基于 SimSwap 换脸模型的分析与优化

521030910414 张博伦 *521030910417* 谢靖宇 *519030910364* 代铭波

1 摘要

在本次实验中，我们从换脸模型 SimSwap[1] 出发，探讨了 SimSwap 存在的问题与缺陷，并对该模型提出的一系列的改进方法。最后，针对换脸这一充满争议的话题，我们进行了伦理道德层面的讨论。

2 引言

随着深度学习技术的不断发展，图像合成和编辑技术也取得了巨大的进步。其中，换脸技术作为图像编辑领域的重要分支之一，受到了广泛关注。近年来，基于对抗生成网络（GAN）的换脸模型已经取得了显著的成就，SimSwap 作为其中的一种代表性模型，在实现高质量的换脸效果方面具有独特的优势。然而，尽管 SimSwap 已经取得了令人瞩目的成绩，但仍存在一些问题和局限性。因此，本文旨在对 SimSwap 模型进行改进，以提高其换脸效果的质量和稳定性。

在本文中，我们首先回顾了当前换脸技术的研究现状和 SimSwap 模型的相关工作。然后，我们提出了针对 SimSwap 存在问题的改进方案，并详细阐述了改进方法的理论基础和技术实现。最后，我们对本文的研究成果进行了总结，并讨论了换脸技术的伦理问题以及外来前景。本文的主要贡献如下所示：

1. 分析并指出了 Simswap 模型存在的问题，模型训练推理的速度较低，生成的人脸图像有时候不够自然，同时在经过测试脚本计算时，生成结果的质量仍有提升空间，此外，生成换脸图片/视频的伦理问题也有待解决。

2. 提出了引入 mask 来提高模型性能的思路，通过使用预训练的人脸检测以及图像分割模型获取 mask，使模型获得更好的效果。

3. 提出了通过优化卷积层网络结构来减少模型推理时间和优化模型生成图片的思路，通过使用动态卷积等结构来提高模型性能。

4. 提出了模仿 diffusion model 改进 id 注入部分的思路，通过提供 id 注入后的 target 来让 id 注入效果更好。

5. 对充满伦理争议的换脸问题发表了新颖的、有建设性的思考，例如为每张换脸后的图片添加 AIGC 水印等措施来防止污名化与侵权问题，并对换脸的未来前景来提出了自己独有的观点。

通过本文的研究，我们旨在为换脸技术的发展贡献一份力量，提高换脸模型的实用性和稳定性，为图像编辑领域的进一步发展提供有益的启示和参考。

3 SimSwap

3.1 简介

SimSwap 是一个设计用于实现通用且高保真面部交换的高效框架。与先前的方法相比，该框架能够将任何源脸的身份转换为任何目标脸，同时保留目标脸的属性。

该网络主要由生成器和判别器组成。生成器（如图 1 所示）由三个部分组成：编码器、身份注入模块和解码器。编码器从 I_T 中提取特征，ID 注入模块将 I_s 的身份信息注入到 I_T 的特征中，解码器将修改后的特征恢复为生成的图像。身份损失用于鼓励生成具有与源脸相似身份的结果。弱特征匹配保留目标脸的属性，而不会显著影响身份修改性能。

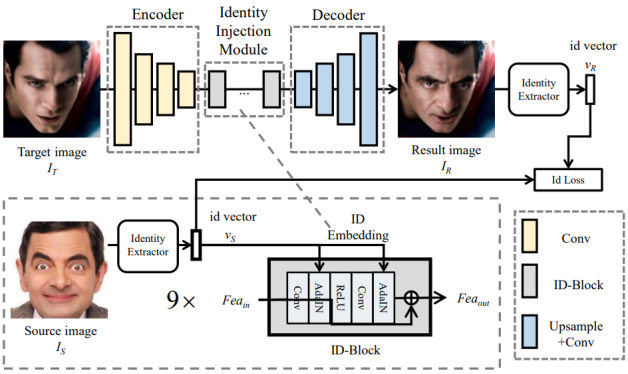

图 1: *SimSwap* 生成器

判别器（如图 2 所示）是一个对抗网络，用于评估生成器图像与真实图像之间的相似性和差异。它依次接受换过脸以及未换过脸的图片，并判断接收到的图片是否经历过换脸处理，从而向生成器提供反馈，以生成更真实和真实的图像。

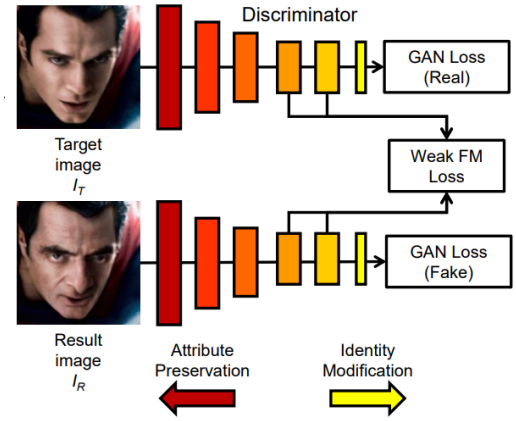

图 2: *SimSwap* 判别器

3.2 问题和缺陷

首先，模型训练推理的速度较低。模型速度的瓶颈主要来自于 id 注入模块。id 注入模块的注入方式和注入次数缺乏理论支撑。我们工作的主要想法是改进该模块。

4 改进

4.1 引入 Mask

4.1.1 思路由来

SimSwap 网络主要由两个大部分组成：生成网络 Net G 以及判别网络 Net D。在实际训练中，生成网络生成的图片直接进入判别网络实现对抗学习，而没有经过任何处理。然而，可以肯定的是，一张图片的背景多多少少会给判别网络的判断带来一些影响。因此，我们考虑在网络的特定部分对图像进行 mask 处理，力争减少背景对整个训练带来的影响。

4.1.2 基于 ResNet 的人脸检测

基于 ResNet 的人脸检测算法是一种利用深度卷积神经网络进行人脸检测的方法，通过引入残差模块训练非常深的网络，获得更好的效果。

在基于 ResNet 的人脸检测算法中，通常会使用预训练的 ResNet 模型作为基础网络，然后在其基础上进行微调，以适应人脸检测任务。微调的过程包括替换网络的最后几层，以及对整个网络进行参数微调，从而使其能够更好地适应人脸检测任务。训练完成后，该算法可以用于检测图像中的人脸，并输出人脸的位置和边界框。

在本次实验中，我们使用了基于 ResNet 的人脸检测预训练模型，来具体勾勒训练图片中的人脸边界框。在获得了边界框后，我们就可以生成这张图片对应的 mask，mask 是一个与图像维度完全相同的张量，边界框外的值为 0，边界框内的值为 1。将 mask 与图片张良按位相乘，就可以消除一部分背景。

4.1.3 基于 Deeplabv3+ 的图像分割

DeepLabv3+[2] 是语义图像分割的先进深度学习模型，旨在准确地为图像中的每个像素分配一个标签，区分不同的对象及其边界。

DeepLabV3+ 结构由图 3 所示，其由编码器和解码器构成，编码器中的空间金字塔池化模块可以捕获丰富的语义信息，通过对不同的分辨率进行特征池化来实现，编解码器的组合可获得更加

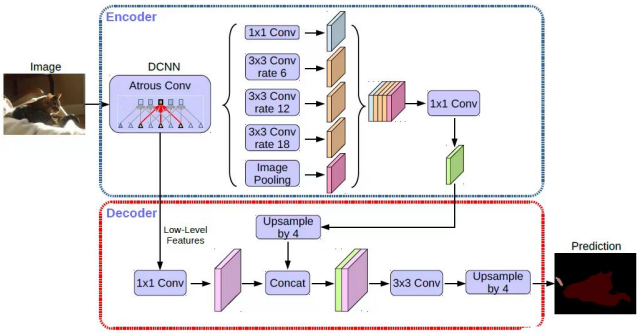

图 3: *Deeplabv3+* 网络

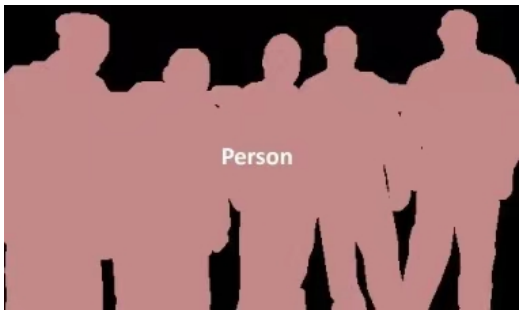

图 4: *Deeplabv3+* 图像分割效果示意

准确的边界。在具体实现中，DeepLabV3+ 应用了几个并行的，伴随不同速率的 Atrous convolution，即 Atrous 空间金字塔池化，同时使用 PSPNet 在不同的网格尺度执行池化运算。

图像分割大致效果由图 4 所示。

4.1.4 改进方法

在实验初期，我们采用了基于 ResNet 的人脸检测网络，但是效果并不理想。后来我们考虑到，人脸检测框内依旧存在背景，并且检测框会将人的头发等身体部位排除在外，因此在实验后期，我们转而使用基于 Deeplabv3+ 的图像分割网络，完全分离背景与人脸。

我们主要考虑在生成网络和判别网络之间增加 mask，旨在提高判别网络判别的准确性。在训练初期，我们并不使用 mask，希望模型能先学到一个正确的背景输出，即先保证模型生成的换脸图片的背景与原图是一样的。随着训练的进行，我们逐渐加入 mask，使得网络完全致力于人脸而不受任何背景的影响。模型效果如下所示，A 表示与 baseline 相比，最终测试图片 ArcFace 余弦距离的降低百分比。

方法	A(%)
人脸检测	-0.04
图像分割	7.54

表 1: *Performance of Different Method*

我们发现，人脸检测方法的效果不尽如人意，但图像分割方法取得了 7.54% 的提升。

实际上，我们还想到了其他加入 mask 的方法。比如，我们可以直接在把图片送入网络前就取 mask，这样进入网络的图片就不存在任何的背景信息。在网络输出图片后，再将背景补回。但由于时间等原因，这些方法并未被付诸实践，将来我们还可以在此想法上进行进一步的研究。

4.2 优化卷积层网络

4.2.1 思路由来

在实验训练过程中，我们深感于时间与资源的紧张，于是针对于前向推理过程，我们设置了一些节点来检测各部分运行时间，最终发现在生成器部分，卷积层的计算相对耗时比较长且有较大优化空间。同时，在测试过程中，我们还发现，更改卷积层对模型输出也有一定影响。因此，我们对卷积层展开优化，以其达到减少模型推理时间与优化模型输出的功能。

4.2.2 深度可分离卷积、动态卷积与调制卷积

综合考虑效果与实现成本，我们选择以下三种改进算法来进行实验

深度可分离卷积（Depthwise Separable Convolution） [3] 是一种卷积操作的改进形式，分为两个步骤：深度卷积和逐点卷积。首先，对每个输入通道进行单独的卷积操作，然后使用 1x1 的逐点卷积将通道间的信息整合。这相对于传统的卷积操作来说，减少了参数数量和计算复杂度，我们期待利用这个网络来加速模型运算

Algorithm 1 深度可分离卷积

输入: 输入特征图 $\mathbf{X}$, 深度可分离卷积核 $\mathbf{K}$, 步长 s
输出: 输出特征图 $\mathbf{Y}$
分解卷积核:
将深度可分离卷积核 $\mathbf{K}$ 分解为深度卷积核 $\mathbf{K_d}$ 和逐点卷积核 $\mathbf{K_p}$
for 每个深度卷积核 $\mathbf{K_d}$ **do**
  初始化输出特征图: $\mathbf{Y_d} = 0$
  for 每个通道 c **do**
    提取输入特征图中的通道: $\mathbf{X_c} = \text{channel}(\mathbf{X}, c)$
    提取深度卷积核: $\mathbf{K_{d_c}} = \text{channel}(\mathbf{K_d}, c)$
    执行深度卷积: $\mathbf{Y_c} = \text{convolution}(\mathbf{X_c}, \mathbf{K_{d_c}}, s)$
    累加到输出特征图: $\mathbf{Y_d} = \mathbf{Y_d} + \mathbf{Y_c}$
  end for
end for
逐点卷积:
执行逐点卷积: $\mathbf{Y_p} = \text{convolution}(\mathbf{X}, \mathbf{K_p}, s)$
融合结果:
输出特征图为深度卷积和逐点卷积的融合: $\mathbf{Y} = \mathbf{Y_d} + \mathbf{Y_p}$

return 输出特征图 $\mathbf{Y}$

动态卷积 (Dynamic convolution) [4] 是指卷积核的权重在运行时动态调整的卷积操作。这种方法使得神经网络能够在处理不同输入时自适应地调整卷积核，从而更好地捕捉数据中的动态模式和特征。具体而言，对于输入序列: $x = [x_1, x_2, \ldots, x_n]$ 与动态卷积核权重: $w = [w_1, w_2, \ldots, w_{k_{\max}}]$ 以及每个位置卷积核大小: s_i 有输出序列

$$y_i = \sum_{j=1}^{s_i} w_j \cdot x_{i+j-1} + b$$

其中，$1 \le i \le n - k_{\max} + 1$

调制卷积 (Modulated Convolution) [5] 调制卷积引入了一个可学习的调制信号，用于动态地调整卷积核的权重。这个调制信号可以根据输入数据的不同部分而变化，从而使网络更灵活地适应不同的输入模式。调制卷积在一些需要对输入进行动态调整的任务中表现出色，例如需要处理不同尺度的目标的任务。在我们的任务中，调制卷积在理论上能够发挥出作用。

Algorithm 2 调制卷积

输入: 输入信号 $\mathbf{x}$, 调制信号 $\mathbf{m}$, 卷积核 $\mathbf{w}$, 步长 s
输出: 输出信号 $\mathbf{y}$
for 每个卷积核的时间步 t **do**
  计算当前调制系数: $\alpha_t = f(\mathbf{m}[t])$ {这里 f 是调制函数}
  初始化输出信号的当前时间步: $y_t = 0$
  for 每个卷积核的滤波器权重 w_i **do**
    计算当前滤波器的输入: $x_{\text{input}} = \mathbf{x}[t \cdot s + i]$
    计算加权输入: $x_{\text{weighted}} = \alpha_t \cdot x_{\text{input}}$
    累加到输出信号: $y_t = y_t + w_i \cdot x_{\text{weighted}}$
  end for
  存储当前时间步的输出: $\mathbf{y}[t] = y_t$
end for

return 输出信号 $\mathbf{y}$

4.2.3 改进方法

针对于前两种方法, 我们直接把原有的上下采样过程中的卷积层进行了替换，而对后一种，因为引入了调制信号，也就是隐藏向量，我们也相应修改了调整了前向传播中隐藏向量的传递过程，来向卷积网络中加入调制信号。

遗憾的是，由于在训练动态卷积层时的显存开销过大，我们未能实现动态卷积模型的完整训练与测试。针对其他两种改进，我们测试了其对一次前向传播时间降低的百分比 T 与最终测试图片 ArcFace 余弦距离降低百分比 A，如下所示

方法	T(%)	A(%)
深度可分离卷积	-0.52	0.01
调制卷积	-4.11	1.83

表 2: *Performance of Different Method*

让人意外的是，深度可分离卷积不仅没有提升速度, 反而带来了极小的降低; 而调制卷积在牺牲了部分运行时间的代价下带来了 1.83% 的提升。

4.3 模仿 diffusion 调整 id 注入

ID 注入模块的工作是将目标身份信息更改为源面部的身份信息。该模块由身份提取部分和嵌入部分组成。在身份提取部分，使用人脸识别网络来提取身份向量。在嵌入部分，重复使用 9 个 id -block 将身份信息注入到特征中。对于这 9 个 block，我们认为可以进行类似 diffusion model 框架的实现，让每个 block 一步接一步地实现换脸。在训练过程中，我们首先使用 1 个 id-block，然后每 100000 个 epoch 就添加一个新的 block，同时固定之前的 block 和 encoder 的参数，仅训练新的 block、decoder 和 discriminator。训练了 900000 个 epoch 之后效果与 baseline 类似，但训练时间有一定的减少。

修改的另一种思路是将整个 id 注入的部分更换为 diffusion model。在这之前，我们修改了 encoder 的最后一层和 decoder 的第一层，认为 decoder 和 encoder 将图片压缩到了一个比较小的维度。然后将中间的 id 注入部分替换为 DDPM 的框架来实现换脸。注意到 DDPM 框架中额外需要在每一层输入添加中间步骤的结果，而我们缺少换脸后的真值，所以无法实现换脸。于是我们将换脸后的图片作为原图，真正的原图作为换脸后的真值生成多个中间结果作为输入进行了探究，但运行时间过长，且在训练时的显存开销过大 (仅能运行 BatchSize=2)，未能实现模型的完整训练与测试。

4.4 VGGFace2 数据清洗

一个好的数据集对模型的性能有直接的影响。充分、准确、多样化的数据集可以提供更多的信息，帮助模型更好地理解问题，并从中学习到更准确的规律和模式，并减少过拟合风险，提高泛化能力，从而提升模型的性能。我们对 VGGFace2 的数据进行了大致清洗，删掉了明显模糊失真和不属于同一个人物类别的图片，并将 sjtu 的数据从 png 格式转换为 jpg 格式一起进行训练，在 baseline 模型上取得了一定的性能提升。

数据集	A(%)
sjtu	-0.03
sjtu + VGGFace2	4.31

表 3: *Performance of Different Method*

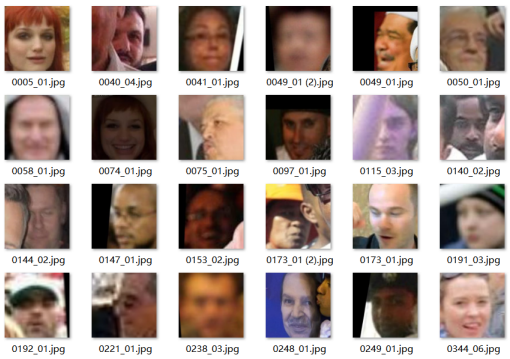

图 5: *VGGFace2* 删除的部分数据

5 AI 换脸的伦理之忧与未来展望

5.1 伦理之忧

AI 换脸引发了严重的伦理问题，涉及隐私侵犯、虚假信息和欺骗、社交工程及网络攻击、法律责任以及信任破坏等方面。其制作虚假视频的潜在风险包括侵犯个人隐私、混淆观众对真实事件的认知，甚至可能导致法律责任问题。为了应对这一挑战，社会需要制定有效法规，对深度伪造技术进行监管，以保护公众的权益和隐私。同时，在技术上，我们也考虑为图片生成 AIGC 水印。

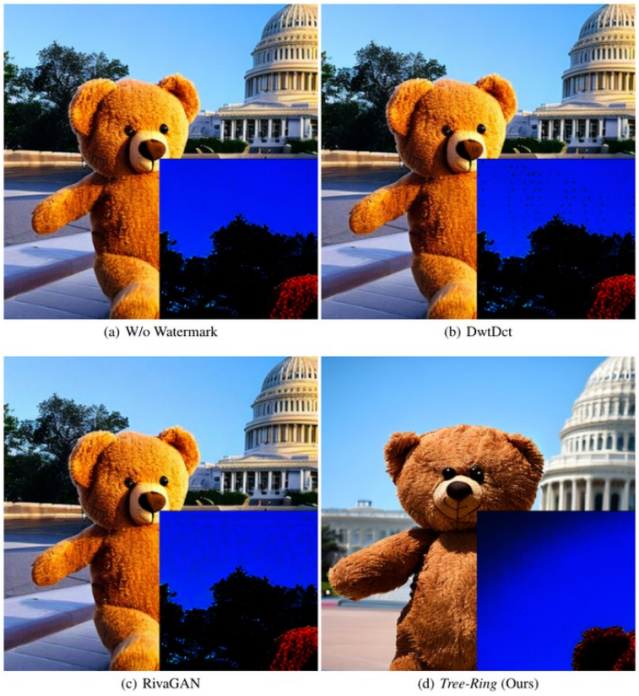

图 6: 不同水印算法的效果 *[6]*

例如，应用 Tree-Ring Watermark 技术 [6]。具体而言，该算法能对扩散模型的输出结果进行稳定识别。与在采样后对图像进行事后修改的现有方法不同，树环水印技术会对整个采样过程产生微妙的影响，从而生成人类无法看到的模型指纹。水印将一种模式嵌入用于采样的初始噪声矢量中。这些图案在傅立叶空间中形成，因此不受卷积、裁剪、扩张、翻转和旋转的影响。图像生成后，通过反转扩散过程来检索噪声矢量，从而检测水印信号，然后检查噪声矢量中是否有嵌入的信号。最终能够获得在每个样本基础上真正隐形的水印。这样一来，针对于换脸后所得到的图片，我们能够用水印技术保护图片中人。

5.2 未来展望

作为相对热门的领域，在 SimSwap 之后已有很多新工作的出现，包括但不限于 Deepfacelab[7]，SimSwap++[8] 等。AI 换脸这项技术本身充满了潜在的挑战和机遇。其在艺术和娱乐领域的广泛应用可能为创作者提供更多创意空间，然而，如上文所述，随着技术的进步，可能引发更深层次的隐私和伦理问题，特别是在社交媒体和网络欺诈方面。在未来，将会有法律和监管的完善，以适应技术的发展，并可能涌现更先进的对抗技术来实现更优秀的生成效果。

6 成员分工

张博伦：初期数据清洗，SimSwap 算法 baseline 运行，SimSwap 算法深度可分离卷积、动态卷积、调制卷积改进的实现与实验，ArcFace 余弦距离测试脚本编写，报告撰写自己工作的部分。

谢靖宇：初期数据清洗，SimSwap 算法 baseline 运行，研究与实现引入 mask 优化 SimSwap，报告撰写自己工作的部分。

代铭波：初期数据清洗，SimSwap 算法 baseline 运行，VGGFace2 数据集清洗，研究与实现模仿 diffusion 优化 id 注入，报告撰写自己工作的部分。

7 References

[1] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2003–2011, 2020.

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[3] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[4] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[5] Xiaodi Wang, Baochang Zhang, Ce Li, Rongrong Ji, Jungong Han, Xianbin Cao, and Jianzhuang Liu. Modulated convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 840–848, 2018.

[6] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*, 2023.

[7] Kunlin Liu, Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Wenbo Zhou, and Weiming Zhang. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *Pattern Recognition*, 141:109628, 2023.

[8] Xuanhong Chen, Bingbing Ni, Yutian Liu, Naiyuan Liu, Zhilin Zeng, and Hang Wang. Simswap++: Towards faster and high-quality identity swapping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.