基于多尺度特征融合的人脸照片 – 素描合成

梁昌城¹, 王楠楠^{1*}, 朱明瑞^{1*}, 杨曦¹, 李洁¹, 高新波²

1. 西安电子科技大学综合业务网理论及关键技术国家重点实验室, 西安 710071

2. 重庆邮电大学重庆市图像认知重点实验室, 重庆 400065

* 通信作者. E-mail: nmwang@xidian.edu.cn, mrzhu@xidian.edu.cn

收稿日期: 2021-03-10; 修回日期: 2021-04-30; 接受日期: 2021-05-28; 网络出版日期: 2022-01-25

摘要 从真实的人脸照片合成面部素描及其逆过程具有广泛的用途, 例如数字娱乐与协助刑事案件的侦查. 但是, 由于照片与素描在纹理上的显著差异, 它们之间的互相转换仍是一个具有挑战性的问题. 最近基于生成对抗网络的方法已在图像间转换问题, 特别是照片到素描的转换方面展现出令人鼓舞的结果, 但它们大多会在面部关键组件产生不同的形变或者模糊, 使得合成图像的真实性受影响. 为了应对这一挑战, 我们提出了一种新颖的基于多尺度特征融合的人脸照片 – 素描合成算法, 来提高合成图像的结构完整性与纹理逼真度. 首先使用编码器提取输入图像的多尺度编码特征, 然后将最底层编码特征经过空洞卷积模块后传入解码器进行解码. 解码过程中将不同尺度的解码特征与对应尺度的编码特征在通道维度上拼接, 从而获得多尺度编解码融合特征. 最后在解码器的输出端将不同尺度的编解码融合特征进一步融合, 并通过一层卷积层产生最终合成结果. 通过这种同时将编码 – 解码过程中不同尺度的特征在通道维度进行拼接的方式, 能够保持较好的图像结构以及纹理细节, 生成逼真的面部素描/照片图像. 我们在多个具有挑战性的数据集中验证了所提方法的有效性. 定量和定性评估表明, 本文模型在生成具有高视觉质量的人脸素描 (或照片) 方面优于其他最新技术.

关键词 人脸照片 – 素描合成, 图像翻译, 生成对抗网络, 多尺度特征融合, 空洞卷积

1 引言

素描是最基本的艺术绘画风格之一, 其中人脸照片 – 素描合成是指在给定一个面部照片的情况下获得合成的素描风格的人脸图像, 它是计算机视觉中的重要任务, 在数字娱乐和刑侦追凶^[1] 等方向具有非常广泛的应用. 例如, 人们将线条优美的人脸素描画作为头像或者个人资料发布在社交媒体中, 素描图像正变得越来越流行. 此外, 在执法和刑事案件中, 监视摄像机有时无法很好地记录嫌疑犯的面部照片. 在这种情况下, 经验丰富的艺术家通常会根据目击者的描述绘制犯罪嫌疑人的面部素描. 但是, 由于面部照片和素描之间的纹理外观存在很大差异, 因此很难根据嫌疑犯的面部素描在保存有

引用格式: 梁昌城, 王楠楠, 朱明瑞, 等. 基于多尺度特征融合的人脸照片 – 素描合成. 中国科学: 信息科学, 2022, 52: 334-347, doi: 10.1360/SSI-2021-0085
Liang C C, Wang N N, Zhu M R, et al. Face photo-sketch synthesis based on multi-scale feature fusion (in Chinese). Sci Sin Inform, 2022, 52: 334-347, doi: 10.1360/SSI-2021-0085

犯罪嫌疑人照片的数据库中准确地检索犯罪嫌疑人^[2]. 因此, 旨在自动从面部照片(素描)合成面部素描(照片)的人脸照片-素描合成已经引起了广泛的关注.

人脸照片-素描合成的难点主要在于合成视觉上逼真的图像以及语义上合理的图像, 而且, 克服照片与素描之间明显的差异(颜色、形状和纹理细节)也具有挑战性. 早期的关于人脸照片-素描合成的研究工作主要是集中在将参考图像对中的样本图像(图像块)进行匹配与组合^[3~5], 以合成目标图像. 这些方法在受约束条件下(例如较少的姿势、照明变化)可以较好地工作, 但遇到复杂的条件时会失败. 此外, 这些方法还具备两个明显的缺陷: (1) 模糊或过于平滑; (2) 耗费时间. 随着深度神经网络(deep neural networks, DNN)的快速发展, 特别是生成对抗网络(generative adversarial nets, GAN)^[6], 启发了近期的工作^[7,8]将人脸照片-素描合成任务视作图像到图像的翻译^[9,10]问题. 基于生成对抗网络的方法能够生成具有逼真纹理的图像, 这都得益于对抗性损失. 但是, 通过仔细检查现有方法合成的图像, 我们可以观察到嘴巴和毛发区域的严重变形和混叠缺陷, 合成的照片/素描通常会具有令人不愉快的模糊, 这样的伪像导致合成的素描不真实. 用现有方法合成的素描(照片)的性能还远远不够理想. 我们发现两个主要因素限制了现有模型的性能: (1) 在图像生成模型中采用的编码器-解码器结构会导致一些结构信息的丢失. (2) 在这些模型的卷积层中实现了单一大小的感受野, 无法捕捉到不同的上下文信息. 结构信息和感受野是深度卷积神经网络实现高性能的关键, 然而, 这两种需求很难同时满足.

为了应对这一挑战, 本文提出了一种基于生成对抗网络的新型高质量人脸照片-素描合成框架, 它通过将网络训练过程中的多尺度特征进行融合来提高人脸照片-素描合成效果. 具体来说, 提出了一种充分利用图像编码-解码过程中多尺度特征的生成器, 通过将解码过程中的不同尺度的特征提取出并在解码器的输出端进行融合, 从而利用了图像解码过程中的全局区域的多尺度特征. 其次, 在生成器的瓶颈层中使用了一种高效的空洞卷积模块来提取局部和全局特征的结构. 我们提出的方法能够合成具有合理的纹理细节的素描(照片), 因此, 与最新的方法相比, 可以获得更佳的视觉效果. 本文工作的主要贡献如下: (1) 基于生成对抗网络框架, 提出了一个充分利用编码-解码过程中的多尺度特征的高效的 U 型网络结构, 利用该网络能合成具有合理的纹理细节的人脸素描(照片); (2) 在生成器的瓶颈层处, 使用了一个能够更好地提取图像局部与全局特征的结构空洞卷积模块; (3) 定量和定性实验表明, 与真实环境中的其他公共基准和人脸图像技术相比, 本文所提出的方法在性能上优于其他最新方法.

本文第 2 节介绍了人脸照片-素描合成的相关工作; 第 3 节详述了本文的基于多尺度特征融合的人脸照片-素描合成网络的框架与技术细节; 第 4 节阐述了网络中使用的损失函数的构建; 第 5 节给出了实验讨论与结果的分析; 最后一节则是对全文的总结.

2 相关工作

现有的人脸照片-素描合成方法主要可以分为两大类. 第 1 类是传统的基于样例的方法, 通过挖掘输入图像(图像块)和参考照片-素描对集合中的图像(图像块)之间的对应关系来重建目标图像. 第 2 类则是试图通过基于学习的方法来解决合成问题, 例如训练深层卷积神经网络从源图像像素预测目标图像像素.

2.1 传统的基于样本的方法

传统的基于样本的方法可以进一步分为 3 类: 基于子空间学习的方法^[3]、基于稀疏表示的方法^[4]

和基于贝叶斯 (Bayes) 推理的方法^[5]. 现有的基于样本的方法的详细描述可以在 Wang 等^[1] 的工作中找到. 传统方法的两个基本问题是: (1) 如何在参考集合中匹配最相似的图像 (图像块); (2) 估计匹配图像 (图像块) 的最准确的重建权重值. 对不同问题的处理方式导致了不同方法的出现. 在这些方法中, 基于贝叶斯推理的方法取得了最好的性能, 因为它们成功地使用概率图模型对局部相似性和空间一致性进行联合建模. Wang 等^[2] 首次将 Markov 随机场 (Markov random field, MRF) 引入人脸素描合成. Zhou 等^[11] 提出了一种 Markov 权重场 (Markov weight fields, MWF) 模型, 将线性组合引入到 Markov 权重场模型中. Zhu 等^[5] 建议使用卷积神经网络特征代替像素灰度值, 观察到卷积神经网络特征对环境噪声的鲁棒性更强, 该方法能够合成噪声小的目标图像, 在现有的基于样例的方法中取得了最好的性能. 然而, 模糊效果和耗时是困扰该方法和其他基于样例的方法的两个主要缺点.

2.2 基于深度学习的方法

最近, 基于卷积神经网络和基于生成对抗网络的方法已经成为很有前途的人脸照片 – 素描合成方法. Zhang 等^[12] 最初尝试训练端到端全卷积神经网络 (fully convolutional networks, FCN), 用于直接建模人脸照片和人脸素描之间的非线性映射. 然而, 由于受到网络层数较浅和像素级损失的限制, 它不能捕捉到纹理细节, 也不能保持合理的结构. Isola 等^[9] 将条件生成对抗网络 (conditional generative adversarial nets, CGAN) 作为统一的解决方案, 用于边缘到照片、标签到街景、白天到夜晚等几种图像到图像的翻译任务. Zhu 等^[10] 通过引入循环一致性损失, 提出了一种非配对图像到图像转换的 CycleGAN 模型, 这两个模型可以直接应用于面向图像合成的任务. 几种方法^[7,13] 遵循图像到图像翻译的思路, 专注于通过添加先验信息来提高人脸照片 – 素描的合成性能. Wang 等^[7] 提出了一种多尺度鉴别器, 实现对不同图像分辨率的对抗性监督. Zhang 等^[8] 提出了一种由粗到精的两阶段合成目标草图的方法. PGAN^[13] 试图通过在合成过程中嵌入关键照片先验并设计一种光照补偿策略在光照不受控制的条件下合成图像. CA-GAN^[14] 引入面部成分信息作为附加输入, 以帮助生成素描肖像, 并提出基于面部成分信息的成分损失. 为了解决缺乏配对训练数据的问题, Chen 等^[15] 研究并提出了一种半监督学习方法, 通过合成附加训练照片的伪草图特征来扩充成对的训练样本, 并学习它们之间的映射函数. 在任务相似的人脸肖像画生成任务上, Yi 等^[16] 提出一种多生成器网络, 分别用于生成人脸不同组件, 再将不同组件融合成完整人脸肖像画. 在任务相似的人脸卡通化任务上, Kim 等^[17] 根据辅助分类器得到人脸的注意力, 通过区分源域和目标域, 帮助模型获得集中转换区域, 并提出了一种新的自适应归一化层, 实现人脸卡通化的效果. Gou 等^[18] 提出了一种面向图像恢复的多尺度神经网络结构搜索方法, 通过多尺度的搜索空间上构建一个超网络, 并利用基于梯度优化的搜索策略, 直接在特定的图像恢复任务上进行网络结构搜索, 这种利用多尺度特征融合的方法在图像恢复任务上取得了令人满意的恢复效果. 虽然上述的人脸照片素描合成方法已经取得了很大的进展, 但仍然存在一些不理想的伪影和扭曲的结构, 特别是在真实场景的结果中.

3 基于多尺度特征融合的人脸照片 – 素描合成网络

受生成对抗网络在人脸照片 – 素描合成领域工作的启发, 本文提出了一种基于生成对抗网络框架的人脸照片 – 素描合成网络结构. 所提出的方法能够处理人脸素描合成和照片合成, 因为这两个过程是对称的. 在本小节中, 以人脸素描合成为例来介绍我们的方法. 该网络生成器部分由基于 U-Net^[19] 的 U 型编码 – 解码器和放置于瓶颈位置的空洞卷积模块构成, 其中生成器的作用是合成逼真的人脸素描合成图像, 判别器的作用则是对输入的人脸素描进行判别, 判断输入图像是生成器合成的人脸素描

描还是真实的人脸素描. 通过生成器持续地试图使判别器无法得出正确的判断, 判别器持续地获得合成的人脸素描与真实的人脸素描并不断地给出判定结果, 从而利用判别结果的误差并通过梯度下降法使神经网络参数得以优化, 最后促使生成器合成逼近真实人脸素描的合成素描图像.

基于多尺度特征融合的人脸照片-素描合成网络中的生成器以真实人脸照片作为输入, 经过编码器编码得到其特征图并将所得特征图输入到空洞卷积模块, 解码器将空洞卷积模块的输出以及编码器中的不同尺度特征一起作为输入, 最终输出合成的人脸素描. 该方法以真实人脸照片作为输入, 采用 U 型的生成对抗网络结构并结合了本文提出的空洞卷积模块作为生成器瓶颈, 同时利用最小二乘生成对抗网络损失作为对抗损失使得网络能够稳定地训练且生成真实的人脸素描. 本文的人脸照片-素描合成网络中的生成器采用编码器-解码器结构, 该结构可以有效提取输入的人脸照片的局部纹理结构信息以及全局轮廓特征信息, 并将所提取的信息进行编码从而获得具有更小规模的精简信息, 在经过空洞卷积模块后, 所需的纹理细节特征信息以及全局轮廓信息得以保留, 最后通过编码器的重构生成合成的人脸素描图像. 该网络的编码器与解码器结构采用的是与 U-Net 相似的 U 型编码-解码结构, 该结构能够较好地保持输入图像的细节纹理信息并同时提取全局特征.

3.1 基于多尺度特征融合的 U 型编码-解码结构

在编码器和解码器的构建中, 本文采用 U-Net 编码-解码结构作为生成器的框架, 编码器与解码器构成一个 U 型结构, 本文在 U-Net 原有基础上作出改进, 使得该编码器-解码器结构能够更加有效地解决编码-解码过程中的结构信息丢失问题以及更好地捕捉不同的上下文信息.

如图 1 所示, U 型网络结构主要由卷积单元组成, 网络左半边为编码器部分即由 1 个 7×7 卷积层以及 5 个 3×3 卷积层组成, 此时进行的是下采样阶段, 通过对输入图像进行下采样获得包含全局结构信息的多尺度特征. 网络的右半边为解码器结构, 解码器部分也是由 1 个 7×7 卷积层以及 5 个 3×3 卷积层组成, 此时进行的是上采样阶段, 上采样阶段将下采样过程中提取的特征与瓶颈输出的深层特征在通道维度拼接, 从而将多尺度的特征融合. 以最后一个上采样为例, 它的特征既来自下采样阶段的第 1 个下采样卷积层的输出, 也来自上采样的输出. 在解码阶段, 将解码过程中的中间特征提取出来, 这些中间特征既包含了浅层特征, 也包含了深层特征, 然后将提取出来的不同尺度的特征上采样到相同的 256×256 大小, 最后将尺寸大小相同的不同深浅层的多尺度特征在最后一个上采样层的输出处按通道维度进行拼接操作, 在经过最后一个 7×7 卷积层后得到合成的人脸素描, 通过同时将编码过程和解码过程的多尺度特征在通道维度上进行融合, 实现了对编码-解码过程中间多尺度特征的充分利用.

3.2 空洞卷积模块

仅通过对图像进行下采样操作难以从图像的浅层获得特征图的全局信息, 为了从图像的浅层获得更多的全局信息, 最直接的想法就是扩大卷积层的感受野. 如图 2 所示, 采用了一个空洞卷积模块作为生成器的瓶颈, 该空洞卷积模块由 8 个感受野大小同为 3×3 但膨胀大小不同的卷积层构成, 分别为: 3 个膨胀大小为 1, 2 个膨胀大小为 2, 2 个膨胀大小为 4, 1 个膨胀大小为 8 的卷积层. 空洞卷积模块内部通过 U 型结构以及首个卷积层至最后一个卷积层之间的残差连接来保持特征图中的精细细节并尽可能使结构信息得以保留. 通过采用空洞卷积模块, 我们能够获得更大范围的感受野和更丰富的局部和全局特征, 这也有益于后续的解码过程.

图 1 (网络版彩图) 生成器网络结构示意图
Figure 1 (Color online) Generator network structure diagram

图 2 (网络版彩图) 空洞卷积模块结构示意图
Figure 2 (Color online) Dilated convolution structure diagram

3.3 判别器结构

在人脸素描合成网络中, 判别器需要尽可能准确地判断输入的图像是合成的人脸素描还是真实的人脸素描. 本文的判别器部分采用的是 PatchGAN^[9], 判别器由 5 个感受野大小为 4×4 的卷积层组成, 通过将合成图像与真实图像分别输入判别器并提取它们的特征, 将所得特征用于对抗性损失的计算, 判别器部分输入的是尺寸大小为 70×70 的补丁块.

4 损失函数

在损失函数的构建中, 考虑到神经网络训练过程中的稳定性以及网络的收敛速度, 我们的全部损失函数由对抗性损失和感知损失两个部分组成.

4.1 对抗性损失

在这项工作中, 使用最小二乘 GAN^[20] 进行稳定训练, 而不是普通的 GAN. 对于映射函数 G 及其鉴别器 D_Y , 将目标函数表示为

$$\mathcal{L}_{\text{adversarial}} = \mathbb{E}_y[(D_Y(y))^2] + \mathbb{E}_x(x)[(1 - D_Y(G(x)))^2]. \quad (1)$$

4.2 感知损失

合成图像和目标图像应该具有相似的高频表示和相同的身份, 这对于人类感知质量是十分重要的. 因此, 我们使用类似于 Johnson 等^[21] 的特征表示法, 将感知损失表示如下:

$$\mathcal{L}_{\text{perceptual}} = \mathbb{E}_x \left[\frac{1}{C_j H_j W_j} \|\phi_j(G(x)) - \phi_j(y)\|_1 \right], \quad (2)$$

其中 ϕ_j 分别表示预先训练的 VGG-19^[22] 模型的第 j 层, C_j , H_j 和 W_j 分别表示通道编号、特征图的高度和宽度. 本文所采用的 VGG-19 模型是在 ImageNet 数据集上预训练后的 VGG-19 模型, ImageNet 数据集包含 21841 个类别的自然图像, 经该数据集预训练后的 VGG-19 模型具有良好的特征提取能力.

VGG 不同层中的特征包含高频和身份信息, 将浅层的感知损失降至最低往往会产生包含与合成素描 (照片) 无法区分的纹理图像. 相比之下, 对较深层特征计算感知相似度会保持面部特征和整体空间结构. 因此, 在我们的实验中, 选择了浅层和深层, 同时使用了浅层和深层特征, 将它们的感知损失降到最低能够促使输出图像与目标图像在感知上类似, 并且具有相同身份.

4.3 总损失函数

通过合并以上损失, 总损失函数表示如下:

$$\mathcal{L}_{\text{full}} = \lambda_1 \mathcal{L}_{\text{adversarial}} + \lambda_2 \mathcal{L}_{\text{perceptual}}. \quad (3)$$

在本工作中, 为了使相应损失函数保持在相同数量级, 设置 $\lambda_1 = 1$, $\lambda_2 = 5$.

5 实验结果与分析

本节首先对实验设置和评估方法进行了详细的讨论. 然后, 在两个公开的数据集上进行了定性和定量实验验证, 这两个公开的数据集分别是: CUFS^[2] 和 CUFSF^[23]. 在人脸照片合成素描和人脸素描合成照片两种情况下都进行了定量比较.

5.1 实验设置

本文在两个公共数据集上进行了实验, 分别为 CUFS 和 CUFSF 数据集. CUFS 数据集由来自香港中文大学 (CUHK) 学生数据库的 188 个人脸^[24], 来自 AR 数据库的 123 个人脸^[25], 和来自 XM2VTS 数据库的 295 个人脸^[26] 组成. 在正常光照条件下, 每个面部都有一对照片 - 素描, 并带有中性表情. CUFSF 数据集由来自费列特数据库的 1194 张人脸组成^[27]. 每张脸都有一张灯光变化的照片和一张艺术家绘制的形状夸张的素描. 因此, 该数据集比 CUFS 数据集更具挑战性. 所有图像都是通过将眼睛中心对准固定位置并裁剪到 200×250 的大小来处理的. 在 CUHK 学生数据库中, 随机选取了 88 对人脸照片 - 素描进行训练, 其余的用于测试. 对于增强现实数据库, 随机选择 80 对人脸照片 - 素描进

表 1 数据集的划分
Table 1 Partitioning of datasets

	Dataset	Training pairs	Testing pairs
CUFS	CUHK student	88	100
	AR	80	43
	XM2VTS	100	195
	CUFSF	250	944

行训练, 其余 43 对用于测试. 对于 XM2VTS 数据库, 随机选择 100 对进行训练, 剩下的 195 对用于测试. 对于 CUFSF 数据集, 选择 250 个人脸照片 – 素描对进行训练, 其余的用于测试.

所有模型均在 NVIDIA GTX 1080-Ti GPU 上进行训练, $\beta_1 = 0.5$, $\beta_2 = 0.99$. 我们以 0.0002 的固定学习率训练所有模型, 直到 300000 次迭代. 所有实验的批大小设置为 1, 权重从均值为 0, 标准差为 0.02 的高斯 (Gauss) 分布中初始化. 我们将输入图像的大小缩放到 256×256 , 并将像素值归一化为 $[-1, 1]$ 间隔, 然后将它们放入模型中. 在训练过程中, 在每次迭代中交替更新 G 和 D .

本文使用 3 种类型的评价指标来评价合成图像的客观质量: 学习感知图像块相似性 (learned perceptual image patch similarity, LPIPS) [28]、弗雷歇起始距离 (Fréchet inception distance, FID) [29] 和特征相似性指数 (feature similarity index mersure, FSIM) [30]. LPIPS 以两幅图像 (图像块) 为输入, 计算其归一化深度特征嵌入之间的 L2 距离, 并通过线性层预测感知判断得分, 分数越低, 表示合成图像的质量越好. FID 旨在捕捉两个高斯分布的弗雷歇差异 (合成图像与真实图像), 较低的 FID 分数表明合成图像的质量较好. FSIM 是用于全参考图像质量评估的常用度量, 它捕捉图像的低层特征之间的相似性. 我们计算了合成图像和真实图像之间的平均 FSIM 分数, FSIM 分数越高, 表示合成图像的质量越好.

5.2 本文人脸照片 – 素描合成方法实验效果

本文基于生成对抗网络提出了一种人脸照片 (素描) 到素描 (照片) 的合成网络结构, 分别在 CUFS 和 CUFSF 数据集上验证了该网络结构在人脸照片 (素描) 到素描 (照片) 合成任务中的有效性和优良的性能表现, 表 1 给出了本文实验中所采用的数据集的划分情况, 这是一种最常用的划分方式 [31]. 对不同的数据集, 输入为人脸照片 (素描), 经过本文神经网络结构处理后, 输出得到合成的人脸素描 (照片).

图 3 给出了本文方法合成的部分人脸素描和人脸照片样例.

5.3 不同人脸照片 – 素描合成方法的比较

为了与本文提出的人脸照片 – 素描合成方法进行比较, 将分别在人脸照片合成人脸素描和人脸素描合成人脸照片两个任务上对本文方法与传统的基于样例的方法以及基于深度学习的方法进行比较, 与本文方法进行比较的方法分别是: LLE [3], MRF [2], MWF [11], SSD [32], Bayesian [33], RSLCR [31], DGFL [5], FCN [12], pix2pix [9], CycleGAN [10], PS2MAN [7], FaceSketchWild [15], SCAGAN [14].

在人脸照片合成人脸素描任务上, 将本文提出的方法与现有的许多高级方法进行了比较, 包括: LLE, MRF, MWF, SSD, Bayesian, RSLCR, DGFL, FCN, pix2pix, CycleGAN, PS2MAN, FaceSketchWild, SCAGAN. 表 2 显示了对比结果, 显而易见, 本文提出的方法在 CUFS 和 CUFSF 两个数据集中表现优异, 在 LPIPS 与 FID 得分上均远低于之前的人脸照片合成人脸素描方法, 这表明了我们提出

图 3 (网络版彩图) 本文方法的部分合成样例

Figure 3 (Color online) Partial synthesis results of the method in this paper

方法的有效性, 本文方法极大地改善了合成的人脸素描的真实性. 此外, 对于 FSIM 指标, 本文所提方法与所比较的方法具有高度的可比性, 值得注意的是, FSIM 指标设计用于评估由于模糊、噪点或压缩导致的质量下降, 不适用于评估素描的视觉质量, 此外, Blau 等^[34] 在数学上证明了失真和感知质量彼此矛盾, 换言之, FSIM 数值的增加, 会导致感知质量的下降, 所以 FSIM 数值不是最高并不意味着感知质量差, 本文使用 FSIM 作为评估指标之一仅是因为它在现有工作中得到了较多的使用. 图 4 和 5 展示了来自 CUFS 数据集和 CUFSF 数据集上不同方法的一些合成人脸素描, 图 4 为本文方法与传统方法在人脸素描合成任务上的效果对比图, 由于受到篇幅限制, 这里仅与几种高级的方法进行对比. 显而易见的是, MRF, MWF 和 SSD 会在多个面部区域产生较大的形状改变和严重的模糊效果. 相比之下, 基于深度学习的方法生成的人脸素描在纹理和阴影上效果更佳, 但是 FCN, pix2pix 和 PS2MAN 会产生难以接受的面部形变; FaceSketchWild 会产生过度平滑的人脸肖像, 缺乏真实感; 目前效果最好的 SCAGAN 生成的人脸素描图仍存在轮廓的细小残缺以及头发区域的模糊效果, 而本文方法几乎消除了上述缺陷并且在肖像风格以及真实性上得到了进一步提升.

在人脸素描合成人脸照片任务上, 我们将本文提出的方法与现有的许多高级方法进行了比较, 包括: pix2pix, CycleGAN, PS2MAN, SCAGAN. 我们将人脸素描作为模型输入, 人脸照片作为模型的输出, 并对模型的人脸照片合成性能进行了评估, 据我们所了解, 目前所提出的人脸照片合成方法并不多. 由表 2 可以看出, 在 CUFS 数据集上, 本文所提出方法在指标 LPIPS 和 FID 上明显优于现有方

表 2 使用 CUFS 数据库和 CUFSF 数据库的合成人脸照片/素描与最新技术方法进行定量比较的结果^{a)}
Table 2 Quantitative comparison of face photo/sketch synthesis results using the latest technology methods on CUFS and CUFSF datasets^{a)}

		LLE	MRF	MWF	SSD	Bayesian	RSLCR	DGFL	FCN	pix2pix	CycleGAN	PS2MAN	FaceSketchWild	SCAGAN	Ours
CUFS	Photo														
	LPIPS (alex) ↓	-	-	-	-	-	-	-	-	0.1993	0.2096	0.2464	-	0.1727	0.1549
	LPIPS (squeeze) ↓	-	-	-	-	-	-	-	-	0.1830	0.2094	0.2158	-	0.1643	0.1301
	LPIPS (vgg) ↓	-	-	-	-	-	-	-	-	0.3525	0.3882	0.3254	-	0.3053	0.2709
	FSIM ↑	-	-	-	-	-	-	-	-	0.7726	0.7450	0.7819	-	0.7937	0.7813
	FID ↓	-	-	-	-	-	-	-	-	73.56	80.44	65.04	-	80.53	28.60
	Sketch														
	LPIPS (alex) ↓	0.3257	0.2771	0.3105	0.3802	0.3264	0.4477	0.3316	0.4517	0.2263	0.2139	0.2961	0.2807	0.2408	0.1822
	LPIPS (squeeze) ↓	0.2491	0.1898	0.2228	0.2886	0.2443	0.3615	0.2635	0.3596	0.1552	0.1529	0.2265	0.2210	0.1722	0.1216
	LPIPS (vgg) ↓	0.3902	0.3789	0.3670	0.4031	0.3696	0.4041	0.3654	0.4350	0.3734	0.3598	0.3707	0.3639	0.3627	0.3005
CUFSF	Photo														
	LPIPS (alex) ↓	-	-	-	-	-	-	-	-	70.81	69.93	44.91	59.26	38.61	20.45
	LPIPS (squeeze) ↓	-	-	-	-	-	-	-	-	0.2463	0.2557	0.3145	-	0.1735	0.2179
	LPIPS (vgg) ↓	-	-	-	-	-	-	-	-	0.2005	0.2002	0.2853	-	0.1469	0.1704
	FSIM ↑	-	-	-	-	-	-	-	-	0.4019	0.3791	0.4237	-	0.3128	0.3444
	FID ↓	-	-	-	-	-	-	-	-	0.7777	0.7645	0.7812	-	0.8395	0.7835
	Sketch														
	LPIPS (alex) ↓	0.3266	0.3122	0.3455	0.3981	0.3875	0.4868	0.3524	0.4793	0.2408	0.2371	0.3288	0.3288	0.2188	0.1968
	LPIPS (squeeze) ↓	0.2414	0.2190	0.2579	0.3234	0.2984	0.4081	0.2794	0.3895	0.1628	0.1589	0.2397	0.2473	0.1500	0.1352
	LPIPS (vgg) ↓	0.4176	0.4769	0.4090	0.4283	0.4195	0.4521	0.3972	0.5305	0.3824	0.3744	0.4170	0.4053	0.3536	0.3368
	FSIM ↑	0.7043	0.6956	0.7029	0.6824	0.6754	0.6650	0.6957	0.6624	0.7283	0.7088	0.7233	0.6821	0.7270	0.7214
	FID ↓	-	-	-	-	-	-	57.33	124.40	35.52	14.62	64.42	59.76	18.32	8.10

a) Alex, squeeze and vgg denote the pretrained AlexNet, SqueezeNet, and VGG, respectively.

图 4 (网络版彩图) 本文方法与传统方法合成的人脸素描对比图

Figure 4 (Color online) Comparison of face sketches synthesized by the method in this paper and other traditional methods

图 5 (网络版彩图) 本文方法与深度学习方法合成的人脸素描对比图

Figure 5 (Color online) Comparison of face sketches synthesized by the method in this paper and other deep learning methods

法. 在 CUFSF 数据集上, 本文所提出方法在 FID 评分上优于所有对比方法且在 LPIPS 评分上仅差于 SCAGAN. 本文所提出方法接近了感知失真范围, 如此显著的性能改进表明我们的方法能够合成真实且自然的人脸照片, 同时人脸素描中的身份也能够得到很好的保持. 图 6 展示了来自 CUFS 数据集和 CUFSF 数据集上不同方法的一些合成人脸照片, 由图 6 可知, pix2pix 和 PS2MAN 这两种方法生成的人脸照片存在严重的模糊区域, 非常影响视觉体验; CycleGAN 则存在着一定的素描风格, 真实感不高; SCAGAN 虽然已经取得很大的效果改进, 但是仍在毛发、嘴巴等区域存在着模糊效果, 合成照片存在的模糊问题未能得到有效解决, 而本文方法较好地解决了合成照片局部模糊问题, 同时很好地保持着人脸素描中的细小纹理细节和整体结构, 视觉上更具真实感.

图 6 (网络版彩图) 本文方法与深度学习方法合成的人脸照片对比图

Figure 6 (Color online) Comparison of face photos synthesized by the method in this paper and other deep learning methods

图 7 (网络版彩图) 本文方法与其他方法在真实场景下合成的人脸素描对比图

Figure 7 (Color online) Comparison of face sketches synthesized by the method in this paper and other methods in real scenes

为验证本文提出模型的泛化能力, 将在 CUFS 训练集上训练好的模型应用于数据集之外的真实人脸, 首先使用互联网上下载的照片对模型进行测试, 与训练集中照片相比, 这些照片包含不同的光照与不同的背景. 其次, 将模型应用于一组具有光照变化的面部照片, 图 7 显示了本文方法与其他高级方法的合成素描对比图. 显然, DGFL, pix2pix 等方法均在黑暗区域存在模糊和严重的伪影, 相比之下, 本文所提出的方法合成的人脸素描具有更少的模糊和伪影, 显示出了最好的质量.

表 3 消融实验: 不同配置下的 LPIPS 得分

Table 3 Ablation study: LPIPS scores in different configurations

Configuration	Photo-LPIPS (squeeze) ↓	Sketch-LPIPS (squeeze) ↓
(a)	0.1239	0.1418
(b)	0.1225	0.1353
(c)	0.1216	0.1301

5.4 消融实验

我们在不同配置下计算 CUFS 数据库上人脸照片到人脸素描合成任务中合成的图像与真实图像之间的 LPIPS (squeeze) 分数, 以量化不同配置对整体有效性的贡献. 消融研究是针对以下 3 种配置进行的: (a) U-Net 网络结构; (b) 增加多尺度特征融合操作; (c) 增加空洞卷积作为瓶颈. 评价得分如表 3 所示, Photo 表示人脸照片合成人脸素描任务, Sketch 表示人脸素描合成人脸照片任务.

6 结束语

针对人脸照片 – 素描合成任务, 本文提出了高效的人脸照片 (素描) 到人脸素描 (照片) 合成网络, 该网络仅以人脸照片 (素描) 作为输入, 通过 U 型生成对抗网络实现人脸素描 (照片) 的合成, 该网络的生成器部分采用编码器 – 解码器结构, 同时在生成器中还使用了一种有效的空洞卷积模块, 充分利用了图像的多尺度特征信息, 使得生成器生成的合成人脸素描 (照片) 结构完整并且细节纹理真实; 判别器通过使用最小二乘 GAN 作为损失函数来优化生成器合成的人脸素描 (照片), 最终实现人脸照片 – 素描合成. 定量和定性实验表明本文提出的基于多尺度特征融合的人脸照片 – 素描合成网络在两个公开的人脸照片 – 素描数据集 CUFS 和 CUFSF 上均取得了优异的合成效果, 该网络能够很好地保持人脸的局部部件以及全局的精细结构信息, 并且对于数据集之外的真实人脸照片 – 素描合成也有着良好的合成效果, 具有很好的鲁棒性.

然而, 本文所提出的方法仅通过利用多尺度特征融合来保持合成图像的结构完整性, 尚未能同时充分利用单个面部组件的局部信息, 这就可能导致在某些真实场景中极端光照情况下进行人脸照片 – 素描合成任务时出现面部某个组件结构缺失的现象, 难以获得理想的合成效果. 未来的工作将考虑在生成器部分融入人脸各部位组件位置信息来解决上述的问题, 充分利用人脸图像的局部部件信息以及全局多尺度特征, 改进真实场景中出现的极端光照情况下的人脸照片 – 素描合成效果.

参考文献

- 1 Wang N N, Tao D C, Gao X B, et al. A comprehensive survey to face hallucination. *Int J Comput Vis*, 2014, 106: 9–30
- 2 Wang X G, Tang X O. Face photo-sketch synthesis and recognition. *IEEE Trans Pattern Anal Mach Intell*, 2009, 31: 1955–1967
- 3 Liu Q S, Tang X O, Jin H L, et al. A nonlinear approach for face sketch synthesis and recognition. In: *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005. 1005–1010
- 4 Zhou M Q, Liang C, Han Y J, et al. Face sketch synthesis via sparse representation. In: *Proceedings of International Conference on Pattern Recognition*, 2010. 2146–2149
- 5 Zhu M R, Wang N N, Gao X B, et al. Deep graphical feature learning for face sketch synthesis. In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017. 3574–3580

- 6 Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2014. 2672–2680
- 7 Wang L D, Sindagi V, Patel V. High-quality facial photo-sketch synthesis using multi-adversarial networks. In: *Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition*, 2018. 83–90
- 8 Zhang M J, Wang N N, Li Y S, et al. Face sketch synthesis from coarse to fine. In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018
- 9 Isola P, Zhu J Y, Zhou T H, et al. Image-to-image translation with conditional adversarial networks. In: *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2017. 1125–1134
- 10 Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of IEEE International Conference on Computer Vision*, 2017. 2223–2232
- 11 Zhou H, Kuang Z H, Kwan-Yee K W. Markov weight fields for face sketch synthesis. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 1091–1097
- 12 Zhang L L, Lin L, Wu X, et al. End-to-end photo-sketch generation via fully convolutional representation learning. In: *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, 2015. 627–634
- 13 Zhang S C, Ji R R, Hu J, et al. Robust face sketch synthesis via generative adversarial fusion of priors and parametric sigmoid. In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018. 1163–1169
- 14 Yu J, Xu X X, Gao F, et al. Toward realistic face photo-sketch synthesis via composition-aided GANs. *IEEE Trans Cybern*, 2020, 51: 4350–4362
- 15 Chen C F, Liu W, Tan X, et al. Semi-supervised learning for face sketch synthesis in the wild. In: *Proceedings of Asian Conference on Computer Vision*, 2018
- 16 Yi R, Liu Y J, Lai Y K, et al. ApdrawingGAN: generating artistic portrait drawings from face photos with hierarchical GANs. In: *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 10735–10744
- 17 Kim J H, Kim M J, Kang H W, et al. U-GAT-IT: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In: *Proceedings of International Conference on Learning Representations*, 2020
- 18 Gou Y B, Li B Y, Liu Z T, et al. CLEARER: multi-scale neural architecture search for image restoration. In: *Proceedings of Advances in Neural Information Processing Systems* 33, 2020
- 19 Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015
- 20 Mao X D, Li Q, Xie H R, et al. Least squares generative adversarial networks. In: *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017. 2813–2821
- 21 Johnson J, Alahi A, Li F F. Perceptual losses for real-time style transfer and super-resolution. In: *Proceedings of European Conference on Computer Vision*, 2016. 694–711
- 22 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv:1409.1556
- 23 Zhang W, Wang X G, Tang X O. Coupled information-theoretic encoding for face photo-sketch recognition. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 513–520
- 24 Tang X O, Wang X G. Face sketch synthesis and recognition. In: *Proceedings of the 9th IEEE International Conference on Computer Vision*, 2003. 687–694
- 25 Martinez A M, Benavente R. The AR Face Database. CVC Technical Report 24, 1998
- 26 Messer K, Matas J, Kittler J, et al. XM2VTSDB: the extended M2VTS database. In: *Proceedings of the 2nd International Conference on Audio and Video-based Biometric Person Authentication*, 1999. 72–77
- 27 Phillips P J, Hyeonjoon M, Rizvi S A, et al. The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans Pattern Anal Mach Intell*, 2000, 22: 1090–1104
- 28 Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric. In: *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 586–595
- 29 Heusel M, Ramsauer H, Unterthiner T, et al. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: *Proceedings of the 31st Conference on Neural Information Processing System*, 2017
- 30 Zhang L, Zhang L, Mou X Q, et al. FSIM: a feature similarity index for image quality assessment. *IEEE Trans Image Process*, 2011, 20: 2378–2386
- 31 Wang N N, Gao X B, Li J. Random sampling for fast face sketch synthesis. *Pattern Recogn*, 2018, 76: 215–227

- 32 Song Y B, Bao L C, Yang Q, et al. Real-time exemplar-based face sketch synthesis. In: Proceedings of European Conference on Computer Vision, 2014. 800–813
- 33 Liu W, Tang X O, Liu J Z. Bayesian tensor inference for sketch-based facial photo hallucination. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence, 2007. 2141–2146
- 34 Blau Y, Michaeli T. The perception-distortion tradeoff. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018. 6228–6237

Face photo-sketch synthesis based on multi-scale feature fusion

Changcheng LIANG¹, Nannan WANG^{1*}, Mingrui ZHU^{1*}, Xi YANG¹, Jie LI¹ & Xinbo GAO²

1. State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China;

2. Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

* Corresponding author. E-mail: nnwang@xidian.edu.cn, mrzhu@xidian.edu.cn

Abstract The synthesis of face sketches from real face photos and its inverse process have a wide range of applications, such as digital entertainment and assisting in the investigation of criminal cases. However, due to the significant discrepancies in texture between photos and sketches, the transformation between them is still a challenging problem. Recently, the methods based on generative adversarial nets (GANs) have shown encouraging results in the transformation between images, especially the transformation from photos to sketches. However, most of them produce different deformations or blur in the key components of the face, which affects the realism of the synthetic image. To address this challenge, we propose a novel face photo-sketch synthesis algorithm based on multi-scale feature fusion to improve the structural integrity and texture fidelity of the synthetic image. Firstly, an encoder is utilized to extract the multi-scale coding features of the input image. Then the bottom coding features are passed by a dilated convolution module and transmitted to the decoder for decoding. In the decoding process, the decoding features of different scales and the coding features of corresponding scales are concatenated in the channel dimension to obtain the multi-scale coding and decoding fusion features. Finally, the coding and decoding fusion features of different scales are further fused at the output of the decoder and the final synthesis result is generated through a convolution layer. By combining the features of different scales in the encoding and decoding processes in the channel dimension at the same time, we can maintain good image structure and texture details and generate realistic face sketch/photo images. We verify the effectiveness of the proposed method in several challenging datasets. The quantitative and qualitative evaluation shows that the proposed method is superior to other state-of-the-art methods in generating face sketches (or photos) with high visual quality.

Keywords face photo sketch synthesis, image-to-image translation, generative adversarial network, multi-scale feature fusion, dilated convolution