学习一种风格纹理的概率分布，在训练生成式对抗网络时需

要更庞大的参数量，以及更多的训练技巧和尝试从而规避无法收敛或学习不出特征的问题，其训练的时间成本也远大于Gram矩阵方法。

（iv）生成式对抗网络在生成高分辨率的图像上效果不如低分辨率图片，或是需要更大的模型和更长的训练周期才能达到较好效果；基于Gram矩阵的方法则在各个分辨率尺度上都有稳定的表现。

自Gatys等[1-2]的工作以来，基于Gram矩阵的风格迁移方法一直是该领域的主流方法，至少现阶段包括基于生成式对抗网络在内的方法都还无法得到这样高效且效果良好的风格迁移结果。但Gram矩阵作为人工设计的统计量，必然受到人们先验知识的限制，通过对生成式对抗网络训练过程的深度挖掘以及与Gram矩阵之间的关联的分析或许能为风格迁移进一步的发展提供动力。

2　基于层间互相关感知损失的风格迁移技术

2.1　层间互相关矩阵

卷积神经网络的卷积核提取的特征等级往往与该卷积核所处的深度有关，即浅层的卷积核提取低级特征，深层的卷积核提取高级特征。理论上，Gram矩阵只能表现同层级的特征之间的相关程度。针对这一问题，我们提出使用层间互相关矩阵来进行补充。

给定图像I在预训练网络的第l_1、l_2层的输出$F^{l_1}(I)$及$F^{l_2}(I)$（$l_1 < l_2$），层间互相关矩阵为一个$N_{l_1} \times N_{l_2}$的矩阵$G^{l_1 l_2}=(G_{jk})_{N_{l_1} \times N_{l_2}}$，

$$G_{jk}=\sum_{s=1}^{M} D(F_{js}^{l_1}(I)) \times F_{ks}^{l_2}(I), 1 \leqslant j \leqslant N_{l_1}, 1 \leqslant k \leqslant N_{l_2}$$

其中N_{l_1}和N_{l_2}分别为预训练网络的第l_1层和第l_2层的通道数，M表示特征映射的长宽乘积，$D(*)$为降采样函数。

由于不同深度的特征映射的长宽不一致（如VGG-16不同层的特征映射的长宽最大相差16倍），我们需要对浅层特征映射使用降采样或对深层特征映射使用升采样，使得用于层间互相关矩阵计算的两个特征映射的长宽一致。考虑到计算成本等原因，我们选择对浅层特征映射进行降采样，降采样函数可使用平均池化或最大池化等。

对于不同的特征映射组，需要使用不同的降采样参数使得两者得感受野对齐。以VGG-16网络为例，通过对感受野的计算可以发现，特征映射relu2_1的每一个元素实际上对应着特征映射relu1_1中一个8 × 8大小的区域。具体地，我们需要在大小为3的填充下使用8 × 8大小的池化滤波器以步长为2的方式对relu1_1进行池化操作。表1中为VGG-16中部分特征映射层使用池化方式进行层间互相关计算应该使用的参数，特征映射层A为池化操作的作用层。

表1　层间互相关矩阵计算池化操作参数

Table 1　pooling parameters for cross-layer correlation matrix computation

特征映射A	特征映射B	滤波器大小	步长	填充
relu4_1	relu5_1	10×10	2	4
relu3_1	relu5_1	10×10	4	12
relu2_1	relu5_1	10×10	8	27
relu1_1	relu5_1	10×10	16	57

2.2　层间互相关矩阵与Gram矩阵对比

2.2.1　*语义特征的登记对比*　Gram矩阵可以理解为层内互相关矩阵，计算同等级语义特征间的相关程度。与之对应，层间互相关矩阵计算不同等级语义特征间的相关程度。从这个意义上来理解，Gram矩阵和层间互相关矩阵在对图像风格的描述上应该是互为补充的。

从直观理解出发，层间互相关矩阵的意义甚至比层内互相关矩阵更重要。举例而言，不同的颜色应该属于同一等级的特征，不同的动物、植物也应该属于同一等级的特征，而对于图像中一块具体的感受野来说，只应该是某个颜色或者某种动植物，而不应该同时具备多个。相比之下，层间互相关矩阵的可解释性更强，如某种动植物或山水的特征与某个颜色或线条纹理的特征相关性强，可以理解为作品中对某种事物的刻画使用

了某种技法，这些相关性共同描述了作品的艺术风格。因此，如果从相同深度的卷积核只提取同等级的语义特征这个前提出发，层间互相关矩阵对于风格差异的描述更加具有可解释性。然而在实际的预训练网络中，相同深度的卷积核提取的特征有时也难以说明是否为同一等级的特征，甚至有许多卷积核提取的特征拿出来单独看无法从人的视觉角度理解，因此无论层间互相关矩阵还是层内互相关矩阵在实际的应用中都表现出相似的效果。

2.2.2 *计算与存储效率对比* 显然，区别于Gram矩阵的对称方阵的特点，层间互相关矩阵是一个$C_1 \times C_2$的矩阵，且每个元素对应的含义都唯一。而相比之下Gram矩阵有将近一半的重复元素，信息的冗余度较高，占用大量内存的同时却没有尽可能精简出不重复的信息。层间互相关矩阵通过融合两个特征映射层使得其可以用单个矩阵对两个特征映射层的信息进行表达，且其存储和计算量都低于Gram矩阵方法。

以relu3_1和relu5_1为例，使用Gram方法进行风格迁移需要存储的Gram矩阵大小为$256^2 + 512^2 = 327\,680$，使用层间互相关矩阵方法需要存储的矩阵大小为$256 \times 512 = 131\,072$，仅为Gram方法存储量的40%，自然，风格损失的计算量也为Gram方法的40%；而计算Gram矩阵本身的成本也比计算层间互相关矩阵要高，Gram矩阵方法需要进行约$2 \times (256^2 \times 64^2 + 512^2 \times 16^2) \approx 6.7 \times 10^8$次运算，层间互相关方法仅需进行约$2 \times (256 \times 512 \times 16^2) \approx 6.7 \times 10^7$次运算，仅为前者的10%。

然而Gram矩阵相对层间互相关矩阵而言，由于不需要关注特征映射大小改变的问题，在训练过程中计算过程更加简单可理解，相比之下层间互相关矩阵的计算不仅要根据特定的两个层的选择来确定采样操作中的参数，对于不同的预训练网络而言也要重新计算，增加了额外的计算且拓展性不如Gram矩阵好。

3 实 验

3.1 实验设计

本文实验采用与Gatys等[1]相似的模型结构，选取多个特征映射层及其组合计算Gram矩阵和层间互相关矩阵，并对比它们在纹理合成及风格迁移中的实际效果。

实验使用python3.6及tensorflow1.13，预训练网络使用matlab平台在ImageNet数据集上预训练的VGG-16网络①，使用一块Tesla K80 GPU加速。所有图片均缩放至256 × 256大小，区别于Gatys等[1]的实验，我们使用Adam优化器[17]，学习率设置为10^{-2}。在纹理合成实验中，损失函数仅使用风格损失，不加入内容损失。

3.2 实验条件

3.2.1 *优化器的选择* 在本文所进行的所有实验中，统一选择了Adam优化器进行模型的训练。在Gatys等[1]最初提出的风格迁移方法中，使用了L-BFGS方法进行梯度求解。我们注意到，从Johnson等[3]开始的一系列风格迁移的研究中，使用Adam优化器已经成为了主流的方法。在大量的研究实验中，Adam优化器证明了其在大规模参数优化当中卓越的性能[18-19]，几乎所有的深度学习框架对其都有良好的支持，更方便结果的横向比对。Adam算法的提出时间和Gatys等[1]提出Gram矩阵的时间相近，在当时仍未普遍使用，但现如今使用Adam算法已经是主流的做法。为了保持实验条件的一致性，我们在复现Gatys等[1]提出的方法时也将优化方法改为了Adam方法，以保证结果对比的公正客观。

3.2.2 *特征提取网络的选择* Gatys等[1]的实验中选择了VGG-19[4]作为特征提取网络，这也是VGG系列中最深且性能最强悍的网络。在本文的所有实验中，特征提取网络都选择了VGG-16网络。这是由于在多个权威的图像分类数据集上VGG-19相比VGG-16在准确率上的提升都不明显，且参数量更大，占用资源更多。图像分类的准确率相当反映出模型在特征提取上的能力相当，而风格迁移中使用预训练的特征网络的核心目的就是借助其特征提取能力得到有意义的特征映射，因此选择VGG-16与VGG-19在结果上的区别并不明显（可以从本文的实验和Gatys等[1]的实验结果对比看出），综合实验的计算资源限制等因素，本文使用VGG-16代替VGG-19。

①MatConvNet Pretrained Models. http://www.vlfeat.org/matconvnet/pretrained/imagenet-ilsvrc-classification.

近年来，一些具有革命性意义的模型改进方法使得更深更大的神经网络模型的训练成为可能，在性能上也大大超越了早期的 VGG 等模型[20-21]。然而，这些方法需要的计算规模也远超早期的方法，且往往层数很大，这会给风格迁移任务带来一个问题，即如何有效地选择适合的特征映射层进行 Gram 矩阵或层间互相关矩阵的计算。由于使用大部分特征映射层计算得到的 Gram 矩阵共同进行风格损失的计算并不现实，而通过实验对比选择适合的特征映射层又有层数过多的问题导致实验成本较大，使用最新的高精度模型作为特征提取网络并不是一个好的选择。这也解释了大部分风格迁移方法的研究中都使用较浅的神经网络模型进行特征提取的原因。

3.3　实验结果

3.3.1　风格纹理学习实验　风格纹理学习实验使用不同的特征映射层计算层间互相关矩阵和 Gram 矩阵，对六幅绘画作品的风格纹理进行学习，得到输出结果（图 2）。层间互相关矩阵方法全部使用平均池化对浅层特征映射做降采样处理。从输出结果可以看出，单纯使用层间互相关矩阵作为损失函数学习到的风格纹理与使用 Gram 矩阵的模型得到的相似，说明单纯使用层间互相关矩阵也可以很好地完成风格迁移的目标；其次，通过对输出结果的观察我们可以看出，层间互相关矩阵方法得到的输出结果的语义等级（纹理的颗粒度、色彩深浅）大概处于其使用的两个特征映射层分别使用 Gram 矩阵方法进行纹理学习得到的输出结果之间，可以看做是两者的一个加权融合。

另外，从风格纹理学习的实验中我们发现了一些值得关注的细节问。

（i）层间互相关矩阵使用的两个特征映射层越深，学到的风格图像中的语义信息越多。这也映证了越深的特征映射层会提取越多高级特征，使得输出结果带有越多风格图像中的画面轮廓。这也是层间互相关矩阵和 Gram 矩阵共同具有的属性，而这点也可以启发我们在风格迁移任务中根据对风格迁移程度的要求对风格损失函数的组合进行选择。

（ii）理论上风格迁移任务中对风格图像的学习并不需要对其进行缩放，因为层间互相关矩阵或者 Gram 矩阵的输出大小都与原始图像输入大小无关。然而由于部分语义相关信息的带入，可能会使得风格迁移结果的纹理大小粗细在输出尺寸下显得突兀，导致风格迁移的结果不佳。因此，对于原始风格图像与风格迁移目标输出尺寸相差较大的需要进行缩放处理。

（iii）从随机得到的噪声图像开始优化图像，单纯使用风格损失很难避免局部失真的现象，即局部色块中出现明显不符合原风格图像特征的噪声点。使用更小的学习率以及更多的迭代次数只能稍微缓解该现象，使用一定的平滑技术才能较好地解决该问题，如在目标损失函数中加入总变分损失。但加入的平滑技术会在一定程度上破坏渲染出的风格纹理，引入局部的条状或块状纹理。因此，如何寻找合适的平滑技巧或其他方法使得输出结果，尤其是在高分辨率输出中避免局部失真现象仍是需要解决的问题。在第二部分的风格迁移实验中我们使用内容图像作为初始化代替了引入平滑损失的做法，并发现具有较好的效果。

（iv）实验部分展示的层间互相关矩阵方法得到的结果均采用降采样的方法。除了 2.1 节中提到的资源消耗原因外，在实际的实验结果中我们也发现基于上采样的层间互相关矩阵方法效果较差。最重要的原因是上采样无法类似降采样通过步长以及卷积核大小的控制进行感受野的对齐，造成了特征相关关系的紊乱；另外，降采样的过程是将特征进行组合精简，是特征的再提取过程，但上采样则试图将精简后特征还原，而这并非一个可逆的过程。

（v）风格纹理学习实验部分展示的结果均使用平均池化，我们在实验中发现使用最大池化得到的结果在纹理的细节和连贯上不如平均池化效果好，即纹理出现局部失真和断层的现象更多。在风格迁移实验部分我们展示了最大池化和平均池化的结果对比。

3.3.2　风格迁移实验

风格迁移实验对使用不同特征映射层进行计算的 Gram 矩阵方法及层间互相关矩阵方法的输出结果进行对比（图 3），用图 2 中的三种风格对图 1 中中山大学怀士堂、中山大学北门牌坊两张图片进行风格迁移。从风格迁移实验的输出图像以及训练时长对比（表 2）中，我们得到以下结论：

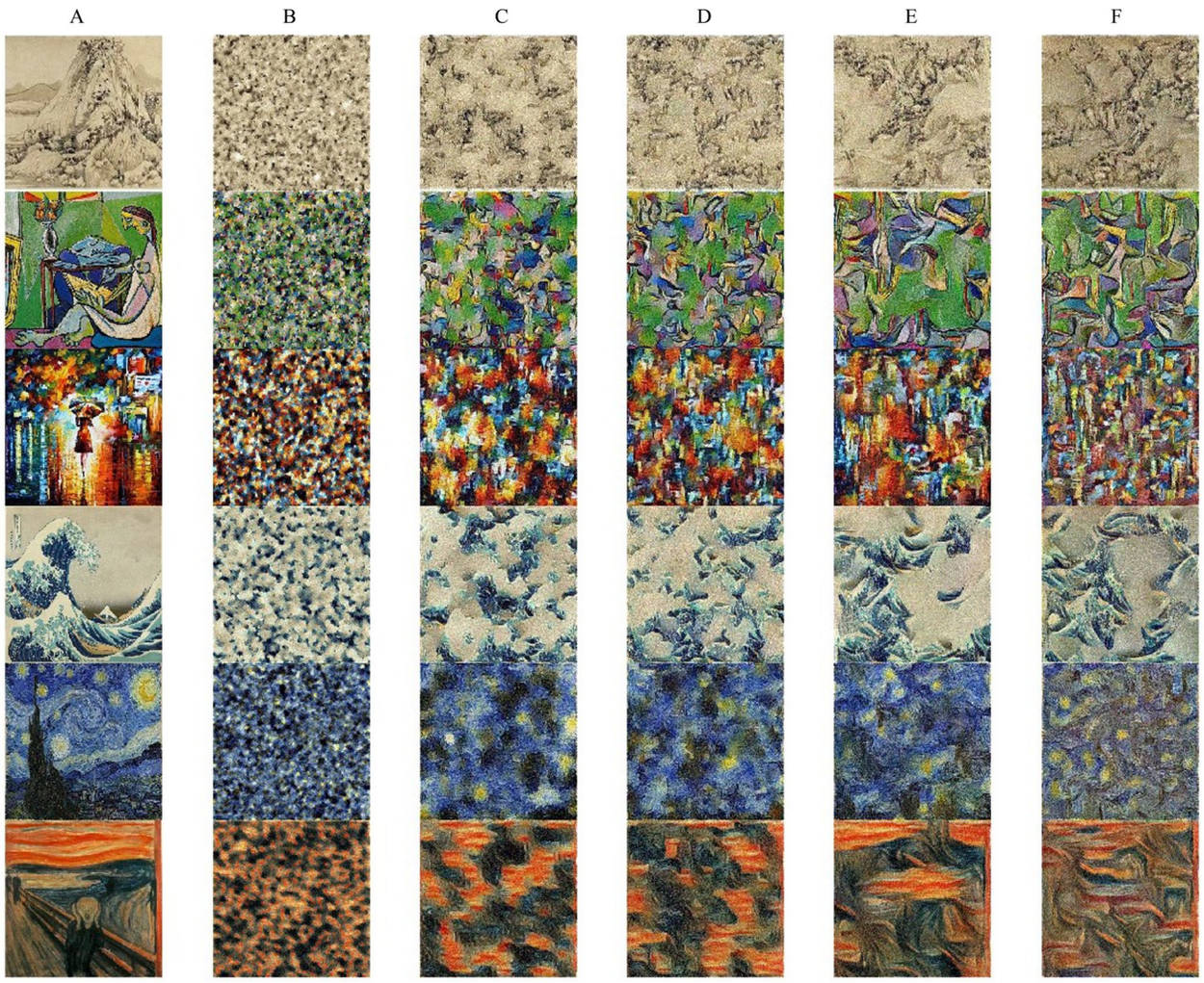

图2 纹理合成实验结果

Fig. 2 Texture synthesis outputs

A：原始风格图像（缩放至256 × 256）；B：relu1_1 Gram矩阵方法结果；C：relu1_1 relu2_1层间互相关矩阵方法结果；D：relu2_1 Gram矩阵方法结果；E：relu2_1 relu3_1层间互相关矩阵方法结果；F：relu3_1 Gram矩阵方法结果

表2 风格迁移方法结果比较

Table 2 comparision of style transfer methods

风格损失组合	采样方法	时间（s/迭代）
（relu1_1，relu2_1），（relu2_1，relu3_1）	平均池化	0.048 3
（relu1_1，relu2_1），（relu2_1，relu3_1）	最大池化	0.046 4
relu1_1，relu2_1，relu3_1	–	0.064 8
（relu1_1，relu3_1），（relu2_1，relu4_1）	平均池化	0.052 5
（relu1_1，relu3_1），（relu2_1，relu4_1）	最大池化	0.048 6
relu1_1，relu2_1，relu3_1，relu4_1	–	0.069 5

（i）选取相同的特征映射层，层间互相关矩阵方法和Gram矩阵方法会得到相似水平的输出，例如图3第3行中，使用relu1_1，relu2_1，relu3_1三个特征映射层的Gram矩阵方法和层间互相关矩阵方法都较好地保留了礼堂整体轮廓形状，而加入了特征映射层relu4_1后的三个输出结果，整体轮廓都在一定程度上被破坏；第5行中，后三组实验的天空部分有明显的黄色块状纹理，而前三组则没有。

（ii）无论是Gram矩阵方法还是层间互相关矩

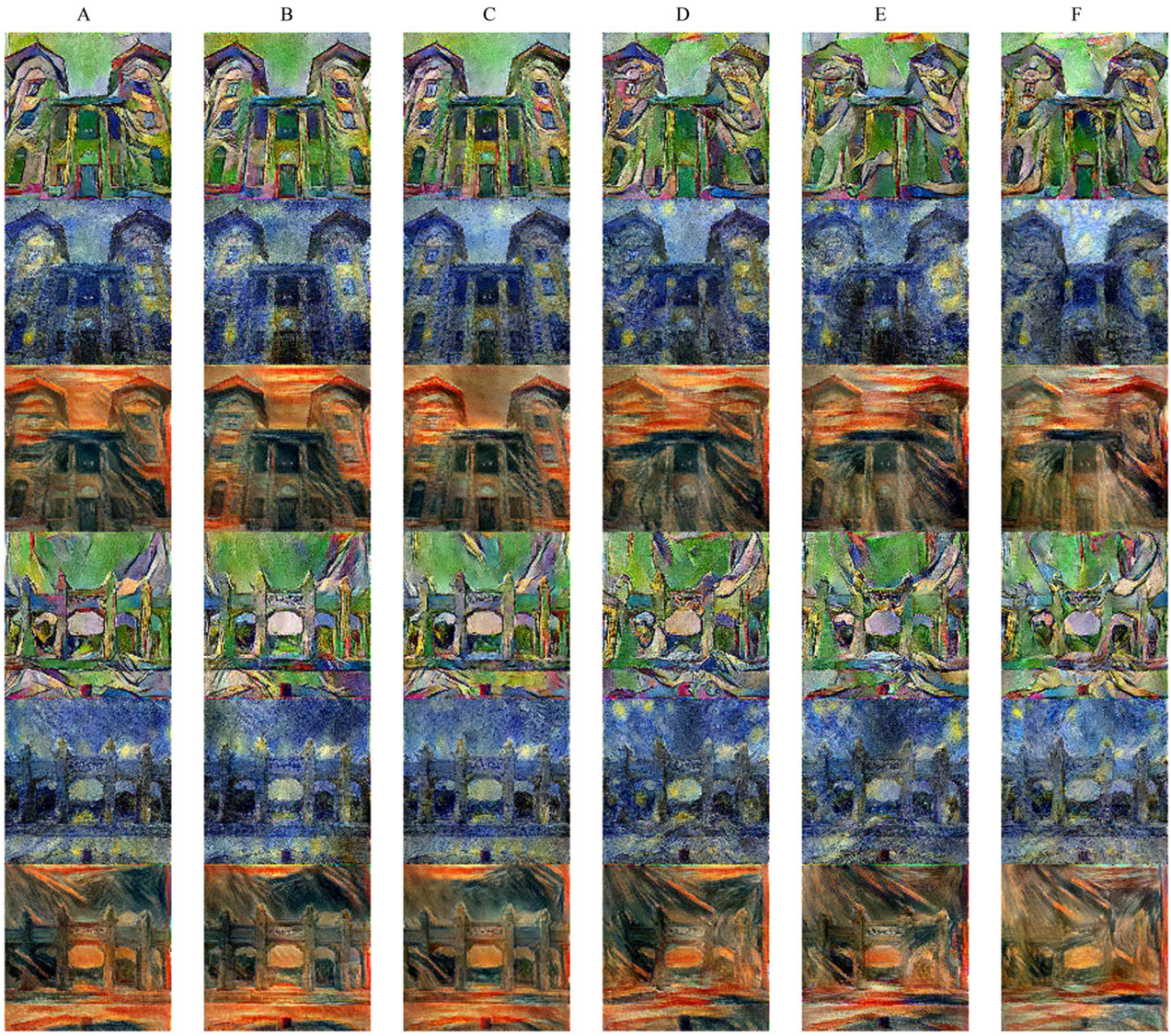

图 3　风格迁移实验结果

Fig. 3　style transfer outputs

A：relu1_1，relu2_1，relu3_1 Gram 矩阵方法结果；B：(relu1_1，relu2_1)，(relu2_1，relu3_1)+平均池化层间互相关矩阵方法结果；C：(relu1_1，relu2_1)，(relu2_1，relu3_1)+最大池化层间互相关矩阵方法结果；D：relu1_1，relu2_1，relu3_1，relu4_1 Gram 矩阵方法结果；E：(relu1_1，relu3_1)，(relu2_1，relu4_1)+平均池化层间互相关矩阵方法结果；F：(relu1_1，relu3_1)，(relu2_1，relu4_1)+最大池化层间互相关矩阵方法结果

阵方法，都存在一定的局部失真。在风格迁移的实际实验过程中，我们首先尝试了内容损失、风格损失和总变分损失三部分进行加权组合的损失函数，但发现尽管总变分项的加入使得最终的输出结果更平滑，但会带来与艺术风格不匹配的局部纹理。另外，总变分损失的加入使得训练过程中内容损失和风格损失的下降变得困难，难以达到令人满意的输出结果，且对于不同的风格和内容，总变分损失部分的损失似乎都需要特殊的调参，否则结果差异较大。基于此事实，以及本文比较不同风格损失函数的核心，最终在实验中我们舍弃了总变分损失的部分，虽然导致风格迁移结果局部失真，但在实际实验中使用内容图像进行初始化的方式很大程度地缓解了问题。

（iii）尽管在使用的特征映射层相同的条件下，使用 Gram 矩阵和层间互相关矩阵方法得到的风格迁移输出结果类似，从个别例子中仍能看出使用最大池化的方法不如使用平均池化方法得到的图像效果稳定，相比之下其差异会更大。如图 3 第 6 行的后三组实验中，使用最大池化得到的结果和

使用Gram矩阵或平均池化的层间互相关矩阵方法得到的结果差异较大。

（iv）在得到相似输出水平的情况下，使用层间互相关矩阵方法比使用Gram矩阵的方法在速度上有着显著优势。在风格迁移的所有实验中我们对输入图像的优化迭代次数都是20 000次，这也是前期实验对比得出的经验值。我们发现无论是Gram矩阵方法还是层间互相关矩阵方法，得到效果良好的输出图像所需要的迭代次数是相当的，这可能是因为相似的损失函数构成和相同的学习率使得每次反向传播过程中给出的梯度值都在基本相同的量级。因此在表2的运行时间对比中我们给出了单次迭代耗时，实际上也是总耗时与迭代次数的比值。在前三组实验中，使用平均池化、最大池化方法的单次迭代耗时约为Gram矩阵方法的74.54%及71.6%；后三组实验中，使用平均池化、最大池化方法的单次迭代耗时约为Gram矩阵方法的75.54%及69.93%，提速均在20%以上，使用最大池化的层间互相关矩阵方法速度最快。

（v）我们在第2节中提到，理论上层间互相关矩阵刻画的是不同等级语义特征之间的相关性。然而，在实际实验结果中我们发现，艺术风格纹理的差异主要是使用了不同的特征映射层造成的，与使用层间互相关矩阵还是Gram矩阵的关系并不显著。这是由于这种理论上的语义等级差异并不完全和实际情况相符。实际上，卷积神经网络的中间隐藏层所提取的很多特征从人的角度是难以理解的，其对语义特征等级的区分也和人的理解有差异。另外，卷积神经网络也难以确保在训练中将同等级的特征提取放在同一层中进行，光依靠卷积层和池化层的结构是无法保证这种限制的。因此，同一层的卷积核提取的特征之间也可能存在等级差异，很多情况下，只要选取的特征映射一致，Gram矩阵方法和层间互相关矩阵方法得到的风格迁移输出结果整体不会有很大差异。但这并不妨碍在细节处层间互相关矩阵带来的纹理多样性。

（vi）本文实验中对比了Gram矩阵和层间互相关矩阵在Gatys等[1]提出的算法中表现的差异，若在基于前馈网络的快速风格迁移系列方法中将Gram矩阵替换为层间互相关矩阵，则无法提升模型使用时的效率，而是提升了模型训练时的效率。

4 结 论

本文提出使用层间互相关矩阵作为Gram矩阵的代替或补充，用于风格迁移任务中风格损失的计算。实验表明，在获得与基于Gram矩阵的神经风格迁移方法相似水平的输出结果的情况下，使用层间互相关矩阵的方法可以在一定程度上提高模型的训练效率。

除了风格迁移任务本身外，层间互相关矩阵和Gram矩阵的有效性也表明深度学习方法在艺术风格的表示、分类、聚类等问题上有着很大的潜力。另外，由于风格迁移任务的特殊性，我们可能需要更多艺术专业领域的专家知识的指导，作为先验知识，这可能为未来风格迁移的效果带来一定的提升。

参考文献：

[1] GATYS L A, ECKER A S, BETHGE M. Image style transfer using convolutional neural networks [C]//Computer Vision and Pattern Recognition, 2016: 2414-2423.

[2] GATYS L A, ECKER A S, BETHGE M. Texture synthesis using convolutional neural networks [C]//International Conference on Neural Information Processing Systems, 2015: 262-270.

[3] JOHNSON J, ALAHI A, FEIFEI L. Perceptual losses for real-time style transfer and super-resolution [C]//European Conference on Computer Vision, 2016: 694-711.

[4] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [J]. International Conference on Learning Representations, 2015.

[5] LI Y, WANG N, LIU J, et al. Demystifying neural style transfer [C]//IJCAI, 2017: 2230-2236.

[6] ULYANOV D, VEDALDI A, LEMPITSKY V. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis [C]// Computer Vision and Pattern Recognition, 2017: 4105-4113.

[7] DUMOULIN V, SHLENS J, KUDLUR M. A learned representation for artistic style [J]. International Conference on Learning Representations, 2017.

[8] HUANG X, BELONGIE S. Arbitrary style transfer in real-time with adaptive instance normalization [C]//International Conference on Computer Vision, 2017: 1510-1519.

[9] WANG H, LIANG X, ZHANG H, et al. Zm-net: real-time zero-shot image manipulation network [C]//Computer Vision and Pattern Recognition, 2017.

[10] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks [J]. Advances in Neural Information Processing Systems, 2014, 3: 2672-2680.

[11] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks [C]//International Conference on Learning Representations, 2016.

[12] ZHU J, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks [C]//International Conference on Computer Vision, 2017: 2242-2251.

[13] YI Z, ZHANG H, TAN P, et al. DualGAN: unsupervised dual learning for image-to-image translation [C]//International Conference on Computer Vision, 2017: 2868-2876.

[14] KIM T, CHA M, KIM H, et al. Learning to discover cross-domain relations with generative adversarial networks [C]//Computer Vision and Pattern Recognition, 2017.

[15] KARRAS T, LAINE S, AILA T, et al. A style-based generator architecture for generative adversarial networks [C]//Computer Vision and Pattern Recognition, 2019: 4401-4410.

[16] SHAHAM T R, DEKEL T, MICHAELI T, et al. SinGAN: Learning a generative model from a single natural image [C]//International Conference on Computer Vision, 2019: 4570-4580.

[17] KINGMA D P, BA J. Adam: a method for stochastic optimization [C]//International Conference on Learning Representations, 2015.

[18] RUDER S. An overview of gradient descent optimization algorithms [J]. arXiv: Learning, 2016.

[19] DOGO E M, AFOLABI O J, NWULU N I, et al. A comparative analysis of gradient descent-based optimization algorithms on convolutional neural networks [J]. International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Belgaum, India, 2018: 92-99.

[20] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]//Computer Vision and Pattern Recognition, 2016: 770-778.

[21] GAO S, CHENG M, ZHAO K, et al. Res2Net: A new multi-scale backbone architecture [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019: 1-1.

（责任编辑　冯兆永）