图像神经风格迁移领域起源、现状与挑战

摘要

图像风格迁移是一种将真实照片的内容与另一张图像的艺术风格相结合以创建新的风格化图像的技术。本文对图像风格迁移领域进行了全面的回顾，追溯了其从植根于纹理模拟数学模型的传统方法到利用深度学习和神经网络的现代方法的发展。该研究将风格迁移的演变分为两个主要阶段：传统风格迁移（依赖于纹理合成和直方图匹配等技术）和神经风格迁移（利用卷积神经网络捕获和应用复杂的艺术风格）。本文还探讨了该领域使用的各种评估参数，比较了代表性成果，并讨论了风格迁移在环境渲染、字体生成和虚拟现实等领域的实际应用。最后，本文强调了风格迁移领域尚未解决的问题和未来研究的潜在方向。

关键词：神经风格迁移，卷积神经网络，生成式对抗网络

1 引言

人们常说，一图胜千言，优秀的艺术作品往往提供不同于真实照片的信息。然而，普通人若不经过长时间的专业训练，则可能无法独立完成一幅令自己与他人满意的艺术作品。同时，训练一位真正的艺术家的需要的时间与成本也是难以估量的。不仅如此，即使是能熟练进行风格作品绘制的艺术家，完成一幅艺术作品也需要较长的周期。为了能高效地将真实照片转换为艺术图像，图像风格迁移任务顺势而生。

图像风格迁移旨在将一张真实照片的内容与一张艺术作品的风格结合起来，形成一张同时具有照片图像内容与艺术作品风格的风格化照片。在风格迁移任务中提供内容的图像被称作内容图像，提供风格的图像被称作风格图像，生成的结果被称作风格化图像。同时，本文主要介绍的内容也是图像风格迁移，若非特殊说明，下文中的风格迁移特指图像风格迁移。

风格迁移的发展主要可以分为两个阶段：20 世纪 90 年代中期[1]到 2016[2]为第一阶段，其主要特征为使用数学模型模拟纹理模拟。2016[2]至今为第二阶段，其主要特征为使用深度学习与神经网络进行风格迁移。二者相比，前者较为传统，后者吸纳新的方法，故而本文将第一阶段称做“传统风格迁移”，将第二阶段称做“神经风格迁移”。

传统风格迁移通常使用数学和信号处理技术，如纹理合成、直方图匹配和滤波等。这些方法涉及对像素的操作，以达成模拟所需风格的目的。例如，可以通过频域滤波来增强或减弱图像的某些频率成分，从而改变其外观。传统风格迁移的优势在于具有更高的计算速度与更低的资源占用，但它们可能无法捕捉到更高级的艺术风格和纹理。

基于神经网络的风格迁移方法则更为灵活与高效。这些方法使用深度学习技术学习和应用图像风格。它们通过训练神经网络捕捉不同艺术风格的特征，随后将这些特征应用于输入图像，以生成具有所需风格的新图像。神经风格迁移的优势在于能够更好地捕捉到艺术风格的细节和复杂性，所需时间与资源占用随网络结构的不同会有较大的差距。

传统风格迁移与神经风格迁移之间不应该是被代替与代替的关系，相反，目前部分神经风格迁移成果（如[3,4]等）的思想来源于传统风格迁移与数字图像处理。同时，基于神经网络的风格迁移技术也有一些缺陷，如伪影、难以控制风格化过程等缺陷，将传统风格迁移与神经风格迁移相结合可能会取得更好的结果。

图像风格迁移在实际生活中具有较多应用场景，如在环境氛围渲染[5]、字体生成[6]、字体识别[7]、肖像编辑[8,9]、辅助设计[10–14]、照片修复[15]、虚拟现实（Virtual Reality，VR）与增强现实（Augmented Reality，AR）[16]等领域中，均有应用。

同时，风格迁移作为计算机视觉的底层任务，可以辅助如对抗样本研究[17,18]、图像生成

研究[19]、领域自适应[20]等其他任务研究的开展。不论从实际应用还是科学研究的角度，风格迁移任务均有广泛的应用。

本文主要介绍图像风格迁移任务，组织架构如下。首先介绍第一阶段的传统风格迁移成果；其次介绍第二阶段神经风格迁移阶段的成果；再次介绍风格迁移领域的评价指标，并比较风格迁移领域中具有代表性的成果；随后介绍风格迁移在其他领域的应用以及非图像的风格迁移任务；最后就风格迁移领域待解决的问题进行讨论。本文的主要贡献如下：

1. 按发展顺序介绍风格迁移部分成果，并完善了一种神经风格迁移的细分方法
2. 比较风格迁移领域部分代表性成果
3. 汇总并分析了风格迁移领域中的客观评价指标。在风格迁移领域存在众说纷纭的客观评价指标，各个文章使用的评价指标各不相同且具有较大差异。据本文所知，本文首次汇总并分析了近年来大部分风格迁移工作中使用的客观评价指标。
4. 介绍风格迁移领域目前存在的问题

下图为本文文章的分类与总览：

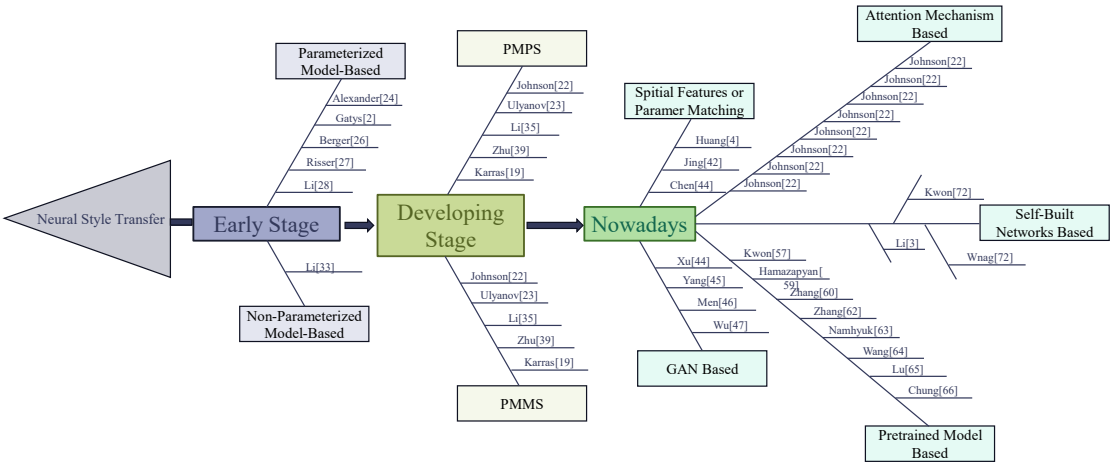

图 1 本文文章总览

2 传统风格迁移

实际上，风格迁移通常被用来待指 2016 年 Gatys 等人发表文章[2]后第二阶段的神经风格迁移工作，在这之前，“风格迁移”一词并为被学术届广泛接受。第一阶段的工作往往被称做非真实感绘制（Non-Photo Realistic Rendering，NPR）或基于图像的艺术渲染（Image-Based Artistic Rendering, IB-AR）。

本文将第一阶段（1990s-2016）工作称作传统风格迁移是为了使读者能更好的理解二者之间一脉相承的关系。

本文接受文章[1]中的建议，采用[21]中传统风格迁移的分类方法进行介绍。但本文的重点不是传统风格迁移，当前同时传统风格迁移已与神经风格迁移工作结合，仅使用传统方法进行风格迁移的工作已不多见，所以本文推荐需要系统性了解传统风格迁移方法的读者阅读这两份成果[21]。

本文在介绍传统风格迁移成果时，将按照特点介绍、优势分析、劣势探讨的方式展开。

基于笔触渲染的风格迁移（Stroke-Based Rendering, SBR）算法的核心是在 2D 画布上覆盖原子级渲染原语，以模拟特定的艺术风格。这些原语通常包括虚拟笔触、图块、点画和阴影标记等。

SBR 中最常见的形式是使用虚拟笔触进行渲染。这些笔触的颜色、方向、大小和顺序可

能是半自动或完全自动确定的。生成的艺术风格化输出既取决于用于渲染每个笔触的媒介仿真，也取决于笔触的放置过程和设置其属性的方法。

SBR 的笔触放置过程可以大致分为局部和全局两种。局部方法通常根据笔触空间邻域内的像素来驱动笔触放置决策。这可以是算法中明确规定的（例如，窗口内的图像矩），也可以是由于先前的卷积操作（例如，Sobel 边缘）而隐含的。

SBR 的一个分支是使用不同于彩色像素或颜料的媒介来填充图像区域，这包括使用小点（点画）进行色调描述，使用线条模式或曲线（阴影标记），以及将小图块组合在一起的马赛克算法。

在处理视频内容时，笔触的运动应与视频内容的运动相匹配。这一点在 SBR 算法中被特别强调，以保证视频中的动态效果和视觉连贯性。

其优势在于能够创造出非常接近传统艺术作品的效果，尤其适用于模仿油画、水彩画、素描等风格。但是在 SBR 方法可能在处理极为复杂或抽象的艺术风格时遇到困难，因为这些风格可能不易通过传统的笔触模拟来实现。

基于示例的风格迁移（Example-Based Rendering，EBR）是一种以学习和模仿特定艺术风格为目的的技术。它通过分析一个示例对（例如，一个原始图像和一个艺术家对该图像的渲染版本）之间的映射关系，然后应用这种映射来风格化其他图像。这类方法通常编码一组启发式规则，以忠实地描绘预定的风格。它们通过学习和模仿艺术家在特定作品中应用的技术和风格，试图捕捉该风格的本质特征。一旦映射被学习，它就可以用来对任意图像进行风格化处理，使得这些图像在视觉上类似于原始的示例图像。这种方法不仅可以模仿特定的艺术风格，还可以复制特定艺术家的独特风格。其优势在于可以产生高度个性化的结果，尤其适用于模仿特定艺术家或作品的风格。但该方法依赖于高质量的示例对，如果缺乏合适的训练数据，可能难以达到理想的风格化效果。

基于图像处理和滤镜的风格迁移（Image Processing and Filtering， IPF）利用各种图像处理滤镜和算法来实现艺术风格化。这包括基于图像金字塔的技术，以及利用交互技术（如人类注视追踪器、重要性图）来探索图像的不同层次。多样性的滤镜技术探索了各种图像处理滤镜，它们被用于艺术风格化，但迄今为止，只有少数成果被认为从艺术角度产生了有趣的结果。这可能是因为这些滤镜通常关注于恢复和恢复摄影真实图像。图像金字塔和交互技术（如人类注视追踪器、重要性图）浏览通过分割源图像的不同分辨率版本构建的区域包含层级，以实现不同层次的抽象表示。高层次抽象通过仅绘制位于金字塔顶部的粗略大区域或特定区域，可以在高层次上对图像进行渲染。这种方法有助于在图像中捕捉更大的形状和构图特征，而不是详细的纹理或线条。与基于图像的艺术渲染（IB-AR）通常追求的简化相比，这些滤镜方法常常与摄影真实图像的恢复和恢复相关联。其优势在于可以快速并且简单地对图像进行风格化处理，适用于创造多样化和抽象的视觉效果。但是该类方法缺乏艺术风格化中所需的细致度和复杂性，难以精确模仿特定艺术家或风格的细微特征。

总结 作为风格迁移领域的先驱者，这些方法启发了此后神经风格迁移的出现与发展，但他们存在一些共同的缺陷。每种不同的风格迁移方法中蕴含了作者对于特定风格的理解，但正因为方法中蕴含了过多作者对风格的理解，导致迁移效果可能与作者的审美水平相关，从而致使风格化质量参差不齐。同时，由于对于相同或类似的纹理或风格设计的算法类似甚至相同，生成的风格化图像中纹理较为死板且无聊。有限的泛化能力也限制传统风格迁移方法的广泛应用，传统方法针对特定类型的风格和图像设计，在泛化到不同类型的风格或图像时效果不佳。

为了解决以上的问题，神经风格迁移顺势而生。

3 神经风格迁移

利用神经网络进行风格迁移的技术一般被称作神经风格迁移。一般认为，在 Gatys 等人[2]使用卷积神经网络进行风格迁移后，神经风格迁移的概念才逐渐兴起。本文借鉴文献[1]中一些对神经风格迁移进行分类的观点，并根据领域发展现状与个人理解进行一定的修改与添加，形成了新的神经风格迁移分类标准。新的分类标准的主要依据有三个：神经风格迁移发展阶段、不同阶段主要探究的问题与解决问题的不同方式。

在该标准下，本文将神经风格迁移划分为三个发展阶段：

1）早期阶段：Gatys 等人[2]引领的基于像素迭代的在线风格迁移、

2）发展阶段：Johnson 等人[22]与 Ulyanov 等人[23]引领的基于模型迭代的离线风格迁移

3）当前阶段：任意且高效的风格迁移。

将神经风格迁移划分成上述三个部分更具有体现风格迁移领域发展脉络的效果，进而使研究者清晰了解各个阶段出现的原因、主要探究的问题与解决方案。值得注意的是，本文按发展顺序的分类可能无法将某些成果准确地划分到精确的阶段，因为总存在一些成果处于两个阶段的交界阶段，同时具有两个阶段方法的特点。对于这种情况，本文将其划分为前一阶段，以显示其与前人成果之间的联系性。

上述三个阶段中的成果均有对应的细分类别。对于基于像素迭代的在线风格迁移而言，可以根据使用的损失函数的类型进一步将其分类两类：基于参数化模型的风格迁移与基于非参数化模型的风格迁移。对于基于模型迭代的离线风格迁移而言， 可以根据网络与网络所能进行迁移的风格数量进行分类，从而得到两个子类：单网络模型生成单风格，单个网络模型生成多风格。对于任意且高效的风格迁移而言，根据文章实现风格迁移技术的不同，可以分成两类：自行搭建网络框架与使用其他新兴技术辅助。

本文将按照上述分类标准，对各阶段神经风格迁移部分成果进行介绍。对于每一项成果，本文将先对其思想进行简单介绍，随后介绍其优势与劣势，进而引出下一项成果。

3.1 基于像素迭代的在线风格迁移

基于像素迭代的在线风格迁移可以追溯到 Alexander 等人[24]对于卷积神经网络（Convolutional Neural Network, CNN）的研究。Alexander 等人[24]试图探究 CNN 为何具有提取图像特征的能力，为此他们将一个 CNN 倒置，利用该倒置 CNN 对 CNN 提取出的特征图像反向重建，从而试图得到与原图类似图像。据 Alexander 等人在文章[24]中介绍，他们通过上述方法生成了具有“一定艺术特征”的图像，如图 2 所示。

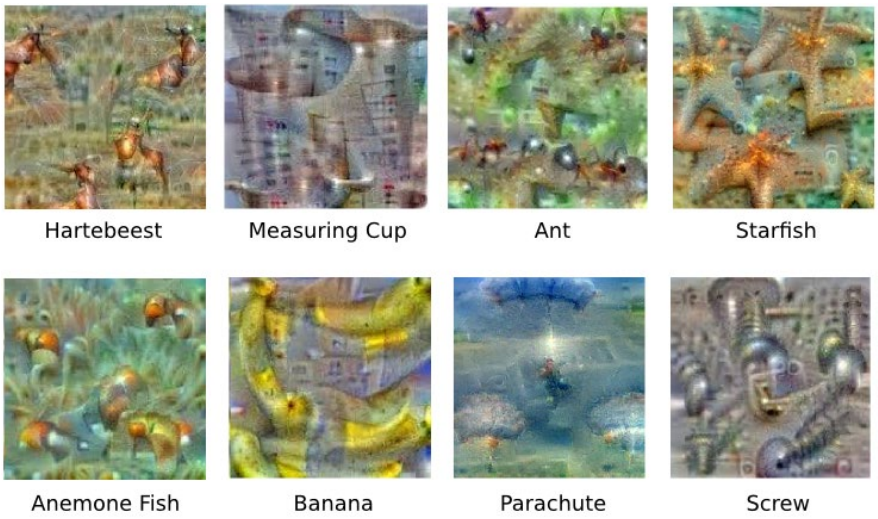

图 2 Alexander 等人[24]生成的艺术图像

通过上述实验，DeepGream 发现 CNN 不仅可以提取图像特征，也可以进行图像生成。这种“利用神经网络进行特征提取，再使用其他方法进行”为后来利用神经网络进行风格迁移打下了一定的基础。

在 DeepDream 之后，Gatys 等人[2]于 2016 年再次将深度学习与风格迁移相结合，作出了突破性的成果，风格迁移的效果达到了一个新的高度。

本文认为，如果一个风格迁移工作以内容图像与风格图像作为输入，并在生成风格化图像时对一张噪声图像进行像素层面的迭代处理，最后输出一张风格化图像，则该风格迁移工作可以被称作基于像素迭代的风格迁移。同时，根据实现方法的不同，可以将该类别进一步细分为基于参数化模型的风格迁移以及基于非参数化模型的风格迁移。

3.1.1 基于参数化模型的风格迁移

Gatys 等人的成果[2]可以视作第一个使用参数化模型进行风格迁移的成果，也是整个神经风格迁移领域的第一个成果。

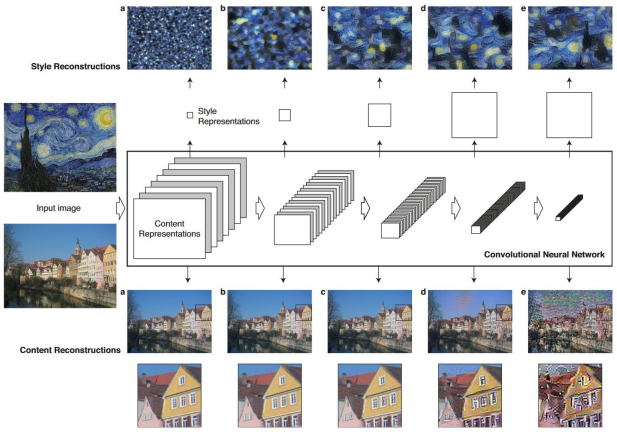

图 3 Gatys 等人网络结构[2]

Gatys 等人在于 2016 年发表的文献[2]中，首次将 CNN 与风格迁移结合。该成果的思想来源于 Gatys 等人在研究 CNN 时的一个发现，即一个使用足够数据训练的 CNN 可以提取提取跨数据集的图像特征[2]。在风格迁移领域中，上述 CNN 提取特征的能力，可用于提取一张照片的内容特征，也可用于提取艺术图像的风格特征，上图 3 展示了基于 CNN 的图像

分类网络 VGG16 在网络不同层具有提取内容特征与风格特征的能力。

在具体实现层面，Gatys[2]等人通过定义一个损失函数，利用 该损失函数对噪声图像进行优化的方式，实现了风格迁移。该损失函数以内容图像的高层特征与正在进行风格化的图像的高层特征之间的差异、以及风格图像与正在风格化的图像的高层特征之间的差异为主要内容。在该损失函数的指导下，将一张噪声图像认定为正在风格化的图像，对该噪声图像进行优化，直到损失函数达到最小值，以获得最终的风格化图像。损失函数的具体形式如下所示：

$$L_{total} = \alpha L_{content} + \beta L_{style} \quad 1$$

其中，是总体的损失函数，L_{total}是内容损失，代表正在风格化的图像与内容图像之间的内容差异程度，L_{style}是风格损失，代表正在风格化的图像与风格图像之间的风格差异程度，α和β是超参数，用于控制生成的风格化图像与内容图像、风格图像之间的相似程度。具体来说，内容损失函数$L_{content}$表示为如下形式：

$$L_{content}(\vec{p}, \vec{x}, l) = \frac{1}{2}\sum_{i,j} \left(F_{ij}^{l} - P_{ij}^{l}\right)^2 \quad 2$$

其中，$\vec{p}$是输入的内容图像的展开，以向量的形式输入网络；$\vec{x}$是正在风格化的图像，同样的，需要对其进行展开成响亮的处理；l代表网络中的第l层，F^l为正在进行风格迁移的图像经过网络中第l层所生成的所有特征图的集合，F_{ij}^{l}表示上述风格化中的图像在 VGG[25]网络第l层生成的特征图集合F^l的第i个特征图的展开向量中位于j位置的值；同理，P^l指的是输入网络的内容图像经过神经网络中的第l层处理后生成的所有内容特征图的集合，P_{ij}^{l}为上述内容特征图集合中的第i张内容特征图的展开向量中位于j位置的值。上述公式的含义在于逐像素计算风格化中的图像的特征图与内容图像的对应的特征图的差值并求和，利用梯度下降等方式对齐进行处理，使其最终稳定在一个较小的值，此时视为优化成功。

另一方面，在给出风格迁移函数L_{style}的具体表达式之前，需要首先介绍其核心——Gram 矩阵。为了提取输入图像的风格特征，Gatys 等人[2]使用了一个旨在捕获纹理信息的特征空间，该特征空间可以建立在网络任意卷积层的输出之上。该特征空间有不同卷积层的特征图之间的相关性构成，而上述相关性用第l层中第i张以及第j张向量化的特征图的内积表示，该内积就是 Gram 矩阵，记作$G^i \in R^{N_l \times N_l}$，Gram 矩阵有公式如下：

$$G_{ij}^{l} = \sum_k F_{ik}^{l} F_{jk}^{L} \quad 3$$

在此基础上，Gatys 等人通过利用梯度下降对噪声图像进行处理，并以最小化原始图像的 Gram 矩阵与风格图像的 Gram 矩阵之间的均方距离为目标。网络中第l层在风格损失函数方面对L_{style}的贡献如下

$$E_l = \frac{1}{4N_l^2 M_l^2}\sum_{ij} \left(G_{ij}^{l} - A_{ij}^{l}\right)^2 \quad 4$$

其中，N_l为网络中一个卷积层中卷积核的个数，也即该层网络能够生成的特征图的个数，M_l是特征图包含的像素的个数，在数值上等于特征图的长度与宽度的乘积，A_{ij}^{l}和G_{ij}^{l}分别表示网络第l层中第i张向量化特征图上位于j位置的像素值。公式5仅计算了网络单层对风格损失L_{style}的影响，将所有层中的损失带权相加即为L_{style}，其具体形式如下：

$$L_{style}(\vec{a}, \vec{x}) = \sum_{l=0}^{L} \omega_l E_l \quad 5$$

其中，w_l是每一层对于总风格损失函数的影响权重，L为网络中卷积层的个数。将式3进行偏导计算，得到结果$\frac{\partial L_{total}}{\partial \vec{x}}$，该结果可以作为作为优化算法的输入，并指导风格化中的图像进行迭代，以完成风格迁移的效果。

Gatys 等人首创的以 Gram 矩阵为核心的参数化模型风格迁移方法具有鲜明的优点。与以往的传统风格迁移方法相比，具有突破传统风格迁移效果笔触呆板、变化度少的优秀的效

果；同时，突破了传统方法仅能对特定风格进行迁移的缺陷，实现了对自然的纹理以及风格化的纹理进行迁移[1]，从而获得良好的迁移效果。但是，此方法存在一些明显的缺陷。由于每次风格迁移都是从一张噪声图像开始，因此在批量进行风格迁移时，需要花费大量的时间，并在实时风格迁移方面效果较差。除此之外，Gram 矩阵更擅长提取特征图的全局信息，这导致对具有长程对称结构的规则纹理的提取效果不能令人满意[1]。同时，与传统风格迁移忽略高层语义信息不同的是，由于 Gatys 等人的方法仅仅考虑了图像中的高层语义信息，而忽略了其中的低层语义信息，导致了合成的风格化图像中精细结构与细节连贯性方面有所缺陷。

为了弥补 Gatys 等人[2]在精细纹理处理上的不足，Berger[26]等人提出了水平垂直像素差异，其主要思想为考察每个像素与其水平与垂直方向上其他像素的差异。在具体实现层面，该方法计算了特征图中位于(i,j)位置的像素与位于$(i,j+\backslash delta)$或位于$(i+\backslash delta,j)$的像素之间特征关系，并将之纳入风格损失的考量。通过这种方式，Berger[26]实现了对具有对称性的长纹理图案的有效迁移，从一定程度上弥补了 Gatys 等人对于精细结构与长程对称结构的规则纹理的模拟效果的不足。同时，由于使用了 Gram 矩阵作为风格迁移的核心，所以本方法依旧保留了一部分 Gatys 等人方法的缺陷，如对细节纹理的模拟不到位，风格迁移质量不稳定。

为了探究使用 Gram 矩阵生成风格化图像的风格不稳定的原因，Risser[27]等人深入研究了 Gatys 等人的方法[2]。Risser 等人认为，生成图像风格不稳定的原因在于具有不同均值和方差的特征图可能具有相同的 Gram 矩阵。基于这个发现，Risser 等人将特征图的直方图纳入风格迁移损失函数的考量，进一步优化了基于 Gram 矩阵的风格迁移方法。其优点在于能够生成更加稳定的风格图像；但由于引入了特征图的直方图，导致该方法的计算更加复杂，风格迁移过程耗时更长，在进行批量迁移时效率更低。同时，由于 Risser 等人仅考虑了风格化过程中生成图像的稳定性问题，不能对纹理精细化描述的问题依旧存在。

Li 等人[28]试图用其他更为成熟的领域知识解释神经风格的原理，他们认为神经风格迁移任务可以视作领域自适应任务的一个特殊的变种任务。以此为切入点，Li 等人得到了不一样的视角。领域自适应任务基于一个事实，即源数据与目标数据的分布不同，其目的是通过在一个带有标签的源域数据集上进行训练，以得到一个能够预测目标领域数据分布情况的模型。领域自适应任务的一种方法是通过最小化源域和目标域中样本的分布差异，从而实现源域与目标域中样本的匹配，其中最大均值差异（Maximum Mean Discrepancy，MMD）是度量两个域之间差异的常用选择。类比到风格迁移任务中，内容图像可以看作是该领域自适应的源域，而风格化的图像即是目标域。Li 等人探究了 Gram 矩阵在风格迁移中的数学作用，证明了对风格图像与风格化中的图像的 Gram 矩阵的匹配过程，即公式4，本质上与最小化一个具有二次多项式核的 MMD 相同。因此，Li 等人认为，最小化具有其他核函数（如线性核、多项式核、高斯核）的 MMD 可能会在风格迁移领域中具有一定的作用。Li 等人的主要贡献在于从理论方面探寻并发现 Gatys 等人方法的原理，使其在原理方面更加清晰。

上述基于参数化模型的神经风格迁移效果存在一些共有的缺陷。由于卷积神经网络丢失了一些图像中的低层信息，导致对于具有规则形状的物体（如人工制造的物件）的迁移结果中往往存在一些不可忽视的扭曲现象，即使 Berger[26]等人在这方面做出了贡献，但仍存在较大的提升空间。同时，风格化的过程会花费大量的时间，在对大量图像进行风格迁移时需要大量的时间，在诸如视频风格迁移等具有大量图像的任务中效率较低，且生成风格不稳定，不具有时间一致性。

3.1.2 基于非参数化模型的神经风格迁移

马尔科夫随机场是用于非参数化模型的神经风格迁移的主要方式，同时使用马尔科夫随

机场（Markov Random Fields，MRF）进行纹理模拟是也传统风格迁移领域中一个较为常见的方法[29–32]。Li 和 Wand 等人[33]选择将传统基于马尔科夫随机场的方法融入神经风格迁移中。他们在文献[33]中将 MRF 与深度卷积神经网络（Deep Concolutional Neueal Networks，dCNN）结合，提出了非参数化的神经风格风格迁移方法。他们认为，使用 Gram 矩阵的参数化风格迁移方法仅考虑了像素和像素之间的差异，没有从空间层次角度对风格化图像进行约束，从而导致了在生成的图像合理性不足。因此，Li 和 Wand 等人将 Gram 矩阵替换为 MRF 正则化器，并引入了一个新的损失函数：

$$L_s = \sum_{l \in l_s} \sum_{i=1}^{m} || \Psi_i \left(F^l(I) \right) - \Psi_{NN(i)} \left(F^l(I_s) \right) ||^2 \qquad 6$$

其中，$\Psi\left(F^l(I)\right)$是特征图$F^(I)$所有局部块的集合；$\Psi_i\left(F^l(I)\right)$为特征图所有局部块集合中的第$i$个；$\Psi_{NN(i)}$是风格化图像$I$中与第$i$个局部块风格最相似的风格块，可以通过计算风格图像中所有风格块的归一化关系从而得到上述最佳匹配的风格化块$\Psi_{NN(i)}$；m为局部块的总数。Li 和 Wand 等人的方法增强了 Gatys 等人风格迁移的效果，使得风格化图像中物体的结构更具合理性，可以更好的保留原图的精细结构，并在合成真实照片方面取得了较大的进步。但同时，该方法缺乏对于图像语义的关注，如果内容与风格图像在结构上存在较大差异时，局部块与风格块的匹配度可能不高，图像块无法正确匹配，最终导致该方法生成的风格图像的效果较差。

3.1.3 总结

基于像素迭代的风格迁移方法可以视作前人在神经风格迁移领域早期的探索，具有开创性的地位，存在一定的优势。

神经风格迁移技术突破了传统风格迁移技术只能迁移一种风格的缺陷，具有较好的泛化能力，一定程度上做到了“一劳永逸”，即构建一个网络结构即可对迁移任意风格与任意分辨率的图像。同时，得益于 CNN 提取特征的能力，神经风格迁移能够捕捉高级艺术风格与纹理，在同一张风格画中具有灵活多变的纹理特征。

虽然基于像素迭代的风格迁移取得了开创性的成果，但同时一些共有问题是不可忽视的，其中最重要的当属资源占用与时间消耗问题。基于像素迭代的风格迁移要求在损失函数的指导下对一张噪声图像进行多轮迭代，每轮迭代均需要经过 CNN 的特征提取与计算，所以大量计算资源与时间被花费在图像生成阶段。在迁移一张512×512的图像风格以生成同样大小的风格化图像时，需要时间高达 51.19 秒[1]，如果试图对一张高分辨率（如 4k）图像进行迁移，则需要的时间往往在实际应用中更是难以接受的。%插入图像表格：基于像素迭代的风格迁移时间占用情况，给出图像分辨率与对应时间。不仅如此，即使投入大量时间与计算资源进行风格迁移，也无法稳定获得超越传统风格迁移方法的效果。

以上两点缺陷导致神经风格迁移技术甚至无法如传统风格迁移技术一般在现实生活中广泛应用，从而导致神经风格迁移的研究成为无源之水。

3.2 基于模型迭代的风格迁移

前文提及，基于像素迭代的风格迁移虽然具备进行任意风格迁移的能力，但是由于其效率问题导致该类方法无法在现实生活中大规模使用，尤其是面对大量图像的任务时，该缺点尤为明显。为了提升神经风格迁移的效率，研究者试图以风格迁移的质量或能迁移风格种类数量为代价实现的快速风格迁移方法被称做基于模型迭代的风格迁移。

一般来说，实现基于模型迭代的风格迁移的主要思路为：将推理时所需时间前置，转换

为训练时的时间。在具体实现层面，其主要实现方式为训练一个特定的风格迁移网络，将风格信息以参数的形式保存在网络中，使其面对不同风格时快速调用对应参数，完成快速风格迁移。

从发展历程角度，基于模型迭代的风格迁移主要可以分成个两阶段，首先出现的是单模型生成单风格，其次出现的是单模型生成多风格。为方便读者对文献采用的方法精确分类，本文给出上述两种类别的详细描述：若一项风格迁移成果接收一张内容图像作为输入，网络能以较快的速度将其转换为特定的风格化图像，则该成果可被称作单模型生成单风格的风格迁移；若一项风格迁移成果接收一张内容图以及最终所需风格的编码为输入，网络能以较快速度将其转换为风格编码对应的风格化图像，则该成果可被称作单模型生成多风格的风格迁移。

3.2.1 单模型生成单风格

利用前馈网络实现单模型生成单风格 Johnson 等人[22]与 Ulyanov[23]等人率先发表了实时风格迁移工作。双方于 2016 年彼此独立地提出了利用前馈神经网络进行实时风格迁移的方法，其主要思想为训练一个前馈神经网络，将风格信息以参数的形式保存在网络中，使得进行风格迁移时无需对噪声图像进行多轮优化即可得到风格化图像。在具体实现层面，需要训练如下的前馈神经网络：

$$\begin{aligned} \theta^* &= \arg\ min L_{total}\left(I_c, I_s, g_\theta * (I_c)\right), \\ I^* &= g_\theta * (I_c), \end{aligned} \tag{7}$$

其中，θ^*为最优的参数，即令损失函数L_{total}取最小值的参数；

$L_{total}\left(I_c, I_s, g_\theta * (I_c)\right)$为整体的损失函数，衡量了风格迁移的质量。该函数的输入包括三个部分：内容图像l_c，风格图像l_s，风格化图像$g_\theta * (I_c)$；该函数仅有一个输出，即I^*，为最终生成的风格化图像。该函数的目标为寻找一个最小化L_{total}的参数θ^*，该参数被用于将内容图像l_c转换为最终的风格化图像I^*。

虽然 Johnson 等人与 Ulyanov 等人同时提出了该方法，但二者的网络结构存在差异：Johnson 等人在 Radford 等人[34]方法的基础上，添加了残差块与分步卷积，并引入的实例归一化层（ Instance Normalization，IN）以加快网络的收敛速度，如图 4 所示：

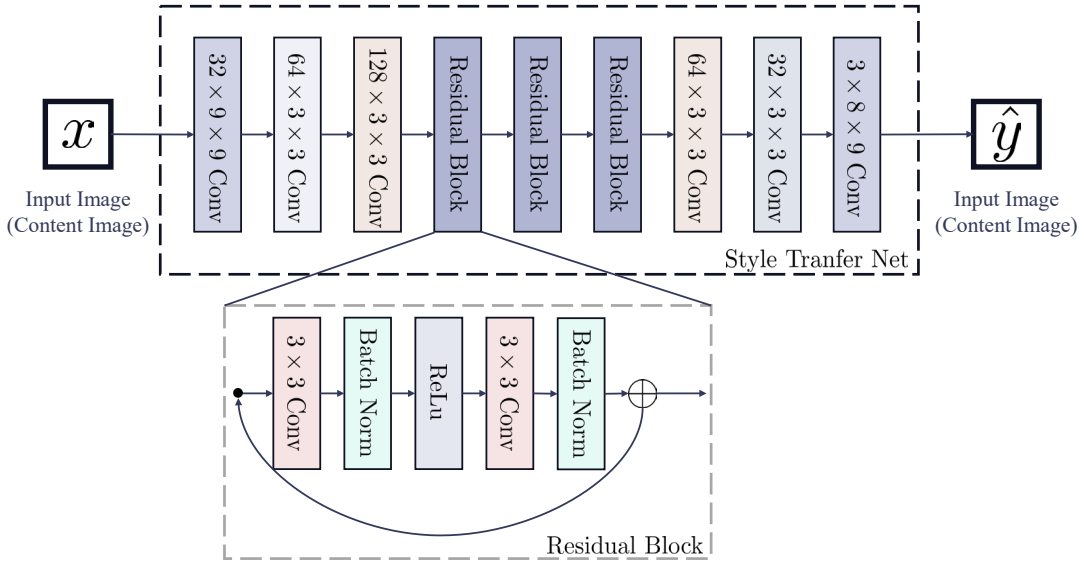

图 4 Johnson 等人的网络结构[23]

而 Ulyanov 等人则以多尺度结构作为生成网络（如图 5 所示），目标函数与 Gatys 等人类似。Johnson 等人与 Ulyanov 等人均以前馈生成网络为基础，达成了实时风格迁移的效果，

风格迁移的速度 Gatys 等人相比，提升了两个数量级。

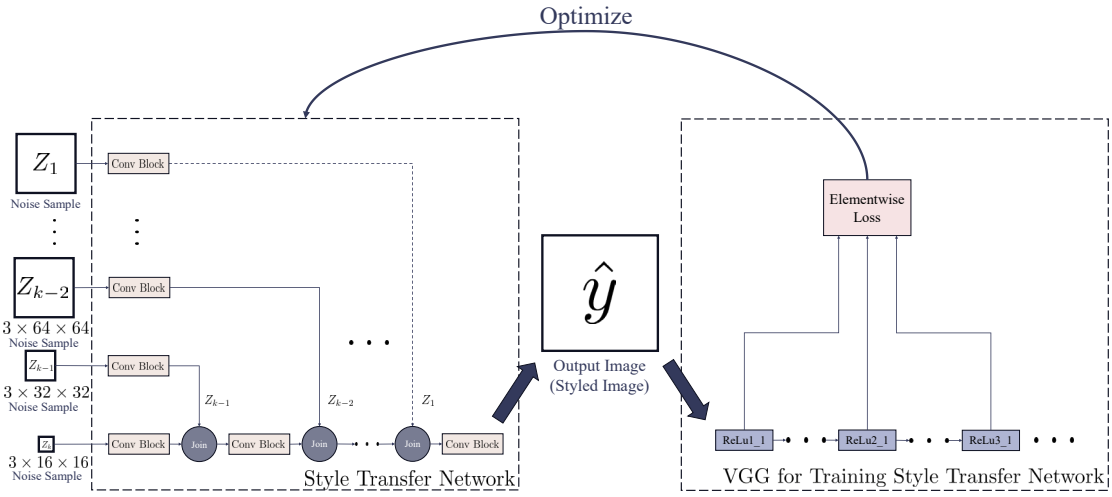

图 5 Ulyanov 等人的网络结构

但由于两人的方法在训练时基本遵循 Gatys 等人[2]提出方法，所以在迁移效果方面存在类似的问题，如在图像细节与结构合理性方面效果不能令人满意等。

利用马尔科夫随机场实现单模型生成单风格 利用马尔科夫随机场同样也可提升风格迁移速度。Li 和 Wand 等人[35]改进自身以往的工作[33]，他们以对抗训练的训练方式获得一个马尔科夫前馈网络（如图 6 所示），从而解决的效率问题。从原理上来说，该方法类似于之前他们自身在文献[33]中的一种基于图像块的非参数风格迁移方法。这使得他们的方法在物体结构的合理性方面具有更好的效果，从而更好地保留原图的精细结构。但同样的，原始方法的一些缺点也被保留，如局部块与风格块的匹配度不高时，生成的风格图像效果较差。

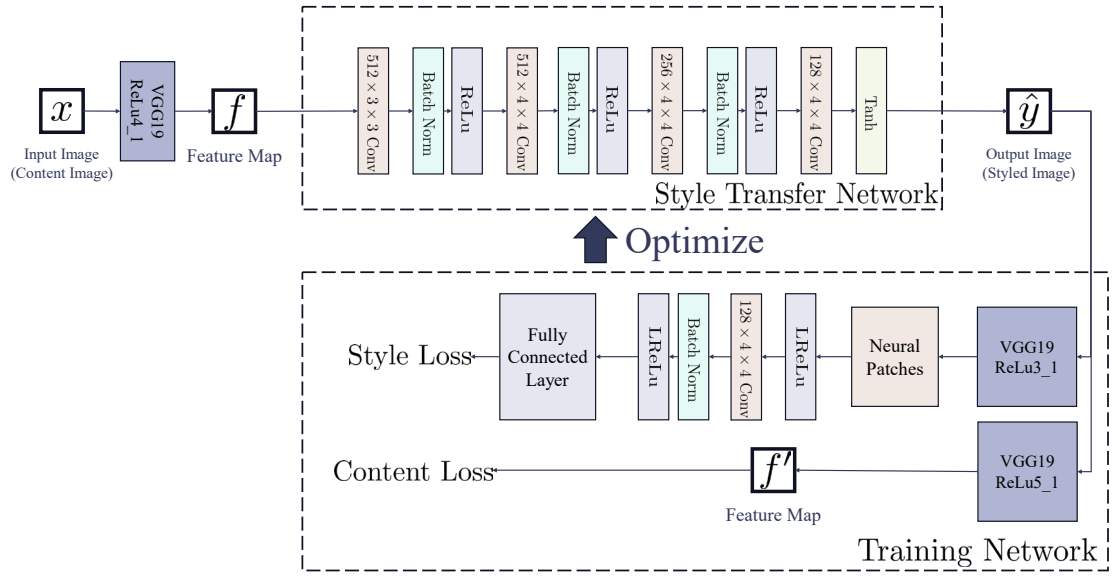

图 6 Li 等人网络结构[35]

利用生成式对抗网络实现单模型生成单风格 利用生成式对抗网络（Generative Adversarial Networks，GAN）进行风格迁移则是另一种基于模型迭代的风格迁移方法。

GAN 由 Goodfellow 等人于 2014 年提出[36]，该模型采用一个生成网络和一个判别网络进行对抗。在训练时，需要先对判别网络进行训练：给定一组数据，判别网络判断该组数据中的每一个数据项是否属于目标域，并根据真实值与网络计算得到的判断值之间的差距进行优化，直到能够判断测试集中的数据项是否属于目标域；随后对生成器进行训练，生成器生

成数据，交由判别器进行判断是否属于目标域，生成器再根据判别器给出的结果进行调整。生成器与判别器二者不断进行对抗，以达到两个网络之间的均衡，从而完成网络的训练，并实现生成数据与真实数据之间分布的相似。GAN 的损失函数如下所示：

$$V(D,G)=\mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)]+\mathbb{E}_{z\sim p_z(z)}\left[\log\left(1-D\left(G(z)\right)\right)\right] \quad 8$$

其中，G与D分别为生成器网络与判别器网络，$p_{\text{data}}(x)$表示数据x所满足的分布，$p_z(z)$表示噪声图像的分布，$D(x)$表示x属于真实数据而非输入生成器生成的数据p_g的概率，$G(z)$表示生成器对噪声图像的处理结果，$\mathbb{E}$表示期望。训练判别器的过程以最大化判别器能否识别来自数据集而非生成器G的图像的概率；生成器的训练过程以最小化$\log\left(1-D\left(G(z)\right)\right)$为目标，$1-D\left(G(z)\right)$表示判别器认为生成器生成的图像不属于真实数据集的概率，最小化该函数及欺骗判别器，使其认为生成器生成的数据来自真实数据集。GAN 优秀的网络结构使其能够胜任风格迁移的任务。

GAN 在实现实时风格迁移的同时存在一些问题与缺陷。首先，GAN 的训练通常比较困难，容易出现训练不稳定的情况，如模式崩溃（mode collapse）等。其次，GAN 模型的损失函数并未给出类似于 Gatys 等人[2]损失函数中的超参数α与β，从而导致了无法控制生成图像与内容图像或风格图像的相似度，在风格迁移任务中难以细粒度地控制两者。GAN 虽然可以用于风格迁移任务，却不是一个专用于风格迁移的网络。

ZHU 等人[37]对 GAN[38]进行改造，使其能更好地适应实时风格迁移任务。ZHU 等人于文献[37]中提出的 CycleGAN 实现了无监督训练的单模型生成单风格。CycleGAN 的独特之处在于它引入了“循环一致性损失”，并通过同时训练两个生成器与判别器实现该损失，从而确保了转换是双向性，即具有从一个域到另一个域并返回的能力，在此过程中信息不会丢失。与 GAN 类似，CycleGAN[37]的生成器旨在将图像从一个域映射到另一个域，判别器则试图区分生成的图像和真实图像（图 7）。

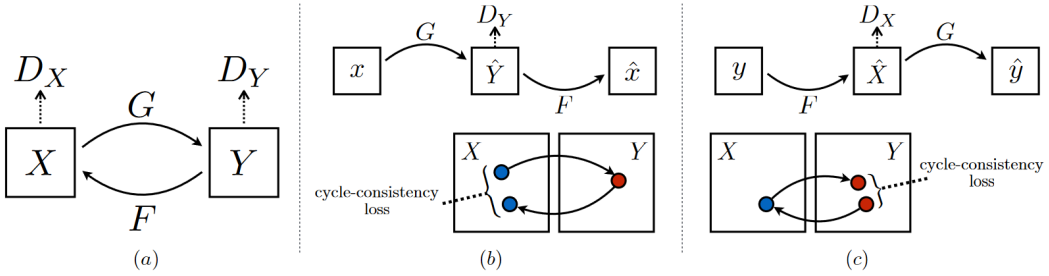

图 7 CycleGAN 工作流程[38]

CycleGAN 的最大优势之一是其能够处理非配对数据从而实现无监督学习，该方法能够学习如何进行跨域图像转换，而无需在训练时提供每个样本的明确匹配。但同时，CycleGAN 在风格迁移领域有一些明显的缺陷，其生成的图像有时可能比真实图像更模糊甚至失真。其次，不当的超参数选择可能降低训练过程的稳定性，或者造成生成图像质量较差的后果。

StyleGAN[19]是使用 GAN 进行风格迁移的另一项杰出成果，本文尤其关注肖像编辑功能，可以生成具有特定特征的肖像，且真实度高。StyleGAN 通过调整输入潜在空间的“风格”参数，从而控制生成图像的各个细节和整体风格。StyleGAN 的核心创新是引入了一种新的风格转移机制，能够在不同层次上分别控制图像的内容和风格，从而在保持图像质量的同时提高了生成图像的多样性和可控性。StyleGAN 的优点包括能生成高分辨率、高质量的图像，并且通过风格控制提供强大的图像编辑能力。它在面部和其他复杂图像生成方面表现出色。缺点是训练过程资源密集且耗时，对计算资源要求高。此外，生成的图像有时可能包

含难以预测的伪影，需要进一步优化以提高稳定性和可靠性。

总结来说，单模型生成单风格通过将原本迁移时需要的时间一定程度上转移到训练阶段，从而解决了前一阶段工作，即基于像素迭代的风格迁移的最主要缺陷——迁移时间长效率低，但代价是只能迁移特定单一风格。若试图以该类方法对多种风格进行迁移，则需要对每种风格进行对应的网络训练，产生大量的参数。在此情况下，大量的参数使其难以部署到一些资源有限的设备（如智能手表等），因此需要进行风格种类多少与性能弱强之间的博弈。

3.2.2 单模型生成多风格

上述单模型生成单风格的方法虽然解决了实时风格化的问题，提升了风格迁移的效率，但是单个模型只能对应某个特定的风格，进行新风格的迁移时，需要花费大量的时间进行新模型的训练，同时难以应用到资源有限的设备上。为此，学术届开始研究在保留研究快速迁移的同时，也具有迁移多种风格能力的网络，即单模型生成多风格网络。实现多风格迁移的一个主要思路是将某一特定风格与少部分网络参数绑定，以减少重复的参数，并结合增加网络中体现风格差异的必要参数的方式实现多风格迁移。

Dumoulin 等人[39]率先将特定风格与网络中参数绑定，从而实现单模型生成多风格的风格迁移。在减少重复参数层面，他们认为部分不同风格的迁移工作中存在相似或者相同的计算部分，因为许多名称不同的艺术风格具有相似或相同的笔触（如印象派绘画具有相似的笔触，区别仅在使用的画面中使用的颜色），将这些具有相似笔触的画作看成不同的风格似乎是很浪费的。而传统一对一的风格迁移模型忽略了这一点，导致在对新风格进行迁移时，造成了训练阶段不必要的时间浪费。

在实验中，Dumoulin 等人[39]发现，只需要将标准化后的参数进行缩放或变换，即可适应某种特定的风格；对于一个卷积神经网络而言，这个发现表示网络中所有卷积核的参数可以在调整后用于迁移不同风格。具体来说，仅需在归一化后对卷积核的某些参数进行调整即可实现不同风格的迁移。在实现层面，Dumoulin 等人[39]以 Ulyanov 等人[23]的方法为基础，在进行实例归一化后继续进行了一次仿射变换，完成了不同风格的迁移工作。这个过程被他们称作条件归一化（Conditional Instance Normalization，CIN），可以用如下公式表示：

$$\mathrm{CIN}\big(\mathcal{F}(I_c,s)\big)=\gamma^s\left(\frac{\mathcal{F}(I_c)-\mu\big(\mathcal{F}(I_c)\big)}{\sigma\big(\mathcal{F}(I_c)\big)}\right)+\beta^s \tag{9}$$

公式的输入为内容图像I_c与可迁移风格的序列号s，$\mathcal{F}(x)$表示图像x的特征图，$\mu(x)$与$\sigma(x)$分别表示图像的均值与标准差。该方法通过缩放或变换参数γ^s与β^s完成不同风格的迁移，即每种风格s均可通过调整仿射变换的参数实现。

其优点在于，通过将近似风格的参数进行仿射变换的方式，Dumoulin 等人实现了重复参数的减少，较单模型生成单风格方法在生成同样数量的风格时，所需参数量大为降低。但在面对风格差异巨大的图像时，也需要重新训练网络以获得对应风格的参数与调整参数的方式。因此随着可迁移风格的增加，该方法中网络参数也随着不断增加。

Chen 等人[40]通过另一种方法实现了参数简化。与 Dumoulin 等人[39]用相似的参数表示相似的风格不同，Chen[40]等人从内容图像处理方面考虑，认为网络中处理内容信息的部分可以相同，因此诞生了将网络中内容处理模块与风格处理模块解耦合的思想。利用解耦合的思想，即采用独立的网络模块学习图像的内容信息和风格信息，使得网络能够更加灵活地处理风格迁移任务。该方法使用中层卷积滤波器（文中[40]称为“StyleBank”层，如图 8 所示）专用于学习不同的风格，“StyleBank”层中包含了多组参数，每组参数与一个特定的风格相关联。

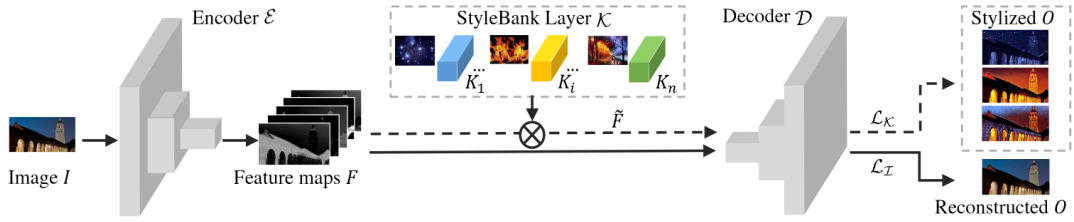

图 8 Chen 等人网络模型[40]

除了“StyleBank”层之外的其他部分用于学习内容信息。由于内容处理模块对不同风格的处理是相同的，故而不同风格图像可以应用相同的内容处理模块，这提高了网络处理不同风格的效率。在实现多种类风格迁移时，仅需进行增量训练即可。具体来说，当需要添加一个新的风格时，可以固定网络中用于学习内容信息的部分，只对新风格的“StyleBank”层进行训练。这种方式使得网络能够在不影响已学习风格的基础上，有效地学习新风格。StyleBank 中一组或多组卷积核代表了一个特定的风格，将不同组的卷积核放入神经网络中，即可完成对不同风格的迁移工作，从而获得了较好的拓展性。

上述思路虽然能使用单个网络迁移多种风格，并且优化了网络中参数数量，但是问题依旧存在。在参数方面，该方法若试图迁移多种风格，则参数数量会随着能迁移的风格数量而逐渐攀升，这导致使用该思路的方法无法做到任意风格的迁移。在生成图像质量方面，由于采用了部分参数共享的方式，因此迁移质量可能无法达到如基于模型迭代的风格迁移的优秀结果。

3.3 任意且实时风格迁移

上述风格迁移的发展历程可以概括为一句话，即迁移数量、迁移质量与迁移时间、资源之间的权衡。但在实际应用中，用户可能对迁移风格与迁移速度同时提出要求。为此，能够迁移任意风格、能较高质量且高效完成风格迁移的方法随之出现。

目前主流实现任意且实时风格迁移的方法按是否使用其他技术辅助可分成三类：

其一为利用图像空域特征或参数匹配进行风格迁移；

其二为使用诸如 GAN[38]、注意力机制、扩散模型、预训练大模型等技术作为辅助；

其三为以深入研究风格迁移为主要特点，通过思考与研究风格迁移新流程与网络新结构实现任意且实时风格迁移任务。

由于基于 GAN[38]的方法尤为众多，因此将其从第二类中分离出来。根据上述描述，本节将从以下 4 个类别介绍任意且实时的风格迁移成果：

1）基于空域特征或参数匹配的风格迁移；

2）基于 GAN[38]的任意且实时风格迁移；

3）基于注意力机制的任意且实时风格迁移；

4）基于预训练大模型的任意且实时风格迁移

5）基于自搭建网络的任意且实时风格迁移。

基于空域特征或参数匹配进行风格迁移的方法出现时间早，且迁移效果质量不佳，作为任意且实时的风格迁移先驱介绍。后三者的发展齐头并进，目前的风格迁移领域的最新成果也大致可以归纳至该三个类别中。

3.3.1 基于空域特征或参数匹配的风格迁移

基于空域特征或参数匹配的风格迁移方法以图像空域中的特征为突破口，试图通过调整

空域特征参数或匹配内容图与风格图中对应区域进行任意风格迁移。

据本文所知，第一个实现任意且同时的风格迁移的工作由Huang等人于2017年提出[4]。Huang 等人受到 CIN[39]等人方法的启发，提出了自适应实例归一化层（Adaptive Instance Normalization， AdaIN）。该层用于实现内容特征的方差与均值与风格特征的均值与方差的对应，通过这种方差、均值的匹配实现了任意风格的迁移，AdaIN 的公式可以描述为如下形式：

$$\mathrm{AdaIN}\big(\mathcal{F}(I_c),\mathcal{F}(I_s)\big)=\sigma\big(\mathcal{F}(I_s)\big)\left(\frac{\mathcal{F}(I_c)-\mu\big(\mathcal{F}(I_c)\big)}{\sigma\big(\mathcal{F}(I_c)\big)}\right)+\mu\big(\mathcal{F}(I_s)\big) \qquad 10$$

与[39]不同，Huang 等人的风格迁移网络中编码器是固定的，包含预训练的 VGG 网络中的前几层，其网络整体结构如图 9 所示。

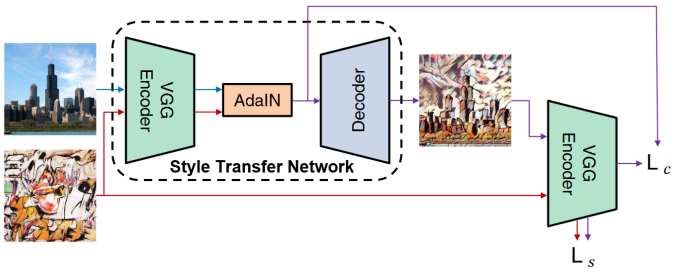

图 9 Huang 等人网络结构[4]

考虑到前人使用归一化进行风格迁移[4,39]时存在的共性问题，Jing 等人[41]修改 AdaIN [4]并提出了动态实例归一化层（Dynamic Instance Normalization，DIN）。DIN 的主要优势在于，在通过对齐内容和风格特征之间的均值和方差（最简单的统计数据）以进行任意风格迁移时，无须手动手动定义计算仿射参数的方法（公式），而是引入了更广义的动态卷积变换，其中参数自适应地改变根据不同风格以可学习的方式，从而更准确地对齐真实复杂的风格特征统计数据。给定一对内容图像 I_c 和风格图像 I_s 作为输入，所提出的 DIN 层可以用以下公式表示：

$$\begin{aligned}\mathrm{DIN}(\mathcal{F}_c,\mathcal{F}_s) &= f[\mathcal{F}_s,\mathrm{IN}(\mathcal{F}_c)],\\ \mathrm{IN}(\mathcal{F}_c) &= \frac{\mathcal{F}_c-\mu(\mathcal{F}_c)}{\sigma(\mathcal{F}_c)}\end{aligned} \qquad 11$$

其中$\mathcal{F}_c$ 和$\mathcal{F}_s$ 是$\mathcal{I}_c$和$\mathcal{I}_s$对应的特征表示，f 是动态卷积运算[42]。与权重和偏差是模型参数的标准卷积不同，DIN 动态卷积 f 中的权重和偏差是通过编码不同的输入样式图像动态生成的。这种动态调整机制使得 DIN 在捕捉和应用细腻风格特征方面更为灵活和精确，从而在保持内容结构的同时丰富风格细节表达。

Chen 等人的工作[43]也是一项早期的任意风格迁移工作。与[4,39,41]一类使用图像空域特征参数匹配进行风格迁移的方法不同，该方法注重内容图像与风格图像之间的差异，该方法先将图像划分成多个图像块（patches），并交换内容图像与风格图像之间最相似的特征块以实现风格迁移实时且任意的风格迁移。模型总体结构如图 10 所示，具体步骤如下：1. 提取内容特征图与风格特征图中的一组图像块，分别记为内容特征块组 $\{\phi_i(C)\}_{i\in n_c}$ 与风格特征块组 $\{\phi_i(S)\}_{j\in n_s}$，其中n_c与n_s分别表示内容特征块与风格特征块的数量。该组图像块应从所有特征图中采样，且之间应该有足够的重叠；2. 对于每个内容特征块，根据归一化互相关度量（公式12）确定最接近匹配的风格特征块；3.将每个内容特征块$\phi_i(C)$与其最接近匹配的风格图像块$\phi^{ss}(C,S)$交换；4.对与重叠区域，如果在步骤 3 中获得不同的值，则将其进行平均化处理。

$$\phi_i^{ss}(C,S) \coloneqq \underset{\phi_j(S),j=1,\dots,n_s}{\operatorname{argmax}} \frac{\langle \phi_i(C),\phi_j(S)\rangle}{\|\phi_i(C)\|\cdot\|\phi_j(S)\|} \tag{12}$$

通过上述方式即可获得风格化特征图，将该风格化特征图进行上采样重建，即可获得风格化图像。该方法虽然实现了任意且实时的风格迁移，且生成质量较上述基于空域特征参数匹配的方法[4,39,41]更优，但在时间与资源消耗方面则是不如[4,39,41]的方法。同时，由于使用了图像块作为风格迁移的单元，从而忽略了全局风格信息，并在相邻图像块的相似性测量上有所欠缺。

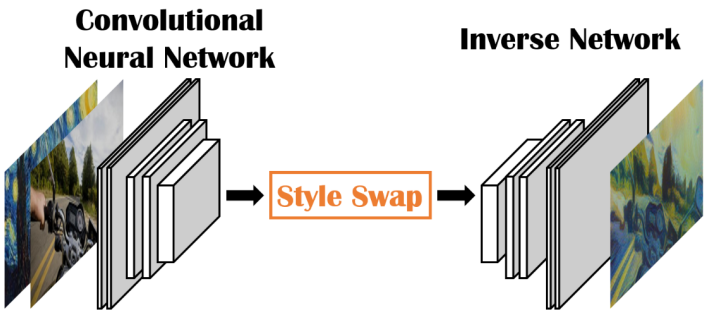

图 10 Chen 等人网络结构[43]

3.3.2 基于 GAN 的任意且实时风格迁移

由于 GAN 具有一系列优点，如强大的无监督学习能力、特征学习能力、数据泛化能力等，在风格迁移发展到任意且实时阶段，研究者也因此考虑对 GAN 进行进一步的调整，使其具有更好的风格迁移能力。

据本文所知，Xu 等人[44]第一个利用 GAN 实现了任意且实时风格迁移，该文提出动态残差块生成对抗网络（DRB-GAN）用于风格迁移。动态残差块（DRBs）的想法受到 DIN 和 StyleGAN 的启发，将“风格代码”建模为动态卷积和 AdaINs 在动态残差块中的共享参数，并在风格迁移网络的瓶颈处设计多个动态残差块（DRBs）。每个 DRB 由卷积层、动态卷积层、ReLU 层、AdaIN 层和具有残差连接的实例归一化层组成，从而达成了调整动态卷积的共享参数，并适应性地调整 AdaINs 的仿射参数的目的，从而确保内容图像和风格图像之间瓶颈特征空间的统计匹配。使用动态残差块的原因在于其能够提供灵活的参数调整，以更好地实现风格和内容之间的统计特征匹配。这种设计使得网络能够在保持内容图像的基本结构的同时，有效地融合各种风格特征。

Yang 等人[45]在 StyleGAN[19]的基础上进行工作，认为 StyleGAN 仅能对特定风格进行快速迁移，而无法对任意风格进行实时迁移，且无法生成真正具有艺术价值的肖像。为了应对这些挑战，Yang 等人提出了 DualStyleGAN（图 11），以实现基于样本的肖像风格迁移。DualStyleGAN 保留了 StyleGAN 的一个内在风格路径来控制原始域的风格，同时添加了一个外在风格路径来建模和控制目标扩展域的风格。此外，外在风格路径继承了 StyleGAN 的分层架构，以调制粗分辨率层中的结构风格和精细分辨率层中的颜色风格，以实现灵活的多级风格操作。DualStyleGAN 虽然可以实现灵活的肖像风格迁移，但是当内容图像背景中存在其他物件时，该网络往往也会将该物件识别为人脸，从而生成难以描述的图案。

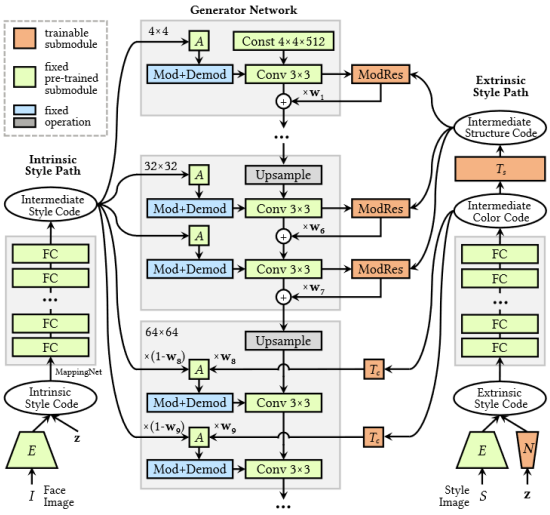

图 11 Yang 等人网络结构[45]

Men 等人[46]改进了 DualStyleGAN[45]无法区分主体与背景的缺陷，实现了前背景分离的风格迁移（图 12）。该方法的核心思想是引入门控循环映射（Gated Cycle Mapping），GCM 利用新颖的门控映射单元来生成特定于类别的风格编码，并将该编码嵌入循环网络中以控制翻译过程。特别的，Men 等人将图像分为不同的类别（例如，4 种类型：照片/卡通肖像/场景），并学习更细粒度的类别转换，而不是两个域（例如照片和卡通）之间的整体映射。Men 等人将前背景分离，取得了更好的卡通风格迁移效果，但在迁移具有对称性的背景物体时，会出现线条的扭曲现象，从而降低了迁移的最终效果。

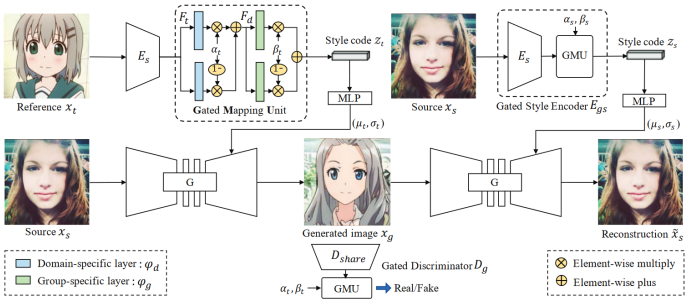

图 12 Men 等人网络结构[46]

Wu 等人[47]从另一个角度考虑风格迁移问题，他们认为即使同一位艺术家，也可能具有风格迥异的艺术作品，因此将同一艺术家的不同艺术风格同质化是不严谨的做法，而且以往方法对未见过的艺术家也缺乏概括性。为了解决个挑战，Wu 等人提出了一种双风格传输模块（Double-Style Transferring Module，DSTM）。该模块从不同的艺术品中提取不同的艺术家风格和艺术品风格，并保留同一艺术家的不同艺术品之间的内在多样性。考虑到从单幅艺术作品中学习风格可能会引起对其的过度适应，从而导致风格图像结构特征的引入，Wu 等人进一步提出了边缘增强模块（Edge Enhancing Module，EEM），该模块从多尺度和多层次特征中提取边缘信息，以增强结构一致性。本文优势在于，使用该方法进行风格迁移而得的风格化作品中，较少地出现风格图像中的内容特征，从而很好的保留了内容图像的内容特征。

3.3.3 基于注意力机制的任意且实时风格迁移

在深度神经网络中使用注意力机制逐渐成为风格迁移任务中的一个常见选择。先前的风格迁移成果在模型设计上往往更关注图像的整体风格，但是忽略了局部的风格对应关系。通过引入注意力机制，借助其捕捉图像的空间布局和语义关系的特点，风格迁移模型能够有效

地识别图像中的关键区域，并将风格特征精准地应用于这些区域，从而能够在一定程度上实现整体与局部风格的兼顾。此外，注意力机制允许跨不同图像的特征交换，这对于实现细腻的风格迁移效果尤为关键。

Liu 等人[48]试图借助上述注意力机制的特点，以提升风格化结果中的局部质量。他们认为现有的解决方案要么仅仅考虑将深层风格特征融入深层内容特征而不考虑特征分布的情况，要么根据风格自适应地对深度内容特征进行归一化，使其全局统计量相匹配[4,39,41]。以上方法虽然有效，但仅仅关注图像的深层特征与全局特征，忽略了浅层特征与局部特征，从而导致了局部失真。为了缓解这个问题，Liu 等人提出了一种新颖的注意力和归一化模块，称为自适应注意力归一化（Adaptive Attention Normalization，AdaAttN），以在每个像素的基础上自适应地执行注意力归一化。AdaAttN 的工作过程分为三个步骤：1）从浅层到深层计算具有内容和风格特征的注意力图；2）计算风格特征的加权均值图和标准方差图； 3）自适应归一化内容特征以进行逐像素特征分布对齐。此后利用解码器处理 AdaAttN 的输出，即可完成风格迁移的工作。该方法利用注意力机制考量图像的局部信息匹配，实现了更细致和个性化的风格转换，增强了图像的局部视觉质量。

与 Liu 等人[48]认为以往工作忽略局部特征不同，Deng 等人[49]则认为，CNN 中卷积的感受野有限，只能感受局部信息，在提取和维护输入图像的全局信息方面存在困难。

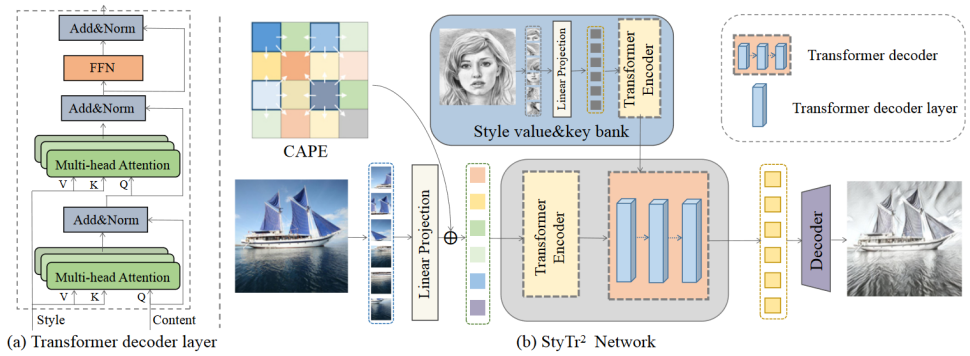

图 13 Deng 等人模型总览[49]

为了解决这个问题，Deng 等人[49]提出了一种名为 StyTr2 的基于 Transformer 的方法（图 13），用以考虑输入图像的远程依赖性。StyTr2 包含两个不同的转换器编码器，分别为内容和风格生成特定于领域的序列。在编码器之后，采用多层转换器解码器根据风格序列对内容序列进行风格化。StyTr2 缓解了基于 CNN 的模型的内容泄漏问题，取得了更好的风格迁移效果。但由于 Transformer 的庞大参数，Deng 等人[49]的方法在运行时间方面相比前人则是稍逊一筹。

Li 等人[50]也注意到基于 CNN 的方法难以获取长距离信息，同时考虑到基于 Transformers 的方法计算成本过大，因此他们试图在长距离信息与大量参数之间实现平衡，并从一定程度上环节依旧存在的内容泄露问题。为了解决计算和泄漏问题，Li 等人设计了一个名为 AdaFormer 的紧凑型变压器。该设计采用图像块投影和位置编码来增强全局交互，并假设内容与风格特征的差异在编码器的高层捕获。AdaFormer 通过共享参数的初始变换器编码层，然后通过各自的编码层分别提取内容和风格特征，从而实现更高效的特征提取。该算法还采用了多层变换器解码器进行特征融合，并通过动态权重选择风格元素。此外，采用自适应实例归一化 AdaIN[4]取代层归一化，以使风格更符合内容。最后，通过上采样解码器生成多样化的风格化作品。相比于 StyTr2[48]，Li 等人的方法在内存占用方面确实有所下降，但以及需要占用约 35G 内存[50]。

Zhang 等人[51]则将注意力机制融入经典的 AdaIN[4]，以改进该类方法。他们认为，当前

基于 AdaIN[4]的风格迁移工作具有缺陷：在生成高质量风格化图像时，容易出现内容与风格不匹配、图像伪影以及风格特征提取不准确等问题。Zhang 等人[51]分析造成该现象的原因有 2 个：（1）使用 VGG 网络进行特征提取，以及（2）在实例归一化时，仅仅使用统计参数调整作为风格迁移的手段。

对于第一个问题，Zhang 等人[51]认为 VGG 是一个用于图像分类的网络，从而导致该网络在风格迁移时，会过分关注无用的图像分类信息。为了解决这个问题，他们提出了一个基于 Transformer 的风格图像特征提取器，并称之为感知编码器 (Perception Encoder, PE)。PE 通过捕捉风格图像的长程依赖信息和高频风格细节，避免了 VGG 只专注于显著分类特征（如边缘或形状）的局限性，从而更准确地提取风格信息。

针对第二个问题，他们提出了风格一致性实例归一化(Style Consistency Instance Normalization, SCIN)。该方法与 AdaIN 仅根据均值和方差进行简单对齐以实现风格迁移的方法不同，SCIN 使用 Transformer 捕捉风格特征图中的长程、非局部依赖关系，提供更丰富的风格信息。此外，SCIN 生成的缩放和平移参数是通过学习得到的，能够更好地适应不同风格的图像分布，而不仅仅依赖于固定的统计特征（如均值和方差）。这一改进使得 SCIN 在风格与内容特征对齐时更加灵活和准确，减少了伪影并提升了风格化图像的质量。

为了进一步提升风格迁移结果质量, 使不同风格的图像具有较大区分度, 文章还提出了基于实例的对比学习 (Instance-based Contractive Learning, ICL). ICL 帮助模型学习风格化图像之间的关系，确保相同内容或风格的图像嵌入更接近，不同风格的图像则远离，从而增强了风格化图像的质量。

通过上述三个方法，文章从整体与局部结合的角度出发，改进了 AdaIN[4]仅考虑全局特征统一的不足，减少了伪影并最终提升了风格化图像的质量。

Wang 等人[52]另辟蹊径，认为当前风格迁移方法产生的纹理是不可预测的，这与不符合艺术创作逻辑，并认为能让使用者参与风格迁移过程的交互是十分重要的。为了能够产生可预测的纹理，Wang 等人提出了交互式图像风格迁移网络。该网络可以生成由涂鸦曲线引导的任意方向的笔触的风格化结果，以确保风格化图像的风格分布更接近现实生活中的艺术品。具体来说，IIST-Net 由两个编码器、一个交互式画笔纹理生成模块（Interactive Brush-texture Generation，IBG）、一个多层风格注意力模块（Multilayer Style Attention，MSA）和一个解码器组成。其中两个编码器用于对内容风格图像和画笔纹理图像进行编码，以分别产生多层风格特征和融合内容特征；IBG 模块根据用户输入交互信息生成可控画笔纹理；MSA 模块可用于进一步提炼多层风格特征并将其与融合后的内容特征融合。本文的方法为从一定程度上弥补了神经风格迁移领域中有关交互性的不足，但是生成的图像过于强调纹理，导致内容特征较难难引起注意。

Zhang 等人[53]意识到了内容特征与风格特征相互平衡的重要性。他们认为如果用足够的风格模式对图像进行风格化，可能会破坏内容细节，有时甚至无法清晰地区分图像的对象。为了解决内容与风格的平衡问题，Zhang 等人提出了一种基于 Transformer 的风格迁移方法，称为 STT（Style Transfer via Transformers）。该方法中，Transformer 主要被用于从内容与风格特征的编码，并在最后完成解码工作。为了使风格化图像的内容结构清晰，Zhang 等人设计了一种新颖的边缘损失以增强输出图像中对象的边缘。与边缘检测或轮廓提取等任务不同，风格迁移中输出的内容细节可能与内容图像的内容细节不同，特别是背景可能具有风格图像的艺术图案。因此，如果直接将内容图像和风格化图像的边缘图之间的相似度作为优化目标，结果将会变得模糊。因此需要解决的问题之一就是过滤掉内容图像的主体结构不存在的地方。Zhang 等人引入掩膜操作来解决这个问题：风格化图像边缘图中不存在于内容图像边缘图中的对应位置的所有边缘将被遮盖掉。同时，Zhang 等人还设置了一个阈值来排除边缘图的弱响应，以防止其可能引起的噪声。Zhang 等人的方法通过突出内容图案达成了风格

图案与内容图案在风格化图像中的平衡，获得了更好的视觉体验，该方法也从一程度上缓解了内容泄露问题。但同样的，由于使用了 Transformer 作为编解码器，该方法相比其他未使用 Transformer 的任意且实时风格迁移方法需要更多时间与资源。

Hong 等人[54]则较为担忧风格迁移中训练数据不匹配的问题。他们认为，任意内容与风格图像之间的低语义对应关系导致注意机制聚焦于风格图像的有限区域。这会阻碍基于注意力的方法准确地捕获和表达参考图像的整个风格，从而导致了不和谐的图案。为了克服上述局限性，Hong 等人专注于增强注意力机制，并致力于捕捉风格图像中纹理的韵律。Hong 等人设计了图案重复率以衡量不同风格特征代表整个风格图像的程度。选择图案重复率最高的风格作为主要风格进行风格化，很好的避免了生成图像中不和谐的图案。

Zhu 等人[55]与 Hong 等人[54]几乎同时注意到风格图像中重复风格特征的问题，通过与 Hong[54]不同的方法实现了最主要风格特征的筛选。Zhu 等人的解决方案是一种新颖的全键注意力（All to Key，A2K）机制，它将每个查询与稳定的“关键”键相匹配。A2k 主要包含两个部分（图 14）。其一为分布式注意力机制。为了缓解全对全（All to ALL）注意力带来的问题，分布式注意力首先学习分布式关键点来描述风格特征的局部区域分布。然后，内容特征的每个查询与这些代表性关键点匹配。由于匹配的关键点是区域风格的表示而非孤立位置，分布式注意力可以提高匹配误差的容忍度。此外，由于这些关键点固定表示几个局部区域，分布式注意力的匹配效果对不同位置的查询具有较高稳定性。其二为渐进式注意力机制。与传统的全注意力（All to ALL）直接聚焦于特定位置不同，渐进式注意力机制首先关注粗粒度区域，然后逐步集中于细粒度位置。这种方法有助于在更大尺度上匹配风格模式，从而在粗粒度风格模式中找到更相似的语义。在此基础上，可以通过点对点注意力进一步定位粗粒度区域内的细粒度位置。此外，由于同一局部区域内的查询匹配到相同的关键点，因此转换后的特征也具有区域稳定性。

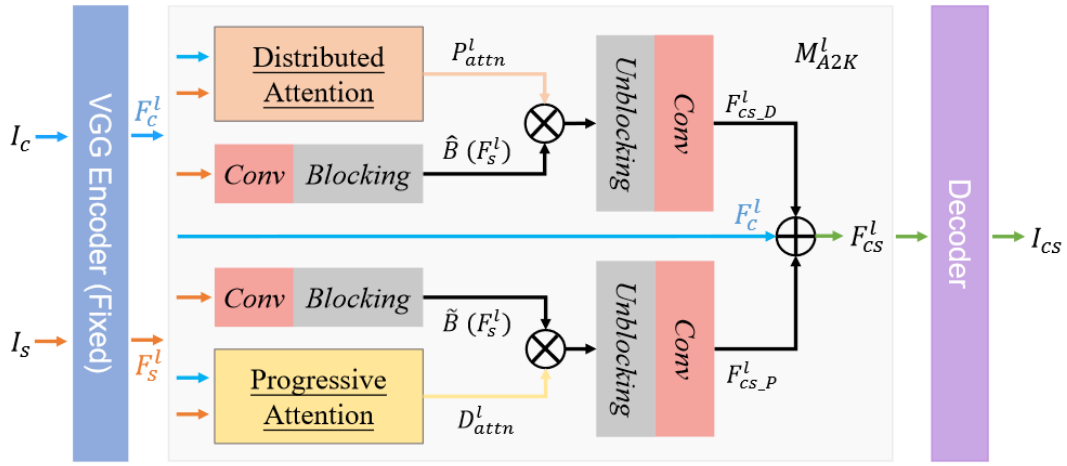

图 14 Zhu 等人模型总览[55]

通过上述两种新颖的注意力机制，Zhu 等人同样改善了由全对全注意力带来的问题，避免了生成图像中不和谐的图案，但过多探讨提高风格表现力的方面探索[55]，在表现力方面与前人工作较为相似。

3.3.4 基于预训练大模型的任意且实时风格迁移

预训练大模型是在大型数据集上预先训练的深度学习模型，具有广泛的知识和强大的学习能力。这些模型可以在特定任务上进行微调，以提高性能和效率。最近，有许多研究试图

利用大模型辅助风格迁移的工作。本节将着重介绍利用大模型的风格迁移成果。

利用 CLIP[56]大模型进行辅助风格迁移是当前的一个研究热点。Kwon 等人[57]认为，在许多实际情况中，用户可能没有参考风格图像，但仍然有兴趣体验风格迁移的成果。为了处理这样的应用，Kwon 提出了一个新的框架，旨在通过预训练的 CLIP 模型将目标文本的语义风格转移到内容图像上。在仅通过 CLIP 的监督来获得语义变换后的图像的过程中，Kwon 发现使用传统的像素优化方法时无法反映出所需的纹理。为了解决这个问题，Kwon 等人引入了一个 CNN 编码器-解码器模型，用于捕捉内容图像的分层视觉特征，并同时在深度特征空间中为图像添加风格，以获得真实的风格化结果。Kwon 等人方法的优势在于，仅需通过更改文本条件而无需任何样式图像即可逼真的样式迁移结果。但是由于使用了大模型，导致本文方法在进行推理时所需时间相较其他方法而言更长。

使用扩散模型进行风格迁移已成为当前研究中的一个常见选择，这主要归功于扩散模型强大的生成能力。其独特的结构使得扩散模型能够有效捕捉复杂的图像特征，从而生成高质量的风格转换结果。研究者们利用这些模型在各种风格之间进行灵活迁移，展现了扩散模型在艺术创作和图像处理领域的巨大潜力。这一趋势推动了更多关于如何优化和应用扩散模型的研究，进一步拓展了其在风格迁移中的应用范围。扩散模型是一种生成模型，最早由 Ho 等人[58]给出了详细的数学证明、推导与可运行的代码。扩散模型可以从噪声中生成目标数据样本。它包括两个过程：前向过程（forward process）和逆向过程（reverse process），其中前向过程又称为扩散过程（diffusion process）。前向过程是加噪的过程，前向过程中图像x_t只和上一时刻的 x_{t-1} 有关, 该过程可以视为马尔科夫过程, 满足:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})q(x_t|x_{t-1}) = N(x_t, \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \qquad 13$$

其中β_t是一组预先定义的超参数，并满足$\beta_1 < \beta_2 < \ldots < \beta_T$。逆向过程是去噪的过程，如果得到逆向过程$q(x_{t-1}|x_t)$，就可以通过随机噪声$x_T$逐步还原出一张图像。

Hamazaspyan 与 Navasardyan[59]将扩散模型与风格迁移任务相结合，提出了扩散增强的块匹配（Diffusion-Enhanced PatchMatch，DEPM）模型。该模型利用 Stable Diffusion 来捕获高级风格特征，同时保留原始图像的细粒度纹理细节。DEPM 允许在推理过程中转移任意样式，而无需任何微调或预训练，从而使过程更加灵活和高效。

Zhang 等人[60]试图从预训练的 Stable Diffusion[61]模型中获取先验知识以进行风格迁移。他们认为，基于小型模型的方法可以保留内容结构，但无法生成高度逼真的风格化图像并引入伪影和不和谐的纹理；预先训练的基于大规模模型的方法可以生成高度逼真的风格化图像，但很难保留内容结构。为了解决上述问题，Zhang 等人提出了 ArtBank ，可以生成高度逼真的风格化图像，同时保留内容图像的内容结构。在具体实现层面，为了充分挖掘预训练的大规模模型中的知识，他们设计了一个隐式风格提示库，该库由一组可训练的参数矩阵组成，用于从艺术品集合中学习和存储知识，并充当视觉提示指导预先训练的大型模型生成高度逼真的风格化图像，同时保留内容结构。在训练阶段，Zhang 等人额外提出了一种新的基于空间统计的自注意力模块用于加速隐式风格提示库的收敛。通过训练隐式风格提示库的方式，Zhang 等人很好的从 Stable Diffusion（Ver 1.4）中提取了相关知识，从而完成了风格迁移。Zhang 等人的方法开创了新的风格迁移思路，即如何快速有效的从预训练大模型中提取所需知识。

Zhang 等人[62]则在扩散模型的基础上提出了具有全新思想的风格迁移——将学习各艺术风格中的隐含文字标签作为风格迁移核心，其网络结构如图 15 所示。

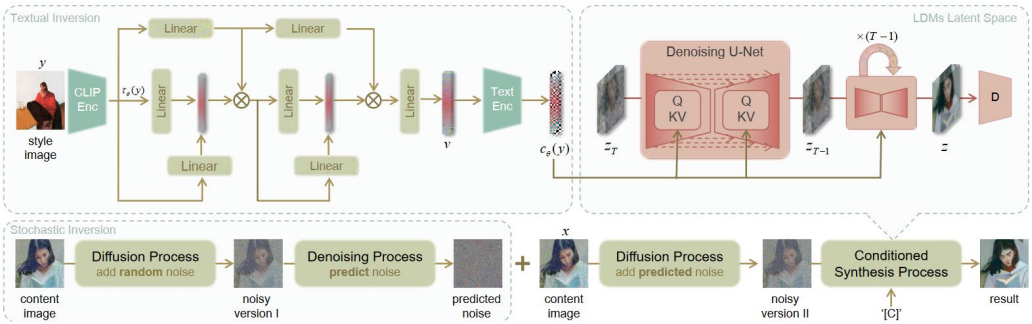

图 15 Zhang 等人模型总览[62]

该方法的主要思想是将艺术画中的风格看作一幅画的可学习的文本描述，并根据风格标签指导扩散模型生成图像。在实现方面，Zhang 等人利用所提出的反衍风格迁移方法（Inversion-Based Style Transfer Method，InST）高效准确的学习图像相关信息。具体来说，使用条件生成模型学习图像和文本之间的对应关系，从而获得图像嵌入；在图像嵌入的基础上，利用注意力导向的反转模块接收到图像嵌入，并利用注意力机制来生成对应的文本嵌入。该模块会关注图像嵌入中的不同特征，例如语义、材质、对象形状、笔触和颜色等，最终得到对应的文本嵌入，及文本标签。利用文本标签即可指导扩散模型进行风格迁移。该文本标签并不一定可以用自然语言描述出来，是一种只有扩散模型可以读懂的对风格进行描述的一串字符或者一个令牌（token）。该文本标签一经学习，Diffusion 即可固定该标签对应风格，在这个层面上，可以视本方法为基于模型迭代的风格迁移方法。本文最大的特色是可以改变风格化过程中图像的形状，这是以往风格迁移模型所不能的。

类似的，Namhyuk 等人[63]观察到，风格信息是难以用语言准确描述，故而考虑将风格图像编码到文本空间，从而为 Stable Duffion 提供文本约束。具体来说，Namhyuk 等人[63]等人将风格迁移与基于 Stable Diffusion 的文生图任务结合，提出了 DreamStyler 框架。该框架能够将图像中的风格信息提取到 CLIP 文本空间中。为了能将文本描述与 Stalbe Diffusion 结合，作者基于文本反转 (Textual Inversion, TI)提出了扩展文本嵌入空间的想法。该想法将扩散模型的时间步分为多个组，每组称为一个时间步块 (A Chunk of Timesteps)，并称一个时间步块与对应文本描述的组合为一个文本反转阶段 (TI Stage)。通过组合多个文本反转阶段的方式，Stable Diffusion 能够在图像合成的不同时间步块中理解相似但有区别的风格描述嵌入。这被作者称为多阶段文本反转 (Muilti-Stage Textual Inversion)。此外，Namhyuk 等人[63]还提出了一种上下文感知的提示增强，可以将风格图像中的将风格和上下文信息解耦合。在解耦合后，风格信息可以被编码为特殊的文本嵌入，提供更为准确的风格描述词供多阶段文本反转使用。

Wang 等人[64]以内容与风格解纠缠的思想为出发点，构建了一种名为 StyleDiffusion 的框架，实现了图像风格与内容的分离，如图 16 所示。该框架基于扩散模型，通过扩散过程分别移除图像中的风格信息和内容信息，并通过协调样式重建先验的基于 CLIP 的样式解缠损失，实现了内容和风格的完全解缠。框架由三个关键组成部分：基于扩散的风格去除模块、基于扩散的风格转移模块以及与风格重建先验协调的基于 CLIP 的风格解缠损失。实验证明，该框架能够生成高质量的风格转换结果，并且相对于其他方法，更好地考虑了内容和风格之间的关系。与之前的方法相比，本文方法通过扩散模型完全解耦了内容（C）和风格（S），从而更好地考虑了它们之间的关系。这样一来，风格转换结果更加自然和和谐（尤其对于具有挑战性的风格，如立体派和油画）。Wang 等人的方法通过扩散模型和基于 CLIP 的风格解耦损失，实现了对风格转换过程的精确控制。通过调整参数，可以灵活地控制风格去除的程度和内容与风格的解耦程度，从而得到更加理想的风格迁移效果。同时，该模式具有较高的

可解释性和可扩展性。通过引入扩散模型和基于 CLIP 的风格解耦损失，风格迁移过程更具可解释性。此外，该方法还可以应用于其他图像转换或操作任务，具有一定的扩展性。

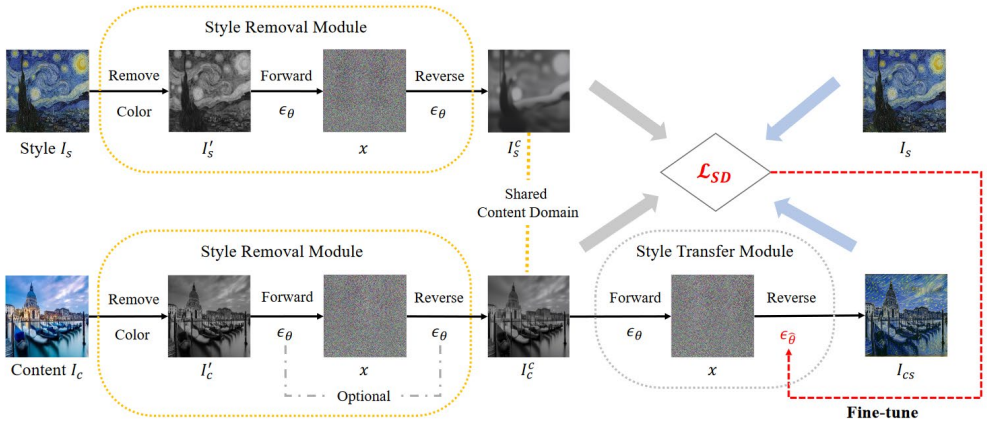

图 16 Wang 等人模型总览[64]

Lu 等人[65]试图利用扩散模型解决如何通过少量图像样本对预训练的扩散模型进行微调，以学习任何未见过的风格的问题。Lu 等人提出了一种名为"Specialist Diffusion"的方法，它可以通过对预训练的扩散模型进行微调来学习任何未见过的风格。通过仅使用少量图像（例如少于 10 张），便可以对预训练的扩散模型进行微调，使其能够以指定风格生成任意对象的高质量图像。为了实现这种极低样本微调，Lu 等人提出了一套新颖的微调技术，包括文本到图像的定制数据增强、内容损失以促进内容和风格的解耦，以及只关注少数时间步骤的稀疏更新。"Specialist Diffusion"方法可以与现有的扩散模型和其他个性化技术无缝集成，并在学习高度复杂的风格时，以超级高效的微调效果胜过最新的少样本个性化扩散模型。此外，"Specialist Diffusion"可以与反转方法相结合，进一步提高性能，甚至在非常不寻常的风格上也能取得成功。但同时，该方法存在一定的局限性。首先，虽然本方法可以通过少量图像进行微调来学习未见过的风格，但是对于一些高度特定和不寻常的风格，仍然可能存在学习不充分的情况。其次，本方法在样本效率上取得了很好的效果，但对于某些复杂的或接近训练数据分布的内容，生成的结果可能不尽如人意。最后，本方法的性能受到预训练模型的限制，如果预训练模型的质量不高，可能会影响到微调后的结果。

Chung 等人[66]试图解决在风格迁移任务中，使用扩散模型时所需的推理时间过长的问题。尽管已有类似的免训练方法，但先前的研究未能有效地将这些方法应用于大型扩散模型(如 Stable Diffusion)。通过回顾现有文献，Chung 等人发现使用大型扩散模型进行图像翻译的两个关键特征：第一，注意力图决定了生成图像的空间布局；第二，调整交叉注意力机制中的查询（Query）与键（Key）可以影响生成图像的内容。基于以上发现，Chung 等人提出了一种无需重新训练扩散模型即可实现风格迁移的策略。该策略的基本思想是使用风格图像的自注意力图中的查询与键来替代内容图像的自注意力图中的查询与键。在实现上述基本思想时, Chung 等人发现了两个主要问题：内容中断和颜色错误。针对内容中断，文章引入了“查询保存”策略，而对颜色错误则采用了初始潜在的自适应实例归一化（Intial Latent AdaIN）技术。通过这三者的结合,该方法实现了免训练的基于大型扩散模型的风格迁移方法。这一方法的优势在于，如果内容图像的查询具有相似语义，它们会使用相似的键，从而在风格迁移后保持内容图像查询之间的关系。此外，迁移结果中的每个内容图像的查询与具有相似纹理和语义的键之间的高度相似性，也使得风格迁移的效果更加自然和和谐。与 Namhyuk 等人[63]将风格信息编码到文本空间不同，Deng 等人[67]则认为使用编码器将风格图像编码为文本特征以指导 Stable Diffusion 模型生成风格图像的方法具有较大的损耗，因为这种基于文本的特征描述并不准确，且生成结果往往无法很好地捕捉图像的细节风格特征。他们指出，

基础的扩散模型在无文本约束情况下即可直接用于提取风格信息，因为该类模型中常用的 U-Net 结构具有该功能，从而达成避免依赖文本嵌入的目的。基于以上想法，Deng 等人提出了一种基于 Stable Diffusion 的双路径去噪模型来实现风格迁移。该模型包括两个独立且相同的扩散模型，分别用于处理内容和风格图像，均采用 U-Net 作为核心网络。为了便于描述，我们将这两个扩散模型称为风格扩散模型和内容扩散模型。风格扩散模型旨在通过 T 步扩散过程，从噪声图像逐步还原出原始风格图像；同样，内容扩散模型则试图还原内容图像。当扩散模型运行到任一时间步 $t \in [0, T]$ 时，U-Net 从内容扩散路径中提取的内容特征图 X^c_t 与从风格扩散路径中提取的风格特征图 X^s_t 结合，通过一种基于交叉注意力的机制生成风格化的潜在特征图 $\hat f_c$。然后，这些潜在特征图通过逆扩散过程生成最终的风格化图像。与传统的基于 Gram 矩阵计算风格特征的方法相比，该方法能够更好地保留内容图像的结构和细节。

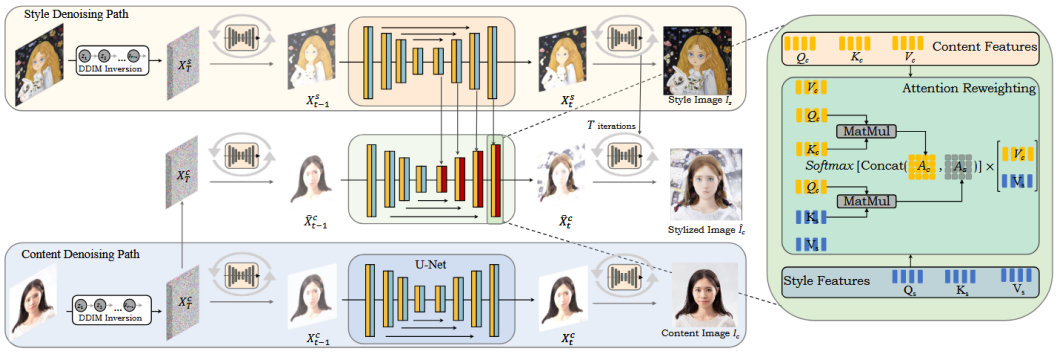

图 17 Deng 等人模型总览[64]

这一融合机制被称为“交叉注意力重构”，核心思想是将内容图像中的不同像素点视为查询(Query)，与风格图像中的特征（Key）进行关联。由于内容像素与风格信息的相关度可能不同，这种差异会影响风格迁移的效果。部分内容像素可能对风格信息的贡献较大，导致这些区域在风格化结果中表现得不够自然，甚至影响内容的保真度。因此，作者提出对这些低相关度的内容像素进行加权抑制，以减少它们在风格化潜在特征图$\hat f_c$中的影响。

然而，由于 Softmax 函数的特性，低相关度的像素点（即QK^T值较小的点）在经过 Softmax 处理后，可能会获得相对较大的注意力权重，导致不理想的风格迁移效果。为了应对这一问题，本文提出了一种重加权的交叉注意力机制，通过在 Softmax 函数中引入一个可调整的权重参数λ，动态调节不同像素的注意力权重。这样既能避免低相关度像素的影响，又能有效增强与风格图像相关性较强的像素的表现。

基于以上方法，Deng 等人实现了一种无需文本嵌入、基于扩散模型的全新风格迁移方法，该方法的风格化结果中既保留了内容细节，又能精准地呈现风格特征。

3.3.5 基于自搭建网络的任意且实时风格迁移

基于图像频域特征 频域处理在计算机图像领域具有重要地位，它通过将图像从空间域转换到频域，从而分析和修改图像的频率特性，以实现多种关键任务，如滤波、压缩和特征提取。频域处理可以有效地降低噪声、增强图像质量、检测边缘和纹理等细节，同时也为图像压缩和识别等应用提供了强大的工具，为图像处理领域的研究和应用提供了丰富的技术手段。

Li 等人[3]的方法与以往从空域考虑的大部分方法[4,40,50,55,68,69]不同，该方法从频域角度考虑图像的内容特征与风格特征。Li 等人认为，内容与风格的有效解耦是合成任意风格图像的关键因素，并表示现有的方法侧重于在空间域中解开内容和风格的特征表示，但其中内容和风格特征本质上是纠缠在一起的，这种纠缠导致了以往方法的局部扭曲、较弱的泛化能力

和不灵活等问题。Li 等人的方法基于一个发现；图像或特征图可以转换到频域，其中低频部分描述平滑变化的部分，高频部分则与快速变化的部分[70]相关，这种特性被他们称为频率可分特性（Frequency Separable Property，FSP）。

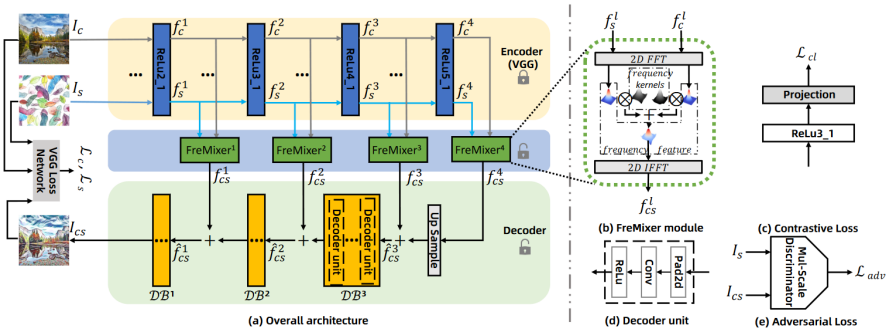

图 18 Li 等人模型总览[3]

基于频率可分特性，Li 等人提出了混频器（FreMixer）模块（图 18），该模块可以在频域中解开和重新缠结内容和风格组件的频谱。由于内容和风格组件具有不同的频域特征（频带和频率模式），因此 FreMixer 可以很好地分解这两个组件。该方法的步骤如下：1）首先，通过沿空间维度执行二维快速傅里叶变换（2-D FFT），将空间特征映射转换为频率域；2）接着，引入了经过网络学习得到的频率核，其作用类似于全局深度卷积层，用于解开频谱图中的内容和风格频率部分；3）在解开内容和风格频率模式后，通过逐元素加法操作将这两个频率谱重新组合；4）最后，将频谱图通过二维逆傅里叶变换（2-D IFFT）转换为空间域中的风格化特征；通过以上步骤，即可将图像的内容和风格分离到频率域中，然后再将它们重新组合到空间域中，从而实现图像的风格迁移。Li 等人的工作[3]从频域角度出发，在频域中实现风格特征与内容特征的解耦合，为神经风格迁移指出了一条新的方向。

Kwon[71]等人与 Li 等人[3]考虑问题的角度不同，但得出了类似的结论。Kwon 认为，当前的风格迁移方法难以有效的迁移具有美感的艺术信息，同时由于使用预训练模型，导致计算成本高且特征解纠缠效果差。为了解决这个问题，这项工作提出了一个轻量级但有效的模型，被 Kwon 等人称做美学特征感知模型（Aesthetic Feature-Aware，AesFA）。与 Li 等人[3]类似，Kwon 等人主要思想也是通过图像的频率来分解图像，以更好地将美学风格与参考图像分开。 在训练时，同时端到端的方式训练整个模型，以在推理时完全排除预训练的模型。同时，为了提高网络提取更独特的表示的能力并进一步提高风格化质量，Kwon 等人在文章中引入了一个新的损失函数：美学特征对比损失。该方法能够生成具有较优秀特征的风格化图像，同时 AesFA 在迁移高分辨率图像时花费时间也较少，具有良好的风格迁移效果。

构建新型特征提取器 Wang 等人[72]从风格迁移算法效率入手，认为现有的任意风格迁移方法难以甚至无法对超高分辨率（如 4K）的图像进行迁移，这严重阻碍了它们的进一步应用。他们重新思考了整个风格迁移流程，认为导致迁移速度慢的原因在于以往的神经风格迁移方法广泛使用了 VGG[25]作为特征提取器，这是由于大型的预训练 DCNN 的全连接层需要大量的计算资源导致的。

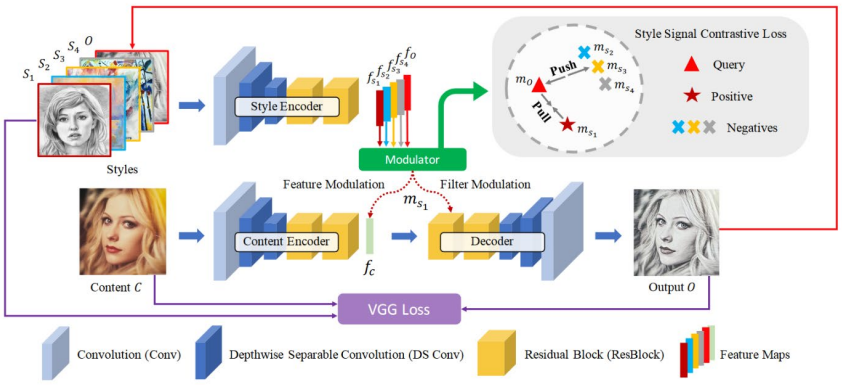

图 19 Zhang 等人模型总览[72]

为了能够实现在高分辨率图像风格迁移任务上也可快速完成工作的风格迁移模型，Wang 等人完全抛弃了以往使用 VGG[25]作为内容与风格特征提取器，而是自行搭建了全新的特征提取器，他们称整个网络模型为 MicroAST（图 19）。同时，Wang 等人还提出了新的损失函数“风格信号损失（Style Signal Contrastive Loss）”以辅助 MicroAST 进行风格迁移工作。MicroAST 包括两个部分——微型编码器与微型解码器。其中，微型编码器可分为微型内容编码器与微型风格编码器，它们具有相同的结构，仅包括 1 个标准 stride-1 卷积层、2 个 stride-2 深度可分离卷积（DS Conv）层和 2 个 stride-1 残差块（ResBlocks）。而微型解码器的结构与微型编码器几乎对称。通过舍弃 VGG[25]特征提取器，Wang 等人大幅降低了高分辨率风格迁移的所需时间与内存占用，并取得了不错的风格迁移结果。

值得注意的是，文章[71]也抛弃了传统的 VGG[25]特征提取器以获得更高的速度与更低的内存占用，但文章使用的其他方法才是核心，故未归纳为该类别。

4 评价指标

由于风格迁移领域发展较晚[2]，且难以客观评价风格化图像的生成质量，因此当前风格迁移领域具有众多评价指标，研究者在试图对模型风格迁移能力评价时往往难以找到被广泛接受的客观指标。因此本文收集了过去三年发表在 CVPR、ICCV、ECCV、AAAI 上风格迁移相关文献中使用的部分评价指标，供各位参考。

Content Loss 与 Style Loss 是第一篇关于神经风格迁移的成果[2]中使用的损失函数，现在有部分学者依旧考虑使用这两个指标作为图像生成的标准。其中，Content Loss 的计算方法如公式2所示，Style Loss 的计算方法如公式5所示。

时间与内存占用 在风格迁移中用于衡量算法的效率和实用性。时间占用反映了算法完成任务所需的时间长度，对于需要快速响应的应用场景尤为重要。内存占用则体现了算法在处理图像时对计算资源的需求，对于资源有限的设备，如移动设备或嵌入式系统，内存效率尤其关键。因此，这两个指标对于评估和选择适合特定需求的风格迁移算法具有重要意义。

Deception Rate 的概念首先由[73]等人于 2018 年提出，其具体表述如下：训练一个 VGG16 网络，使其能够区分 WikiArt 数据集上 624 个艺术家的不同艺术作品。在此基础上，Deception Rate 被定义为：

$$\text{Deception Rate} = \frac{F_{cs}^{t}}{F_{cs}} \tag{14}$$

其中，F_{cs}代表所有风格化图像，F_{cs}^{t}表示风格化图像中，被 VGG16 认为是艺术家创作的艺术作品的数量。在这样的定义下，Deception Rate 表示了风格化图像被认为是艺术家创作

出、而非生成的图像的比率

FID（Frechet Inception Distance）是一种评估生成模型，特别是在图像生成领域中输出质量的常用指标，该指标最先由 Heusel 等人于 2017 年的工作[74]中提出。该方法通过比较生成图像与真实图像在特征空间中的分布来进行评估。具体来说，FID 计算了在 Inception 网络某层输出的特征空间中，生成数据分布与真实数据分布之间的 Fréchet 距离（又称 Wasserstein-2 距离）。FID 的计算流程如下：

1. 特征提取：使用 Inception-v3 网络的某一层（通常是池化层之后的层）来提取生成图像和真实图像的特征。

2. 计算统计量：对于两组特征（生成图像特征集和真实图像特征集），分别计算它们的均值和协方差矩阵。

3. 计算 Fréchet 距离：利用以下公式12计算两个多维高斯分布之间的 Fréchet 距离：

$$\mathrm{FID}(x,g) = \| \mu_x - \mu_g \|^2 + Tr\left((\Sigma_x + \Sigma_g - 2(\Sigma_x\Sigma_g)^{\frac{1}{2}})\right) \quad 15$$

其中， μ_x, Σ_x是真实图像特征的均值和协方差矩阵μ_g, Σ_g是生成图像特征的均值和协方差矩阵。

FID 值越低表示生成图像的质量越高，即生成图像的分布与真实图像的分布越接近。FID 的优点在于它不仅考虑了单个图像的质量，还考虑了生成图像集的整体多样性。与其他评估指标如 Inception Score 相比，FID 更能准确反映人类视觉感知的差异，因此在图像生成领域得到了广泛应用。然而，FID 评分也存在一定的局限性，比如对于不同的数据集和不同的模型架构，其敏感性可能会有所不同，且计算过程需要较大的计算资源。

LPIPS（Learned Perceptual Image Patch Similarity）主要用于测量两张图像之间的视觉相似度，最先由 Zhang 等人于 2018[75]提出。与传统的像素级别比较方法不同，LPIPS 更加注重于图像在感知层面的差异。它通过利用深度学习模型，尤其是预训练的神经网络，来模拟人类视觉系统对图像差异的感知。具体来说，LPIPS 的计算方法可以通过以下步骤描述：

1. 特征提取：对于两张图像 x 和 y，使用预训练的深度神经网络（如 AlexNet、VGG 等）提取它们在多个层级l的特征，表示为 $F^l(x)$和 $F^l(y)$。

2. 计算归一化特征差异：对于每一层的每个空间位置 i，计算 x 和 y 在该位置的特征差异，并对特征进行归一化。差异计算公式如下：

$$d_i^l = \frac{1}{N_l} \|_i^l (x) - F_i^l(y)|Vert_2^2 \quad 16$$

其中，N_l 是第 l 层的通道数，$||\cdot||_2$表示 L2 范数。

3. 应用学习到的权重和汇总：每一层的特征差异通过一个学习到的权重 w_l 进行加权，然后对所有层次和位置的加权差异进行求和，得到最终的 LPIPS 距离：

$$LPIPS(x,y) = \sum_l \sum_i w_l \cdot d_i^l \quad 17$$

其中，w_l 是通过训练过程学习到的第l层的权重。

LPIPS 分数越低，表示两张图像在感知层面上越相似。这个指标通过考虑网络多个层次的特征，并通过学习得到的权重来调整不同层次特征的重要性，从而更好地模拟人类视觉系统的感知特性。

SSIM（结构相似性指数，Structural Similarity Index Measure）是一种评估图像质量的指标，用于量化两张图像在视觉感知上的相似度。与传统基于像素的误差度量不同，SSIM 考虑了图像的结构信息，更好地模拟了人类的视觉感知系统。SSIM 的核心思想是基于这样的观察：人类视觉系统高度适应于从视觉场景中提取结构信息。因此，SSIM 考虑了亮度、对比度和结构三个比较维度来评估图像质量。具体来说，对于两张图像 x 和 y，SSIM 被定义为：

$$SSIM(x,y)=\frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)} \quad 18$$

其中μ_x,μ_y是图像x,y的平均亮度，σ_x^2,σ_y^2是图像x,y的方差，σ_{xy}是图像x和y的协方差，c_1,c_2是为了避免分母为零而加入的小常数，通常$c_1=(k_1L)^2,c_2=(k_2L)^2$，L是像素值的动态范围，$k_1$和$k_2$是常数，取值较小。

SSIM 的值范围在 0 到 1 之间，1 代表两张图像完全相同。SSIM 通过综合考虑亮度、对比度和结构这三个方面，使得其比单纯基于像素差异的度量（如 MSE）在评价图像质量时更加准确、更符合人类视觉感知特性。

PD（Pixel Distance）由 Wang 等人[76]首先提出。该评价指标通过计算像素空间和深度特征空间中样本对的平均距离来衡量多样性。具体来说，会计算生成图像对之间的平均像素距离。对于像素空间的距离，直接在 RGB 通道上计算两幅图像间的平均像素距离，其计算公式如下：

$$d_{pixel}(\tilde{x}_1,\tilde{x}_2)=\frac{||\tilde{x}_1-\tilde{x}_2||_1}{W\times H\times 255\times 3} \quad 19$$

其中$\tilde{x}_1$和$\tilde{x}_2$表示要计算像素距离的两幅图像，W和H是图像的宽度和高度。这种方法能够量化图像对在视觉上的差异程度。

用户调研 在评价风格迁移方法方面至关重要，因为它提供了对风格迁移结果的主观评价。技术指标虽然可以客观衡量一些方面，但艺术风格迁移的效果往往更依赖于人的视觉和感知。通过问卷调查，研究者可以收集用户对风格迁移结果的满意度、视觉吸引力和真实感等方面的反馈，从而更全面地评估和改进风格迁移方法。在 2019 年后的每篇神经风格迁移均进行了用户调研，以证明自身方法在生成风格图方面的优秀表现。

5 领域前沿与挑战

通过将一幅图像的艺术风格与另一幅图像的内容结合，风格迁移技术已经在图像编辑、电影制作和计算机艺术等领域产生了深远的影响（如[15,20,77]）。

然而，尽管已取得了令人印象深刻的进展，风格迁移领域依然面临一系列复杂而具有挑战性的问题。这些问题不仅涉及到图像质量和视觉真实性的改进，还包括算法的效率、通用性和可解释性等方面的挑战。本文将关注仍然需要解决的问题，以期为未来的研究提供有价值的指导和启发。

5.1 评价指标

当前风格领域中，在客观评价指标层面具有多样且不统一的标准。以 2023 年部分文章为例，Wu 等人[47]使用 Deception Rate、FID、LPIPS 作为评价指标；

Wang 等人[72]使用时间与内存占用、SSIM、Style Loss 等作为评价指标；Li 等人[78]使用 SSIM 、 LPIPS 、 Content Loss 、 Style Loss 作为评价指标； Cheng 等人[79]\cite{chengUserControllableArbitraryStyle2023}使用 Pixel Distance、LPIPS、Deception Rate 作为评价指标；这些多样且无法统一的评价指标使得研究者在评估模型质量时存在较大的困难，往往难以比较两个模型之间在不同方面的性能。

同时，由于风格迁移往往产生具有艺术感的图像，所以研究者试图从问卷调研中获得较为客观的评价。据本文统计，自 2019 年以来的几乎所有风格迁移领域工作，均设计了问卷

以获取对当前工作与以往工作之间的比较。

为了讨论该方法的有效性，Jing[1]等人调查了年龄与职业对审美能力的影响，他们将每幅风格化图像交由 8 位相同职业但年龄不同的受试者（男 4 人,女 4 人）进行评分。实验结果（插入图像）发现，即使给定相同的风格化结果，具有相同职业和年龄的不同观察者仍然具有截然不同的评分。

故而目前风格迁移领域并未有一个获得众人认可的评价指标。考虑到艺术图像的评价受到审美能力的影响，可能需要在艺术从业者的协助下指定令人信服的评价指标。

5.2 可解释性

目前的风格迁移任务往往是基于深度学习与神经网络；同时部分成果更像是通过“发现”而非构建一个可解释的过程进行风格迁移[23]。这使得风格迁移的过程不可控，从而导致了结果图像可能无法满足人员的预期。

5.3 变形问题

目前的风格迁移算法仅考虑了在纹理与颜色上将内容图像转换为风格图像。然而部分风格画是对现实世界的抽象与简化（如动画风格与抽象派风格等），因此仅仅将纹理进行迁移是不够的。在迁移时需要对目标风格进行一定的探究，并通过设计实现在风格转换时将图像形变纳入考量。

5.4 对纹理和颜色进行迁移

有时，人们希望保留原始图像的颜色，仅将风格图像的纹理迁移至原始的内容图像上，然而当前算法往往同时将纹理与颜色同时迁移至内容图像上。因此仅对图像或仅对纹理进行迁移也是当前需要解决的问题之一。

5.5 人机交互

目前风格迁移的发展多在考虑使用单一模型实现任意风格迁移，但实现任意风格迁移并不意味着可以利用该模型进行生产活动。在生产过程中，往往需要根据需求进行定制化的过程，所以对能够对生成过程加以干涉，从而实现根据目标生成对应风格的图像是一个很重要的问题。

6 总结

神经风格迁移自 2016 年以来\cite{gatysImageStyleTransfer2016}逐渐走向兴盛。从发展历史来看，风格迁移在迁移速度方面从慢速走向实时，在风格实现方面从任意迁移到特定风格迁移再到任意迁移，在结构方面由简单走向复杂。由于各研究者对于风格这一概念理解上的差异，导致具体实现方法层面存在较大差异，但当前改进风格迁移的思路主要有三类：改进风格迁移的质量、优化风格迁移的运行时间、增加风格迁移的交互性。三者相比，交互性是相对研究较少的。

同时，随着风格迁移的发展，也有其他领域的任务使用了风格迁移的思想。有研究者致力于在生活中应用风格迁移，如[5,15]；也有研究者将风格迁移应用到艺术设计领域，如[10–14]；

也有研究者将风格迁移的方法应用到其他研究领域，如对抗样本研究[8,9,17,18]、图像生成研究[19]、领域自适应[20]、字体生成[6]、字体识别[7]。

目前看来，风格迁移已经不再是一个自娱自乐的研究领域，而是一个能够为其他研究提供新视角的研究方向。虽然风格迁移领域已经取得了不小的进展，但是其问题依旧存在，这些问题的存在将促使研究者们继续深入探索，提出更为先进的模型和算法，以应对这些挑战并推动风格迁移技术的进一步发展。未来，风格迁移的研究方向可能会更加注重与其他技术的融合，如结合自监督学习和多模态学习等，以开拓更多创新的应用场景。

引用

[1] JING Y, YANG Y, FENG Z, et al. Neural Style Transfer: A Review[J/OL]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(11): 3365-3385. DOI:10.1109/TVCG.2019.2921336.

[2] GATYS L A, ECKER A S, BETHGE M. Image Style Transfer Using Convolutional Neural Networks[C/OL]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 2414-2423[2023-04-13]. http://ieeexplore.ieee.org/document/7780634/. DOI:10.1109/CVPR.2016.265.

[3] LI D, LUO H, WANG P, et al. Frequency Domain Disentanglement for Arbitrary Neural Style Transfer[J/OL]. AAAI Conference on Artificial Intelligence, 2023, 37(1): 1287-1295. DOI:10.1609/aaai.v37i1.25212.

[4] HUANG X, BELONGIE S. Arbitrary Style Transfer in Real-Time With Adaptive Instance Normalization[C/OL]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 1501-1510[2023-10-30]. https://openaccess.thecvf.com/content_iccv_2017/html/Huang_Arbitrary_Style_Transfer_ICCV_2017_paper.html.

[5] Zhanghan Ke, LIU Y, ZHU L, et al. Neural Preset for Color Style Transfer[C/OL]//Computer Vision and Pattern Recognition. 2023: 14173-14182[2023-12-05]. https://openaccess.thecvf.com/content/CVPR2023/html/Ke_Neural_Preset_for_Color_Style_Transfer_CVPR_2023_paper.html.

[6] FU B, HE J, WANG J, 等. Neural Transformation Fields for Arbitrary-Styled Font Generation[C/OL]//Computer Vision and Pattern Recognition. 2023: 22438-22447[2023-12-11]. https://ieeexplore.ieee.org/document/10204947/. DOI:10.1109/CVPR52729.2023.02149.

[7] TANG L, CAI Y, LIU J, 等. Few-Shot Font Generation by Learning Fine-Grained Local Styles[C/OL]//Computer Vision and Pattern Recognition. 2022: 7885-7894[2023-12-11]. https://ieeexplore.ieee.org/document/9878987/. DOI:10.1109/CVPR52688.2022.00774.

[8] LIU S, JIANG W, GAO C, 等. PSGAN++: Robust Detail-Preserving Makeup Transfer and Removal[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(11): 8538-8551. DOI:10.1109/TPAMI.2021.3083484.

[9] XU Y, YIN Y, JIANG L, et al. TransEditor: Transformer-Based Dual-Space GAN for Highly Controllable Facial Editing[C/OL]//Computer Vision and Pattern Recognition. 2022: 7683-7692[2023-12-08]. https://openaccess.thecvf.com/content/CVPR2022/html/Xu_TransEditor_Transformer-Based_Dual-Space_GAN_for_Highly_Controllable_Facial_Editing_CVPR_2022_paper.html.

[10] LIU B, ZHU Y, SONG K, et al. Self-Supervised Sketch-to-Image Synthesis[J/OL]. AAAI

Conference on Artificial Intelligence, 2021, 35(3): 2073-2081. DOI:10.1609/aaai.v35i3.16304.
[11] BAE K, KIM H I, KWON Y, et al. Unsupervised Bidirectional Style Transfer Network Using Local Feature Transform Module[C/OL]//Computer Vision and Pattern Recognition. 2023: 740-749[2023-12-05].
https://openaccess.thecvf.com/content/CVPR2023W/GCV/html/Bae_Unsupervised_Bidirectional_Style_Transfer_Network_Using_Local_Feature_Transform_Module_CVPRW_2023_paper.html.
[12] HÖLLEIN L, JOHNSON J, NIESSNER M. StyleMesh: Style Transfer for Indoor 3D Scene Reconstructions[C/OL]//Computer Vision and Pattern Recognition. 2022: 6198-6208[2023-12-05].
https://openaccess.thecvf.com/content/CVPR2022/html/Hollein_StyleMesh_Style_Transfer_for_Indoor_3D_Scene_Reconstructions_CVPR_2022_paper.html.
[13] YIN K, GAO J, SHUGRINA M, 等. 3DStyleNet: Creating 3D Shapes with Geometric and Texture Style Variations[C/OL]//International Conference on Computer Vision. 2021: 12436-12445[2023-12-11]. https://ieeexplore.ieee.org/document/9710137/. DOI:10.1109/ICCV48922.2021.01223.
[14] YANG J, GUO F, CHEN S, et al. Industrial Style Transfer With Large-Scale Geometric Warping and Content Preservation[C/OL]//Computer Vision and Pattern Recognition. 2022: 7834-7843[2023-12-05].
https://openaccess.thecvf.com/content/CVPR2022/html/Yang_Industrial_Style_Transfer_With_Large-Scale_Geometric_Warping_and_Content_Preservation_CVPR_2022_paper.html.
[15] GUNAWAN A, KIM S Y, SIM H, et al. Modernizing Old Photos Using Multiple References via Photorealistic Style Transfer[C/OL]//Computer Vision and Pattern Recognition. 2023: 12460-12469[2023-12-08].
https://openaccess.thecvf.com/content/CVPR2023/html/Gunawan_Modernizing_Old_Photos_Using_Multiple_References_via_Photorealistic_Style_Transfer_CVPR_2023_paper.html.
[16] MU F, WANG J, WU Y, 等. 3D Photo Stylization: Learning to Generate Stylized Novel Views from a Single Image[C/OL]//Computer Vision and Pattern Recognition. 2022: 16252-16261[2023-12-11]. https://ieeexplore.ieee.org/document/9878882/. DOI:10.1109/CVPR52688.2022.01579.
[17] NASEER M, KHAN S, HAYAT M, 等. Stylized Adversarial Defense[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(5): 6403-6414. DOI:10.1109/TPAMI.2022.3207917.
[18] CAO Y, XIAO X, SUN R, 等. StyleFool: Fooling Video Classification Systems via Style Transfer[C/OL]//2023 IEEE Symposium on Security and Privacy (SP). 2023: 1631-1648[2023-12-05]. https://ieeexplore.ieee.org/abstract/document/10179383. DOI:10.1109/SP46215.2023.10179383.
[19] KARRAS T, LAINE S, AILA T. A Style-Based Generator Architecture for Generative Adversarial Networks[C/OL]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4401-4410[2023-10-07].
https://openaccess.thecvf.com/content_CVPR_2019/html/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.html.
[20] GUAN M, XIANG S, LIU T, 等. CDTNET: Cross-Domain Transformer Based on Attributes for Person Re-Identification[C/OL]//International Conference on Multimedia and Expo Workshops. 2022: 1-6[2023-12-08]. https://ieeexplore.ieee.org/abstract/document/9859330. DOI:10.1109/ICMEW56448.2022.9859330.
[21] KYPRIANIDIS J E, COLLOMOSSE J, WANG T, 等. State of the "Art": A Taxonomy of

Artistic Stylization Techniques for Images and Video[J/OL]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(5): 866-885. DOI:10.1109/TVCG.2012.160.
[22] JOHNSON J, ALAHI A, FEI-FEI L. Perceptual Losses for Real-Time Style Transfer and Super-Resolution[A/OL]. arXiv, 2016[2023-04-10]. http://arxiv.org/abs/1603.08155. DOI:10.48550/arXiv.1603.08155.
[23] ULYANOV D, LEBEDEV V, VEDALDI A, et al. Texture Networks: Feed-forward Synthesis of Textures and Stylized Images[A/OL]. arXiv, 2016[2023-10-29]. http://arxiv.org/abs/1603.03417. DOI:10.48550/arXiv.1603.03417.
[24] MORDVINTSEV A, OLAH C, TYKA M. Inceptionism: Going Deeper into Neural Networks[Z/OL]. (2015)[2023-11-12]. https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.
[25] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition[A/OL]. arXiv, 2015[2024-01-29]. http://arxiv.org/abs/1409.1556. DOI:10.48550/arXiv.1409.1556.
[26] BERGER G, MEMISEVIC R. Incorporating long-range consistency in CNN-based texture generation[A/OL]. arXiv, 2016[2023-10-28]. http://arxiv.org/abs/1606.01286. DOI:10.48550/arXiv.1606.01286.
[27] RISSER E, WILMOT P, BARNES C. Stable and Controllable Neural Texture Synthesis and Style Transfer Using Histogram Losses[A/OL]. arXiv, 2017[2023-10-29]. http://arxiv.org/abs/1701.08893. DOI:10.48550/arXiv.1701.08893.
[28] LI Y, WANG N, LIU J, et al. Demystifying Neural Style Transfer[A/OL]. arXiv, 2017[2023-09-18]. http://arxiv.org/abs/1701.01036. DOI:10.48550/arXiv.1701.01036.
[29] LI S Z. Markov random field models in computer vision[C/OL]//EKLUNDH J O. Computer Vision — ECCV '94. Berlin, Heidelberg: Springer, 1994: 361-370. DOI:10.1007/BFb0028368.
[30] CROSS G R, JAIN A K. Markov Random Field Texture Models[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1983, PAMI-5(1): 25-39. DOI:10.1109/TPAMI.1983.4767341.
[31] CHELLAPPA R, CHATTERJEE S, BAGDAZIAN R. Texture synthesis and compression using Gaussian-Markov random field models[J/OL]. IEEE Transactions on Systems, Man, and Cybernetics, 1985, SMC-15(2): 298-303. DOI:10.1109/TSMC.1985.6313361.
[32] BENNETT J, KHOTANZAD A. Multispectral random field models for synthesis and analysis of color images[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(3): 327-332. DOI:10.1109/34.667889.
[33] LI C, WAND M. Combining Markov Random Fields and Convolutional Neural Networks for Image Synthesis[C/OL]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2479-2486[2023-10-07]. https://openaccess.thecvf.com/content_cvpr_2016/html/Li_Combining_Markov_Random_CVPR_2016_paper.html.
[34] RADFORD A, METZ L, CHINTALA S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks[A/OL]. arXiv, 2016[2023-10-29]. http://arxiv.org/abs/1511.06434. DOI:10.48550/arXiv.1511.06434.
[35] LI C, WAND M. Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks[C/OL]//LEIBE B, MATAS J, SEBE N, et al. Computer Vision – ECCV 2016. Cham: Springer International Publishing, 2016: 702-716. DOI:10.1007/978-3-319-46487-9_43.

[36] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, 等. Generative adversarial networks[J/OL]. Communications of the ACM, 2020, 63(11): 139-144. DOI:10.1145/3422622.
[37] ZHU J Y, PARK T, ISOLA P, et al. Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks[C/OL]//CVPR. 2017: 2223-2232[2023-10-29]. https://openaccess.thecvf.com/content_iccv_2017/html/Zhu_Unpaired_Image-To-Image_Translation_ICCV_2017_paper.html.
[38] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, 等. Generative Adversarial Nets[C/OL]//Advances in Neural Information Processing Systems: 卷 27. Curran Associates, Inc., 2014[2023-10-29]. https://proceedings.neurips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.
[39] DUMOULIN V, SHLENS J, KUDLUR M. A Learned Representation For Artistic Style[A/OL]. arXiv, 2017[2023-10-29]. http://arxiv.org/abs/1610.07629. DOI:10.48550/arXiv.1610.07629.
[40] CHEN D, YUAN L, LIAO J, et al. StyleBank: An Explicit Representation for Neural Image Style Transfer[C/OL]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1897-1906[2023-09-19]. https://openaccess.thecvf.com/content_cvpr_2017/html/Chen_StyleBank_An_Explicit_CVPR_2017_paper.html.
[41] JING Y, LIU X, DING Y, et al. Dynamic Instance Normalization for Arbitrary Style Transfer[J/OL]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(04): 4369-4376. DOI:10.1609/aaai.v34i04.5862.
[42] JIA X, DE BRABANDERE B, TUYTELAARS T, 等. Dynamic Filter Networks[C/OL]//Advances in Neural Information Processing Systems: 卷 29. Curran Associates, Inc., 2016[2024-01-29]. https://proceedings.neurips.cc/paper/2016/hash/8bf1211fd4b7b94528899de0a43b9fb3-Abstract.html.
[43] CHEN T Q, SCHMIDT M. Fast Patch-based Style Transfer of Arbitrary Style[A/OL]. arXiv, 2016[2023-10-30]. http://arxiv.org/abs/1612.04337. DOI:10.48550/arXiv.1612.04337.
[44] XU W, LONG C, WANG R, et al. DRB-GAN: A Dynamic ResBlock Generative Adversarial Network for Artistic Style Transfer[C/OL]//International Conference on Computer Vision. 2021: 6383-6392[2023-10-30]. https://openaccess.thecvf.com/content/ICCV2021/html/Xu_DRB-GAN_A_Dynamic_ResBlock_Generative_Adversarial_Network_for_Artistic_Style_ICCV_2021_paper.html.
[45] YANG S, JIANG L, LIU Z, et al. Pastiche Master: Exemplar-Based High-Resolution Portrait Style Transfer[C/OL]//Computer Vision and Pattern Recognition. 2022: 7693-7702[2023-12-05]. https://openaccess.thecvf.com/content/CVPR2022/html/Yang_Pastiche_Master_Exemplar-Based_High-Resolution_Portrait_Style_Transfer_CVPR_2022_paper.html.
[46] MEN Y, YAO Y, CUI M, et al. Unpaired Cartoon Image Synthesis via Gated Cycle Mapping[C/OL]//Computer Vision and Pattern Recognition. 2022: 3491-3500[2023-12-11]. https://ieeexplore.ieee.org/document/9879244/. DOI:10.1109/CVPR52688.2022.00349.
[47] Jingyu Wu, Lefan Hou, Zejian Li, et al. Preserving Structural Consistency in Arbitrary Artist and Artwork Style Transfer[J/OL]. AAAI Conference on Artificial Intelligence, 2023, 37(3): 2830-2838. DOI:10.1609/aaai.v37i3.25384.
[48] LIU S, LIN T, HE D, et al. AdaAttN: Revisit Attention Mechanism in Arbitrary Neural Style

Transfer[C/OL]//International Conference on Computer Vision. 2021: 6629-6638[2023-12-11]. https://ieeexplore.ieee.org/document/9710208/. DOI:10.1109/ICCV48922.2021.00658.
[49] DENG Y, TANG F, DONG W, et al. StyTr2: Image Style Transfer With Transformers[C/OL]//Computer Vision and Pattern Recognition. 2022: 11326-11336[2023-12-05]. https://openaccess.thecvf.com/content/CVPR2022/html/Deng_StyTr2_Image_Style_Transfer_With_Transformers_CVPR_2022_paper.html.
[50] LI Y, XIE X, FU H, et al. A Compact Transformer for Adaptive Style Transfer[C/OL]//2023 IEEE International Conference on Multimedia and Expo (ICME). 2023: 2687-2692[2023-12-08]. https://ieeexplore.ieee.org/abstract/document/10219607. DOI:10.1109/ICME55011.2023.00457.
[51] Zhanjie Zhang, Jiakai Sun, Guangyuan Li, et al. Rethink arbitrary style transfer with transformer and contrastive learning[J/OL]. Computer Vision and Image Understanding, 2024, 241: 103951. DOI:10.1016/j.cviu.2024.103951.
[52] WANG Q, REN Y, ZHANG X, et al. Interactive Image Style Transfer Guided by Graffiti[C/OL]//ACM International Conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2023: 6685-6694[2023-12-07]. https://dl.acm.org/doi/10.1145/3581783.3612203. DOI:10.1145/3581783.3612203.
[53] ZHANG C, DAI Z, CAO P, et al. Edge Enhanced Image Style Transfer via Transformers[C/OL]//Proceedings of the 2023 ACM International Conference on Multimedia Retrieval. New York, NY, USA: Association for Computing Machinery, 2023: 105-114[2023-12-04]. https://dl.acm.org/doi/10.1145/3591106.3592257. DOI:10.1145/3591106.3592257.
[54] HONG K, JEON S, LEE J, et al. AesPA-Net: Aesthetic Pattern-Aware Style Transfer Networks[C/OL]//International Conference on Computer Vision. 2023: 22758-22767[2023-12-05]. https://openaccess.thecvf.com/content/ICCV2023/html/Hong_AesPA-Net_Aesthetic_Pattern-Aware_Style_Transfer_Networks_ICCV_2023_paper.html.
[55] ZHU M, HE X, WANG N, et al. All-to-Key Attention for Arbitrary Style Transfer[C/OL]//Computer Vision and Pattern Recognition. 2023: 23109-23119[2023-12-05]. https://openaccess.thecvf.com/content/ICCV2023/html/Zhu_All-to-Key_Attention_for_Arbitrary_Style_Transfer_ICCV_2023_paper.html.
[56] RADFORD A, KIM J W, HALLACY C, et al. Learning Transferable Visual Models From Natural Language Supervision[C/OL]//Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021: 8748-8763[2023-09-26]. https://proceedings.mlr.press/v139/radford21a.html.
[57] KWON G, YE J C. CLIPstyler: Image Style Transfer With a Single Text Condition[C/OL]//Computer Vision and Pattern Recognition. 2022: 18062-18071[2023-09-21]. https://openaccess.thecvf.com/content/CVPR2022/html/Kwon_CLIPstyler_Image_Style_Transfer_With_a_Single_Text_Condition_CVPR_2022_paper.html.
[58] HO J, JAIN A, ABBEEL P. Denoising Diffusion Probabilistic Models[C/OL]//Advances in Neural Information Processing Systems: 卷 33. Curran Associates, Inc., 2020: 6840-6851[2023-10-20]. https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.
[59] HAMAZASPYAN M, NAVASARDYAN S. Diffusion-Enhanced PatchMatch: A Framework for Arbitrary Style Transfer With Diffusion Models[C/OL]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 797-805[2023-10-30]. https://openaccess.thecvf.com/content/CVPR2023W/GCV/html/Hamazaspyan_Diffusion-

Enhanced_PatchMatch_A_Framework_for_Arbitrary_Style_Transfer_With_Diffusion_CVPRW_2023_paper.html.
[60] ZHANG Z, ZHANG Q, LI G, 等. ArtBank: Artistic Style Transfer with Pre-trained Diffusion Model and Implicit Style Prompt Bank[A/OL]. arXiv, 2023[2024-01-23]. http://arxiv.org/abs/2312.06135. DOI:10.48550/arXiv.2312.06135.
[61] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-Resolution Image Synthesis With Latent Diffusion Models[C/OL]//Computer Vision and Pattern Recognition. 2022: 10684-10695[2023-10-20]. https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.
[62] ZHANG Y, HUANG N, TANG F, et al. Inversion-Based Style Transfer With Diffusion Models[C/OL]//Computer Vision and Pattern Recognition. 2023: 10146-10156[2023-11-02]. https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_Inversion-Based_Style_Transfer_With_Diffusion_Models_CVPR_2023_paper.html.
[63] AHN N, LEE J, LEE C, et al. DreamStyler: Paint by Style Inversion with Text-to-Image Diffusion Models[J/OL]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(2): 674-681. DOI:10.1609/aaai.v38i2.27824.
[64] Zhizhong Wang, Lei Zhao, Wei Xing. StyleDiffusion: Controllable Disentangled Style Transfer via Diffusion Models[C/OL]//Computer Vision and Pattern Recognition. 2023: 7677-7689[2023-12-05].
https://openaccess.thecvf.com/content/ICCV2023/html/Wang_StyleDiffusion_Controllable_Disentangled_Style_Transfer_via_Diffusion_Models_ICCV_2023_paper.html.
[65] LU H, TUNANYAN H, WANG K, et al. Specialist Diffusion: Plug-and-Play Sample-Efficient Fine-Tuning of Text-to-Image Diffusion Models To Learn Any Unseen Style[C/OL]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 14267-14276[2023-11-02].
https://openaccess.thecvf.com/content/CVPR2023/html/Lu_Specialist_Diffusion_Plug-and-Play_Sample-Efficient_Fine-Tuning_of_Text-to-Image_Diffusion_Models_To_CVPR_2023_paper.html.
[66] Jiwoo Chung, HYUN S, HEO J P. Style Injection in Diffusion: A Training-free Approach for Adapting Large-scale Diffusion Models for Style Transfer[C/OL]//Computer Vision and Pattern Recognition. 2024: 8795-8805[2024-09-04].
https://openaccess.thecvf.com/content/CVPR2024/html/Chung_Style_Injection_in_Diffusion_A_Training-free_Approach_for_Adapting_Large-scale_CVPR_2024_paper.html.
[67] DENG Y, HE X, TANG F, et al. Z*: Zero-shot Style Transfer via Attention Reweighting[C/OL]//Computer Vision and Pattern Recognition. 2024: 6934-6944[2024-09-19].
https://openaccess.thecvf.com/content/CVPR2024/html/Deng_Z_Zero-shot_Style_Transfer_via_Attention_Reweighting_CVPR_2024_paper.html.
[68] ZITNICK C L, VEDANTAM R, PARIKH D. Adopting Abstract Images for Semantic Scene Understanding[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(4): 627-638. DOI:10.1109/TPAMI.2014.2366143.
[69] ZHANG L, JI Y, LIN X, 等. Style Transfer for Anime Sketches with Enhanced Residual U-net and Auxiliary Classifier GAN[C/OL]//2017 4th IAPR Asian Conference on Pattern Recognition (ACPR). 2017: 506-511[2024-01-22]. https://ieeexplore.ieee.org/abstract/document/8575875. DOI:10.1109/ACPR.2017.61.

[70] CHEN Y, FAN H, XU B, 等. Drop an Octave: Reducing Spatial Redundancy in Convolutional Neural Networks With Octave Convolution[C/OL]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 3435-3444[2024-01-29]. https://openaccess.thecvf.com/content_ICCV_2019/html/Chen_Drop_an_Octave_Reducing_Spatial_Redundancy_in_Convolutional_Neural_Networks_ICCV_2019_paper.html.
[71] KWON J, KIM S, LIN Y, et al. AesFA: An Aesthetic Feature-Aware Arbitrary Neural Style Transfer[A/OL]. arXiv, 2023[2024-01-23]. http://arxiv.org/abs/2312.05928. DOI:10.48550/arXiv.2312.05928.
[72] Zhizhong Wang, Lei Zhao, Zhiwen Zuo, et al. MicroAST: Towards Super-fast Ultra-Resolution Arbitrary Style Transfer[J/OL]. AAAI Conference on Artificial Intelligence, 2023, 37(3): 2742-2750. DOI:10.1609/aaai.v37i3.25374.
[73] SANAKOYEU A, KOTOVENKO D, LANG S, 等. A Style-Aware Content Loss for Real-time HD Style Transfer[C/OL]//Computer Vision and Pattern Recognition. 2018: 698-714[2024-01-17]. https://openaccess.thecvf.com/content_ECCV_2018/html/Artsiom_Sanakoyeu_A_Style-aware_Content_ECCV_2018_paper.html.
[74] HEUSEL M, RAMSAUER H, UNTERTHINER T, 等. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium[C/OL]//Advances in Neural Information Processing Systems: 卷 30. Curran Associates, Inc., 2017[2024-01-17]. https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.
[75] ZHANG R, ISOLA P, EFROS A A, 等. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric[C/OL]//Computer Vision and Pattern Recognition. 2018: 586-595[2024-01-17]. https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.
[76] WANG Z, ZHAO L, CHEN H, 等. Diversified Arbitrary Style Transfer via Deep Feature Perturbation[C/OL]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 7789-7798[2024-01-17]. https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_Diversified_Arbitrary_Style_Transfer_via_Deep_Feature_Perturbation_CVPR_2020_paper.html.
[77] WANG Y, WANG Y, YU L, et al. DeepVecFont-v2: Exploiting Transformers To Synthesize Vector Fonts With Higher Quality[C/OL]//Computer Vision and Pattern Recognition. 2023: 18320-18328[2023-12-08]. https://openaccess.thecvf.com/content/CVPR2023/html/Wang_DeepVecFont-v2_Exploiting_Transformers_To_Synthesize_Vector_Fonts_With_Higher_Quality_CVPR_2023_paper.html.
[78] Tianwei Lin, Honglin Lin, Fu Li, et al. AdaCM: Adaptive ColorMLP for Real-Time Universal Photo-Realistic Style Transfer[J/OL]. AAAI Conference on Artificial Intelligence, 2023, 37(2): 1613-1621. DOI:10.1609/aaai.v37i2.25248.
[79] CHENG J, WU Y, JAISWAL A, et al. User-Controllable Arbitrary Style Transfer via Entropy Regularization[J/OL]. AAAI Conference on Artificial Intelligence, 2023, 37(1): 433-441. DOI:10.1609/aaai.v37i1.25117.