基于多尺度 Scale-Unet 的单样本图像翻译

周蓬勃[1],冯　龙[2*],寇宇帆[2]
(1. 北京师范大学 艺术与传媒学院,北京 100032;
2. 西北大学 信息科学与技术学院,陕西 西安 710127)

摘　要:随着生成对抗网络(GAN)的发展,基于单样本的无监督图像到图像翻译(UI2I)取得了重大进展。然而,以前方法无法捕获图像中的复杂纹理并保留原始内容信息。为解决这个问题,提出了一种基于尺度可变 U-Net 结构(Scale—Unet)的新型单样本图像翻译结构 SUGAN。所提出的 SUGAN 使用 Scale—Unet 作为生成器,利用多尺度结构和渐进方法不断改进网络结构,以从粗到细地学习图像特征。同时,提出了尺度像素损失 scale-pixel 来更好地约束保留原始内容信息,防止信息丢失。实验表明,与 SinGAN、TuiGAN、TSIT、StyTR2 等公共数据集 Summer↔Winter、Horse↔Zebra 上的方法相比,该方法生成图像的 SIFID 值平均降低了 30%。所提方法可更好地保留图像内容信息,同时生成详细逼真的高质量图像。

关键词:单样本图像翻译;Scale-Unet;多尺度结构;渐进方法;尺度像素损失
中图分类号:TP394.1;TH691.9　　**文献标识码**:A　　**文章编号**:1673-629X(2024)04-0055-07
doi:10.20165/j.cnki.ISSN1673-629X.2024.0009

Single-sample Image Translation Based on Multi-scale Scale-Unet

ZHOU Peng-bo[1],FENG Long[2*],KOU Yu-fan[2]
(1. School of Art and Media,Beijing Normal University,Beijing 100032,China;
2. School of Information Science and Technology,Northwest University,Xi'an 710127,China)

Abstract:Single-sample unsupervised image-to-image translation (UI2I) has made significant progress with the development of generative adversarial networks (GANs). However,previous methods cannot capture complex textures in images and preserve original content information. We propose a novel one-shot image translation structure SUGAN based on a scale-variable U-Net structure (Scale—Unet). The proposed SUGAN uses Scale—Unet as a generator to continuously improve the network structure using multi-scale structures and progressive methods to learn image features from coarse to fine. Meanwhile,we propose the scale-pixel loss to better constrain the preservation of original content information and prevent information loss. Experiments show that compared with SinGAN,TuiGAN,TSIT,StyTR2 and another methods on public datasets Summer↔Winter,Horse↔Zebra,the SIFID value of the generated image is reduced by 30%. The proposed method can better preserve the content information of the image while generating detailed and realistic high-quality images.

Key words:single-sample image translation;Scale-Unet;multi-scale structure;progressive approach;scale-pixel loss

0　引　言

无监督图像到图像的翻译(UI2I)旨在将图像从源域映射到目标域。Ul2I 在保留源域图像内容的同时,将源域图像翻译成目标域风格。已有的 UI2I 方法极度依赖海量的成对训练数据,而海量的成对标注数据常常难于获取。因此,利用少样本数据甚至是单样本数据来实现无监督图像到图像翻译,日益成为一个热点。

针对单样本的无监督图像翻译问题,SinGAN[1]首次提出了一个无条件生成模型的方法,实现了无监督图像到图像的翻译任务,取得了不错的效果。但是,SinGAN 是一个串行的多阶段模型结构,它的阶段 N 以上一阶段 $N-1$ 的生成图像作为模型输入,所以训练速度非常缓慢,并且图像生成结果与源图像之间缺少内容和目标风格的一致性约束,导致生成图像发生扭曲和模糊。与 SinGAN 不同,InGAN[2]使用单层生成

收稿日期:2023-03-11　　**修回日期**:2023-07-13
基金项目:国家自然科学基金项目(62271393);国博文旅部重点实验室开放课题(CRRT2021K01);陕西省重点研发计划(2019GY-215,2021ZDLSF06-04)
作者简介:周蓬勃(1984-),男,高级工程师,博士,研究方向为数字艺术与虚拟现实;通讯作者:冯　龙(1996-),男,博士研究生,研究方向为计算机视觉、图像生成。

对抗网络学习单一图像的内部结构,生成新的不同大小、形状、比例甚至矩形的图像,但不能用于图像翻译工作。Shuffle-SinGAN[3]将随机像素洗牌添加到多尺度训练过程中,从而生成略微精细的翻译图像,因此翻译后的图像仍然模糊不清。虽然 ConSinGAN[4]通过并行的方式,提高了模型训练的速度,但是它也存在 SinGAN 相同的问题,即模型中没有风格约束,所以翻译图像模糊的问题仍然存在。结合 SinGAN 的思想,Lin J 等人提出了 TuiGAN[5]方法。它也是一个多阶段的图像翻译模型,通过循环一致性损失[6]来约束源图像、图像翻译结果和重建的图像,解决了翻译图像模糊的问题。但是,和 SinGAN 一样,受限于模型的串行结构,TuiGAN 的模型训练成本也非常高。

已有的图像多尺度[7]研究表明,低尺度图像包含的细节特征内容较少,而高尺度图像包含的图像纹理和结构信息更完整。除此之外,TransferI2I[8]也表明低维特征对于保持图像全局结构贡献较大,高维特征对于保持局部信息(比如图像纹理和颜色)非常重要,换句话说,图像分辨率较低的训练阶段对全局结构很重要,而分辨率较高的训练阶段对最终图像的纹理和颜色很重要。为得到更好的图像翻译结果,TuiGAN 模型需要更多卷积层去学习高尺度的图像特征。而 TuiGAN 多阶段训练中,每个阶段的尺度都需要重新初始化卷积参数,这也导致高尺度图像的特征学习时间复杂度很高。

Unet[9]结构是一个图像分割的标准方法,它可以对单样本图像进行语义分割,同时可以保留图像内容特征和防止图像语义信息丢失。受此启发,该文提出了一个随图像尺度变化的 Unet 结构(Scale-Unet)。它以 Unet 结构和注意力结合作为生成器,通过不断增加卷积块来学习更高尺度图像的特征。

总而言之,该文的主要贡献如下:

(1)引入单次图像翻译网络 SUGAN,它引入了渐进式增长生成器和多尺度训练方法,能有效捕捉源域和目标域中两种不同分布之间的关系,从全局信息中学习图像目标风格。

(2)设计了 Scale-Unet 结构作为生成器,以充分提取不同尺度图像的全局和局部特征,防止信息丢失。

(3)设计了一种尺度像素损失,以确保源图像的图像内容特征和经过 Scale 图像的图像内容特征和 Scale-Unet 后的图像转换结果是一致的。

1 文中方法

1.1 SUGAN 网络结构

SUGAN 整体框架如图 1 所示。SUGAN 由金字塔(G_{AB} 和 G_{BA})及判别器金字塔(D_A 和 D_B)组成。生成器 G_{AB} 用于将 I_A 翻译成 I_B (图 1(a)),生成器 G_{BA} 用于将 I_B 翻译成 I_A (图 1(b)), D_A 用于区分输入图像来自真实图像 I_A 还是生成图像 I_{AB} , D_B 用于区分输入图像来自真实图像 I_B 还是生成图像 I_{BA} 。生成器 G_{AB}^n 和 G_{BA}^n 的网络结构相同,都是由 Scale-Unet 网络和硬注意力机制 HA 组成,在具有不同权值参数的网络中形成对称结构,以便更好地提取图像内容特征 Scale-Unet 模块对源图像提取图像分割特征,硬注意机制 HA 获取图像重要信息。

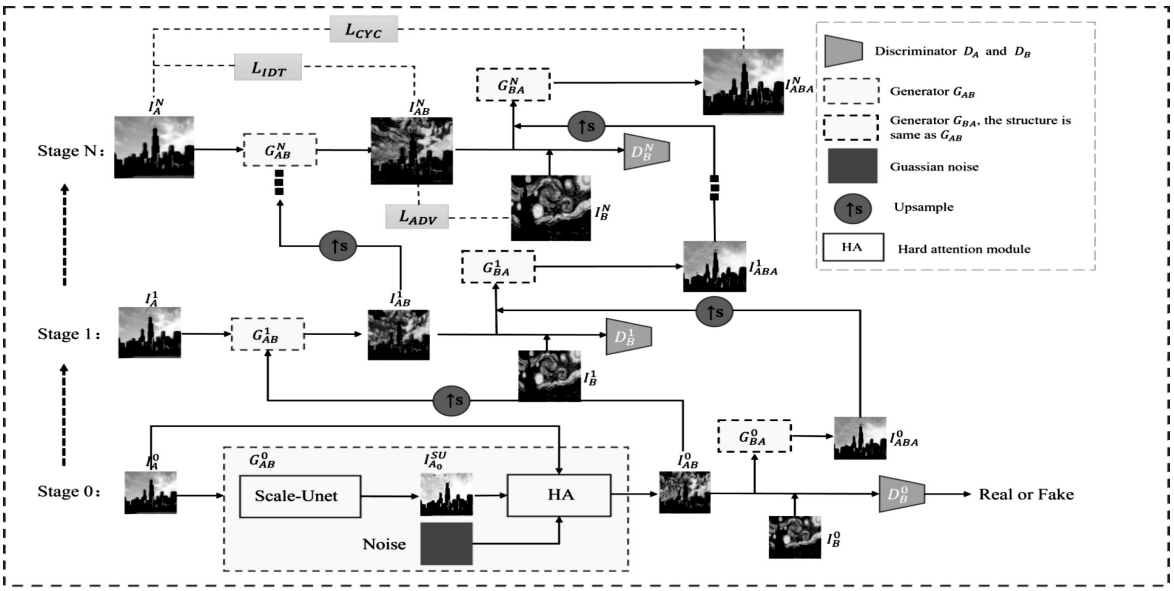

图 1 SUGAN 网络结构

该文在阶段 n 从源图像 I_A 到目标图像 I_B 的图像翻译过程如图 2 所示。I_A^n 的图像分割特征、$I_{A_n}^{SU}$ 用 G_{AB}^n 中的 Scale -Unet 提取, $n-1$ 阶段上采样后的图像翻译结果 I_{AB}^{n-1} ,与当前阶段源图像 I_A^n 通过硬注意机制得

到当前翻译结果 I_{AB}^{n}。图像分割特征 $I_{AB_n}^{SU}$ 的图像翻译结果在对称 G_{BA}^{n} 的 Scale-Unet 中提取,前一阶段(stage $n-1$)图像重建结果 I_{ABA}^{n-1} 上采样后,与当前阶段 stage n 图像翻译结果 I_{AB}^{n} 通过硬注意机制 HA 重建图像翻译结果 I_{ABA}^{n}。判别器 D_B^n 用于鉴别 I_{AB}^{n} 是否与目标图像 I_B 相似。

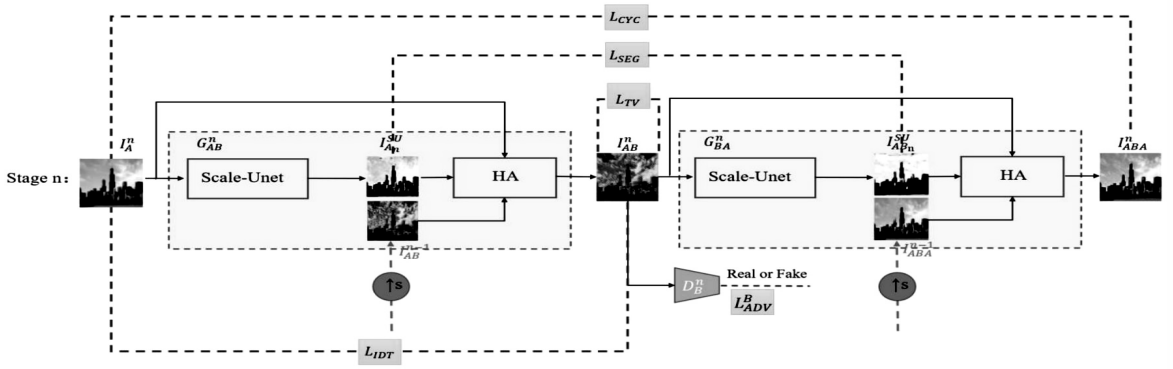

图 2　文中方法在阶段 n 上源图像 I_A 到目标图像 I_B 的图像翻译示意图 $(0 \leqslant n \leqslant N)$

整个网络分为 N 个阶段,即 Stage 0, Stage 1, …, Stage N,如图 1 所示。利用尺度可变 Unet 结构(Scale-Unet)来充分提取图像分割特征,以减少迭代次数。

$$I_{A_n}^{SU} = \mathrm{SU}_n(I_A^n) \tag{1}$$

其中, SU_n 代表第 n 阶段 $(0 \leqslant n \leqslant N)$ 的 Scale-Unet$_n$ 网络, I_A^n 为输入的源图像, $I_{A_n}^{SU}$ 为图像分割后的结果。

利用硬注意机制获取图像重要信息。

$$I_{AB}^{n-1\uparrow} = \mathrm{HA}(I_{AB}^{n}, I_{A_n}^{SU}, I_{AB}^{n-1\uparrow}) \tag{2}$$

$$I_{ABA}^{n-1\uparrow} = \mathrm{HA}(I_{AB}^{n}, I_{AB_n}^{SU}, I_{ABA}^{n-1\uparrow}) \tag{3}$$

其中,HA 代表一种硬注意机制, I_A^n 为输入源图像, $I_{A_n}^{SU}$ 为源图像 I_A^n 通过当前阶段 n 的图像分割特征, $I_{AB}^{n-1\uparrow}$ 为前阶段 $n-1$ 的图像翻译结果, I_{AB}^{n} 为当前阶段 n 的图像翻译结果, $I_{AB_n}^{SU}$ 为当前阶段图像翻译分割特征, I_{ABA}^{n-1} 为前一阶段图像重构结果, $\uparrow$ 为上采样过程。

随着级数的增加,生成器的尺寸和图像的尺寸也随之增加。对于任意阶段 n,SUGAN 有两个生成器 G_{AB}^{n}、G_{BA}^{n} 和两个判别器 D_A^n、D_B^n。D_A^n 和 D_B^n 确保转换后的图像属于正确的图像域, G_{AB}^{n} 将图像域 I_A 映射得到图像域 I_B 并得到翻译后的图像 I_{BA}^{n}, G_{BA}^{n} 通过反向映射得到翻译后的图像 I_{BA}^{n}。

$$I_{AB}^{n} = G_{AB}^{n}(I_A^n), I_{BA}^{n} = G_{BA}^{n}(I_B^n) \tag{4}$$

重复这个过程 N 次,直到达到想要的输出尺度。

1.2　Scale-Unet$_n$ 网络结构

图 3 是 Scale - Unet$_n$ 的网络结构。网络由三个模块(下采样、硬注意力机制和上采样)组成,下采样和上采样呈对称结构,下采样模块连续地对源图像进行下采样,并将残差连接起来,得到最小尺度 100×100 的分割特征 $I_{A_n}^{SU_d}$。为了更好地学习局部信息,Scale-Unet 中的硬注意机制模块 HA 采用最低尺度的 100×100 图像 $I_{A_n}^{0}$ 和高斯噪声通过级联作为输入,输出 $I_{A_n}^{SU_{cat}}$, $I_{A_n}^{SU_{cat}}$ 将被上采样以使上采样模块能够学习图像更精细的特征。上采样模块获取 $I_{A_n}^{SU_{cat}}$ 并连续上采样,残差跳跃连接是为了获得当前阶段 n 的图像分割特征,以便学习更精细的图像内容特征。

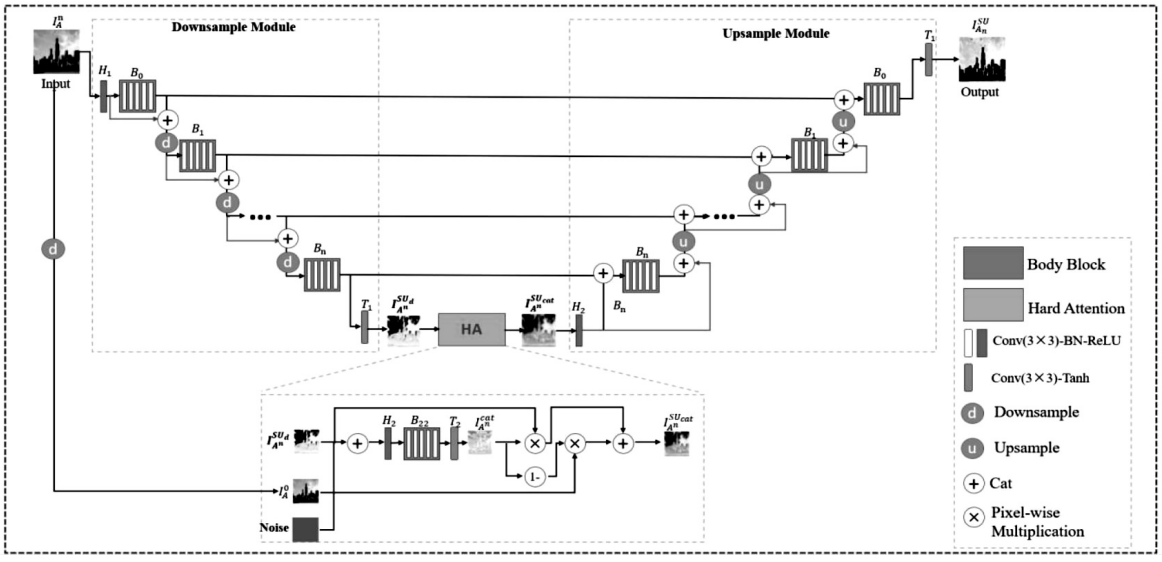

图 3　Scale - Unet$_n$ 的网络结构

该文采用了残差连接防止梯度消失使图像特征信息更加完善。

$$I_{A_n}^{\mathrm{SU}_d}=\begin{cases}T_1(d(B_0(H_1(I_A^0)))),n=0\\ T_1(d(\mathrm{cat}(d(B_0(H_1(I_A^0))),B_1))),n\geqslant 1\end{cases}\tag{5}$$

其中，$I_{A_n}^{\mathrm{SU}_d}$ 为下采样图像分割特征提取结果，T_1 是 Tail 层，d 代表下采样，n 代表当前训练阶段，H_1 是 Head 块。

为了更好地提取图像内容特征，增强图像尺度信息。

$$I_{\mathrm{cat}}=\mathrm{cat}(I_{A_n}^{\mathrm{SU}_d},I_A^0,\mathrm{noise})\tag{6}$$

$$I_{A_n}^{\mathrm{cat}}=T_2(B_{22}(H_2(I_{\mathrm{cat}})))\tag{7}$$

$$I_{A_n}^{\mathrm{SU}_{\mathrm{cat}}}=(I_{A_n}^{\mathrm{SU}_d}\otimes I_{A_n}^{\mathrm{cat}})\oplus(I_A^0\otimes(1-I_{A_n}^{\mathrm{cat}}))\tag{8}$$

利用 cat 作为硬注意机制模块 HA 的输入，高斯噪声，$I_{A_n}^{\mathrm{SU}_d}$ 和 I_A^0 信道连接，生成图像合并区域代码 $I_{A_n}^{\mathrm{Cat}}$。输出的 $I_{A_n}^{\mathrm{Cat}}$ 是由 $I_{A_n}^{\mathrm{SU}_d}$ 和 I_A^0 线性结合得到。

在 Scale-Unet 的上采样过程中，采用跳跃连接和残差块获取图像局部信息，以保持图像内容特征。

1.3 损失函数

该文设计了一个尺度像素损失来保持源图像的内容信息。此外，结合对抗性损失、循环一致性损失、身份损失、总变差损失，形成新的损失函数，用于实现单样本图像翻译，防止信息丢失。

对于 $\forall n\in\{0,1,\cdots,N\}$，第 n 个阶段的总损失定义如下：

$$L_{\mathrm{ALL}}^n=\lambda_{\mathrm{ADV}}L_{\mathrm{ADV}}^n+\lambda_{\mathrm{CYC}}L_{\mathrm{CYC}}^n+\lambda_{\mathrm{SEG}}L_{\mathrm{SEG}}^n+\lambda_{\mathrm{IDT}}L_{\mathrm{IDT}}^n+\lambda_{\mathrm{TV}}L_{\mathrm{TV}}^n\tag{9}$$

其中，L_{ALL}^n 是总损失，L_{ADV}^n 是对抗性损失，L_{CYC}^n 是循环一致性损失，L_{SEG}^n 是尺度-像素损失，L_{IDT}^n 是身份损失，L_{TV}^n 是总变差损失，λ_{ADV}，λ_{CYC}，λ_{SEG}，λ_{IDT}，λ_{TV} 是平衡各项损失函数的超参数。

尺度-像素损失：基于循环一致性损失在迭代次数少的情况下约束强度不够，该文通过尺度-像素损失约束模型，以保留图像翻译后输入图像的内容特征（n 表示训练阶段数）。

$$\mathcal{L}_{\mathrm{SEG}}^n=\|I_{AB_n}^{\mathrm{SU}}-I_{A_n}^{\mathrm{SU}}\|_1+\|I_{BA_n}^{\mathrm{SU}}-I_{B_n}^{\mathrm{SU}}\|_1\tag{10}$$

其中，

$$I_{A_n}^{\mathrm{SU}}=\mathrm{Scale-Unet}_n(I_A^n)$$

$$I_{B_n}^{\mathrm{SU}}=\mathrm{Scale-Unet}_n(I_B^n)$$

$$I_{AB_n}^{\mathrm{SU}}=\mathrm{Scale-Unet}_n(I_{AB}^n)$$

$$I_{BA_n}^{\mathrm{SU}}=\mathrm{Scale-Unet}_n(I_{BA}^n)$$

对抗性损失：使用对抗性损失来支持生成器生成视觉上与目标域图像相似的图像。有两个判别器 D_A^n 和 D_B^n，分别以 I_{BA}^n 和 I_{AB}^n 作为输入，选择 WGAN-GP 损失，其中可以通过权值裁剪和梯度惩罚有效增加训练的稳定性：

$$\begin{aligned}\mathcal{L}_{\mathrm{ADV}}^n=&D_B^n(I_B^n)-D_B^n(G_{AB}^n(I_A^n))+\\&D_A^n(I_A^n)-D_A^n(G_{BA}^n(I_B^n))-\\&\lambda_{\mathrm{PEN}}(\|\nabla_{\hat{I}_B^n}D_B^n(\hat{I}_B^n)\|_2-1)^2-\\&\lambda_{\mathrm{PEN}}(\|\nabla_{\hat{I}_A^n}D_A^n(\hat{I}_A^n)\|_2-1)^2\end{aligned}\tag{11}$$

其中，$\hat{I}_B^n=\alpha I_B^n+(1-\alpha)I_{AB}^n$，$\hat{I}_A^n=\alpha I_A^n+(1-\alpha)I_{BA}^n$，$\alpha\sim U(0,1)$，$\lambda_{\mathrm{PEN}}$ 为惩罚系数。

循环一致性损失：为了缓解模型模式崩溃问题，在生成器上加了循环一致性损失，这可以约束模型保留固有的输入图像翻译后的属性：

$$\mathcal{L}_{\mathrm{CYC}}^n=\|I_A^n-I_{ABA}^n\|_1+\|I_B^n-I_{BAB}^n\|_1\tag{12}$$

其中，$I_{ABA}^n=F_{BA}^n(I_{AB}^n)$，$I_{BAB}^n=G_{AB}^n(I_{BA}^n)$。

身份损失：为保证输入图像和输出图像的颜色分布相似，该文对生成器施加身份一致性约束[4]。例如，给定图像 $I_A^n\in A$，经过生成器 G^n 翻译后，图像的颜色和纹理不应发生变化：

$$\mathcal{L}_{\mathrm{IDT}}^n=\|I_A^n-I_{AA}^n\|_1+\|I_B^n-I_{BB}^n\|_1\tag{13}$$

其中，$I_{AA}^n=F^n(I_A^n)$，$I_{BB}^n=G^n(I_B^n)$。

总变差损失：为了避免噪声对图像的影响，引入总变差（TV）损耗[5]，有助于去除翻译图像中的粗糙纹理。设 $x[i,j]$ 表示图像 x 位于第 i 行、第 j 列的像素，第 n 级 TV 损失计算如下：

$$\mathcal{L}_{\mathrm{TV}}^n=L_{\mathrm{tv}}(I_{AB}^n)+L_{\mathrm{tv}}(I_{BA}^n)\tag{14}$$

其中，

$$L_{\mathrm{tv}}(x)=\sum_{i,j}\sqrt{(x[i,j+1]-x[i,j])^2+(x[i+1,j]-x[i,j])^2}$$

$$x\in\{I_{AB}^n,I_{BA}^n\}$$

2 实验结果与分析

训练细节、数据集、评估指标、所有基线以及实验结果描述如下。

2.1 实施细节

训练设定：使用 Adam 训练网络，其中 $\beta_1=0.5$，$\beta_2=0.999$。初始学习率为 0.000 5，默认训练 6 个阶段，每阶段有 100 个 epochs。此外，使用基于 Resnet 的生成器和 PatchGAN 判别器，批处理大小为 1，图像最大分辨率为 250×250，最小分辨率为 100×100。所有实验的权重参数设置为：$\lambda_{\mathrm{CYC}}=1$，$\lambda_{\mathrm{SEG}}=0.5$，$\lambda_{\mathrm{IDT}}=1$，$\lambda_{\mathrm{TV}}=0.1$，$\lambda_{\mathrm{PEN}}=0.1$，No_pos = 3。文中模型在单个 2080-Ti GPU 上训练 15 ~ 30 分钟。

2.2 基线模型

将文中方法与最新的 UI2I 方法进行了定性和定量比较，并选择了以下基线：SinGAN[1] 是一个金字塔式无条件生成模型，只对目标域的一幅图像进行训练。

TuiGAN[5]采用多级训练结构,使用两幅未配对的图像进行图像转换。CycleGAN[6]引入了周期一致性损失,以学习从目标域到源域的反向映射。TSIT[10]为图像翻译提供了一个精心设计的双流生成模型。StyTR2[11]是一个使用变压器作为编码器的风格转换模型。Qs-Attn[12]设计了一个查询选择注意模块,确保源图像在图像转换的相应位置学习目标图像特征。

2.3　数据集

Summer↔Winter 包含 1 540 幅夏季图片和 1 200 幅冬季图片。Van Gogh↔Photo 包含 400 幅梵高画作和 6 287 张照片。Ukiyoe↔Photo 包含 562 幅浮世绘和来自 Flick 的 6 287 张照片。Cezanne↔Photo 包含塞尚画作和来自 Flickr 的 6 287 张照片。Monet↔Photo 包含 1 072 幅莫奈的画作和来自 Flickr 的 6 287 张照片。Horse↔zebra 是 CycleGAN 已发布的数据集,包含 1 067幅马图像和 1 334 幅斑马图像。

2.4　实验分析

2.4.1　定性分析

与基线模型进行对比,结果如图 4 所示。SinGAN[1]在翻译结果中都改变了源图像的全局颜色,但没有传递高层语义结构,在图像补丁块差异较大时无法学习到较好的图像分布,容易导致生成不真实的图像。CycleGAN[6]无法学习目标图像的纹理和颜色(例如,图 4 中 Zebra→Horse 翻译生成的斑马纹理不完整)。TSIT[10]是一种风格转移方法,能较好地保留源图像的内容特征,但会丢失目标样本的风格(例如,图 4 中 Zebra→Horse 中生成的马的纹理丢失)。StyTR2[11]在 Summer→Winter 中的传输效果更好,因为它使用变换器结构保留了图像的全局信息,但在 Zebra→Horse 中无法传输彼此的纹理。Qs-Attn[12]在 Zebra→Horse 的图像翻译过程中会丢失目标图像的样式信息。与上述方法相比,TuiGAN[5]使用两个不配对的数据进行训练取得了更真实的结果,但翻译后的图像通常不清晰,质量较差,存在噪声、失真等与人类视觉不相符的部分(例如,图 4 中的 Winter→Summer 生成的图像源内容信息丢失)。文中方法通过增加 Scale—Unet 模块和尺度像素损失,有利于学习低尺度图像和高尺度图像中的全局信息和局部信息,在保留源图像的同时提高图像质量[13]。

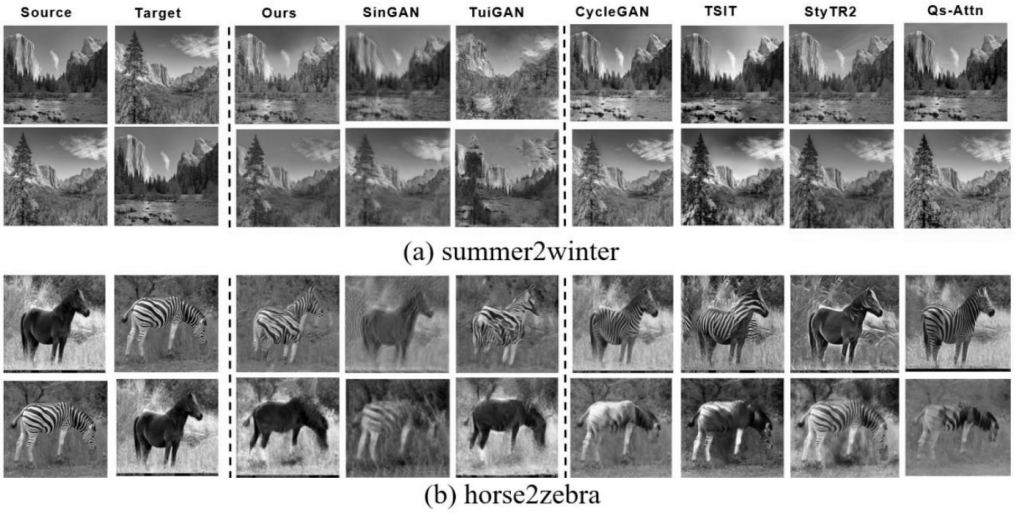

图 4　文中方法与不同基线在 Summer↔Winter 和 Horse↔Zebra 的翻译结果

2.4.2　定量分析

SIFID 指标:使用 Single Image Fréchet Inception Distance (SIFID)[14]评估翻译图像的质量。SIFID 通过计算两幅图像的深度特征之间的 Fréchet Inception Distance (FID)[15]来捕捉两幅图像内部分布的差异。FID 与人类感知高度对应,它基于 Inception Score (IS)。FID 越低,表示真实图像和生成图像之间 Fréchet 距离越小。也就是说,较低的 FID 意味着翻译图像更真实。

PIPS 指标:使用学习感知相似度 Learned Perceptual Image Patch Similarity(LPIPS)[16]。该度量标准学习生成图像到 Ground Truth 的反向映射强制生成器学习从假图像中重构真实图像的反向映射,并优先处理它们之间的感知相似度。LPIPS 的值越低表示两张图像越相似。

表 1 显示了不同方法在 Summer↔Winter、Horse↔Zebra 上 SIFID 值和 LPIPS 值。可以发现,文中方法的 LPIPS 值最小,说明可以在学习目标样式的同时保留内容特征,这是因为该方法使用了缩放像素损失,可以更好地保留源图像的语义信息;SIFID 值最小说明文中方法生成的图像更符合真实分布。在单样本图像翻译中,SinGAN 生成图像的图像学习分布不符合真实分布,TuiGAN 使用两幅图像学习图像内部分布,生成的图像更真实,学习目标风格也会更相似,但不能更好

地保留源图像内容信息。在多样本图像翻译中，CycleGAN、TSIT、StyTR2、Qs-Attn 能较好地保留内容信息，但与目标样本风格差距较远。

表 1 Summer↔Winter 和 Horse↔Zebra 不同方法的 SIFID 值和 LPIPS 值

Method	Summer↔Winter		Horse↔Zebra	
	SIFID	LPIPS	SIFID	LPIPS
SinGAN	1.817	0.761 0	0.768	0.577 4
TuiGAN	0.768	0.744 3	2.211	0.575 9
CRPGAN	0.733	0.744 3	1.057	0.575 9
CycleGAN	0.991	0.795 9	1.164	0.670 7
TSIT	0.905	0.590 1	5.833	0.682 5
StyTR2	2.564	0.623 5	3.803	0.716 3
Qs-Attn	1.582	0.851 2	1.183	0.601 2
SUGAN(ours)	**0.650**	**0.586 7**	**0.622**	**0.561 0**

注:最好的分数用黑体字表示。

2.4.3 消融实验

为了评估各个损失函数以及 Scale-Unet 层中注意力对图像翻译结果的影响，基于 Photo↔Van Gogh 设计了两组消融实验，如图 5 所示。

(1)固定 $N=6$，去除内容分割损失(SUGAN w/o L_{SEG})，循环一致性损失(SUGAN w/o L_{CYC})，身份损失(SUGAN w/o L_{IDT})，总变差损失(SUGAN w/o L_{TV})并比较不同。

(2)固定 $N=6$，去掉 Scale-Unet 中的硬注意力机制 HA 结构，来验证这个结构的作用。

定性结果如图 5 示。如果没有尺度像素损失 L_{SEG}，生成的结果没有正确的颜色和纹理(例如，第 1 列中的图像翻译和图像重建后的 Van Gogh 上的灰色)。如果没有循环一致性损失 L_{CYC}，该模型将无法保证物体形状的完整性和重建的准确性(第 2 列中重建后 Van Gogh 和 Photo 都与源图像差距较远)。如果没有 L_{TV}，模型生成的结果会出现伪影(第 3 列中图像翻译后的 Photo 的周边会出现空洞)。如果没有 L_{IDT}，模型生成的结果会出现错误的图像内容结构(第 4 列图像翻译后结果周边会无缘无故出现黑色区域)。如果 Scale-Unet 中没有硬注意力机制 HA 结构，生成的结果会虚化，与真实结果有些差距(第 5 列图像翻译后的 Photo 会有些虚化)。文中模型可以在保存源图像的内容特征的同时，迁移目标图像的风格特征。

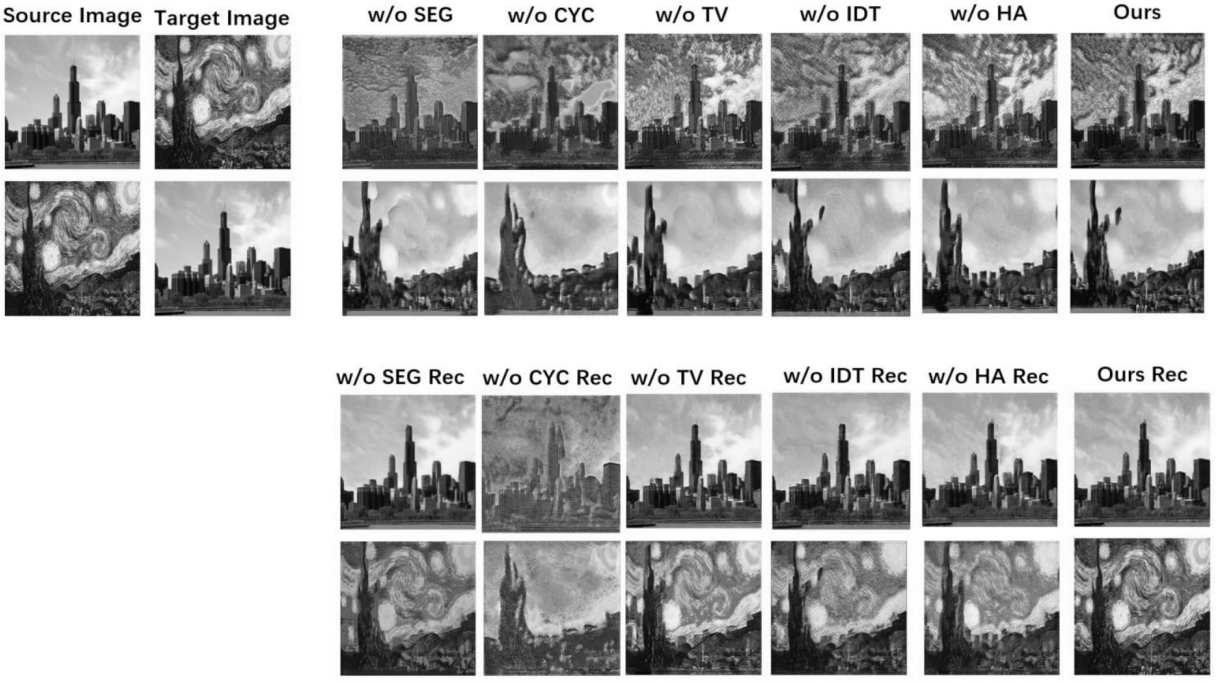

图 5 消融实验定性结果

3　结束语

该文提出了一种新的单样本图像到图像翻译网络 SUGAN,利用两个不配对的图像进行图像翻译。SUGAN 采用多尺度多阶段的训练方式,利用渐进式生成器不断优化网络,并提出尺度可变的 Unet 结构(Scale-Unet)作为生成器,从粗到细学习图像的全局和局部结构(纹理和风格特征)。并提出了尺度像素损失来保留图像的原始特征防止内容信息丢失。实验结果表明,在数据极其有限的图像翻译任务中,该方法仅需 15 ~ 30 分钟即可训练,能够更好地利用图像信息生成详细、真实的图像翻译结果,该框架可广泛应用于图像领域翻译任务。通过消融实验验证了不同损失函数对该方法的影响和不同尺度对该方法的影响,验证了参数的合理性。该方法的 SIFID 在多个数据集上相较于 TuiGAN、SinGAN 等平均下降了 30% 。

参考文献:

[1] SHAHAM T R, DEKEL T, MICHAELI T. Singan: learning a generative model from a single natural image[C]//Proceedings of the IEEE/CVF international conference on computer vision. Seoul: IEEE, 2019: 4570-4580.

[2] SHOCHER A, BAGON S, ISOLA P, et al. Ingan: capturing and retargeting the "dna" of a natural image[C]//Proceedings of the IEEE/CVF international conference on computer vision. Seoul: IEEE, 2019: 4492-4501.

[3] ZHENG M, ZHANG P, GAO Y, et al. Shuffling-singan: improvement on generative model from a single image[J]. Journal of Physics: Conference Series, 2021, 2024: 012011.

[4] HINZ T, FISHER M, WANG O, et al. Improved techniques for training single-image gans[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. Waikoloa: IEEE, 2021: 1300-1309.

[5] LIN J, PANG Y, XIA Y, et al. Tuigan: learning versatile image-to-image translation with two unpaired images[C]//Computer vision - ECCV 2020: 16th European conference. Glasgow: Springer, 2020: 18-35.

[6] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE international conference on computer vision. Venice: IEEE, 2017: 2223-2232.

[7] ISMAEL S F, KAYABOL K, APTOULA E. Unsupervised domain adaptation for the semantic segmentation of remote sensing images via one-shot image-to-image translation[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 20: 2502605.

[8] WANG Y, LARIA H, VAN DE WEIJER J, et al. Transferi2i: transfer learning for image-to-image translation from small datasets[C]//Proceedings of the IEEE/CVF international conference on computer vision. Montreal: IEEE, 2021: 14010-14019.

[9] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation[C]//Medical image computing and computer-assisted intervention - MICCAI 2015: 18th international conference. Munich: Springer, 2015: 234-241.

[10] JIANG L, ZHANG C, HUANG M, et al. Tsit: asimple and versatile framework for image-to-image translation[C]//Computer vision - ECCV 2020: 16th European conference, Glasgow: Springer, 2020: 206-222.

[11] DENG Y, TANG F, DONG W, et al. Stytr2: image style transfer with transformers[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Changchun: IEEE, 2022: 11326-11336.

[12] HU X, ZHOU X, HUANG Q, et al. Qs-attn: query-selected attention for contrastive learning in i2i translation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. New Orleans: IEEE, 2022: 18291-18300.

[13] FENG L, GENG G, LI Q, et al. CRPGAN: learning image-to-image translation of two unpaired images by cross-attention mechanism and parallelization strategy[J]. Plos One, 2023, 18(1): e0280073.

[14] ARORA R, LEE Y J. Singan-gif: learning a generative video model from a single gif[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. Waikoloa: IEEE, 2021: 1310-1319.

[15] SHEN F, YAN S, ZENG G. Neural style transfer via meta networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City: IEEE, 2018: 8061-8069.

[16] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City: IEEE, 2018: 586-595.