计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202220039
60(4): 930−946, 2023

基于自相似与对比学习的图像跨域转换算法

赵 磊[1] 张慧铭[1] 邢 卫[1] 林志洁[2] 林怀忠[1] 鲁东明[1] 潘 洵[3] 许端清[1]
[1](浙江大学计算机科学与技术学院 杭州 310027)
[2](浙江科技学院信息与电子工程学院 杭州 310023)
[3](浙江大学外语学院 杭州 310027)
(cszhl@zju.edu.cn)

Image Cross-Domain Translation Algorithm Based on Self-Similarity and Contrastive Learning

Zhao Lei[1], Zhang Huiming[1], Xing Wei[1], Lin Zhijie[2], Lin Huaizhong[1], Lu Dongming[1], Pan Xun[3], and Xu Duanqing[1]
[1](*School of Computer Science and Technology*, *Zhejiang University*, *Hangzhou* 310027)
[2](*School of Information and Electronic Engineering*, *Zhejiang University of Science and Technology*, *Hangzhou* 310023)
[3](*School of International Studies*, *Zhejiang University*, *Hangzhou* 310027)

Abstract Image cross-domain transformation, also known as image translation, is a technology that aims to transform the images of the source domain into the ones of the target domain. Specifically, the converted images have the style of the target domain images (contour, posture, etc.) while maintaining the structure of the source domain images (texture, color, etc.). Image cross-domain transformation technology is widely used in the field of vision, such as photo editing and video special effects production. In recent years, this technology has developed rapidly based on deep learning, especially the generation of adversarial networks, and achieved impressive results. However, there are still problems, including the collapse of color mode and the inability to maintain the content structures in the transformed images. To solve the above problems, we propose an image cross-domain transformation algorithm based on self-similarity and contrastive learning. The algorithm uses the pre-trained deep neural network model to extract the content and style features of the images and takes the perceptual loss and the loss based on self-similarity as the image content loss function. At the same time, a loose optimal transport loss and the moment matching loss are used as the image style loss function to train the proposed neural network, and the transformed images and the target domain images are marked as positive sample pairs, and the translated images and the source domain images are marked as negative samples for contrastive learning. The proposed algorithm is verified by experiments on four data sets. The results show that the proposed method maintains the content structure of the source domain images, reduces the mode collapse of color, and makes the style of the translated images more consistent with that of the

收稿日期: 2022-01-09; **修回日期**: 2022-06-07
基金项目: 国家重点研发计划项目(2020YFC1522704); 国家自然科学基金项目(62172365); 浙江省自然科学基金项目(LY21F02005, LY19F020049); 国家社科基金重大项目(19ZDA197); 浙江省文物保护科技项目(2019011); 浙江省尖兵计划项目(2022C01222); 石窟寺文物数字化保护国家文物局重点科研基地项目; 浙江大学教育部脑与脑机融合前沿科学中心项目(2021008)
This work was supported by the National Key Research and Development Program of China (2020YFC1522704), the National Natural Science Foundation of China (62172365), the Natural Science Foundation of Zhejiang Province (LY21F02005, LY19F020049), the Key Program of the National Social Science Foundation of China(19ZDA197), the Zhejiang Cultural Relics Protection Science and Technology Project (2019011), the Zhejiang Elite Program(2022C01222), the Key Scientific Research Base Project for Digital Conservation of Cave Temples of State Administration for Cultural Heritage, and the Project of MOE Frontier Science Center for Brain Science & Brain-Machine Integration (Zhejiang University) (2021008).
通信作者: 邢卫(wxing@zju.edu.cn)

guidance images.

Key words cross-domain image transformation; self-similarity; contrastive learning; color mode collapse; style transfer

摘 要 图像跨域转换，又称图像翻译，是一种旨在将源域的图像转换为目标域的图像的技术，具体来说是使生成图像在保持源域图像的结构（轮廓、姿态等）的同时具有目标域图像的风格（纹理、颜色等）. 图像跨域转换技术在视觉领域有着广泛的应用，如照片编辑和视频特效制作. 近年来，该技术在深度学习尤其是生成对抗网络的基础上得到了飞速发展，也取得了令人印象深刻的结果，但是迁移后的生成图像仍然存在颜色模式坍塌、内容结构无法保持等问题，针对这些问题，提出了一种基于自相似性与对比学习的图像跨域转换算法. 该算法利用预先训练的深度神经网络模型提取图像的内容特征和风格特征，将感知损失和基于自相似性的损失作为图像内容损失函数，同时使用一种宽松的最优传输损失和基于矩匹配计算的损失作为图像风格损失函数对提出的神经网络进行训练，并通过将生成图像和目标域图像标记为正样本对，将生成图像和源域标记为负样本进行对比学习. 在 4 个数据集上对提出的算法进行了实验验证，结果表明提出的算法在生成的结果图像上较好地保持了源域图像的内容结构，同时减少颜色的模式坍塌，且使生成的图像风格与引导图像的风格更加一致.

关键词 跨域图像转换；自相似；比对学习；颜色模式坍塌；风格迁移

中图法分类号 TP391

跨域图像转换的目标主要是学习能够将源域的图像映射到目标域对应图像的函数，在跨域图像转换的研究场景下，源域和目标域往往具有相同的内容结构和语义相关性，比如人脸图像中的年轻域和年老域、室外场景的白天域和夜晚域. 跨域转换后生成的图像保持输入的源域图像的内容结构，同时其图像风格应具有目标域特有的风格属性（如白天图像转换成夜晚图像，转换后的图像应该具有夜晚图像的风格）. 跨域图像转换方向得到了深度学习和计算机视觉领域研究人员的广泛关注，因为它已广泛应用于图像风格转换[1-2]、图像编辑[3-5]、图像超分辨率[6]和图像彩色化[7]. 早期的跨域图像转换算法使用成对样本以有监督的方式训练条件深度神经网络模型[8-10]或简单回归模型[11-12]. 这些算法在许多应用场景中都是不切实际的，因为它们需要成对的数据. 在没有成对样本可用的场景中，许多算法[13-19]以无监督的方式成功地使用潜变量编码（latent coder）和循环一致性约束来实现图像跨域转换. 文献 [1−19] 算法虽然取得了直观、逼真的转换结果，但只能生成与实际情况不符的单一转换结果. 当给定源域图像时，有许多相应的目标域图像满足跨域图像转换要求. 为了生成多样化的转换结果，最近研究者已经提出了许多算法，包括 MUNIT[20] 和 DRIT[21]. 这些算法通常设计不同的网络框架和损失约束来分离图像的内容和风格，并将图像内容（来自源域）和参考图像风格（来自目标域）结合起来形成不同的转换结果. 不同的图像转换任务在源域图像和对应的目标域图像之间具有广泛的形状和纹理变化. 图像转换任务（如 photo2vangogh 和 photo2portrait）的源域图像和对应的目标域图像之间形状变化较小，而转换任务（如 selfie2anime, apple2orange, cat2dog）的源域图像和对应的目标域图像之间形状变化较大.

尽管目前的多样化跨域图像转换算法在许多图像转换任务数据集上取得了令人印象深刻的结果，但它们很难同时考虑这 2 种类型的图像转换任务. 这些算法根据源域图像和对应的目标域图像之间形状和纹理的变化量不同而呈现出不同的转换性能.在本文中，并不试图提高多样化跨域图像转换算法的泛化性，以便它们能够更好地执行不同的跨域图像转换任务. 本文提出的算法适用于电影后期制作、图像风格编辑等特殊领域，其要求转换后的图像内容对转换前的图像内容保持高度一致性，这就要求在图像转换过程中，源域图像的内容结构（形状）应尽可能少地改变. 对于这种形状变化很小或没有变化的跨域图像转换任务，目前的多样化跨域转换算法还存在 2 个问题：

1）转换结果图像的内容结构与输入的源域图像的内容结构存在显著差异，无法满足对图像编辑前后结构保持要求严格的应用要求.

2）转换结果图像和参考图像（来自目标域）之间的风格差异导致颜色模式崩溃（仅学习一些显著的颜色模式），这意味着转换结果图像的颜色内容不够

丰富，没有将参考图像的色彩空间全部学习到.

为了解决这2个问题，受自注意力机制[22]的启发，提出了一种称为SSAL-GAN的新算法，其中SSAL代表自结构注意力损失.自结构注意力损失函数确保转换图像的内容与源域图像的内容高度一致.此外，还设计了一个基于统计的颜色损失函数，以提高转换图像的色彩丰富性.在多个数据集上的实验结果表明，SSAL-GAN算法能够保持源图像的内容结构，并成功地将参考图像的颜色空间映射到转换图像中.综上所述，提出的算法有3方面的贡献：

1）提出了一种自结构注意力损失函数来改进图像跨域转换中的内容结构保持.该损失函数可以充分利用局部结构之间的长程依赖性，保持微结构之间的相对关系，实现源域图像与相应转换结果图像内容结构的一致性.

2）提出了一种基于统计的颜色损失函数，主要统计参考图像的颜色信息.通过颜色损失函数约束，可以使转换结果图像的颜色分布与参考图像的颜色分布保持一致，从而显著缓解图像转换过程中的颜色模式崩溃问题.

3）提出的框架可以实现图像内容和风格的分离学习.与MUNIT[20]，DRIT[21]，DSMGAN[23]等现有最先进的算法相比，所提算法不需要循环一致性损失函数和其他复杂的损失函数约束.这些复杂的损失函数需要在训练过程中具有熟练的超参调优能力.

1 相关工作

1.1 有监督的图像跨域转换算法

对于有监督学习算法而言，它所使用的训练集中的图像是成对的，即对于每一个输入图像都有一个真实的输出图像与之对应，而有监督的图像跨域转换算法的目标即是学习将输入的图像转换为与其对应的真实图像的映射函数.Isola等人[24]于2017年提出条件对抗生成网络模型，又称为Pix2Pix，用以在成对图像数据集之间进行跨域转换，该算法在从地图标签生成真实图像、边缘图生成真实图等任务上都较为有效，成为了后续多种算法的基础模型.Pix2Pix[24]以源域的图像x作为条件输入，即生成器将源域图像x和噪声z作为输入而生成目标域图像y.该算法的另一个贡献是在生成器中使用了“U-Net”结构，将下采样时的特征通过“跳跃连接”与上采样时的特征拼接在一起，能够在一定程度上克服瓶颈的限制，以更多的特征信息保持生成图像的结构.此外在判别器中使用了“PatchGAN”，使判别器对$N\times N$大小的图像块(image patch)进行判别，在最后取图像块分类结果的平均值作为最终结果，从而使判别器关注到图像的局部信息以使生成图像更为清晰.

Pix2PixHD[25]在Pix2Pix[24]的基础上使用了多尺度生成器和判别器来实现高分辨率的图像生成.Pix2PixHD算法主要针对成对的语义图到真实图的转换，采用了一种由粗到细(coarse-to-fine)方式将生成器分为全局生成网络和局部增强网络，实现了低分辨率到高分辨率的递进式图像生成.由于高分辨率图像的生成往往需要判别器拥有较大感受域，这会导致网络容量增加进而产生过拟合与内存超限问题，为了解决这个问题，该算法使用3个具有相同网络结构但在不同图像尺度下工作的判别器，以引导生成器能够生成全局一致且局部清晰的图像.

在语义图生成真实图时使用传统的归一化层(normalization layer)会倾向于“抹除”语义信息，导致生成模型得到的是局部最优解.当输入的语义图存在大片同样标签值时，极端情况下当整张语义图为天空或草地标签，现有的归一化层将会使该层的输入数据转化为0.为了解决这个问题，SPADE(GauGAN)[26]使用语义图对归一化层的输入激活值进行建模，即提出空间自适应归一化层，由语义图经由2层卷积层学习到归一化参数，将学习到的归一化参数应用于原始的批归一化层，能够有效利用语义信息，提高生成图像质量.

BicycleGAN[27]是在Pix2Pix[24]的基础上提出的一种为了实现输出图像多样性的算法，其核心思路是加强隐变量z和生成图像之间的联系.BicycleGAN通过将cVAE-GAN[28]和cLR-GAN[29]结合起来实现隐变量和输出之间的转换，其中cVAE-GAN(条件变分自编码GAN)是一种通过VAE[30]学习输出的隐变量分布进而实现多模式输出的算法.cVAE-GAN利用KL散度对隐变量分布进行正则化，使其接近标准正态分布，从而在推理过程中可以对隐变量进行采样；而cLR-GAN是从随机采样的隐变量开始，由生成器产生一个输出，这个输出输入编码器应该得到输入的隐变量，从而实现隐变量的循环一致性.这种双向训练的方式能够避免将多个隐变量映射为同一个输出导致的模式坍塌问题，从而实现了跨域的多样性.

有监督的图像跨域转换算法目前已经能取得令人较为满意的效果，但这类算法必须使用成对的数据集，而构建这种数据集是极为耗时耗力，甚至是不可能的，因此无监督跨域转换算法是目前主要的研究方向.

1.2 无监督的图像跨域转换算法

CycleGAN[31] 提出了一种能够解决无配对图像数据集的图像跨域转换问题，它利用了一个假设：将一个输入图像经由源域到目标域的转换后再由目标域转回源域得到的输出应该和输入图像是一致的. CycleGAN 在无成对图像间的跨域转换中主要使用双生成器和双判别器来实现循环一致性. 生成器 G, F 分别是 X 到 Y 和 Y 到 X 的映射，2 个判别器 DX, DY 则对转换后的图像进行判别. CycleGAN 将 2 个 GAN 损失和循环一致性损失结合作为总的损失函数，实现了在“无监督”情况下的图像跨域转换，该算法能够广泛应用于无成对图像数据集的图像跨域转换任务，但当需要进行几何变化时，该算法表现较差.

UNIT[32] 从概率建模的角度分析图像跨域转换问题，最关键的挑战在于学习不同域图像的联合分布. 在无监督的情况下，2 个数据集包含了来自不同域的只有边缘分布的图像，目标是使用这些图像推测联合分布，但从边缘分布推测联合分布是一个高度欠定问题. 为了解决这个问题，UNIT 做了一个共享隐变量的空间假设，即假设一对来自不同域的图像可以被映射为共享隐变量空间中的同一个表示. UNIT 可广泛应用于多种无监督的图像跨域转换任务，能够获得较高质量的图像生成结果，但无监督图像跨域转换本质上是多模态的，UNIT 对其做了过于简单的假设，将其建模为确定的 1 对 1 映射，因此它无法从给定的源域图像中生成不同的输出.

Huang 等人[20] 在 UNIT[32] 的基础上提出了多模态无监督图像跨域转换算法 MUNIT[20]，该算法假设可以将图像表示分解为共享的内容空间和域特定的风格空间，在进行图像跨域转换时，可以将源域的内容隐变量与目标域的风格空间中采样出的随机风格重新组合. 虽然风格的先验分布是单模态的，但借助解码器的非线性可以实现多模态的输出，MUNIT 除了使用对抗生成网络（GAN）损失函数和像素级的循环一致性损失函数之外，还使用内容编码和风格编码的重构损失.

DRIT[21] 与 MUNIT[20] 有着类似的实现思想，在提取 2 个域所共有的内容空间时，DRIT 除了使用权重共享之外还使用了一个内容判别器，该内容判别器对 2 个域的编码器获得的内容特征进行判别（实现时将域 A 的内容编码判别为 0，域 B 的内容编码判别为 1），以使 2 个域获得的内容编码的分布尽可能地接近.

尽管 MUNIT 和 DRIT 都实现了一定程度上生成结果的多样性，但由于 GAN 本身会倾向于忽略噪声，导致多样性受损. 为了解决这个问题，MSGAN[33] 提出在损失函数上加入一个简单正则项，其主要思想是最大化输出图像之间的距离与对应的隐变量之间的距离的比值. 网络的输入噪声 z_i 采样于隐变量分布 Z，一般为高斯分布. 当网络发生模式坍塌时，说明输出图像的多样性相比真实数据降低了. MSGAN 的关键点在于计算输出图像之间的距离与对应噪声之间的距离的比值. 在发生模式坍塌的地方，图片的多样性较低，即图像的距离会变小，导致这一比值会很小，而 MSGAN 通过人为地加入一个正则项使这一项比值保持最大，能够有效提高生成图像的多样性. 这一正则项可以轻易地加入到现有的许多框架中以提高网络生成的多样性，如 Pix2Pix[24]，BicycleGAN[27]，DRIT[21] 等.

在图像跨域转换任务中生成图像会因为背景的干扰而产生一些伪影，为了能够在不改变背景的情况下将注意力集中到需要进行改变的对象上去，AGGAN[34] 使用注意力机制对跨域进行指导，具体是在 CycleGAN[31] 的基础上增加了一个注意力网络，负责产生前景的掩膜. 注意力网络生成的掩膜是 [0,1] 之间的连续值，这样可以使损失函数是连续可微的，从而进行训练，同时允许网络学习如何组合边缘，否则可能会使前景对象看起来“粘在”背景上. U-GAT-IT[35] 也考虑使用注意力机制，与 AGGAN[34] 不同，它通过一个额外的分类器对特征进行权重调整，该算法能够在需要较大几何变化的跨域任务中表现良好. 此外 U-GAT-IT 还提出一种 AdaLIN（adapt layer-instance normalization），对 LN（layer normalization）[36] 和 IN（instance normalization）[37] 进行结合来帮助其灵活地控制形状和纹理的变化.

CUT[38] 指出在图像跨域转换任务中使用循环一致性损失是一个过强的约束，它引入对比损失取代了循环一致性损失，实现了单边的图像跨域转换. CUT 将输入图像与生成图像对应区域的图像块作为正样本，而将输入图像的其他部分的图像块作为负样本，通过计算样本之间的互信息作为对比损失函数. CUT 只使用单张图像对，克服了对比损失需要较大显存的限制，且能够实现单边的图像跨域转换，有效节省了训练时长.

以上模型均只能解决一对一的问题，即一个域到另一个域的转换. 当有很多域需要互相转换时，对于每一对域转换，都需要重新训练一个模型去解决，即现有的图像跨域转换模型为了实现在 k 个不同的风格域上进行转换，需要构建 $k\times(k-1)$ 个生成器. StarGAN[39] 首先提出多域间的转换，它将域信息和图

像一起作为输入对网络进行训练，并在域标签中加入掩膜向量，可以对不同的训练集进行联合训练. StarGAN 提出的多域转换模型与之前的 2 域模型进行多域转换的模型对比. 为了实现多域间的图像跨域转换，StarGAN 加入一个域的控制信息，类似 Pix2Pix 的形式，判别器不仅需要学习判别样本是否真实，还需要判断该样本来自哪个域. StarGAN 的创新之处在于提出了一个域分类损失，即对于给定的输入图像 x 和目标域标签 c，网络的目标是将 x 转换成输出图像 y，而 y 能够被分类为目标域标签 c，为了实现这一点就需要判别器有判别域的功能，所以 StarGAN 在判别器的顶端额外加入了一个分类器，域分类损失函数在优化生成器和判别器时都会用到它. 同时为了联合训练多个数据集，StarGAN 加入一个掩膜向量，在训练时将该向量也输入到生成器，即如果图像来源于数据集 S，则数据集 T 中的标记全部设为 0. StarGAN 使用单组 GAN 实现了多域间的图像跨域转换和多域数据集联合训练，极大地提升了图像跨域转换算法的泛化性. StarGANv2[40] 对域的概念进行了延伸，认为每个域中包含多种风格，提出使用风格标签代替域标签实现生成图像的多样性.

尽管 StarGAN[39] 成功实现了多域间的跨域任务，但图像跨域转换算法仍然受限于训练时需要大量的数据集以及训练好的模型不能很好地适用在其他数据集上，即之前的图像跨域转换算法需要在训练网络时准备含有充足图像的数据集，同时在测试时只能使用与测试数据集同类的图像，而不能测试一些未出现过的图像. FUNIT[41] 提出一种旨在基于少量样本的数据集上进行的跨域转换，能够实现将源域中的图像转换为一些模型从未见过的目标域中的图像. FUNIT 的网络结构可以分为 3 个部分：信息提取器、生成器和判别器，其中信息提取器由源域图像的内容编码器和目标域图像的类编码器组成. 与之前的跨域转换算法不同，FUNIT 将 1 张内容图像和 1 组包含 K 张图片的目标域图像作为输入，从而生成相应的跨域图像. 判别器则设计为同时在多个类别上进行对抗训练，用于判别当前图像是源域中的真实图像还是生成的目标域图像. 当存在 S 个源域图像类别时，判别器将对应生成 S 个输出，实验表明，该算法比利用判别器在 S 个多分类任务上表现更好.

2 基于自相似性与对比学习的图像跨域转换算法

在本节中，基于自相似性与对比学习的图像跨域转换算法主要包括模型结构、损失函数.

2.1 模型结构

本文介绍的算法网络结构如图 1 所示. 网络整体可以分为判别器和生成器 2 个部分，其中判别器判断当前的输入是来自真实数据还是生成数据. 而生成器可以细分为 3 个网络模块：编码器模块、模块 *AdaIN* 和解码器模块. 生成器的输入是 1 张内容图像 x 和 1 张风格图像 y，生成器首先利用编码器(VGG 网络)提取 x、y 图像的特征，分别对应于内容特征 $\boldsymbol{F}_x$ 和风格特征 $\boldsymbol{F}_y$，$\boldsymbol{F}_x$、$\boldsymbol{F}_y$ 经由 *AdaIN* 模块进行融合，其后通过解码器将融合后的特征解码至图像空间. 最终的输出可以表示为：

$$\tilde{y} = D(AdaIN(\Phi(x), \Phi(y))), \tag{1}$$

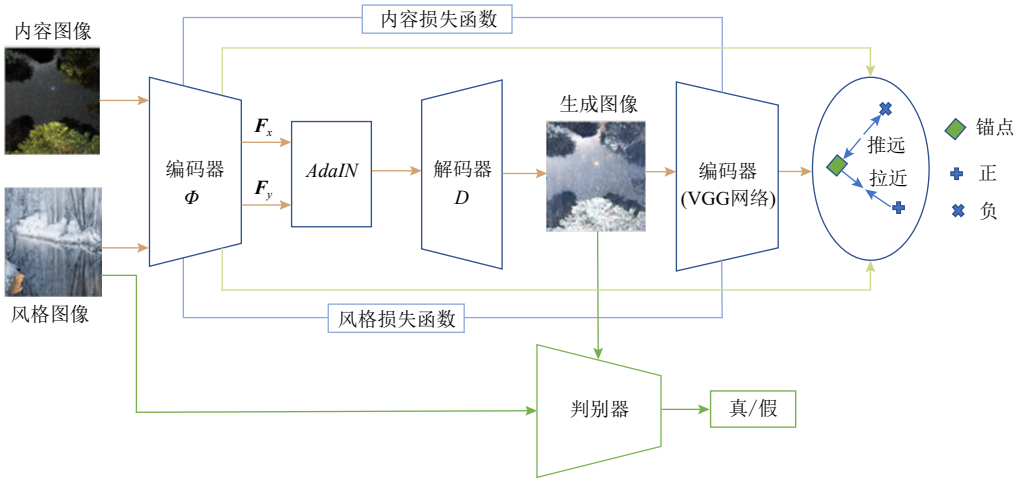


Fig. 1 Network model diagram

图 1 网络模型图

其中 Φ 表示 VGG 编码器，D 表示解码器.生成的图像经由编码器(VGG 网络)提取的特征作为锚点，而将风格特征作为它的正样本，内容特征作为它的负样本，从而使用对比学习损失[42]进一步提升生成效果.除生成对抗损失和对比损失之外，该算法使用感知损失和自相似性损失对内容进行约束，同时使用宽松的最优传输距离和基于统计量的损失对风格进行约束.下面对每个部分进行详细介绍.

2.1.1 编码器

本文算法使用预训练好的 VGG 网络作为编码器，具体是使用在目标检测和定位任务上预训练的 VGG-19 归一化模型，该模型的网络结构如表 1 所示.

Table 1 Normalized VGG-19 Network Structure

表 1 归一化的 VGG-19 网络结构

网络层	卷积核的尺寸	跨度	填充位数	输入图像尺寸	输出图像尺寸
Conv0_1	1×1	1	3	3×256×256	3×256×256
Conv1_1	3×3	1	1	3×256×256	64×256×256
ReLU1_1				64×256×256	64×256×256
Conv1_2	3×3	1	1	64×256×256	64×256×256
ReLU1_2				64×256×256	64×256×256
Pool1	2×2	2	0	64×256×256	64×128×128
Conv2_1	3×3	1	1	64×128×128	128×128×128
ReLU2_1				128×128×128	128×128×128
Conv2_2	3×3	1	1	128×128×128	128×128×128
ReLU2_2				128×128×128	128×128×128
Pool2	2×2	2	0	128×128×128	128×64×64
Conv3_1	3×3	1	1	128×64×64	256×64×64
ReLU3_1				256×64×64	256×64×64
Conv3_2	3×3	1	1	256×64×64	256×64×64
ReLU3_2				256×64×64	256×64×64
Conv3_3	3×3	1	1	256×64×64	256×64×64
ReLU3_3				256×64×64	256×64×64
Conv3_4	3×3	1	1	256×64×64	256×64×64
ReLU3_4				256×64×64	256×64×64
Pool3	2×2	2	0	256×64×64	256×32×32
Conv4_1	3×3	1	1	256×32×32	512×32×32
ReLU4_1				512×32×32	512×32×32
Conv4_2	3×3	1	1	512×32×32	512×32×32
ReLU4_2				512×32×32	512×32×32
Conv4_3	3×3	1	1	512×32×32	512×32×32
ReLU4_3				512×32×32	512×32×32
Conv4_4	3×3	1	1	512×32×32	512×32×32
ReLU4_4				512×32×32	512×32×32
Pool4	2×2	2	0	512×32×32	512×16×16
Conv5_1	3×3	1	1	512×16×16	512×16×16
ReLU5_1				512×16×16	512×16×16

本文算法使用 VGG-19 的 14 个卷积层和 5 个最大池化层，每层的卷积核大小都固定为 3×3，步长为 1，最大池化层的核大小为 2×2，步长为 2，即利用最大池化层进行下采样.分别提取内容图像和风格图像的 ReLUX_L(X=1, 2, 3, 4, 5, L=1, 2, 3, 4, 5)层的特征，根据 ReLUX_L 不同的特点，应用于后续不同的模块中.Gatys 等人[43]指出 VGG 提取的浅层特征包含图像较精确的像素信息，而深层特征包含图像的内容语义信息，所以本文算法中使用内容图像在 ReLU3_1, ReLU4_1, ReLU5_1 层的激活输出作为内容特征，而使用风格图像在 ReLU1_1, ReLU2_1, ReLU3_1, ReLU4_1, ReLU5_1 层的激活输出作为风格特征，用于后续的模块 *AdaIN* 进行操作以及计算各项损失函数.为方便下文说明，此处将提取的内容特征命名为 $\boldsymbol{X}_3, \boldsymbol{X}_4, \boldsymbol{X}_5$，同时将提取的风格特征命名为 $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Y}_3, \boldsymbol{Y}_4, \boldsymbol{Y}_5$.在将特征解码至图像空间之前，首先将提取的图像内容特征和图像风格特征使用 *AdaIN* 模块进行融合.为了能使生成图像能够更好地符合目标域图像的特点且更匹配引导(参考)图像的风格，将使用多层 *AdaIN* 融合获得融合后的特征 $\boldsymbol{F}_{y^{\sim}}$，如式(2)所示.

$$\boldsymbol{F}_{y^{\sim}} = UP(UP(AdaIN(\boldsymbol{X}_5, \boldsymbol{Y}_5)) + AdaIN(\boldsymbol{X}_4, \boldsymbol{Y}_4)) + AdaIN(\boldsymbol{X}_3, \boldsymbol{Y}_3), \tag{2}$$

其中 *UP* 表示使用最近邻算法进行 2 倍上采样.*AdaIN* 操作为：

$$AdaIN(\boldsymbol{X}_i, \boldsymbol{Y}_i) = \delta(\boldsymbol{Y}_i)\left(\frac{x-\mu(\boldsymbol{X}_i)}{\delta(\boldsymbol{X}_i)}\right) + \mu(\boldsymbol{Y}_i), \tag{3}$$

其中 $i \in \{3, 4, 5\}$，$\mu(\boldsymbol{X}_i)$和$\delta(\boldsymbol{X}_i)$分别是特征 $\boldsymbol{X}_i$的均值和方差，同样地有$\mu(\boldsymbol{Y}_i)$和 $\delta(\boldsymbol{Y}_i)$.将融合后的特征通过卷积层解码至图像空间，获得最后输出：

$$y^{\sim} = D(\boldsymbol{F}_{y^{\sim}}). \tag{4}$$

为了使特征之间更灵活地进行组合，本文算法在解码器中首先使用了 2 个残差模块，如图 2 所示.其中 $\oplus$ 表示矩阵相加.解码器中使用“最近邻上采样+卷积”对特征进行 2 倍上采样，取代转置卷积所带来的“棋盘格”效应[44]，以提升算法的生成效果.解码器的网络结构如表 2 所示.

2.1.2 判别器

算法使用 Pix2Pix[24]提出的 PatchGAN 作为判别器，它能够克服 L_1 损失或 L_2 损失对图像进行平滑从而丢失高频信息的缺点.PatchGAN 将图像分为 $N\times N$ 个图像块，分别判断图像块是生成图像还是真实图像，最后对这 $N\times N$ 个结果求平均作为最后的结果，能够鼓励生成器生成具有锐利边缘的高清图像.为了方便实现，本文算法并不直接切割图像以使其变为

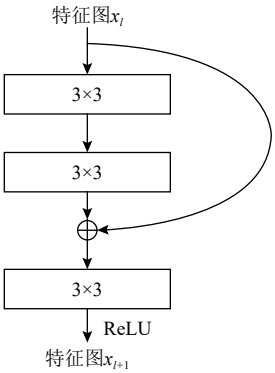


Fig. 2 Residual block model

图 2 残差块模型

$N\times N$ 个图像块, 而是使用 3 个完全相同的判别器, 对图像下采样从而使每个判别器的输入为不同分辨率的图像, 故在判别器相同的情况下, 图像分辨率越小, 判别器的感受域就相对越大, 从而实现在不同尺度的图像块上进行判别的效果. 其中第 1 个判别器的输入图像分辨率为 256×256, 第 2 个判别器为 128×128, 第 3 个判别器为 64×64. 篇幅所限, 表 3 只列出了最大尺度上的判别器的结构, 另外 2 个判别器在网络组成上与表 2 相同, 但每次的输入和输出尺寸分别需要除以 2 和除以 4.

2.2 损失函数

Table 2 Network Structure of Decoder

表 2 解码器的网络结构

网络块	网络层	卷积核的尺寸	跨度	填充位数	输入图像的尺寸	输出图像的尺寸
ResBlock	Conv1	3×3	1	1	512×32×32	512×32×32
	Conv2	3×3	1	1	512×32×32	512×32×32
ConvBlock	Conv1	3×3	1	1	512×32×32	256×32×32
	ReLU				512×32×32	256×32×32
UpBlock	Upsample	2×2			256×32×32	256×64×64
ResBlock	Conv1	3×3	1	1	256×64×64	256×64×64
	Conv2	3×3	1	1	256×64×64	256×64×64
ConvBlock	Conv1	3×3	1	1	256×64×64	128×64×64
	ReLU				256×64×64	128×64×64
UpBlock	Upsample	2×2			128×64×64	128×128×128
ConvBlock	Conv1	3×3	1	1	128×128×128	128×128×128
	ReLU				128×128×128	128×128×128
ConvBlock	Conv1	3×3	1	1	128×128×128	64×128×128
	ReLU				128×128×128	64×128×128
UpBlock	Upsample	2×2			64×128×128	64×256×256
ConvBlock		3×3	1	1	64×256×256	3×256×256

Table 3 Network Structure of Discriminator

表 3 判别器的网络结构

网络块	网络层	卷积核的大小	跨度	填充位数	输入图像尺寸	输出图像尺寸
ConvBlock1	Conv	4×4	2	1	3×256×256	64×128×128
	LeakyReLU				64×128×128	64×128×128
ConvBlock2	Conv	4×4	2	1	64×128×128	128×64×64
	LeakyReLU				128×64×64	128×64×64
ConvBlock3	Conv	4×4	2	1	128×64×64	256×32×32
	LeakyReLU				256×32×32	256×32×32
ConvBlock4	Conv	4×4	2	1	256×32×32	512×16×16
	LeakyReLU				512×16×16	512×16×16
ConvLayer	Conv	1×1	1	0	512×16×16	1×16×16

本文算法中使用了内容损失、风格损失、对比损失和对抗损失，下文将对每种损失函数进行介绍.

2.2.1 内容损失

由于生成图像需要保持源域图像的内容结构，本文算法首先使用 VGG-19 提取生成图像 $y^{\sim}$和源域图像 x 的 ReLU4_1 和 ReLU5_1 特征计算感知损失，即

$$l_{\mathrm{p}}=\sum_{i=4,5}\|norm(\boldsymbol{X}_i)-norm(\boldsymbol{Y}_i^{\sim})\|_1, \tag{5}$$

其中 $norm$ 表示将特征归一化为标准正态分布，如式(6)所示.

$$norm(x)=\frac{x-\mu(x)}{\delta(x)}, \tag{6}$$

$$\mu(x)=\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}x_{nchw}, \tag{7}$$

$$\delta(x)=\sqrt{\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(x_{\mathrm{nchw}}-\mu(x))^2+\varepsilon}. \tag{8}$$

同时，本文算法还使用基于自相似性[45]的内容损失函数. 该损失函数也可称为自结构注意力损失函数，人们通常通过图像中一个物体周围的外观来确定该物体，也即是说图像中物体的相对关系比它的绝对外观更重要. 相比于对像素(特征)绝对值之间的约束，自相似性通过对像素(特征)之间相关性的约束可以更好地保持空间结构和语义信息，也就是说自相似性在保持结构的同时允许生成图像的像素值与源域图像中的像素值发生极大变化，从而提升跨域的效果. 本文算法中使用余弦距离衡量特征之间的相似性，取生成图像的 ReLU3_1 和 ReLU4_1 层的特征，记为$\boldsymbol{Y}_3^{\sim}$和$\boldsymbol{Y}_4^{\sim}$，类似地取内容特征 $\boldsymbol{X}_3$ 和 $\boldsymbol{X}_4$，首先计算这些特征之间的余弦距离矩阵$\boldsymbol{D}_l$，其中 $l\in(3,4)$，如式(9)和式(10)所示.

$$\boldsymbol{D}_l^{Y_{ij}^{\sim}}=\boldsymbol{I}-\frac{\boldsymbol{Y}_3^{\sim}\cdot\boldsymbol{Y}_4^{\sim}}{\|\boldsymbol{Y}_3^{\sim}\|\times\|\boldsymbol{Y}_4^{\sim}\|}, \tag{9}$$

$$\boldsymbol{D}_l^{X_{ij}}=\boldsymbol{I}-\frac{\boldsymbol{X}_l^{i}\cdot\boldsymbol{X}_l^{j}}{\|\boldsymbol{X}_l^{i}\|\times\|\boldsymbol{X}_l^{j}\|}, \tag{10}$$

其中‖‖为取模，$\boldsymbol{I}$为单位矩阵. 然后对生成图像和内容图像之间的余弦距离矩阵计算L_1损失得到自相似性损失函数，即

$$L_{\mathrm{ss}}=\sum_{l\in(3,4)}\left(\frac{1}{n^2}\sum_{i,j}\left|\frac{\boldsymbol{D}_l^{Y_{ij}^{\sim}}}{\sum_i\boldsymbol{D}_l^{Y_{ij}^{\sim}}}-\frac{\boldsymbol{D}_l^{X_{ij}}}{\sum_i\boldsymbol{D}_l^{X_{ij}}}\right|_1\right), \tag{11}$$

其中n为第l层的特征数量. 式(11)表明在生成图像和源域图像从相同的位置提取出的特征的归一化余弦距离应该是一个常数.

2.2.2 风格损失

最优传输[46]用于描述分布变换的问题，即给定 2 个度量空间 X_{S}, Y_{S},以及它们对应的空间分布u, v，期望能找到一种传输变换 T: $X_{\mathrm{S}}\to Y_{\mathrm{S}}$ 将服从于u分布的随机变量转换为服从于v分布的随机变量，其数学表达式为

$$\min_T\int_{X_{\mathrm{S}}}\boldsymbol{C}(x,T(x))u(x)\mathrm{d}x, \tag{12}$$

$$\text{s.t.}\int_A v(y)\mathrm{d}y=\int_{T^{-1}(A)}u(x)\mathrm{d}x,\ \forall A\subset Y_{\mathrm{S}}, \tag{13}$$

其中$\boldsymbol{C}$为传输所需要进行的消耗矩阵，$T^{-1}(A)=\{x|x\in X_{\mathrm{S}},\ T(x)\in A\}$.

本文算法中将“风格”定义为图像的特征分布，则对内容图像进行跨域转换可以看作是将服从于内容图像的“风格”分布的变量转换为服从于风格图像的“风格”分布. 首先提取生成图像和风格图像在 ReLU1_1, ReLU2_1, ReLU3_1, ReLU4_1, ReLU5_1 层的输出特征，分别记作$\boldsymbol{F}_{Y^{\sim}}=(\boldsymbol{Y}_1^{\sim},\boldsymbol{Y}_2^{\sim},\cdots,\boldsymbol{Y}_n^{\sim})$, $\boldsymbol{F}_Y=(\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_n)$，则风格损失函数可以由$\boldsymbol{F}_{Y^{\sim}}$和$\boldsymbol{F}_Y$进行最优传输所耗费的损失进行度量，但由于原始的最优传输损失的计算要求 2 种分布的总质量相同，为了简化起见，假定所有特征的质量均是相同的，可得式(14).

$$L_{\mathrm{s}}(\boldsymbol{F}_{Y^{\sim}},\boldsymbol{F}_Y)=\min_{|T|\geqslant 0}\sum_{ij}T_{ij}C_{ij}, \tag{14}$$

$$\text{s.t.}\sum_j T_{ij}=\frac{1}{m}, \tag{15}$$

$$\text{s.t.}\sum_j T_{ij}=\frac{1}{n}, \tag{16}$$

其中 $\boldsymbol{T}$ 是 2 种分布进行传输的转换矩阵，C_{ij}是将$\boldsymbol{F}_{Y^{\sim}}$中每个特征转换为$\boldsymbol{F}_Y$中的特征所需要的消耗，$m$和$n$是特征的尺寸. 根据式(14)计算最优传输损失的时间复杂度为 $O(\max(m,n)^3)$，为了减少该损失函数的计算量，算法使用更为宽松的一种最优传输形式，即只需满足式(17)和式(18)中的一个约束.

$$L_{\mathrm{s}_1}=\min_{|T|\geqslant 0}\sum_{ij}T_{ij}C_{ij},\ \text{s.t.}\sum_j T_{ij}=\frac{1}{m}, \tag{17}$$

$$L_{\mathrm{s}_2}=\min_{|T|\geqslant 0}\sum_{ij}T_{ij}C_{ij},\ \text{s.t.}\sum_j T_{ij}=\frac{1}{n}, \tag{18}$$

则最优传输损失可以使用式(17)和式(18)的较大值，即

$$L_{\mathrm{s}}=\max(L_{\mathrm{s}_1},L_{\mathrm{s}_2})=\max\left(\frac{1}{n}\sum_i\min_j C_{ij},\frac{1}{m}\sum_j\min_j C_{ij}\right), \tag{19}$$

本文算法同样使用特征图之间的余弦距离计算最优传输中的消耗$\boldsymbol{C}$，即

$$C_{ij}=1-\frac{\boldsymbol{F}_{\boldsymbol{Y}^{\sim}}^{i}\cdot\boldsymbol{F}_{\boldsymbol{Y}^{\sim}}^{j}}{\|\boldsymbol{F}_{\boldsymbol{Y}^{\sim}}^{i}\|\times\|\boldsymbol{F}_{\boldsymbol{Y}^{\sim}}^{j}\|}. \tag{20}$$

尽管使用基于最优传输的风格损失能够实现较好的转换效果，但该损失忽略了特征值的大小以至于会产生一些伪影，为了解决这个问题，本文算法使用矩匹配约束，同时作为风格损失，具体使用了一阶矩和二阶矩的L_{m}损失，即

$$L_{\mathrm{m}}=\frac{1}{d}\|\mu\boldsymbol{F}_{\boldsymbol{Y}^{\sim}}-\mu\boldsymbol{F}_{\boldsymbol{Y}}\|_1+1/d^2\left\|\sum\boldsymbol{F}_{\boldsymbol{Y}^{\sim}}-\sum\boldsymbol{F}_{\boldsymbol{Y}^{\sim}}\right\|, \tag{21}$$

其中$\mu\boldsymbol{F}_{\boldsymbol{Y}^{\sim}}$，$\sum\boldsymbol{F}_{\boldsymbol{Y}^{\sim}}$表示生成图像特征的均值和协方差，$\mu\boldsymbol{F}_{\boldsymbol{Y}}$，$\sum\boldsymbol{F}_{\boldsymbol{Y}^{\sim}}$为目标域图像特征的均值和协方差.

2.2.3 对比损失

对比学习是一种应用于自监督学习领域的重要算法，其特点在于不需要利用额外的人工标注，而是直接利用数据本身作为监督信息学习数据的特征表示. 即对于任意数据x，对比学习期望能学习一个编码器f使式(22)成立.

$$score(f(x),\ f(x^{+}))\gg score(f(x),\ f(x^{-})), \tag{22}$$

其中x^{+}是x的正样本，二者应该是相似的，而x^{-}是x的负样本，二者应该是不相似的，$score$是用于度量样本之间相似度的函数. 本文算法将生成图像标记为锚点，将风格图像标记为正样本，将内容图像标记为负样本，使用对比学习损失[44]拉近锚点和正样本之间的距离，同时推远和负样本之间的距离.在实现中使用VGG在ReLU1_1, ReLU2_1, ReLU3_1层的激活特征以L_1距离计算对比损失L_{contra}：

$$L_{\mathrm{contra}}=\sum_{l=1}^{3}\psi_l\frac{|\boldsymbol{Y}_l^{\sim}-\boldsymbol{Y}_l|_1}{|\boldsymbol{Y}_l^{\sim}-\boldsymbol{X}_l|_1}, \tag{23}$$

其中ψ_i为每层特征之间对比损失的权重，是一个超参数.

2.2.4 对抗损失

传统的GAN训练不稳定，且易产生模式坍塌问题，LSGAN使用最小二乘损失函数能够有效避免此类情况，同时通过将距离策边界较远的生成图像拉向决策边界而提高图像的生成质量，因此本文算法使用LSGAN计算对抗损失，其中判别器的对抗损失函数如式(23)所示，生成器的对抗损失函数如式(24)和式(25)所示：

$$L_{\mathrm{d}}=\frac{1}{2}E_{y\sim P_{\mathrm{r}}}[D(y)-1]^2+\frac{1}{2}E_{y^{\sim}\sim P_{\mathrm{g}}}[D(y^{\sim})]^2, \tag{24}$$

$$L_{\mathrm{g}}=\frac{1}{2}E_{y^{\sim}\sim P_{\mathrm{g}}}[D(y^{\sim})-1]^2, \tag{25}$$

其中$y^{\sim}=G(x,y)$，G为生成器，D为判别器，P_{r}为真实数据分布，即目标域图像的分布，P_{g}为生成数据分布. 综上，生成器总的损失函数为式(26)，判别器损失函数为式(27)：

$$L_{\mathrm{gen}}=\lambda_1L_{\mathrm{p}}+\lambda_2L_{\mathrm{ss}}+\lambda_3L_{\mathrm{s}}+\lambda_4L_{\mathrm{m}}+\lambda_5L_{\mathrm{g}}+\lambda_6L_{\mathrm{contra}}, \tag{26}$$

$$L_{\mathrm{dis}}=L_{\mathrm{d}}. \tag{27}$$

3 实验结果与分析

3.1 数据集

本文算法针对跨域任务中无形变数据集进行研究，分别在summer2winter[47]，monet2photo[47]，night2day[48]，MWI (multi-class weather image)[49]进行. summer2winter[47]是拍摄于约塞米蒂国家公园不同季节的风景图像数据集，图像的分辨率为256×256，分为2个域A、B，分别对应于夏季和冬季，其中每个域中包含训练集和测试集. 域A的训练集包含1 232张图像，测试集包含310张图像；域B的训练集包含963张图像，测试集包含239张图像.

monet2photo[47]为莫奈绘画和自然图像的数据集，该数据集中的图像分辨率不固定，在跨域任务中，首先将其缩放并裁剪至256×256. 该数据集分为2个域A、B，分别对应于莫奈画作和自然图像，每个域中都包含训练集和测试集. 域A的训练集包含1 072张图像，测试集包含121张图像；域B的训练集包含6 287张图像，测试集包含751张图像.

night2day[48]为同一场景下夜间和白天的图像，图像的分辨率均为256×256. 该数据集是一个成对(paired)数据集，即对于每一张夜间的图像，都存在一张内容相同但光照不同的白天图像，但本文算法忽略该数据集的对应关系，即在该数据集上训练时不关心2张图像是否内容相同. 该数据集分为训练集、验证集和测试集，每个子集中的夜间图像和白天图像拼接在一起，为了将其更方便地应用于无监督算法中，将所有图像进行切割，使其包含夜间图像子集和白天图像子集，对应于A, B这2个域，同时合并原始的训练集和验证集使其成为新的训练集，最后得到的数据集为域A的训练集包含17 833张图像，测试集包含2 287张图像，域B的训练集包含17 833张图像，测试集包含2 287张图像. MWI[49]是一个包含cloudy, foggy, rain, snow, sunny这5种类别的数据集，MWI不区分训练集和测试集，分别包含45 662, 357, 1 359, 1 252, 70 501张图像，图像的分辨率不固定. 因为该数据集中cloudy和rain的特点难以分辨，因此本文算法中只使用sunny, foggy, snow，并将MWI的

90% 作为相应的训练集，10% 作为测试集. 各个数据集样例如图 3 所示.

Fig. 3 Example images of each data set

图 3 各数据集样例图

3.2 实验训练过程

实验分为训练阶段和测试阶段. 在训练阶段，首先将源域图像x和目标域图像y都缩放至 286×286；然后再随机裁剪至 256×256 以使模型更加鲁棒. 将x和y同时作为生成器的输入，可以获得生成图像$y^{\sim}$. 将$y^{\sim}$和$y^{\sim}$分别输入判别器，由判别器判断该输入是虚假数据还是真实数据，判别器和生成器交替进行训练. 实验中使用 Adam 优化器对生成器和判别器进行优化，学习率为 0.000 1，Momentum 参数分别为 0.5 和 0.999，由于显存限制，批尺寸设置为 2，损失函数中的超参数依次设置为 1.0，1.0，1.0，0.1，1.0，0.1.在测试阶段，只需要使用训练好的生成器模型，同样地将源域图像x和目标域图像y输入生成器，即可获得生成图像$y^{\sim}$.

3.3 实验结果

首先在 summer2winter 和 monet2photo 数据集上对算法进行了实验，如图 4 所示，其中每行包含对应于 2 个数据集的 1 组结果，每组结果从左至右依次为内容图像、风格图像和生成图像. 从生成图像的视觉效果上来看，本文算法在无形变的图像跨域任务中有较为优越的表现，主要包含 2 个方面：1）生成图像的内容结构保持完整，即生成图像相对于内容图像具有对应的物体和正确的相对位置，同时各个物体的锐度较高，在 monet2photo 中，莫奈的画作相比自然界的真实图像而言缺乏清晰的边缘信息和各类细节信息，而本文算法能够完成边缘和细节的生成，以至于人眼无法区别生成图像和真实图像. 2）生成图像的风格与风格图像一致. 这种一致性除了指整体色调相同之外，还包括其他少量但丰富的色彩信息，而且这些信息不是简单地杂糅，而是带有语义的进行转换，如图中 monet2photo 数据集上的第 1 行，生成图像的天空为蓝色且有少量白云，而草地为鲜绿色，树木为深绿色，帆船为白色，与风格图像呈现出较强的语义相关性.

Fig. 4 Experimental results of our proposed algorithm on summer2winter and monet2photo

图 4 本文算法在 summer2winter 和 monet2photo 上的实验结果

为了评估本文算法的泛化能力，使用夏天到冬天数据集上训练的模型在 night2day 和 MWI 上直接进行测试，生成结果如图 5 所示，可以看出本文算法在未见过的风景数据集上有较好的泛化能力，但生成

效果有所降低，一方面是因为模型在 summer2winter 数据集上训练，另一方面则由数据集本身的特点所决定：night2day 数据集中现代建筑物拥有较复杂的结构，但 night 域图像的内容结构不清晰，故转换后的图像结构较模糊；在 MWI 数据集上展示了 sunny→foggy 和 sunny→snow 的结果，由于这 2 个子域与 summer2winter 数据集在一定程度上相似，故生成效果较好. 如果在这 2 个数据集上同时训练，其效果比单纯在单一数据集上的效果略差一些，但是比在一个数据集上训练而在另外一个数据集上测试的效果要好一些.

综上可知，该算法在源域图像具有清晰的内容结构和在目标域具有丰富的色彩信息的情况下表现最好，它能够如实地维持源域图像的内容且将其转换为与目标域图像一致的风格，且该风格与样例引导图具有较强的语义联系.

(a) night2day (b) MWI

Fig. 5 Experimental results of our proposed algorithm on night2day and MWI

图 5 本文算法在 night2day 和 MWI 上的实验结果

3.4 对比实验

为了更全面和准确地评估本文算法的跨域生成效果，首先选取了近年来在多样化的图像跨域任务中表现较为优秀的算法包括 DRIT, MUIT, MSGAN, 在 2 个数据集 summer2winter 和 monet2photo 上进行比较. 在对比实验中，首先展示了在样例图引导的情况下的生成结果，如图 6 和图 7 所示. 可以看出，本文算法相对于基线(baseline)算法能够捕捉到风格图像

(a) 内容图像 (b) 风格图像 (c) 本文算法 (d) DRIT算法 (e) MUNIT算法 (f) MSGAN算法

Fig. 6 Comparison of experimental results of different algorithms on summer2winter

图 6 不同算法在 summer2winter 上的实验结果对比

(a) 内容图像 (b) 风格图像 (c) 本文算法 (d) DRIT算法 (e) MUNIT算法 (f) MSGAN算法

Fig. 7 Comparison of experimental results of different algorithms on monet2photo

图 7 不同算法在 monet2photo 上的实验结果对比

更多的颜色模式，即 DRIT, MUNIT, MSGAN 均只能学到风格图像整体的风格，而忽略了其余少量却丰富的风格模式，且其中 MUNIT 的生成结果存在“水洗”现象，即图像的锐度较低，DRIT 和 MSGAN 则存在更多的“伪影”(artifacts).

为了进一步评估本文算法对于多样性的提升，展示了本文算法与 baseline 在单张图像的不同风格结果，如图 8 和图 9 所示，分别对应于 summer2winter 数据集和 monet2photo 数据集. 可知，在多张拥有同样内容的生成图像中，本文算法的多样性最明显，且相对于 baseline 算法更加自然和真实，比如在图 8(b) 的倒数第 2 列，4 个算法均生成了更加符合秋季特征.

此外，计算了各个算法在这 2 个数据集上的量化指标，包括弗雷谢开端距离(FID)，感知图像块相似性(LPIPS)，不同容器的数量(NDB)，简森·香农散度距离(JSD)，…如表 4 和表 5 所示，表中结果为 10 次计算的均值和方差. 量化结果与各个算法在视觉上的表现保持了一致，证明了本文算法在无形变的图像跨域转换中在不损失图像质量的前提下提升了生成结果的多样性.

3.5 消融实验

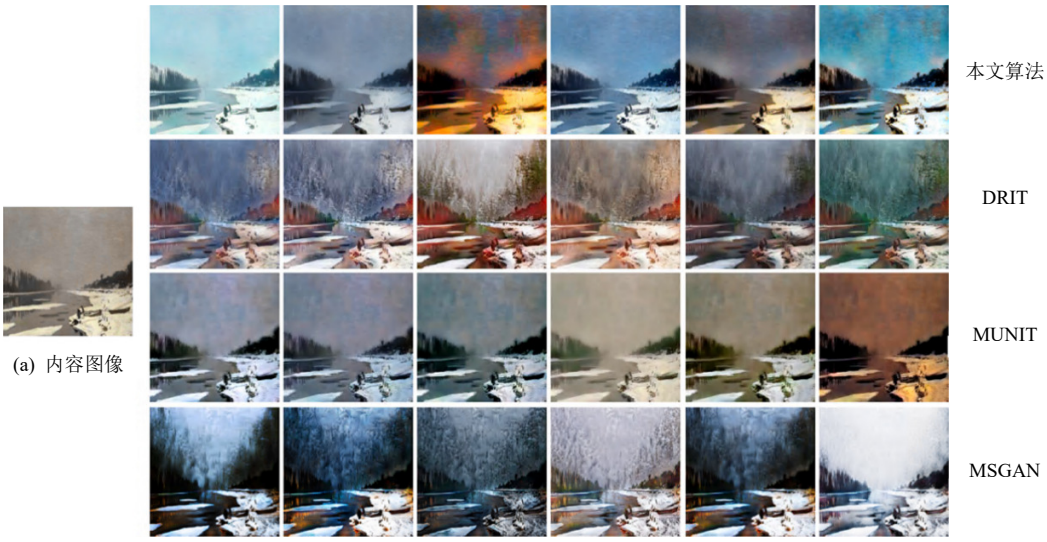
为了证明本文算法中所使用的各个损失函数的有效性，分别对内容损失、风格损失和对比损失在 summer2winter 数据集上做了消融实验，其中内容损

(a) 内容图像

(b) 生成图像

Fig. 8 Comparison images of diversity results of different algorithms on summer2winter

图 8 不同算法在 summer2winter 上的多样性结果对比图



(a) 内容图像 (b) 生成图像

Fig. 9 Comparison images of diversity results of different algorithms on monet2photo

图 9 不同算法在 monet2photo 上的多样性结果对比图

Table 4 Quantitative Metrics of Different Algorithms on summer2winter

表 4 不同算法在 summer2winter 上的量化指标

量化指标	本文算法	DRIT	MUNIT	MSGAN
FID↓	54.64±1.83	58.22±3.03	60.14±2.10	55.47±1.37
LPIPS↑	0.144 7±0.001 7	0.122 5±0.001 6	0.090 9±0.001 0	0.130 3±0.002 1
NDB↓	4.95±1.21	5.47±1.34	5.44±0.87	5.16±1.18
JSD↓	0.038 8±0.002 7	0.053 3±0.002 4	0.054 9±0.004 8	0.042 2±0.003 5

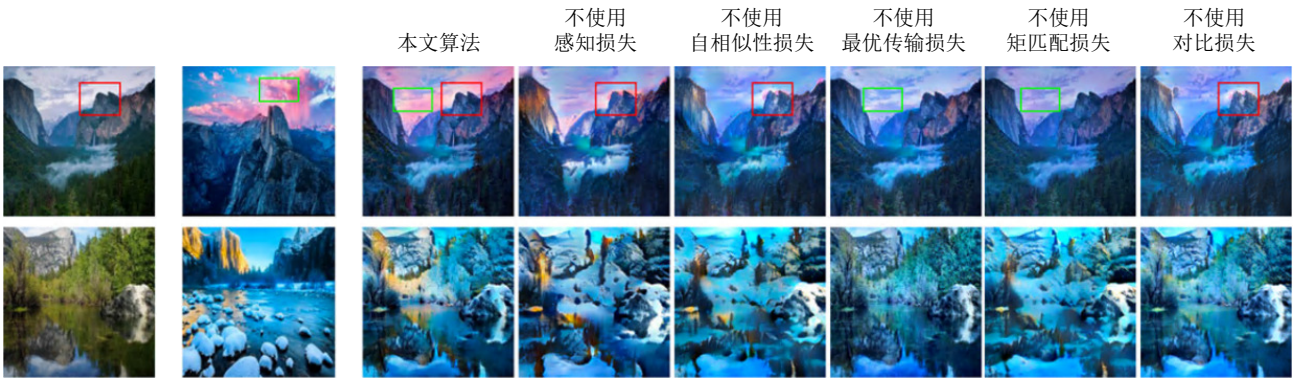
注：“↓”表示数值越小越好，“↑”表示数值越大越好.

Table 5 Quantitative Metrics of Different Algorithms on monet2photo

表 5 不同算法在 monet2photo 数据集上的量化指标

量化指标	本文算法	DRIT	MUNIT	MSGAN
FID↓	84.14±1.73	87.65±2.34	92.66±1.23	85.27±1.55
LPIPS↑	0.215 6±0.001 6	0.169 1±0.001 4	0.161 3±0.001 5	0.192 2±0.001 6
NDB↓	4.99±1.52	6.23±1.05	6.41±1.77	5.54±1.32
JSD↓	0.046 5±0.006 6	0.057 7±0.002 1	0.058 2±0.003 2	0.051 5±0.001 9

注：“↓”表示数值越小越好，“↑”表示数值越大越好.



(a) 内容图像 (b) 风格图像 (c) 生成图像

Fig. 10 Result images of ablation experiment

图 10 消融实验结果图

失包括感知损失和自相似性损失，风格损失包括最优传输损失和矩匹配损失，实验结果如图 10 所示. 为了更直观地对比，图 10(a)(b)(c)中的第 1 行标出了效果比较明显的部位，其中红框表示内容的对比，绿框表示风格对比. 首先观察内容损失的 2 个对比实验：由红框中的内容可知，不使用感知损失的生成结果中的山尖背后存在一些内容图像中未出现的云彩，在 6 组生成图像中不使用感知损失的内容保持效果最差，而不使用自相似性损失的生成图像在生成时忽略了内容图像中的精细细节. 然后观察风格损失的 2 个对比实验：由绿框中的内容可知，不使用最优传输损失的生成结果只学到了风格图像的整体色调，存在严重的风格模式缺失；不使用矩匹配损失的生成图像中学到了多种风格模式，但忽略了每种风格在风格图像中的量的大小. 最后观察不使用对比损失的效果，对比损失实际是风格损失和内容损失的一种变形组合，能够加强二者在模型上的表现，故在经过同样轮次的训练后，不使用对比损失的模型效果较弱.

为了更准确地评估每个损失对算法的提升程度，计算了每种情况下模型在 summer2winter 数据集上的量化指标，如表 6 所示，为了方便展示，只记录了 10 次计算结果的均值，而省略了方差.

Table 6　Quantitative Metrics of Ablation Experiment

表 6　消融实验的量化指标

量化指标	本文算法	无感知损失	无自相似性损失	无最优传输损失	无矩匹配损失	无对比损失
FID↓	54.64	55.12	56.04	56.73	58.44	55.32
LPIPS↑	0.14	0.14	0.14	0.12	0.13	0.14
NDB↓	4.95	5.17	5.04	5.63	5.45	5.08
JSD↓	0.39	0.39	0.40	0.47	0.45	0.40

注：“↓”表示数值越小越好，“↑”表示数值越大越好.

4　结　　论

跨域图像转换过程中为了更好地保留源域图像的内容结构以及引导(参考)图像的风格，减少源域图像内容结构的损失以及引导图像的颜色模式坍塌问题，本文提出了一种自结构注意力损失函数来改进图像跨域转换中的内容结构保持. 该损失函数可以充分利用局部结构之间的长程依赖性，保持微结构之间的相对关系，实现源域图像与相应转换结果图像内容结构的一致性. 利用一种基于统计的颜色损失函数，主要统计参考图像的颜色信息. 通过颜色损失函数约束，可以使转换结果图像的颜色分布与参考图像的颜色分布保持一致，从而显著缓解图像转换过程中的颜色模式崩溃问题. 本文提出的框架可以实现图像内容和风格的分离学习. 与现有最先进的算法相比，本文提出的算法不需要循环一致性损失函数和其他复杂的损失函数约束. 未来，将引入 Transformer 的架构来更好地学习引导图像和源域图像的长范围的内容和风格依赖，进一步提高跨域图像转换的质量和效果.

作者贡献声明：赵磊负责论文思路的提出、整体架构设计；张慧铭负责整个模型实现与代码编写、系统的调参与优化、论文初稿的撰写；邢卫负责论文整体思路的讨论提升；林志洁负责论文中的实验结果与分析部分的提升修改；林怀忠负责论文整体修改与优化；鲁东明负责整体的架构调整与优化；潘洵参与了实验结果数据讨论与分析；许端清负责对论文写作优化.

参　考　文　献

[1] Gatys L, Ecker A, Bethge M. Image style transfer using convolutional neural networks [C] //Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2016: 2414–2423

[2] Huang Xun, Belongie S. Arbitrary style transfer in real-time with adaptive instance normalization [C] //Proc of the 16th IEEE Int Conf on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2017: 1501–1510

[3] Iizuka S, Serra E, Ishikawa H. Globally and locally consistent image completion[J]. ACM Transactions on Graphics, 2017, 36(4): 1−14

[4] Zhao Lei, Mo Qihang, Lin Sihuan, et al. UCTGAN: Diverse image inpainting based on unsupervised cross-space translation [C] //Proc of the 33rd IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2020: 5741–5750

[5] Zhao Lei, Lin Sihuan, Li Ailin, et al. SpatialGAN: Progressive image generation based on spatial recursive adversarial expansion [C] //Proc of the 28th ACM Int Conf on Multimedia (ACM MM). New York: ACM, 2020: 2336–2344

[6] Zhang Kai, Gool L V, Timofte R. Deep unfolding network for image super-resolution [C] //Proc of the 33rd IEEE Conf on Computer

Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2020: 3217–3226

[7] Shamsabadi A S, Matilla R, Cavallaro A. Colorfool: Semantic adversarial colorization [C] //Proc of the 33rd IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2020: 1151–1160

[8] Isola P, Zhu Junyan, Zhou Tinghui, et al. Image-to-image translation with conditional adversarial networks [C] //Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2017: 1125–1134

[9] Wang Tingchun, Liu Mingyu, Zhu Junyan, et al. High-resolution image synthesis and semantic manipulation with conditional GANs [C] // Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2018: 8798–8807

[10] Li Chunyuan, Liu Hao, Chen Changyou, et al. Alice: Towards understanding adversarial learning for joint distribution matching [C] // Proc of the 31st Advances in Neural Information Processing Systems (NIPS). La Jolla, CA: NIPS, 2017: 5495–5503

[11] Larsson G, Maire M, Shakhnarovich G. Learning representations for automatic colorization [C] //Proc of the 15th European Conf on Computer Vision (ECCV). Berlin: Springer, 2016: 577–593

[12] Zhang R, Isola P, Efros A. Colorful image colorization [C] //Proc of the 15th European Conf on Computer Vision (ECCV). Berlin: Springer, 2016: 649–666

[13] Zhu Junyan, Park T, Isola P, et al. Unpaired image-to-image translation using cycle consistent adversarial networks [C] //Proc of the 16th IEEE Int Conf on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2017: 2223–2232

[14] Zhang Hao, Tan Ping, Gong Minglun. DualGAN: Unsupervised dual learning for image-to-image translation [C] //Proc of the 16th European Conf on Computer Vision (ECCV). Berlin: Springer, 2017: 649–666

[15] Anoosheh A, Agustsson E, Timofte R. ComboGAN: Unrestrained scalability for image domain translation [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2018: 783–790

[16] Choi Y, Choi M, Kim M, et al. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2018: 8789–8797

[17] Kim T, Cha M, Kim H, et al. Learning to discover cross-domain relations with generative adversarial networks [C] // Proc of the 34th Int Conf on Machine Learning (ICML). New York: ACM, 2017: 1857−1865

[18] Liu Mingyu, Breuel T, Kautz J. Unsupervised image-to-image translation networks [C] //Proc of the 31st Advances in Neural Information Processing Systems (NIPS). La Jolla, CA: NIPS, 2017: 700–708

[19] Taigman Y, Polyak A, Wolf L. Unsupervised cross-domain image generation [J]. arXiv preprint, arXiv: 1611.02200, 2016

[20] Huang Xun, Liu Mingyu, Belongie S. Multimodal unsupervised image-to-image translation [C] //Proc of the 16th European Conf on Computer Vision (ECCV). Berlin: Springer, 2018: 172–189

[21] Lee H, Tseng H, Huang Jiabin, et al. Diverse image-to-image translation via disentangled representations [C] //Proc of the 16th European Conf on Computer Vision (ECCV). Berlin: Springer, 2018: 35–51

[22] Zhang Han, Goodfellow I, Metaxas D, et al. Self-attention generative adversarial networks [C] //Proc of the 36th Int Conf on Machine Learning (ICML). New York: ACM, 2019: 7354–7363

[23] Chang H, Wang Zhixiang, Chuang Yunyu. Domain-specific mappings for generative adversarial style transfer [C] //Proc of the 19th European Conf on Computer Vision (ECCV). Berlin: Springer, 2020: 573−589

[24] Isola P, Zhu Junyan, Zhou Tinghui, et al. Image-to-image translation with conditional adversarial networks [C] //Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2017: 1125−1134

[25] Wang Tingchun, Liu Mingyu, Zhu Junyan, et al. High-resolution image synthesis and semantic manipulation with conditional gans [C] // Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2018: 8798−8807

[26] Park T, Liu Mingyu, Wang Tingchun, et al. Semantic image synthesis with spatially-adaptive normalization [C] //Proc of the 32nd IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2019: 1−12

[27] Zhu Junyan, Zhang R, Pathak D, et al. Toward multimodal image-to-image translation [C] //Proc of the 31st Advances in Neural Information Processing Systems (NIPS). La Jolla, CA: NIPS, 2017: 1−12

[28] Sohn K, Lee H, Yan Xinchen. Learning structured output representation using deep conditional generative models [C] //Proc of the 29th Advances in Neural Information Processing Systems (NIPS). La Jolla, CA: NIPS, 2017: 3483−3491

[29] Donahue J, Krähenbühl P, Darrell T. Adversarial feature learning [J]. arXiv preprint arXiv: 1605.09782, 2016

[30] Kingma D P, Welling M. Auto-encoding variational Bayes [J]. arXiv preprint arXiv: 1312.6114, 2013

[31] Zhu Junyan, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks [C] //Proc of the 16th IEEE Int Conf on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2017: 2223−2232

[32] Wang Tingchun, Liu Mingyu, Zhu Junyan, et al. High-resolution image synthesis and semantic manipulation with conditional GANs [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE Computer Society, 2018: 8798−8807

[33] Mao Qi, Lee H Y, Tseng H Y, et al. Mode seeking generative adversarial networks for diverse image synthesis [C] //Proc of the 32nd IEEE Conf on Computer Vision and Pattern Recognition

(CVPR). Los Alamitos, CA: IEEE Computer Society, 2019: 1429−1437

[34] Mejjati Y A, Richardt C, Tompkin J, et al. Unsupervised attention-guided image to image translation [C] // Proc of the 32nd Advances in Neural Information Processing Systems (NIPS). La Jolla, CA: NIPS, 2018: 1−11

[35] Kim J, Kim M, Kang H, et al. U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation [J]. arXiv preprint arXiv: 1907.10830, 2019

[36] Ba J, Kiros J R, Hinton G E. Layer normalization[J]. arXiv preprint arXiv: 1607.06450, 2016

[37] Ulyanov D, Vedaldi A, Lempitsky V. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis [C] //Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2017: 6924−6932

[38] Park T, Efros A A, Zhang R, et al. Contrastive learning for unpaired image-to-image translation [C] //Proc of the 17th European Conf on Computer Vision (ECCV). Berlin: Springer, 2020: 319−345

[39] Choi Y, Uh Y, Yoo J, et al. StarGAN v2: Diverse image synthesis for multiple domains[C] //Proc of the 33rd IEEE Conf on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE Computer Society, 2020: 8188−8197

[40] Choi Y, Uh Y, Yoo J, et al. StarGAN v2: Diverse image synthesis for multiple domains [C] //Proc of the 33rd IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2020: 8188−8197

[41] Liu Mingyu, Huang Xun, Mallya A, et al. Few-shot unsupervised image-to-image translation [C] //Proc of the 17th IEEE Int Conf on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2019: 10551−10560

[42] Wu Haiyan, Qu Yanyun, Lin Shaohui, et al. Contrastive learning for compact single image dehazing [C] //Proc of the 34th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2021: 10551−10560

[43] Gatys L A, Ecker A S, Bethge M. Image style transfer using convolutional neural networks [C] //Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2016: 2414−2423

[44] Odena A, Dumoulin V, Olah C. Deconvolution and checkerboard artifacts[J]. Distill, 2016, 1(10): 1−12

[45] Kolkin N, Salavon J, Shakhnarovich G. Style transfer by relaxed optimal transport and self-similarity [C] //Proc of the 32nd IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2019: 10051−10060

[46] Goodfellow I, Pouget J, Mirza M, et al. Generative adversarial nets [C] //Proc of the 31st Advances in Neural Information Processing Systems (NIPS). La Jolla, CA: NIPS, 2017: 102−120

[47] Zhu J Y, Krähenbühl P, Shechtman E, et al. Generative visual manipulation on the natural image manifold[C] //Proc of the 14th European Conference of Computer Vision. Berlin: Springer , 2016: 597−613

[48] Laffont P Y, Ren Zhile, Tao Xiaofeng, et al. Transient attributes for high-level understanding and editing of outdoor scenes[J]. ACM Transactions on Graphics, 2014, 33(4): 1−11

[49] Zhang Zheng, Ma Huadong, Fu Huiyuan, et al. Scene-free multi-class weather classification on single images[J]. Neurocomputing, 2016, 207: 365−373

Zhao Lei, born in 1975. PhD, associate professor. Member of CCF. His main research interests include image restoration, deep learning, and intelligent image generation.

赵　磊, 1975 年生. 博士, 副研究员. CCF 会员. 主要研究方向为图像修复、深度学习、图像智能生成.

Zhang Huiming, born in 1994. Master candidate. Her main research interest includes image cross domain translation. (qinglanwuji@zju.edu.cn)

张慧铭, 1994 年生. 硕士研究生. 主要研究方向为图像跨域迁移.

Xing Wei, born in 1967. PhD, associate professor. His main research interest includes image intelligent processing.

邢　卫, 1967 年生. 博士, 副教授. 主要研究方向为图像智能处理.

Lin Zhijie, born in 1980. PhD, associate professor. His main research interest includes image intelligent processing. (bytelin@qq.com)

林志洁, 1980 年生. 博士, 副教授. 主要研究方向为图像智能处理.

Lin Huaizhong, born in 1970. PhD, associate professor. His main research interest includes image intelligent processing. (linhz@zju.edu.cn)

林怀忠, 1970 年生. 博士, 副教授. 主要研究方向为图像智能处理.

Lu Dongming, born in 1967. PhD, professor. His main research interest includes image intelligent processing. (ldm@zju.edu.cn)

鲁东明, 1967 年生. 博士, 教授. 主要研究方向为图像智能处理.

Pan Xun, born in 1969. PhD, associate professor. His main research interest includes art image style expression. (px-415@163.com)

潘　洵, 1969 年生. 博士, 副教授. 主要研究方向为艺术图像风格表达.

Xu Duanqing, born in 1966. PhD, professor. His main research interest includes image intelligent processing. (xdq@zju.edu.cn)

许端清, 1966 年生. 博士, 教授. 主要研究方向为图像智能处理.