无监督图像转换之UNIT和cycleGAN的比较和分析

姚心然
（21821046）
计算机科学与技术学院
21821046@zju.edu.cn

摘要（Abstract）

图像转换，能将处在两个图像域的图片进行相互映射转换，即：输入域的图片转换为目标域的图片，反之亦然。有监督的图像转换往往要求训练集中处于两个图像域的图片一一对应[4]，这样的数据集通常难以获得。而无监督的图像转换没有这个要求，也不需要过多的人工标注，其数据可得性更广。本文介绍了两个在方法上具有一定相似性的典型无监督图像转换网络——UNIT[1] 和cycleGAN[2]，通过实验分析其异同、优缺点和适用范围。

1. 简介

图像转换是计算机视觉领域里一个有趣的问题，其核心方法是找到两个图像域之间的映射关系。两个图像域的定义范围可以很广——就像素角度而言，可以是低精度图像和高精度图像之间的转换；就颜色角度而言，可以是黑白图像和彩色图像的相互转换；就纹理角度而言，可以是斑马纹和马纹的转换，等等。

生成对抗式网络（GAN）[3] 的出现，将图像生成引入了一个新的篇章。近期涌现的许多图像转换方法，不论是有监督的[4] 还是无监督的[1, 2]，或多或少都应用了GAN的思想。相比于有监督的训练，无监督的图像转换对数据集要求更低，从某种程度上来说，适用性更广。

UNIT和cycleGAN都属于无监督的图像转换方法，这二者之间有诸多相似性，例如：都应用了GAN和循环一致性（cycle consistency）的思想。而不同是，前者提出了一个共享隐空间的假设，后者是对两个图像域的直接转换。

本文在第二章中，将介绍无监督图像转换技术的相关工作和技术比较；第三章重点介绍UNIT和cycleGAN网络的核心和特点；第四章对上述两个网络进行实验设计，并比较分析其异同、优缺点和适用范围；最后，第五章对两个网络特点和实验结果进行总结，并分享本人对其的评述。

2. 相关工作

针对UNIT和cycleGAN网络的工作，主要围绕图像转换、生成对抗式网络、循环一致性等关键词展开。

2.1. 图像转换

图像转换技术可以追溯到 2001 年提出的图像模拟（Image analogies）技术[5]。其核心是将一个非参数纹理模型应用到“输入-输出”图像对中训练。尔后，更多的训练模型被应用到图像对的训练中，通过CNN提取并学习图像间的特征和关系，从而建立参数转换模型[6]。但是，这些方法都是有监督、针对像素的转换。

无监督的方法不需要图像对也能够学习到两个图像域之间的映射转换关系。CoGAN[7] 和corss-modal[8] 通过共享权重的策略，建立两个图片域的共同表达方式；基于分享权重的概念，又有人引入可变自编码（auto-encoder）结构[9]；为了使得结果包含的图片内容和输入尽可能相似，而纹理、色彩等不同，也有前人做了一些预定义的度量机制或假设[10, 11]，如：分类标签、图像特征度量约束等。这些思想在UNIT和cycleGAN网络中或多或少都有体现。

另一方面，风格迁移问题和图像转换类似，但又有区别。诚然，风格迁移确实也能够根据输入的参考图片，提取图像高维特征，生成纹理、色彩等风格类似的目标图片[12]。但，其主要致力于从一幅图像到另一幅图像的纹理和色彩迁移，而不是针对某一图像域到另一图像域的转换。换言之，选取同一图像域的不同图片作为参考，可以生成不同的目标结果，而图像转换不是这样的。

2.2. 生成对抗式网络 (GAN)

随着生成对抗式网络（GAN）的提出，图像生成[13]、图像编辑[14]等广泛应用该方法，均获得了较好的结果。除此，生成对抗的思想在视频[15]、三维模型[16]等方面亦有应用。

GAN的核心思想是零和游戏的利益最大化。通过设置一个生成器G和一个判别器D，二者进行博弈、相互遏制，随着训练迭代次数的增加，使得生成器G和生成器D之间达到某种平衡。生成器G根据输入的图像域A

图片和噪声Z生成一个模拟图像域B的结果，判别器D判断输入的图片是否属于图像域B，是得一分，否则不得分。因此，D会尽可能地降低G生成图片的得分率，而G则尽可能地提高生成图片的得分率，最终二者达到平衡。

该博弈是通过生成对抗损失的设置获得的。损失函数设置如下：

$$\max_{D}\{\mathbb{E}_{x\sim P_{data}}\log D(x|y)+\mathbb{E}_{x\sim P_G}\log(1-(D(x|y))\} \quad (1)$$

其中，P_{data}是图像域B的图片概率分布，P_G是生成器G生成结果的概率分布。该损失函数描述的是判别器D的准确次数，因此，训练过程中应尽可能使该函数值最大。

在UNIT和cycleGAN中，均应用了GAN网络的对抗生成思想。针对图片域A和图片域B分别设置生成器G_A、G_B和判别器D_A、D_B，在训练过程中，使得GA和GB尽可能生成和对应图像域相似的结果，而DA和DB则尽可能区分真实图像和对应生成图像。即，G_A和D_A、G_B和D_B单独来看，是一个完整的GAN网络。

2.3. 循环一致性 (Cycle Consistency)

循环一致性，用通俗的语言描述，是指从初始状态经过一系列组成闭环的映射转换后，最终回到初始状态。对于图像转换的图片域而言，即为将图片域A的图片a通过“A→B”的映射转化为域B图片后，再通过“B→A”映射转换回来，记为a’，此时a’与原始的a应尽可能相似。

由于映射函数的构造可以是多样的，单纯地给出一组“输入-输出”对数据集，可能有成百上千种映射关系均满足从输入到输出的映射。因此，考虑循环一致性能够有效地约束映射方向，避免映射的多样性导致不需要的结果。

循环一致性的思想由来已久，其应用广泛涉及各领域。视觉追踪的“前向-后向”约束[17] 正是循环一致性的体现；在语音翻译领域，反向翻译和核对技术常被用于人工翻译和机器翻译的校对中[18]；此外，在三维模型[19]、深度检测[20]等领域都有对循环一致性的应用。

在网络设置中，通过设置损失函数来保证其循环一致性。以cycleGAN为例，其损失函数设置为：

$$\mathcal{L}_{\text{cyc}}(G,F)=\mathbb{E}_{x\sim p_{\text{data}}(x)}[\|F(G(x))-x\|_1] + \mathbb{E}_{y\sim p_{\text{data}}(y)}[\|G(F(y))-y\|_1]. \quad (2)$$

其中，x、y分别表示不同的图片域图片，G表示为“x→y”的生成器，F表示“y→x”的生成器。目标是使得重新生成的x/y与原始的x/y尽可能相似，即，损失函数值尽可能小。

3. 技术介绍

下面将详细介绍UNIT和cycleGAN网络。

3.1. UNIT

无监督的图像转换（UNsupervise Image-to-image Translation，UNIT）网络，最早便是通过UNIT提出并解决的。

UNIT致力于通过不同图像域的边缘分布数据，去学习它们的联合分布函数，从而实现图像域之间的转换。由于该联合分布具有多种可能，UNIT的网络设置参考了CoupleGAN[7]，并作出两个图片域共享隐空间的假设，引入循环一致性约束，同时以此将两个图片域的图片特征通过一种统一的方式表达。

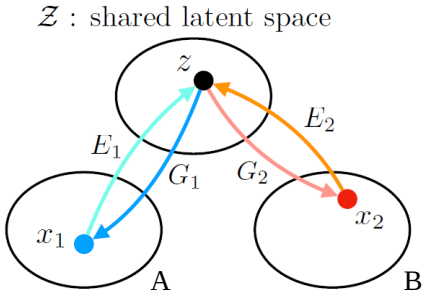

图 1：共享隐空间假设示意图

图 1 显示了共享隐空间的假设理念。其中，A、B表示两个图像域，x_1、x_2 分别表示图像域中的图片数据，z为共享隐空间中图片的统一特征表达向量，E_1、E_2 为将图片从图片域转换到隐空间的编码方式，G_1、G_2 为从隐空间生成图片的生成器。图像转换过程中，域A图片x_1 首先由E_1 编码为隐空间的特征向量z，再通过生成器G_2 将z生成为域B图片x_2，反之亦然。

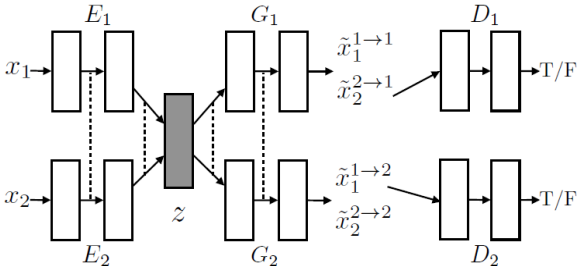

图 2：UNIT网络构架

UNIT整体网络设置参考了CoupleGAN的结构，引入两组GAN网络构成，并增加了编码器用于将图片编码到隐空间。如图 2 所示，图像域的图片通过编码器E转化为隐空间中的向量表示z，之后分为两个方向生成图片——生成另一图像域的图像和重新生成原始图像域图像，后者体现了循环一致性的约束。最后，将生成的图片输入判别器D，利用判别结果反馈更新生成器G和判别器D的参数。图中虚线部分框出编码器E和生成器G共享权重的部分。

在实际应用中，UNIT网络能够有效地进行图像域之间纹理、色彩相关的转换，例如：夏冬变换、街景的日夜变换等。但是，由于其优化过程需要寻找恰当的鞍点，

而鞍点寻找是个困难且不确定的任务，因此生成结果可能会有光斑、模糊、不自然等问题。

3.2. cycleGAN

与UNIT网络类似，cycleGAN也采用了一对GAN网络为基础，并引入循环一致性的约束，但没有做隐空间的假设。

cycleGAN构建了一组从图像域A转换到图像域B和从B转换到A的GAN网络，直接学习相互之间的映射关系，通过图片经过“A→B”和“B→A”的结果和原图片比较进行循环一致性的约束。

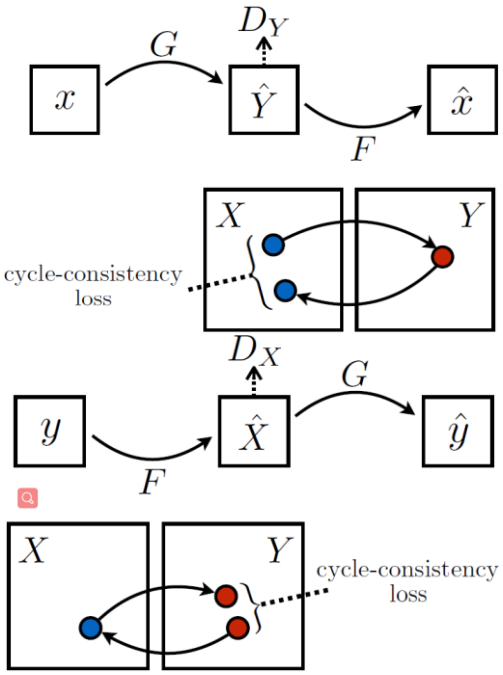

图 3：cycleGAN网络架构

图 3 展示了cycleGAN的网络设置。此处，X和Y表示两个不同的图像域，x、y对应表示所属图像域的图片，G和F分别表示“X→Y”和“Y→X”的映射（生成器），D_X、D_Y表示对应图像域的判别器。图像x经由G映射至图像域Y，得到图片$\hat{Y}$，该图片经过判别器D_Y判断并计算生成对抗损失（1），之后将$\hat{Y}$通过F映射回图像域X，计算循环一致性损失（2）。对图片y的处理同理。

相比UNIT，cycleGAN的结构更为简单，同等训练数据所消耗的训练时间更少。通过原文实验结果表明，cycleGAN可用于风格转换、色彩/纹理迁移等问题处理，但是对于几何变换并不敏感，如：猫狗之间的转换，生成结果极不自然。另，与UNIT相似，cycleGAN也面对着鞍点寻找问题，具有不稳定的特点。

4. 分析与比较

4.1. 实验设置

本实验构建两个自定义数据集，一个为十字绣和平面图相互转换数据集，另一个为机器绣花和平面图相互转换数据集。各数据集的两个类别均有 1000 张图片，按照 8:2 的比例随机分配训练集和测试集。

将两个数据集在UNIT和cycleGAN的经典网络上训练，每个网络均训练 200 个epoch，学习率设置为 0.002。通过分析其结果质量、运行时间等比较二者异同和优劣。

训练所用机器配置为：Intel(R) Core(TM) i7-7700K CPU @ 4.20GHz，内存 32G， NVIDIA GeForce GTX 1060 6GB GPU.

4.2. 结果与分析

1) 图像质量

训练结果详见附录A.

从主观视觉体验来说，cycleGAN的整体效果由于UNIT的训练结果。主要原因在于UNIT的训练结果有明显的色差，如：红色转换为黄色，而cycleGAN的结果色差变化比较小。经分析，cycleGAN加入了identity loss，用于风景照片和经典优化的转换训练中约束色彩改变，而我们的绣花图案和平面图转换训练与该训练目的类似——转换纹理而不转换色彩。实验证明，该约束有一定的效果。

从纹理转换的直观效果上看，cycleGAN的结果略优于UNIT。其纹理清晰度和分布的均匀程度在感官上都比UNIT的结果更好。究其原因，可能是因为cycleGAN的生成器采用了风格迁移的风格感知网络，利用感知损失提取图像特征，比UNIT使用传统的GAN生成器在本例中更具适应性。

但实际上，二者训练的结果都不尽人意。其可能的原因包括：一是循环一致性要求域A和域B的数据分布具有一定相似度，这一点在我们的数据集上体现得并不好，比如：图片内容多样，包含动植物、卡通等多种类，且两个图像域不同种类的分布比例不同；二是GAN网络训练自身的不稳定性限制了图像质量。GAN网络的训练是在生成器和判别器之间寻找平衡，其既要求判别器不能训练太好(判别器太好，生成器无法获得有效梯度更新)，又不能训练太差(判别器太差，训练结果振荡严重)，要找到中间的平衡，无异于大海捞针。

从客观数值分析来说，我们计算了各训练结果图的PSNR值和SSIM值。其中，PSNR为峰值信噪比，是使用最广泛的画质客观评判测量方法，但

其仅通过分析两张图片的数值差别进行评判，并不能很好地契合人眼对图片质量的评判，在本实验中仅作为一个参考项；而SSIM值为结构相似分析，主要分析图像之间内容结构的异同，其在亮度、对比度、结构三个方面建模评判，并根据人眼的误差敏感度在不同亮度中的差异进行建模计算，相比PSNR更具可信度。具体计算结果如下：

	cycleGAN – 十字绣	cycleGAN – 机器绣花	UNIT - 十字绣	UNIT - 机器绣花
PSNR	68.0051	**68.9325**	62.6154	58.9162
SSIM	0.999183	**0.999265**	0.999257	0.998535

表 1：cycle GAN和UNIT网络分别生成十字绣和机器绣花结果的PSNR分析和SSIM分析。

测评数据来看，cycle GAN的训练结果普遍由于UNIT。尽管在十字绣的SSIM得分上不相上下（一定程度上受到数据集质量不够好的影响），在机器绣花上cycle GAN有明显优势。

2) 时间分析

从整体时间来说，cycleGAN平均训练时间为30h上下，一个epoch需要约 9min的时间；而UNIT的平均训练时间为 45h，一个epoch需要约 13min的训练时间。从收敛速度来说，cycleGAN和UNIT均在 100 个epoch左右达到较好的视觉效果，由于UNIT单个epoch消耗时长大于cycleGAN，因此其收敛耗时较长。

可以理解，在单次迭代中，UNIT的网络结构设置更为复杂，计算量更大，相应地，其耗时也比cycleGAN大。由于UNIT在网络设置中，不是直接将图片域A映射至图片域B，而是设置了一个隐空间，将域A图片转换到隐空间，再转换到域B，无形中增加了转换步骤。因此，其耗时多余cucleGAN。

另一方面，由于UNIT和cycleGAN均利用了循环一致性和生成对抗的特点，其在迭代收敛的次数上具有不相上下的表现。

5. 总结

cycleGAN和UNIT都利用了生成对抗和循环一致性的思想，在两个不同图像域之间进行图像转换。在一定程度上，它们均能实现转换效果，但cycleGAN更偏向于风格化的纹理转换，UNIT更倾向于在色彩或几何改变的图像域之间进行转换。

在训练时间上，cycleGAN略优于UNIT，但它们的收敛所需迭代次数和训练效果相差不大。

此外，继承了GAN网络生成优点的同时，他们也有GAN网络不稳定的通病。虽然可以在无监督条件下训练不成对的数据集，但实际上分布更统一，更具约束性的数据集能够获得更好的训练结果。

参考文献

[1] Liu M Y, Breuel T, Kautz J. Unsupervised image-to-image translation networks[C]//Advances in Neural Information Processing Systems. 2017: 700-708.

[2] 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. 2017: 2223-2232.

[3] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Advances in neural information processing systems. 2014: 2672-2680.

[4] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. Conference on Computer Vision and Pattern Recognition, 2017.

[5] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin. Image analogies. In SIGGRAPH, 2001. 2, 3

[6] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015. 2, 3, 6

[7] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In NIPS, 2016. 3, 6, 7

[8] Y. Aytar, L. Castrejon, C. Vondrick, H. Pirsiavash, and A. Torralba. Cross-modal scene networks. PAMI, 2016. 3

[9] D. P. Kingma and M. Welling. Auto-encoding variational bayes. ICLR, 2014. 3

[10] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In CVPR, 2017. 3, 5, 6, 7

[11] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. In ICLR, 2017. 3, 8

[12] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. CVPR, 2016. 3, 8, 9, 14, 15

[13] E. L. Denton, S. Chintala, R. Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In NIPS, 2015. 2

[14] J.-Y. Zhu, P. Kr¨ahenb¨uhl, E. Shechtman, and A. A. Efros. Generative visual manipulation on the natural image manifold. In ECCV, 2016. 2

[15] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In NIPS, 2016. 2

[16] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In NIPS, 2016. 2

[17] Z. Kalal, K. Mikolajczyk, and J. Matas. Forwardbackward error: Automatic detection of tracking failures. In ICPR, 2010. 3

[18] R. W. Brislin. Back-translation for cross-cultural research. Journal of cross-cultural psychology, 1(3):185–216, 1970. 2, 3

[19] Q.-X. Huang and L. Guibas. Consistent shape maps via semidefinite programming. In Symposium on Geometry Processing, 2013. 3

[20] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In CVPR, 2017. 3

附录 A

图 4：机器绣花和平面图数据集。图像域 A 为机器绣花，图像域 B 为平面图。每张图片的第 1 行第 1、2 列和第 2 行第 3、4 列均为生成图像，其余为真实图像。左边为 cycleGAN 生成结果，右边为 UNIT 生成结果。

图 5：十字绣和平面图数据集。图像域 A 为十字绣，图像域 B 为平面图。每张图片的第 1 行第 1、2 列和第 2 行第 3、4 列均为生成图像，其余为真实图像。左边为 cycleGAN 生成结果，右边为 UNIT 生成结果。