Computer Science and Application 计算机科学与应用, 2022, 12(4), 934-940
Published Online April 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.124095

Hans汉斯

基于自适应归一化的语义图像合成

徐文锐，谭台哲

广东工业大学，计算机学院，广东 广州

收稿日期：2022年3月10日；录用日期：2022年4月12日；发布日期：2022年4月19日

摘　要

本文提出了一种可以实现风格控制的自适应归一化。它是一个简单但有效的模块，应用于以分割掩膜为条件的生成对抗网络。以前的方法将风格图像作为输入，输入到深度网络中。本文方法通过在归一化层输入风格信息来学习参数，以此调节归一化层的激活。本文在两个数据集上进行实验，并展示了部分结果。结果表明，本文方法可以根据语义分割掩膜合成符合语义布局和视觉逼真度高的图像，并以同一的模型实现不同风格的转换。

关键词

生成对抗网络，图像合成，风格转换

Semantic Image Synthesis with Adaptive Normalization

Wenrui Xu, Taizhe Tan

Department of Computer, Guangdong University of Technology, Guangzhou Guangdong

Received: Mar. 10th, 2022; accepted: Apr. 12th, 2022; published: Apr. 19th, 2022

Abstract

This paper presents an adaptive normalization method which can realize style control. It is a simple but effective module, which is applied to generate countermeasure network under the condition of segmented mask. The traditional method takes the style image as the input and inputs it into the depth network. In this paper, the method learns parameters by inputting style information in the normalization layer, so as to adjust the activation of the normalization layer. In this paper, experiments are carried out on two data sets, and some results are shown. The results show that this method can synthesize images with high semantic layout and visual fidelity according to the semantic segmentation mask, and realize the transformation of different styles with the

文章引用: 徐文锐, 谭台哲. 基于自适应归一化的语义图像合成[J]. 计算机科学与应用, 2022, 12(4): 934-940.
DOI: 10.12677/csa.2022.124095

same model.

Keywords

Generative Adversarial Networks, Image Synthesis, Style Transfer

1. 引言

视觉数据在日常生活中有着十分重要的角色。我们每天都要面对各种各样的视觉内容。学习生成真实的视觉数据将帮助我们理解和开发机器学习模型来感知视觉世界。此外，由于视觉数据的复杂性和多样性，真实视觉数据的自动生成仍然是一个重大挑战。图像等视觉数据的真实感合成在广告、游戏、虚拟现实等领域有着广泛的应用，长期以来一直是计算机视觉和计算机图形学的研究热点。随着对深度学习应用需求的增加，图像合成也显示出巨大的潜力。

在生成对抗网络(GANs) [1]提出以来，图像合成领域取得了许多令人印象深刻的进展。除了从随机潜在变量生成高质量图像的随机方法外，条件图像合成也因其可控制性的实际优势而受到同等甚至更多的关注。条件图像合成是以一定的输入数据作为条件，从数据集中生成新图像的任务。引导合成的条件输入可以有多种形式，包括 RGB 图像、语义标签、边缘/梯度图等。本文关注的是一种特殊形式的条件图像合成。具体来说是将语义分割掩膜转换为逼真的图像，这一过程被称为语义图像合成。语义图像合成有着广泛的应用，如内容生成和图像编辑[2] [3]。

之前的工作[2]提出了通用的图像到图像翻译框架内的解决方案，该框架直接将语义掩码提供给编码器－解码器网络。为了提高质量，最近的基于空间自适应归一化的语义图像合成(SPADE) [4]采用了空间变化的条件归一化，以避免传统归一化层造成的语义信息丢失。通过研究 SPADE 的模型结构并分析实验结果，本文发现了模型中的不足之处，SPADE 仅仅在网络最开始的部分输入风格信息。而在最近的方法[4]中已经证实将风格信息作为归一化参数输入到网络的多个层中可以起到更好的效果。本文的主要想法是将由风格信息获得的特征输入到网络的多个层中，并由此在不同的层中获得不同的归一化参数。为了评价所提出的方法，本文在人脸数据集 CelebAMask-HQ 和场景数据集 ADE20K 进行了实验。结果表明，本文的方法获得了高质量的结果。

2. 相关工作

2.1. 生成对抗网络

最新的深度生成模型包括变分自动编码器(VAE) [5]和生成对抗网络(GAN) [1]。GAN 由生成器和鉴别器组成，其目标是生成真实的图像，从而使鉴别器无法区分合成的图像和真实的图像。随着 GAN 结构、正则化和损失函数的不断改进，GAN 合成的图像越来越逼真。例如，由 StyleGAN [6]生成的人脸图像质量非常高，不仔细对比几乎无法与真实图片区分开来。最初，GANs 只能生成从随机分布中抽取的样本，因此缺少用户控制能力，但很快出现了能够进行条件图像合成的模型。用户可以通过向生成器提供条件信息来控制合成。本文的方法建立于 GANs 上。

2.2. 图像到图像翻译

图像到图像翻译是一种学习数据分布以生成新样本的方法，以图像作为条件的 GAN 被视作各种图像到图像转换问题的通用解决方案。语义图像合成是众多图像到图像的翻译问题中的一种特殊类型，它可以通过修改输入的语义布局图像来方便用户控制[7]。针对这一任务，迄今为止已有许多出色的方法。其中最具代表性的是 Pix2Pix [2]，它采用编码器-解码器结构进行统一的图像到图像的转换。Pix2pixHD [3]通过提出从粗到细的生成器和判别器改进了 Pix2Pix。随后的方法[8]进一步探索了如何从语义掩膜中合成高质量的图像，并取得了显著的改进。最近的 SPADE [4]通过改进归一化层，在语义图像合成中得到了比以往更高质量的合成图片。本文的想法是通过改变风格信息的输入方式来改进 SPADE。

2.3. 风格转换

图像到图像翻译问题的一种变体是引入附加的指导图像，由此获得更多的用户控制能力。这种指导图像可以采取多种形式，可以是风格转换问题中的风格图像。目前的方法主要分为三处地方进行风格编码：1) 图像特征的统计[9]；2) 网络权值(如快速样式传递[10])；3) 归一化层参数)。第一类方法对图像分类网络提取的图像特征进行匹配统计，优化过程缓慢，时间开销大。第二类方法的泛化能力弱，需要为每一种风格的图像训练一个单独的神经网络。第三类方法不存在前两类方法的缺点，此类方法如 StyleGAN 和 SPADE 可以实现任意风格转换。因此本文的风格转换也建立在归一化层的基础上。

3. 语义图像合成

给定一个输入风格图像及其对应的分割掩膜，本节展示如何根据提取风格图像中的风格信息与分割掩膜中的语义信息来合成逼真图像。

3.1. 自适应归一化

设 Ireal 为真实图像，Iz 为从 Ireal 经过 Encode 所得的隐含量分布采样而来的特征图。本文提出一种类似批归一化的自适应归一化。与批归一化类似，自适应归一化以通道方式进行归一化，然后学习尺度参数 γ 和偏差参数 β 进行调整。图 1 展示了自适应归一化的主要构成。自适应归一化层的激活值为

$$\gamma \frac{h-\mu}{\sqrt{\sigma^2+\varepsilon}}+\beta \tag{1}$$

其中，μ 和 σ 为各个通道激活的平均值和标准差，h 为前一层的激活值，ε 为一个给定小值。

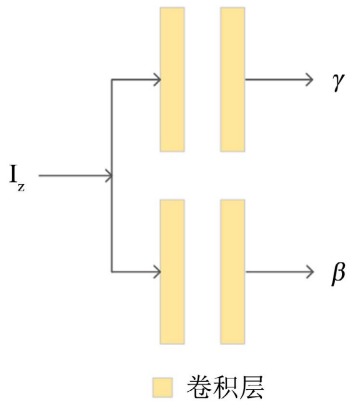

Figure 1. Adaptive normalization
图 1. 自适应归一化

图 2 展示了自适应归一化残差层的整体结构，残差层基本遵循方法[11]和方法[12]的设计，由卷积层、归一化层和激活层组合构成。图 2 下半部分表明残差层也学习了跳过连接。

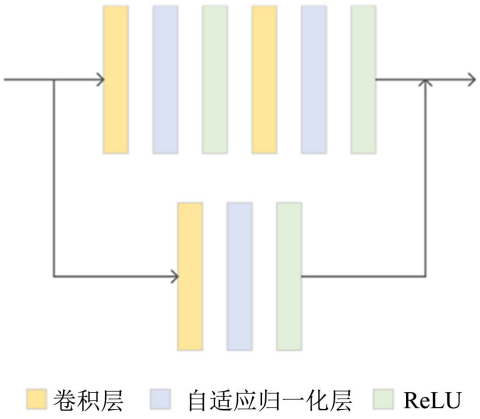

Figure 2. Residual layer with adaptive normalization
图 2. 自适应归一化残差层

3.2. 生成器

如图 3，本文方法的生成器采用编码器-解码器结构，对方法[10]中提出的生成器进行了调整。调整包括两个下采样层、一个由多个残差层组成的语义核心、两个上采样层和一个 Encoder 部分。残差层分为自适应归一化残差层和 SPADE 残差层。自适应归一化残差层与 Encoder 组合，以真实图像作为额外信息学习归一化层的调整参数。在进行风格转换任务时，作为额外信息的真实图像就是风格图像。对 SPADE 残差层添加掩膜图像作为额外信息，SPADE 残差层中的归一化层的参数都使用 SPADE 进行学习。传统归一化层倾向于消去输入中的语义信息，SPADE 的目的是尽可能保留语义信息。

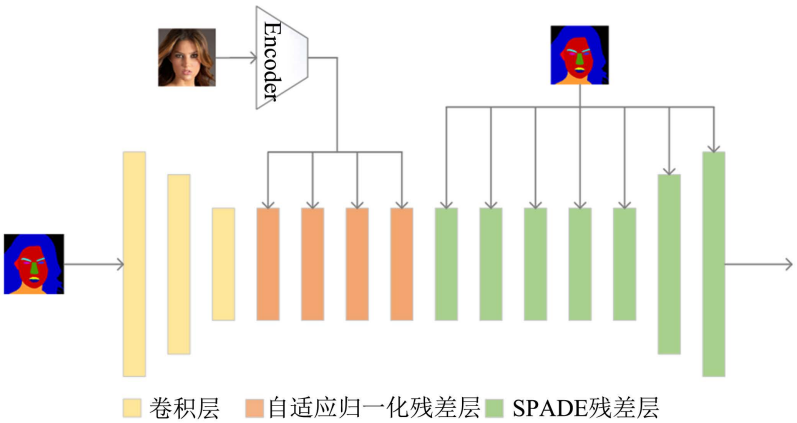

Figure 3. Generator
图 3. 生成器

3.3. 判别器

本文使用与 pix2pixHD [3]相同的多尺度判别器，它将掩膜图像和需要判别图像拼接后作为输入。每个判别器都基于 PatchGAN [2]，判别器的最后一层是卷积层，所以最后的输出并不是一个标量，而是一个矩阵。这样有利于实现更高分辨率的图像生成。

3.4. Loss

本文使用以下损失以对抗的方式训练生成器：作为对抗损失的铰链损失[13] [14] [15]以及特征匹配损失[3]、感知损失[10]和 KL 散度损失。其中，使用 KL 散度损失训练 Encoder 部分，目的是用于进行风格引导图像合成训练任务。

4. 实验

4.1. 实现细节

继 SPADE [4]之后，本文将谱范数[14]应用到生成器和判别器的所有层中。基于方法[15]的研究结果，本文将生成器的学习速率设置为 0.0001，判别器的学习速率设置为 0.0004。对于优化器，本文选择 $\beta_1 = 0$，$\beta_2 = 0.999$ 的 ADAM。所有实验都是在 2 张 NVIDIA RTX 3090 GPUs 上进行的。此外，本文使用的是批归一化的同步版本。

4.2. 数据集

本文在实验中使用了以下的数据集：1) CelebAMask-HQ 包含了用于 CelebAHQ 人脸图像数据集的 30,000 个分割掩码，有 19 个不同的地区类别。2) ADE20K 由 20210 个训练图像和 2000 个验证图像组成，包括 150 个不同语义类的场景。

4.3. 评价指标

本文采用以下指标对本文方法进行评价：

1) Frechet Inception 距离(FID)，FID 用于计算生成器生成分布与真实图像分布之间的差异，FID 越小，则图像多样性越好，质量也越好。计算 FID 需要使用 inception network。可通过以下的公式计算：

$$\mathrm{FID}(x,g) = \left\|\mu_x - \mu_g\right\| + Tr\left(\Sigma_x + \Sigma_g - 2\sqrt{\Sigma_x\Sigma_g}\right) \tag{2}$$

其中，μ_x 和 Σ_x 分别为真实图像在 inception network 输出特征的均值和协方差矩阵，μ_g 和 Σ_g 分别为合成图像在 inception network 输出特征的均值和协方差矩阵，Tr 为矩阵的迹。

2) 通过平均交并比(mIoU)和像素精度(accu)测定的分割精度，mIoU 和 accu 越大图像质量越高。为了计算分割精度，本文使用了目前主流的分割网络 DeepLabV2，用其生成的分割结果计算 mIoU 和 accu 两个性能指标。mIoU 计算公式为：

$$\mathrm{mIoU} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij} + \sum_{j=0}^{k}p_{ji} - p_{ii}} \tag{3}$$

其中 i 为真实值，j 为预测值，p_{ij} 为将 i 预测为 j，k 为类别数。此外，accu 的计算公式为：

$$\mathrm{accu} = \frac{\sum_{i=0}^{k}p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k}p_{ij}} \tag{4}$$

与式(3)类似，k 为类别数，p_{ij} 为真实像素类别为 i 的像素被预测为类别 j 的总数量。

4.4. 定性结果与定量结果

表 1 展示了本文方法与最近的方法在 CelebAMask-HQ 数据集上进行语义图像合成的实验数据对比。可以观察到，本文方法在平均交并比和像素精度比 Pix2PixHD 和 SPADE 两种方法略低，但 FID 低于这两种方法。所以本文方法的合成图像质量与 Pix2PixHD 和 SPADE 相差无几，但多样性要优于这两种方法。

Table 1. CelebAMask-HQ experimental data
表 1. CelebAMask-HQ 实验数据

方法	FID	mIoU	accu
Pix2PixHD	24.37	74.17	95.05
SPADE	22.78	76.58	95.87
本文	21.11	73.57	94.96

本文在图 4 中展示了本文方法用于 CelebAMask-HQ 数据集的结果可视化对比。可以观察到，对网络输入标签图像后输出的合成图像具有较高的图像保真度，合成图像的脸部细节质量高，只在占比很少的图像会存在与真实情况不太符合的部分。

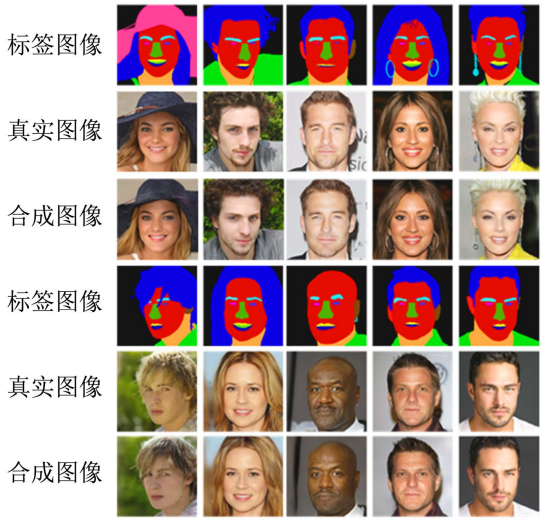

Figure 4. Synthetic image of CelebA-HQ experiment
图 4. CelebA-HQ 实验的合成图像

图 5 展示了使用 ADE20K 数据集训练风格转换任务的部分结果。(a)是真实图像，为了更直观的对比，真实图像内还加上了对应的分割掩膜。(b)~(f)是在输入不同风格图像时输出的合成图像。可以观察到，通过输入不同的风格图像，本文模型的合成图像具有不同外观，但在输入掩膜中都具有相同的语义布局。作为参考，输入分割掩膜显示在真实图像之内。

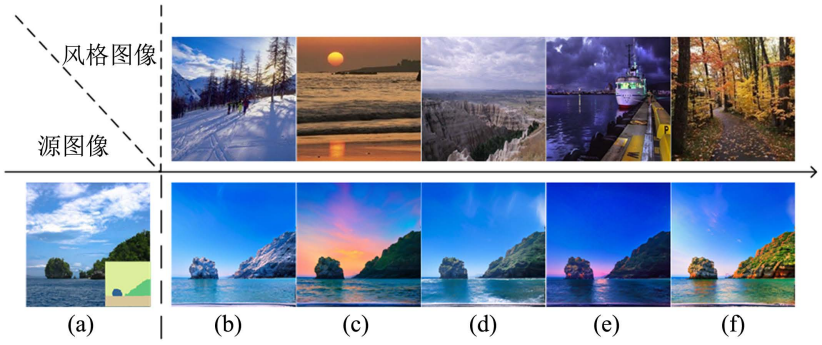

Figure 5. Style transfer
图 5. 风格转换

5. 结论

本文提出了一种用于风格控制的自适应归一化，这是一种简单而有效的生成对抗网络(GANs)构建块，它基于对部分模块的归一化层输入风格信息。主要想法是扩展最近的 SPADE 网络，以多个通道的方式控制语义的风格。该模型可以产生人脸和景观的逼真图像，并且还可以对图像进行风格控制。在公开数据集上进行的实验结果表明，本文提出的语义图像合成方法相比于目前的主流算法能够生成更精准、更多样性的合成图像。在保留用户控制能力的基础上，生成的图像与真实图像更为接近。

参考文献

[1] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., *et al*. (2014) Generative Adversarial Networks. *Advances in Neural Information Processing Systems*, **3**, 2672-2680.

[2] Isola, P., Zhu, J.Y., Zhou, T., *et al*. (2016) Image-to-Image Translation with Conditional Adversarial Networks. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 5967-5976. https://doi.org/10.1109/CVPR.2017.632

[3] Wang, T.C., Liu, M.Y., Zhu, J.Y., *et al*. (2017) High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 8798-8807. https://doi.org/10.1109/CVPR.2018.00917

[4] Park, T., Liu, M.Y., Wang, T.C., *et al*. (2019) Semantic Image Synthesis with Spatially-Adaptive Normalization. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 2332-2341. https://doi.org/10.1109/CVPR.2019.00244

[5] Kingma, D.P. and Welling, M. (2014) Auto-Encoding Variational Bayes. arXiv:1312.6114.

[6] Karras, T., Laine, S. and Aila, T. (2019) A Style-Based Generator Architecture for Generative Adversarial Networks. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 4396-4405. https://doi.org/10.1109/CVPR.2019.00453

[7] Lee, C.H., Liu, Z., Wu, L., *et al*. (2019) MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 5548-5557. https://doi.org/10.1109/CVPR42600.2020.00559

[8] Qi, X., Chen, Q., Jia, J., *et al*. (2018) Semi-Parametric Image Synthesis. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 8808-8816. https://doi.org/10.1109/CVPR.2018.00918

[9] Gatys, L.A., Ecker, A.S. and Bethge, M. (2015) A Neural Algorithm of Artistic Style. *Journal of Vision*, **16**, 326. https://doi.org/10.1167/16.12.326

[10] Johnson, J., Alahi, A. and Li, F.F. (2016) Perceptual Losses for Real-Time Style Transfer and Super-Resolution. *European Conference on Computer Vision*, Amsterdam, 11-14 October 2016, 694-711. https://doi.org/10.1007/978-3-319-46475-6_43

[11] Mescheder, L., Geiger, A. and Nowozin, S. (2018) Which Training Methods for GANs Do Actually Converge? arXiv:1801.04406.

[12] Miyato, T. and Koyama, M. (2018) cGANs with Projection Discriminator. arXiv:1802.05637.

[13] Lim, J.H. and Ye, J.C. (2017) Geometric GAN. arXiv:1705.02894.

[14] Miyato, T., Kataoka, T., Koyama, M., *et al.* (2018) Spectral Normalization for Generative Adversarial Networks. arXiv:1802.05957.

[15] Zhang, H., Goodfellow, I., Metaxas, D., *et al.* (2018) Self-Attention Generative Adversarial Networks. arXiv:1805.08318.