电 子 科 技 大 学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

学士学位论文

BACHELOR THESIS

论文题目 基于图像的虚拟试衣

学　　院 信息与通信工程学院

专　　业 电子信息工程

学　　号 2015020902009

作者姓名 沈凯越

指导教师 胡洋

摘 要

本文从虚拟试装任务出发，将任务转化为时尚图像补全任务，即：给定一张缺失某件服装单品的人像，我们要生成兼具真实性、多样性、兼容性的完整时尚图像。我们采用了一个二阶时尚图像补全网络，通过拆解生成过程为两个子过程：形状生成和纹理合成，实现了形状和纹理的分级控制，提高了生成图像的真实性。在每个生成阶段，借鉴变分自编码器的思想，将控制信息映射到隐空间，利用随机采样实现生成图像的多样性，同时利用两个相互关联的编码器显性地控制服装的兼容性。此外，我们采用对抗生成训练、注意力机制、谱归一化等方法进一步提高了图像的生成质量和模型训练的稳定性。我们在 DeepFashion 数据集上的测试结果，以及与其他模型的定性定量比较也证实了我们模型在时尚图像补全任务上具有优良的性能。

关键词：虚拟试装，图像补全，真实性，多样性，兼容性

ABSTRACT

In this work, we transform the virtual try-on mission into a fashion image completion task. That is, given a portrait missing a piece of clothing, we want to generate a complete fashion image with high realism, diversity and compatibility. By proposing a two-stage fashion image completion network, we divide the generation process into two sub-processes: shape generation and texture synthesis, which realize hierarchical control of shape and texture, and improve the realism of the generated image. In each sub-process, we use the idea of Variational Autoencoder: map the control information to the latent space and then utilize the random sampling to realize the variety of generated images. At the same time, we introduce two interrelated encoders to explicitly control the compatibility among clothing items. In addition, we use methods such as generative adversarial networks, attention mechanism, and spectral normalization to further improve the quality of generated images and stabilize the training process. We test our model on DeepFashion dataset and compare our model with other similar work, the result of which confirm that our model has excellent performance in the fashion image completion task.

Keywords: Virtual try-on, Image completion, Realism, Diversity, Compatibility

目 录

第一章 绪论

1.1 课题背景

试想一下，你打开一个时尚购物网站，从相册挑选一张最近拍摄的全身照，用手机自带的编辑工具抹去你想要购买的某类服装单品，如上衣，处理完照片后将其拖到搜索框内，网页便开始陆陆续续地显示你穿着各种不同且合身的上衣的全身照。难以置信？这可能是时尚购物网站的未来。

近年来互联网在线购物平台日益兴起，去年双十一中国最大的购物网站天猫商城更是凭借 1 小时 47 分 1000 亿的交易额成为其中的典范。一个出色的时尚购物网站首先得具备良好的产品推荐算法，即在考虑不同类别服装单品间的兼容性的同时尽可能得做到多样化；其次，“量体试衣”这一时尚服装所独有的属性决定了它与传统产品相比，更强调与人的兼容性：同一服装穿着在不同体型的人身上应该呈现不同的形态，以解决买家秀与卖家秀存在巨大差异这一现实问题。

在以上需求的推动下，一些公司如 triMirror，Fits Me 等相应地提出了三维服装设计与虚拟试装方案。这些虚拟试装系统的成功都建立在庞大的三维人体数据的基础上，充足的数据使得建立精确的三维模型、操纵模型进行几何变换成为可能，但数据的标注与存储所倚赖的高昂人力物力成本阻碍了其被投入大规模的工业应用。

深度学习网络强大的学习能力使得基于二维图像的虚拟试装成为可能。其中，变分自编码器、对抗生成网络作为最经典的两大生成模型，早已被广泛应用于一系列图像生成问题，如基于类别标签、基于文本、基于条件图像的图像合成、风格迁移等。虚拟试装问题相较于生成其他具有更加严格化结构的物体而言，难度更大，因为它要考虑不同类别服装单品间的兼容性。而如何衡量兼容性并非易事，因为不同类别的服装单品单看每件可能都具有不同的纹理与颜色，它们组合到一起却构成风格一致的套装。因此，在生成虚拟试装结果时，我们更应该侧重于整体风格的一致性而非单品间像素级别的相似性。

直接生成一张服装风格统一的全身照是极具挑战性的任务，因其需要同时生成不同形状、纹理的服装单品并准确贴合到人体上。因此，我们将虚拟试装任务简化为图像补全任务。正如第一段描述的那样，给定一张用户穿着服装的图片，我们的模型能够补全其中对应缺失部分类别的服装单品，它应该同时满足真实、与其他可见服装单品风格兼容的特性。这样的模型不仅适用于我们的任务，还能拓展到服装推荐、服装迁移、时尚设计等其他任务。

过去关于图像补全的研究工作已经证实，深层生成网络能够有效利用关联信息即未被遮挡部分的图像来补全缺失区域，生成的补全结果能较好的与其周围区域保持一致性。但这样的图像补全思路并不能直接应用到时尚合成任务，因为图像补全往往只能提供唯一解，而时尚合成是一个一对多映射的问题：给定一张有一件缺失服装单品的图像，应该有多个形状、纹理不同的可行解，比如，与一条裤子搭配的上衣不只有一件。

综上，我们的课题目标是设计一个兼具真实性、多样性、兼容性的虚拟试装系统，通过图像补全加以实现。

1.2 国内外研究现状

1.2.1 虚拟试装

大多数关于虚拟试装的研究都是基于计算机图形学理论建立的图形学模型。Sekine et al.[1]引入的虚拟试装系统是通过深度图获取三维人体信息进而调整二维服装图像，Chen et al.[2]利用 SCAPE[3]人体模型来生成人体图像，PonsMoll et al.[4]使用 3D 扫描仪自动捕获真实服装以及估计人体体型及姿势。相较于复杂的图形学模型，基于图像的生成模型显然计算成本更低，因此逐步有研究者尝试利用此类模型实现虚拟试装。Jechev 和 Bergmann[5]提出的条件类比对抗生成网络能够在不需要人体特征的前提下，实现人体穿着时尚服装的交换，但是测试时同时需要原始和目标服装产品图这一要求，使得该模型难以被广泛应用。Han et al.[6]提出 VITON 模型，利用 TPS 变化实现服装的形变并利用两阶段生成网络达到最终的换装效果；Wang et al.[7]在 VITON 基础上将 TPS 变化部分利用几何匹配网络实现并对其余网络框架做了微调进一步提高了生成图像的视觉效果；Chou et al.[8]把虚拟试装问题从之前的上衣扩展到了鞋子，并将试装问题转换为图像补全问题。最近，Han et al. [9]也将虚拟试装任务转换为图像补全问题，并提出了服装兼容性的概念，通过同时训练两个编码器显式地约束了待补全服装与已知服装之间的视觉兼容性，该模型的任务与我们的相一致，我们的网络结构在一定程度上参考了该方法。

1.2.2 图像生成

图像生成任务自对抗生成网络[10]和变分自编码器[11]提出后取得了长足的进步，两者以及变种被广泛应用在带条件的图像生成任务上，如图像翻译、未来预测、三维建模等。为了控制待生成图像满足想要的属性，各种不同的先验知识或者监督条件先后被使用，如类别标签[12]、属性[13]、文本[14]、图像[15]。以往的时尚服装

合成方法主要是基于不同姿势、文本描述、产品服装图像的图像生成，而很少会关注服装间的兼容性。

1.2.3 图像补全

传统的图内补全方法建立在“待补全部分与可见部分具有内容相似性”的假设上，这样的方法有基于扩散[16]、补块[17]的方法，常见操作是将可见部分拷贝、匹配、重对齐到待补全部分，显然这样的方法不能生成给定图中没有的物体，也就不适用于我们的任务。

图像补全方法是从整个数据集中学习到有用信息，从而实现单张图像的补全。Hays 和 Efros[18]从数百万张图像中检索出与可见部分最接近的图像作为输出。自从卷积神经网络的引入，之后的方法能够处理的待补全区域的尺寸越来越大，基于对抗生成网络的上下文编码器（CE）[19]已经能够实现 64*64 像素大小破洞的补全。然而此类方法的缺点是它们的生成结果往往趋于模糊乃至扭曲，这在补全较大区域时尤为明显。

为了克服以上问题，许多研究者将两种方法结合。Yang et al.[20]提出多尺度神经局部块合成，该模型通过拷贝神经网络中间特征层的局部块来实现高频信息的合成。最近一些工作[21-23]尝试利用空间注意力机制来获取高频细节。这些方法都是通过比较待补全区域与可见区域的特征来识别、利用近似特征，然而这是违背认知的，因为当两个特征非常接近时，特征迁移是没必要的，反之，不同的特征因为难以匹配而被弃置，从而待补全区域仅能学习到可见区域的相似信息。我们的模型通过自注意力机制[24]来充分利用上下文图像信息。

1.3 本文的主要贡献与创新

1、我们将虚拟试装任务转换为图像补全任务，整个训练过程采用自监督学习，不需要监督学习中的手动标注标签。

2、我们将图像生成过程拆解为两个递进的阶段：形状生成和纹理合成，从而提升最终补全图像的质量。我们的整体网络架构包含一个形状生成网络和一个纹理合成网络。

3、我们引入两个相互关联的编码器来显式地表征及控制服装的视觉兼容性，关于服装兼容性的已知研究仍较欠缺。同时，我们的网络能够做到一定的生成结果多样性，通过控制隐空间向量的不同维度能够实现不同的补全结果。

4、我们在网络中加入注意力机制，从而有效的利用长短期上下文信息、确保补全部分与整体图像在图像域的视觉一致性，将此机制应用到对抗生成网络上能

够提升其性能。

5、我们在已有工作的基础上，对网络模型进行压缩修改，在相同训练次数的前提下实现了相当乃至更好的视觉效果，并在常用对抗生成网络评估指标 Fréchet Inception Distance（FID）上显示了更好的结果（数值更小）。

1.4 本论文的结构安排

第二章：阐述模型涉及的基本算法及原理，包括变分自编码器、对抗生成网络、自注意力机制、残差网络等。

第三章：时尚图像补全网络的模型建立、损失构建和模型分析。

第四章：模型的实验部分，包括数据集介绍、网络结构细节、模型优化细节、对比方法介绍、评估指标介绍、生成结果的定性定量分析。

第五章为全文总结。

第二章 时尚图像补全网络基本理论

2.1 变分自编码器

2.1.1 问题描述

已知一个包含 N 个由独立同分布随机变量 x 生成的样本集合 $\mathrm{X}=\{\mathrm{x}^{(i)}\}_{i=1}^{N}$ ，我们假设数据是由一些随机过程生成的，且其中涉及到一个连续随机隐变量 z。如图 2-1 的实线所示，整个生成过程可以分为两个阶段：（1）由先验分布 $p_\theta(\mathrm{z})$ 生成 $\mathrm{z}^{(i)}$；（2）由条件分布 $p_\theta(\mathrm{x}|\mathrm{z})$ 生成 $\mathrm{x}^{(i)}$。一旦我们确知了参数 θ，对于任何一个观测数据 x，我们就能根据贝叶斯定理计算后验概率，即 $p_\theta(\mathrm{z}|\mathrm{x})=p_\theta(\mathrm{x}|\mathrm{z})p_\theta(\mathrm{z})/p_\theta(\mathrm{x})$，从而有效估计隐空间变量 z。然而，绝大多数情况下，这个生成过程是未知的，即我们无法获悉真实的参数 θ 以及隐变量 $\mathrm{z}^{(i)}$。于是引入一个模型 $q_\phi(\mathrm{z}|\mathrm{x})$ 来近似真实的后验概率 $p_\theta(\mathrm{z}|\mathrm{x})$，即图 2-1 中的虚线部分，继而将原问题转换为模型参数求解问题。

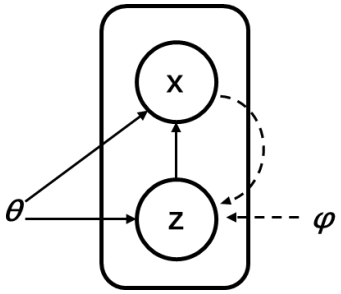

图 2-1 变分自编码器的图模型

以上问题从编码理论的角度看，不可观测的变量 z 可以理解为编码结果，模型 $q_\phi(\mathrm{z}|\mathrm{x})$ 是一个概率编码器，给定一个数据样本 x，它能够产生一个关于编码 z 的分布（比如高斯）。类似地，模型 $p_\theta(\mathrm{x}|\mathrm{z})$ 是一个概率解码器，给定一个编码 z，它能够产生一个关于 x 的分布。将两者结合起来，就是要将一个数据先编码再解码，模型的优化目标应该是最大化的将原数据恢复回来，即，使得真实的 x 对应的 $p(\mathrm{x})$ 最大。

2.1.2 问题分析

边缘似然函数可以表示为单个数据样本的边缘似然函数的总和，即 $\log p_\theta(\mathrm{x}^{(1)},\cdots,\mathrm{x}^{(N)})=\sum_{i=1}^{N}\log p_\theta(\mathrm{x}^{(i)})$，其中每个边缘似然函数都可以写做：

$$\log p_\theta(\mathrm{x}^{(i)})=D_{KL}(q_\phi(\mathrm{z}\,|\,\mathrm{x}^{(i)})\,||\,p_\theta(\mathrm{z}\,|\,\mathrm{x}^{(i)}))+\mathrm{L}(\theta,\phi;\mathrm{x}^{(i)}) \tag{2-1}$$

等式右边第一项为估计后验概率与真实后验概率之间的 KL 散度。由于 KL 散度是非负的，等式右边第二项 $\mathrm{L}(\theta,\phi;\mathrm{x}^{(\mathrm{i})})$ 被称作数据点 i 的边缘似然的可变下界，由公式(2-1)可得，

$$\log p_\theta(\mathrm{x}^{(\mathrm{i})}) \geq \mathrm{L}(\theta,\phi;\mathrm{x}^{(\mathrm{i})})=\mathbf{E}_{q_\phi(\mathrm{z}|\mathrm{x}^{(\mathrm{i})})}[-\log q_\phi(\mathrm{z}\,|\,\mathrm{x}^{(\mathrm{i})})+\log p_\theta(\mathrm{x}^{(\mathrm{i})},\mathrm{z})] \tag{2-2}$$

又因为 $\mathrm{L}(\theta,\phi;\mathrm{x}^{(\mathrm{i})})$ 可以进一步写做：

$$\mathrm{L}(\theta,\phi;\mathrm{x}^{(\mathrm{i})})=-D_{KL}(q_\phi(\mathrm{z}\,|\,\mathrm{x}^{(\mathrm{i})})\,||\,p_\theta(\mathrm{z}))+\mathbf{E}_{q_\phi(\mathrm{z}|\mathrm{x}^{(\mathrm{i})})}[\log p_\theta(\mathrm{x}^{(\mathrm{i})},\mathrm{z})] \tag{2-3}$$

于是，想要最大化 $\log p_\theta(\mathrm{x}^{(\mathrm{i})})$ 就需要最小化 $\mathrm{L}(\theta,\phi;\mathrm{x}^{(\mathrm{i})})$ 的第一项 KL 散度、最大化第二项的平均联合概率。这里的先验分布 $p_\theta(\mathrm{z})$ 常取标准正态分布。结合到我们的模型中，前一项表示为编码网络输出 z 与标准正态分布的交叉熵损失，后一项则用重构损失表示。

我们想通过将 $\mathrm{L}(\theta,\phi;\mathrm{x}^{(\mathrm{i})})$ 关于变分参数 ϕ 和生成参数 θ 做微分来对其进行优化，但普通的蒙特卡罗梯度估计法在这里行不通，因其表现出非常大的波动。于是想到利用重参数技巧，使得整个生成模型变成全局可微，具体过程如下所示。

首先，我们利用可微变换 $g_\phi(\varepsilon,\mathrm{x})$ 将随机变量 $\tilde{\mathrm{z}}\sim q_\phi(\mathrm{z}|\mathrm{x})$ 重参数化，ε 是一个辅助噪声变量：

$$\tilde{\mathrm{z}}=g_\phi(\varepsilon,\mathrm{x}) \quad \text{其中，} \varepsilon\sim p(\varepsilon) \tag{2-4}$$

接着，我们就可以用蒙特卡罗来估计特定函数 $f(\mathrm{z})$ 关于 $q_\phi(\mathrm{z}|\mathrm{x})$ 的期望，

$$\mathbf{E}_{q_\phi(\mathrm{z}|\mathrm{x}^{(\mathrm{i})})}[\mathrm{f}(\mathrm{z})]=\mathbf{E}_{p(\varepsilon)}[f(g_\phi(\varepsilon,\mathrm{x}^{(i)}))]\simeq\frac{1}{L}\sum_{l=1}^{L}f(g_\phi(\varepsilon^{(l)},\mathrm{x}^{(i)})),\text{其中}\varepsilon^{(l)}\sim p(\varepsilon) \tag{2-5}$$

最后，把这个重参数技巧运用到公式(2-2)中的可变下界，就可以得到一般化的随机梯度变分贝叶斯估计 $\mathrm{L}^{\mathrm{A}}(\theta,\phi;\mathrm{x}^{(\mathrm{i})})$：

$$\begin{aligned}&\tilde{\mathrm{L}}^{\mathrm{A}}(\theta,\phi;\mathrm{x}^{(\mathrm{i})})=\frac{1}{L}\sum_{l=1}^{L}p_\theta(\mathrm{x}^{(\mathrm{i})},\mathrm{z}^{(i,l)})-\log q_\phi(\mathrm{z}^{(i,l)}\,|\,\mathrm{x}^{(\mathrm{i})}),\\&\text{其中，}\quad \mathrm{z}^{(i,l)}=g_\phi(\varepsilon^{(i,l)},\mathrm{x}^{(i)})\text{且}\varepsilon^{(l)}\sim p(\varepsilon)\end{aligned} \tag{2-6}$$

在模型的实际实现中，重参数技巧指的是，将编码器输出编码进行重构的第一步不是直接从编码器输出分布中采样，而是利用一个服从某个分布（如标准正态分布）的辅助随机变量，从该分布中进行采样。编码器的输出不再是采样值，而是所需分布的参数（如正态分布的均值和方差）。从而，当进行损失的反向传播时，损失对应生成网络的可训练参数是可导的，优化器能够自然地通过更新网络参数对网络进行优化。

2.2 对抗生成网络

2.2.1 原始问题描述

类似于 2.1 节变分自编码器中所述，已知一个数据集合 $\mathrm{X}=\{\mathrm{x}^{(i)}\}_{i=1}^{N}$，包含 N 个独立同分布样本。为了实现其生成过程，我们想要从数据集 X 中学习到随机变量 x 的分布，于是问题转化为如何学习这个分布。同样，假设存在一个服从某个先验分布的变量 z，我们需要构建一个生成器将变量 z 映射到变量 x。那么如何去学习这个生成器的参数，或者说如何去衡量这个映射关系的正确性？这里便引入了博弈论中零和博弈的思想：为了实现生成器 G 的参数学习，加入了一个判别器 D 来配合其训练，于是整个对抗生成网络由 G、D 两个网络构成，通过交替训练实现各自性能的提升，即解释了对抗生成网络中对抗二字的由来。

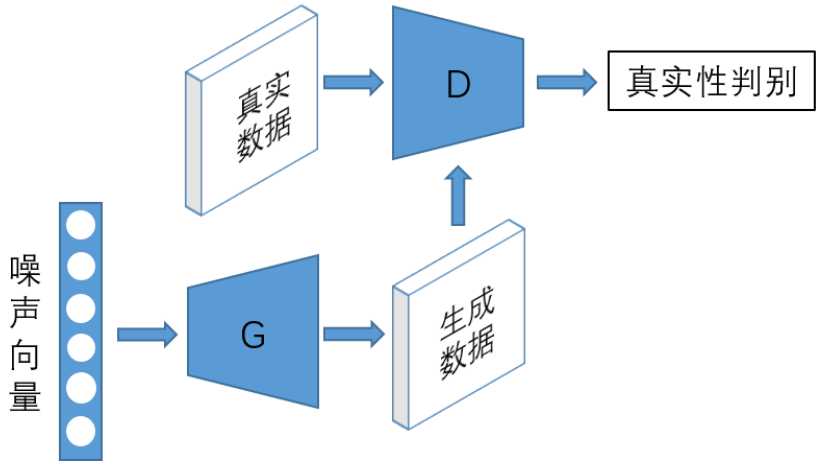

图 2-2 对抗生成网络整体架构

2.2.2 原始问题分析

上文中提到的 G 网络的目标是逼近数据 x 的真实分布 $p_{\mathrm{data(x)}}$，其学习到的分布记做 p_g，D 网络的目标是区分来自真实分布与 G 网络学习到的分布的数据。具体如图 2-2 所示，一方面，生成器先从服从特定分布（如均匀分布、标准正态分布）的变量 z 中随机采样，然后将采样值输入通过一个可微网络 $G(\mathrm{z};\theta_g)$ 映射到数据空间；另一方面，判别器 $D(\mathrm{x};\theta_d)$ 要尽可能的将来自训练数据的真实样本判别为真，并将来自 G 网络生成的数据样本识别为假。于是，对抗生成网络的优化目标函数可以表示为：

$$\min_{G}\max_{D}\mathrm{V}_{\mathrm{GAN}}(D,G)=\mathbf{E}_{\mathrm{x}\sim p_{\mathrm{data(x)}}}[\log D(\mathrm{x})]+\mathbf{E}_{\mathrm{z}\sim p_{\mathrm{z(z)}}}[\log(1-D(G(\mathrm{z})))] \tag{2-7}$$

整个对抗生成网络的训练流程可以理解为先将两个网络 G 和 D 初始化，然后交替的固定其中一个网络的参数，通过优化损失梯度反向传播更新另一个网络的参数。假设变量 z 服从均匀分布，网络初始化后的结果如图 20-3（a）所示，最下方的水平线表示 z 的分布，在这里是均匀分布，z 从中随机采样，向上的箭头表示 G 网络将 z 映射到 x 的过程。此时，生成器 G 输出的分布与真实数据分布存在较大差异，如果将 G 网络参数固定，单独训练 D 网络，判别器能够逐步提升判别性能，直到如图 20-3（b）所示，判别器会收敛到 $D^*(x)=\frac{p_{\text{data}}(x)}{p_{\text{data}}(x)+p_g(x)}$；接着，将 D 网络参数固定，仅仅更新 G 网络，D 网络能够引导 G 网络输出最大概率被判别为真的数据，如图 20-3（c）所示。如此循环反复，每当 D 网络性能提升到最佳即能够轻易区别真假样本时，就开始更新 G 网络提高其合成假样本的能力以尽可能的蒙骗 D 网络，两个网络各自朝着更好的方向优化，理论上它们最终会达到一个纳什均衡点，如图 20-3（d）所示，此时 $p_g=p_{\text{data}}$，即 G 网络完全学习到了真实数据的分布，D 网络再也不能区分这两个分布，$D(x)=\frac{1}{2}$，对抗生成网络的任务得到解

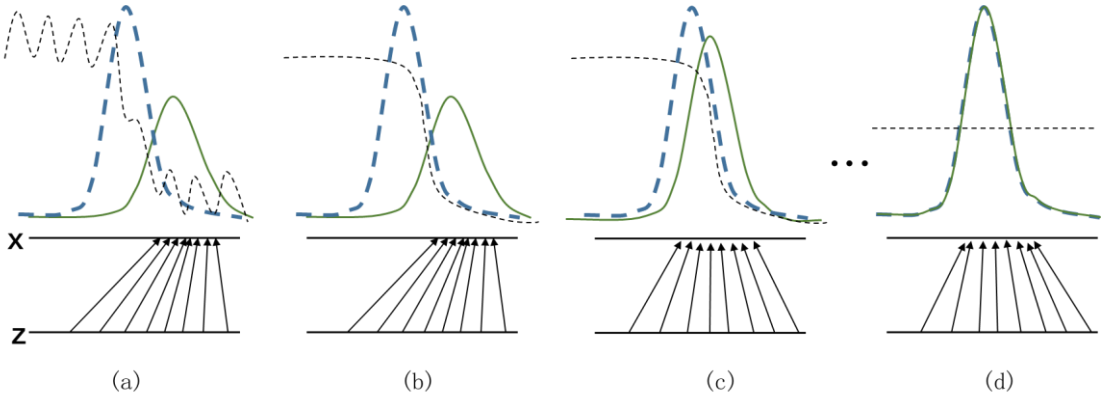

图 20-3 对抗生成网络随训练过程分布变换图[10]

（实线表示生成器 G 输出分布 p_g，细虚线表示判别器输出分布，粗虚线表示真实数据分布）

决。

2.2.3 问题扩展 1：条件对抗生成网络

对抗生成网络可以通过给生成器和判别器加上一些额外的条件 y 而扩展为条件对抗生成网络，这些额外的条件可以是任何额外的信息，如类别标签、待补全图像等。具体而言，可以通过将条件信息与原网络各自的输入组合（如通道拼接），一同作为生成器和判别器的输入。对于生成器而言，就是 y 与噪声变量 z 组合；对于判别器而言，即为 y 与生成器输出 x 组合。

类似式(2-7)，条件对抗生成网络的优化目标函数可以写做：

$$\min_G \max_D \mathrm{V}(D,G) = \mathbf{E}_{\mathrm{x}\sim p_{\mathrm{data(x)}}}[\log D(\mathrm{x}|\mathrm{y})] + \mathbf{E}_{\mathrm{z}\sim p_{\mathrm{z(z)}}}[\log(1-D(G(\mathrm{z}|\mathrm{y})))] \tag{2-8}$$

通过在生成器和判别器上加条件，能够引导数据的生成方向。

2.2.4 问题扩展 2：最小二乘对抗生成网络

在实际训练中，原始对抗生成网络的优化目标函数会导梯度消失的问题。因为在更新生成器时，如果生成器生成的假样本被判别器判别为真，这时根据交叉熵损失计算得到的损失会很小，反向传播的梯度也很小，更新缓慢。实际上这时的假样本与真实样本间的距离仍然较大，还有很大的提升空间。最小二乘对抗生成网络将判别器的损失函数从交叉熵损失改为最小二乘损失函数，有效的稳定了网络的训练过程，同时改善了网络输出图像的质量。

假定 a 和 b 分别是对于判别器 D 而言假样本和真实样本的标签，c 是生成器 G 想要判别器 D 认为的假样本的值，最小二乘对抗生成网络的优化目标函数可以写做：

$$\begin{aligned}
&\min_D \mathrm{V}_{\mathrm{LSGAN}}(D) = \frac{1}{2}\mathbf{E}_{\mathrm{x}\sim p_{\mathrm{data(x)}}}[(D(x)-b)^2] + \frac{1}{2}\mathbf{E}_{\mathrm{z}\sim p_{\mathrm{z(z)}}}[(D(G(\mathrm{z}))-\mathrm{a})^2] \\
&\min_G \mathrm{V}_{\mathrm{LSGAN}}(G) = \frac{1}{2}\mathbf{E}_{\mathrm{z}\sim p_{\mathrm{z(z)}}}[(D(G(\mathrm{z}))-c)^2]
\end{aligned} \tag{2-9}$$

2.3 自注意力机制

2.3.1 问题描述

基于深度卷及网络的对抗生成模型在图像生成任务上已经取得了不错的效果，但是进一步分析这些模型生成的图像样本仍然可以发现其存在的问题。卷积对抗生成网络在多类别数据集（如 ImageNet）上训练时，相对其他类别的生成结果，它在生成某些类别时更加糟糕。例如，当前性能最好的 ImageNet GAN 模型，它能够很好地合成带有很少结构约束的类别的图像，如天空、大地、海洋等，这些类别没有确定的形状约束，往往通过纹理来加以区分，但是它很难学习到几何结构模式，这就使得生成具有特定结构的类别效果较差，如生成的狗狗图像的毛发能够做到很真实但是狗爪的轮廓没有很清晰的界定。一个可能的原因是之前的模型都严重依赖于卷积操作来学习不同图像区域间的相关性。由于卷积操作只有固定的局部感受野，要想学习到大范围的相关性只能通过增加几层卷积层来扩大感受野。这就带来了一个矛盾：小模型不能很好学习图像区域间的关联性，因为优化算法很难寻找到每个层合适的参数值来理想的表征相关性；而增加卷积核的尺寸一方面确实提高了网络的表示能力，但另一方面丢失了卷积结构本身的计算效率。

2.3.2 问题分析

在上一小节中，我们讲到现在广泛使用的卷积操作在实现大范围相关性学习和计算开销之间存在不可协调的矛盾。于是这里引入自注意力机制，它在学习图像区域的大范围相关性和计算效率之间取得了一个良好的平衡。在该自注意力机制下，图像每个像素点上的响应都是所有位置的特征的加权和，这里的权重，或者叫做注意力向量，只需要耗费很少的计算量。

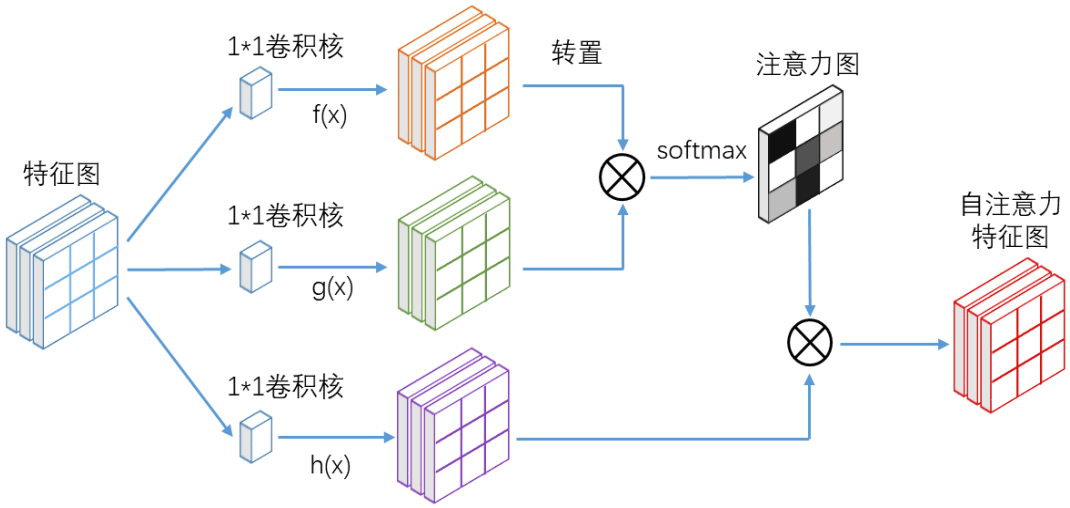

图 2-4 自注意力机制整体架构
（这里的⊗指的是矩阵相乘操作，softmax 对每一行操作）

自注意力机制整体架构如图 2-4 所示。前一隐藏层的特征图作为输入 $\mathrm{x}\in\mathbb{R}^{C\times N}$，这里的 N 代表像素点个数，即，特征图的长乘以宽，C 代表特征图通道数。首先，它分别通过三个卷积核大小为 1*1 的卷积操作变换到三个特征空间 f、g 和 h，其中，$\mathrm{f(x)}=W_f\mathrm{x}$，$\mathrm{g(x)}=W_g\mathrm{x}$，$\mathrm{h(x}_i)=W_h\,\mathrm{x}_i$，$W_f\in\mathbb{R}^{\bar{C}\times C}$，$W_g\in\mathbb{R}^{\bar{C}\times C}$，$W_h\in\mathbb{R}^{C\times C}$，在我们的模型中 $\bar{C}=C/4$。然后根据前两个特征空间来计算注意力图，注意力图中的每个元素的计算方式可以写为：

$$\beta_{j,\mathrm{i}}=\frac{\exp(s_{ij})}{\sum_{i=1}^{N}\exp(s_{ij})},\text{其中，}\ s_{ij}=\mathrm{f}(\mathrm{x}_i)^T\mathrm{g}(\mathrm{x}_j) \tag{2-10}$$

$\beta_{j,\mathrm{i}}$ 指的是在生成第 j 个位置时关注第 i 个位置的程度，换句话讲，是第 i 个位置的值对第 j 个位置的影响程度。紧接着，将得到的注意力图与第三个特征空间做运算可以得到注意力层输出 $o=(o_1,o_2,\ldots,o_j,\ldots,o_N)\in\mathbb{R}^{C\times N}$，其中，每个元素的计算方式如式(2-11)所示。

$$o_j=\sum_{i=1}^{N}\beta_{j,\mathrm{i}}\mathrm{h}(\mathrm{x}_i) \tag{2-11}$$

最后，我们将注意力层输出乘上一个尺度变换参数γ，并与最初输入的特征图求和得到自注意力机制的输出。于是，最终的输出可以表示为：

$$y_i = \gamma o_i + x_i \tag{2-12}$$

这里的γ初始化为 0。因为让网络一开始就学习到每一个像素点与除它之外的所有像素之间的关系是很难的，这样可以使得网络最初只依赖于邻近的像素信息，显然，这更容易一些，然后通过学习，网络会逐步增大γ的值，开始加大非局部区域的权重，每个像素点开始建立起与所有像素点的联系。这样的想法与日常中人们的行为模式是一致的：人们倾向于先学习简单的任务，在掌握这些技能的基础上再去拓展更难、更复杂的事情。

通过这种网络参数学习的方式，可以做到特征图不同位置对某一位置像素点的影响各不相同，且任何一个位置都能对其他位置产生影响，而不像传统卷积操作，每个点只受卷积核大小控制的周边几个像素点影响。

2.4 残差网络

2.4.1 问题描述

自深度卷积神经网络出现后，它就展现出了强大的学习能力。它能有效结合底层（接近像素级别）、中层、上层特征，并且随着层数的增加，其学习到的特征越加丰富。那么，这是否说明网络的层数越多，网络的学习能力就越强呢？答案显然是否定的。伴随着网络的深入，第一个出现的问题就是梯度消失或是梯度爆炸阻碍网络的收敛，因为在进行梯度反向传播时一旦某个神经元节点的梯度过大或者过小，这个效应就会随着层数进行积累，层数越多，问题越明显。虽然这个问题可以通过初始归一化、网络中间加上归一化层来解决，但是另一个问题却逐渐浮现了出来：模型退化，意思是说随着网络层数增加，网络的性能不升反降。相关研究者在一个训练好的、性能良好的网络上又增加了若干层，实验结果却显示模型的性能下降了，这就说明模型退化的问题不是由过拟合造成的。

模型退化的问题说明不是所有模型的优化难度都是相似的。理论上讲，在一个浅层网络 A 上再加上若干层变成深层网络 B，网络 B 的性能或许不能提升，但至少能做到与浅层网络 A 相持平，性能相等的情况出现在 B 前面层的参数与 A 一致，后面增加的几层则为恒等映射，这个理想解是存在的，但是现实情况却很难做到，网络常常趋于退化。

2.4.2 问题分析

为了解决上一小节中引出的模型退化的问题，这里引入残差学习模块。其核心思想是：我们不再要求网络 B 新加的层直接去学习潜在的映射关系，而是让其学习该映射关系与输入之间的差异。

令 $H(x)$ 为需要网络 B 新加的若干层去学习的一个映射关系，x 指的是这几层中第一层的输入。如果我们假设这几个非线性层能够无限逼近函数 $H(x)$，那么它们也一定能逼近残差函数 $H(x)-x$（在 $H(x)$ 与 x 维数相同的情况下），这是等价命题，但是前者的难度大于后者。因为正如上一小节所述，当网络 B 的前面数层已经能够很好地学习数据的特征时，后面增加的层的综合效应理应是恒等映射，前者让若干非线性层去学习恒等映射 $H(x)=x$，而后者只需要学习 $F(x):=H(x)-x=0$，即让输出恒为零，这通过调整这几层的可学习参数为 0 就能实现。如此改造网络的还有一个好处是跳跃连接的 x 既没有引入额外的网络参数也没有增加计算的复杂度。因此，我们选择让网络去拟合 $F(x):=H(x)-x$ 而非 $H(x)$，原本需要学习的映射函数就变成了 $F(x)+x$。

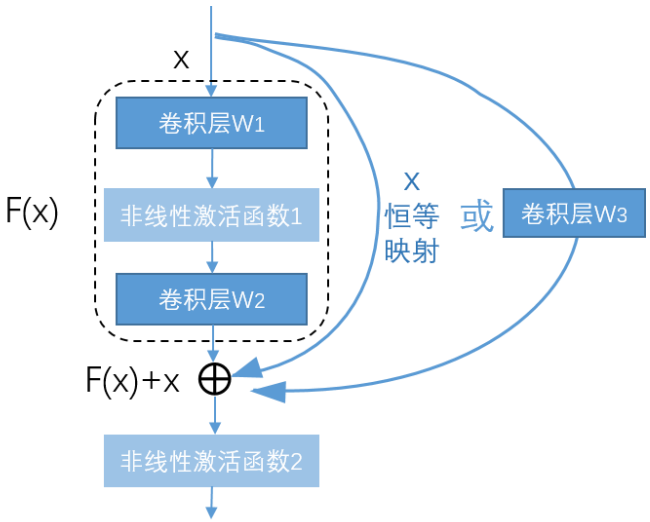

图 2-5 残差学习模块架构示例

残差学习模块的架构示例如图 2-5 所示。我们定义：

$$y=F(x,\{W_i\})+x \tag{2-13}$$

这里的 x、y 分别表示模块的输入和输出，$F(x,\{W_i\})$ 是要学习的残差映射，在此示例中 F 是两个卷积层外加一个非线性激活函数 σ，公式化为 $F=W_2\sigma(W_1 x)$，F+x 是跳跃连接过来的 x 与残差映射输出 F 的对应元素求和，我们在求和过后又加了一个非线性激活函数 σ。如果 F 的维度与 x 不一致，还可以适当改造网络的跳跃连接部分，将恒等映射 x 换成一个线性映射（如卷积层）来匹配维度，残差学

习模块的输出就变成了式(2-14)。

$$F=W_2\sigma(W_1 x)+W_3 x \tag{2-14}$$

2.5 本章小结

本章针对本文的时尚图像补全网络所涉及到的模型、算法作了一一介绍，并针对每个模型进行了详细的问题描述和问题分析。首先对两大生成网络：变分自编码器和对抗生成网络进行了原理解释，特别地，仔细讲解了前者的重参数技巧和后者的两个衍生模型：条件对抗生成网络和最小二乘对抗生成网络，因这些模型和技巧会在本文提出的模型中用到。其次介绍了自注意力机制和残差网络的原理及应用示例，它们是本文模型借鉴到的模型构建思路，能够有效改善训练过程和提升生成质量。

第三章 时尚图像补全网络的模型建立

我们的时尚图像补全网络的任务是，给定一张缺失了一件服装单品的全身人像，也就是将对应部分的像素值全置为零，它需要通过学习其他可见类别的服装单品的风格来补全被遮挡区域，合成真实同时外形、纹理多样的服装。补全部分不仅要与可见区域衔接自然而且要具有良好的视觉兼容性。因此，生成的图像还可以被应用到时尚服装推荐的任务上。但是，由于服装图像一旦变形就会严重影响图像的真实性，外加不同服装单品间纹理的糅合也会大大降低图像生成质量，我们采用了一个二阶生成模型来解决这个问题，将整个图像补全网络拆分为一个形状生成网络和一个纹理合成网络，从而将生成图像的难度降低，尽可能的避免在纹理合成时不同类别服装单品间的干扰。

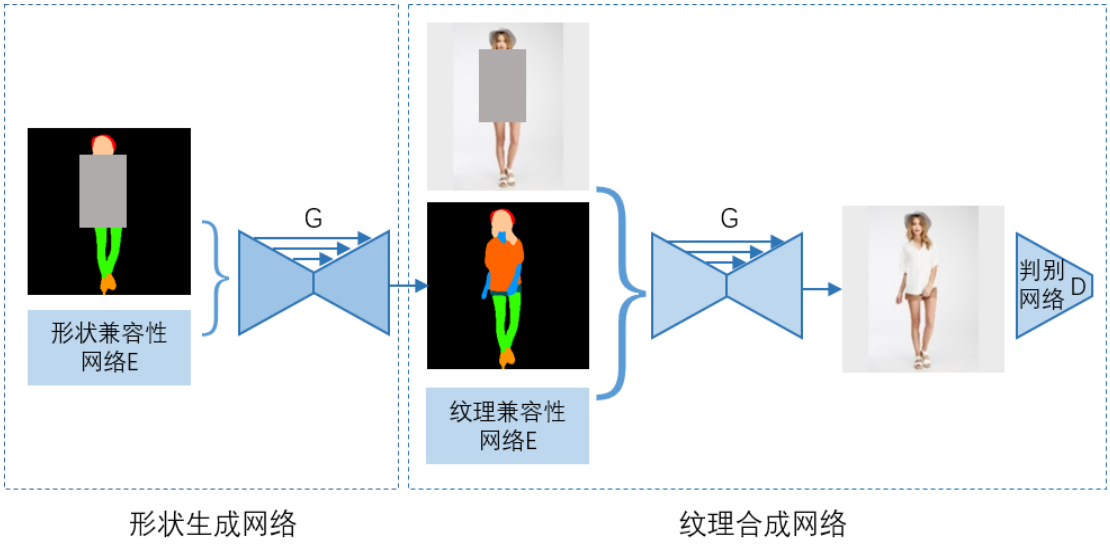

图 3-1 二阶时尚图像补全网络整体结构[9]

我们整体的二阶时尚图像补全网络整体架构如图 3-1 所示，该网络在 FiNet[9] 的基础上，调整了网络的具体结构，并引入了一些训练技巧。左边虚线框内是形状生成网络，包含一个形状兼容性网络和一个语义分割图生成网络，右边虚线框内是纹理合成网络，包含一个纹理兼容性网络、一个真实图像生成网络和一个真假判别网络。接下来的两小节会分别具体介绍两个阶段的网络结构以及损失函数。

3.1 形状生成网络

图 展示了形状生成网络的整体架构，它包含了一个基于编码器-解码器结构的生成器 G_s，以及两个编码器 E_s 和 E_{sc}，前者的作用是通过重构的方式合成完整的人体语义分割图，后两个共同作用于 G_s，约束 G_s 网络的生成过程，使其合成既具备视觉兼容性又多样化的结果。更具体的说，形状生成网络的目标是用 G_s 网络去学习一个映射关系，即，将一张缺失某个服装类别的人体语义分割图 $\hat{S}$ 外加人体表征 p_s 映射到完整的人体语义分割图 S，形状编码器 E_s 的输出作为 G_s 的生成条件。

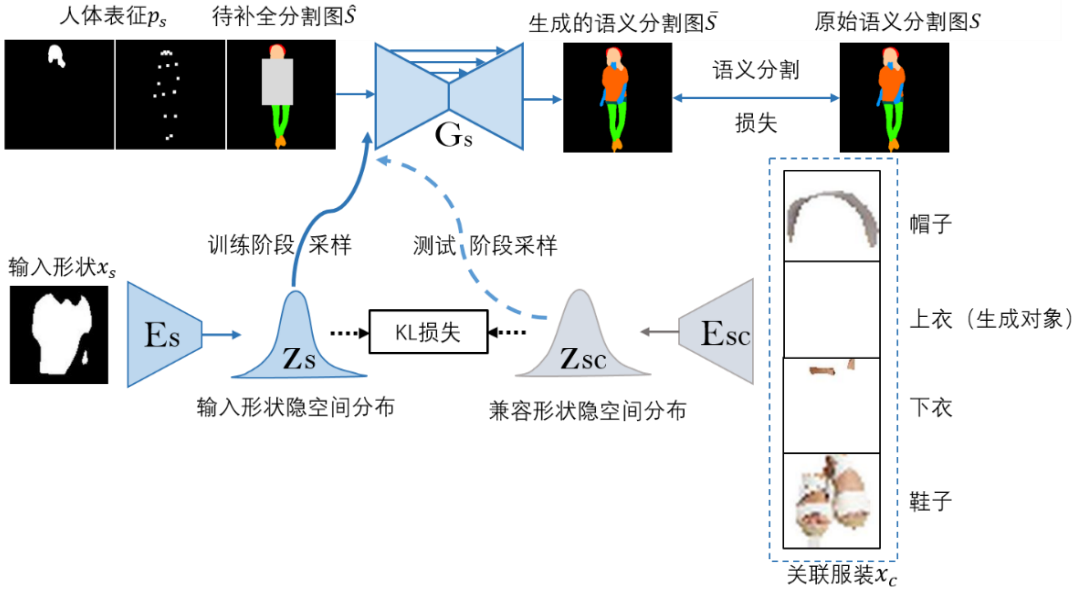

图 3-2 形状生成网络整体架构[9]

为了获取用于训练网络 G_s 的真实人体语义分割图，我们利用了 Look Into Person 数据集上预训练好的人体语义分割模型[25]。具体来说，给定一张输入图像 $I \in R^{H\times W\times 3}$，我们首先借助训练好的人体解析器获取对应的人体语义分割图，由于该网络输出的总分割类别数为 20，为了简化问题，我们将其重新整合为 8 大类：1、脸和头发，2、上半身皮肤（躯干和胳膊），3、下半身皮肤（腿），4、帽子，5、上半身衣服（上面的衣服和外套），6、下半身衣服（裤子、短裙和连衣裙），7、鞋子，8、背景（其他类别）。接着将这八类分割结果转换为 8 通道的二值图 $S \in \{0,1\}^{H\times W\times 8}$，就得到了重构的人体语义分割图的真实值。同时，通过遮挡该真实值的特定类别区域，能得到 G_s 所需的输入 $\hat{S}$，比如说在图 3-2 中，当我们想要合成上半身衣服时，我们就可以利用二值图 S 得到上衣的大致区域，进一步得到能够框住上衣区域的最小边界框，为了确保覆盖的完整性，将边框进行适当放大，然后将二值图对应的该选定框内的值全部重置为零，就只剩下了可见区域 $\hat{S}$。

网络 G_s 的输入除了可见区域 $\hat{S}$ 以外，还要用到服装不可知的人体表征 p_s，以保证形状重构过程中姿势和人体特征的一致性。这里采用类似于[6]中用到的人体表征 p_s，包括人体姿势表征、头发和脸部的分割图。具体来说，人体姿势表征是一个 18 通道的热度图，我们利用了一个在 COCO 关键点检测数据集上预训练好的姿势检测模型[26]来提取人体关键点。头发和脸部的分割图则是一个单一通道的二进制掩膜，通过判断此前得到的人体语义分割图的类别是否为头发或脸部即可获得，若是该两类中任意一类，则像素点值置为 1，反之为 0。将人体姿势表征、头发和脸部的分割图在通道上进行拼接就得到了人体表征 $p_s \in \mathbb{R}^{H\times W\times C_s}$，其中通道数 $C_s=18+1=19$。

利用以上预处理得到的待补全图像 $\hat{S}$ 和不含有服装信息的人体表征 p_s 来重构真实的人体语义分割图 S，一种直接的思路就是把它作为一个图像到图像的翻译任务，即，生成器的编码器部分将输入 $\hat{S}$ 和 p_s 映射到隐空间，解码器输出将隐空间变量映射到输出，使其尽可能的逼近 S，这样的方法能够实现图像生成，但是会导致输出结果是唯一确定的，这不满足我们任务所需的多样性。因此，我们借鉴上一章中介绍的变分自编码器的思想，将编码器输出的隐空间变量 $z_s \in \mathbb{R}^z$ 当做约束条件加到生成器上，每次生成时都从变量 z_s 服从的分布中进行采样，利用采样值的不同实现生成结果的多元化。

为了获取生成器的条件 z_s，我们引入一个形状编码器 E_s。又因为我们的目标是生成不同形状的服装来补全缺失区域，我们希望编码器在训练时能有效的学习到待生成服装的形状信息。具体操作是，我们给定这个网络的输入为待生成服装的形状 x_s，类似于头发和脸部的分割图的获取方式，x_s 也是通过人体语义分割图 S 得到的单通道二进制掩膜，属于待生成服装类别的像素点标记为 1，反之记为 0。然后，我们的形状编码器 E_s 要输出 z_s，表示为 $z_s \sim E_s(x_s)$，这里用到了上一章提到的重参数技巧。通过重参数技巧，我们使得整个网络的损失函数可微，利用梯度的反向传播即可更新网络参数，实现模型的端到端训练。类似于变分自编码器，我们希望 z_s 在训练时能够逼近标准正态分布 $N(0,\mathrm{I})$，这样我们在测试时，也就是 x_s 未知的情况下，就能够从标准正态分布中进行随机采样，采样值再作为生成器的条件 z_s。KL 散度常用来度量两个分布的相似程度，因此，我们通过计算编码器输出分布与标准正态分布之间的 KL 散度作为网络的 KL 损失，具体表达式如式(2-15)所示。

$$\begin{aligned} L_{KL} &= D_{KL}(E_s(x_s)\,\|\,N(0,\mathrm{I})), \\ D_{KL}(p\,\|\,q) &= -\int p(z)\log\frac{p(z)}{q(z)}dz \end{aligned} \tag{2-15}$$

有了形状编码器学习到的 z_s，再加上之前处理得到的待补全图像 $\hat{S}$ 和服装未知的人体表征 p_s，我们可以将这三者一同作为生成器 G_s 的输入来补全完整的人体语义分割图 $\bar{S}$，可以表示为 $\bar{S}=G_s(\hat{S},p_s,z_s)$。为了优化生成器 G_s，我们需要构造一个评判语义分割图质量的损失函数。因为交叉熵损失常在分类任务中用于衡量分类错误，我们这里选择计算合成的人体语义分割图 $\bar{S}$ 与真实的人体语义分割图 S 之间的交叉熵损失，具体表达式如式(2-16)所示。

$$L_{seg}=-\frac{1}{HW}\sum_{m=1}^{HW}\sum_{c=1}^{C}S\log(\bar{S}) \tag{2-16}$$

其中，C=8，C 表示人体语义分割图的通道数。

整个网络的损失函数可以表示为 KL 损失和语义分割交叉熵损失的加权和，即：

$$L=L_{seg}+\lambda_{KL}L_{KL} \tag{2-17}$$

λ_{KL} 是平衡两项损失的权重，在训练时，形状编码器 Es 和生成器 G_s 能够通过最小化损失 L 同时进行优化，在测试时，我们只要从标准正态分布中随机采样得到 z_s，然后联合已知的 $\hat{S},p_s$ 得到不同的重构结果 $\bar{S}=G_s(\hat{S},p_s,z_s)$。

尽管此时我们的形状生成网络已经能合成不同的服装形状，但是它却没有考虑到不同类别服装之间的兼容性。因此，我们利用可见区域 $\hat{S}$ 中的其他类别服装来显式的约束采样过程，使其生成过程做到视觉的兼容性，这里我们将可见的其他类别服装单品称作关联服装，用符号 x_c 表示。为了学习到服装间的风格兼容性，我们又引入了一个形状兼容性编码器 E_{sc}，该编码器联合之前的形状编码器 E_s 能够学习到待合成服装单品与关联服装之间的相关性。这样做的原因是我们认为一件服装的形状会受其关联服装的影响，这是很容易理解的，因为日常生活中，我们在穿搭时总是会考虑到不同服装间的视觉兼容性，即使服装的纹理不同，它们的风格还是相一致的，比如说一条男式长裤比一条女士长裙更搭配一件男式长袖外套，那么在形状生成时，若给定的关联服装中有男式长袖上衣，我们就想让补全部分出现长裤的概率高于长裙的概率。这个思想在概念上同自然语言处理领域的两个主流模型：skip-gram 和 continuous bag-of-words（CBOW）模型[27]非常相像，它们都是利用某个位置的上下文文本的特征表征来预测该位置的词的特征。

为了获取关联服装 x_c，我们首先利用人体语义分割图 S 提取每一类服装单品的实际图像，通过补零、尺寸缩放变成 256*256 大小，这里为了简化问题，我们又从 8 大类别中提取了 4 类具有代表性的服装单品：1、帽子，2、上衣，3、下衣，4、鞋子，然后将所有类别的图像按照以上顺序在通道上进行拼接，合成关联服装 x_c，注意，这里我们将待补全的服装类别对应的三个通道的值赋为 0，因为关联服装里不应出现指定生成的服装类别本身的图像。

获取关联服装 x_c 后，形状兼容性编码器 E_{sc} 就将 x_c 映射到了关联服装的隐空间 $z_{sc} \sim E_{sc}(x_c)$。因为在实际测试时待生成服装形状是不可知的，为了能用关联服装作为引导，更进一步的说，将关联服装的隐空间分布作为先验分布，待生成服装形状的隐空间分布作为后验分布，又由于我们认为符合视觉兼容性的一整套服装中的每件服装单品和它的关联服装是共享同一个隐空间的，那么形状编码器 E_s 的输出分布和形状兼容性编码器 E_{sc} 的输出分布应该相似才对，式(2-15)中的 KL 损失计算方式可以修改为：

$$\hat{L}_{KL} = D_{KL}(E_s(x_s) \| E_{sc}(x_c)) \tag{2-18}$$

该损失使得形状编码器 Es 的输出分布 z_s 和形状兼容性编码器 Esc 输出分布 z_{sc} 尽可能的共享一个隐空间，这同度量学习中的成对兼容性学习[28]非常相似，只不过因为之后需要通过从分布中采样获得随机性，我们是通过最小化两个分布之间的距离而非两个样本之间的距离来实现的。

如此一来，通过优化式(2-18)，整个生成过程 $\bar{S} = G_s(\hat{S}, p_s, z_{sc})$ 就不仅能够考虑到关联服装潜在的兼容性信息，还能在测试过程中、x_s 未知的情况下实现具备兼容性的采样，从而使得合成的服装形状与其他可见区域的服装风格相统一。最终的形状生成网络的损失函数也相应的修改为以下形式：

$$L = L_{seg} + \lambda_{KL}\hat{L}_{KL} \tag{2-19}$$

3.2 纹理合成网络

正如图 3-1 中所示，形状生成网络输出的补全人体语义分割图又作为了纹理合成网络的输入，来引导纹理合成、补全整张图像。如图 3-3 所示，纹理合成网络和形状生成网络非常相似，它们都有一个基于编码器-解码器结构且用于合成图像的生成器、两个用于学习视觉兼容性的编码器（纹理编码器 E_a 将待生成服装纹理图像 x_a 映射到隐空间 z_a，纹理兼容性编码器 E_{ac} 将关联服装 x_c 映射到隐空间 z_{ac}），只不过纹理合成网络相比形状生成网络，又多了一个用于分析合成图像真实性的判别网络 D_a。

同时，网络在细节上也发生了一些变化，具体如下所述：

1、纹理编码器 E_a 的输入变成了待生成服装的真实图像而非二值语义分割图；

2、纹理生成器 G_a 的任务不再是通过最小化交叉熵损失重构人体语义分割图，而是要实现原始的 RGB 图像 I 的重构；

3、纹理生成器 G_a 的输入发生改变，其中，待补全图像 $\hat{I}$ 从人体语义分割图变成 RGB 图像，人体表征 p_a 从头发、脸部的语义分割外加人体关键点热度图变成头发、脸部的 RGB 图像以及人体语义分割图 $S \in \mathbb{R}^{H \times W \times 8}$（训练时为真实的人体语义

分割图，测试时为形状生成网络的测试结果），相对于第一阶段的关键点热度图，人体语义分割图携带了更加丰富的人体信息，头发、脸部的图像使得在重构图像 $\bar{I}=G_a(\hat{I},p_a,z_a)$ 时能够尽可能的保持这些部分不变。

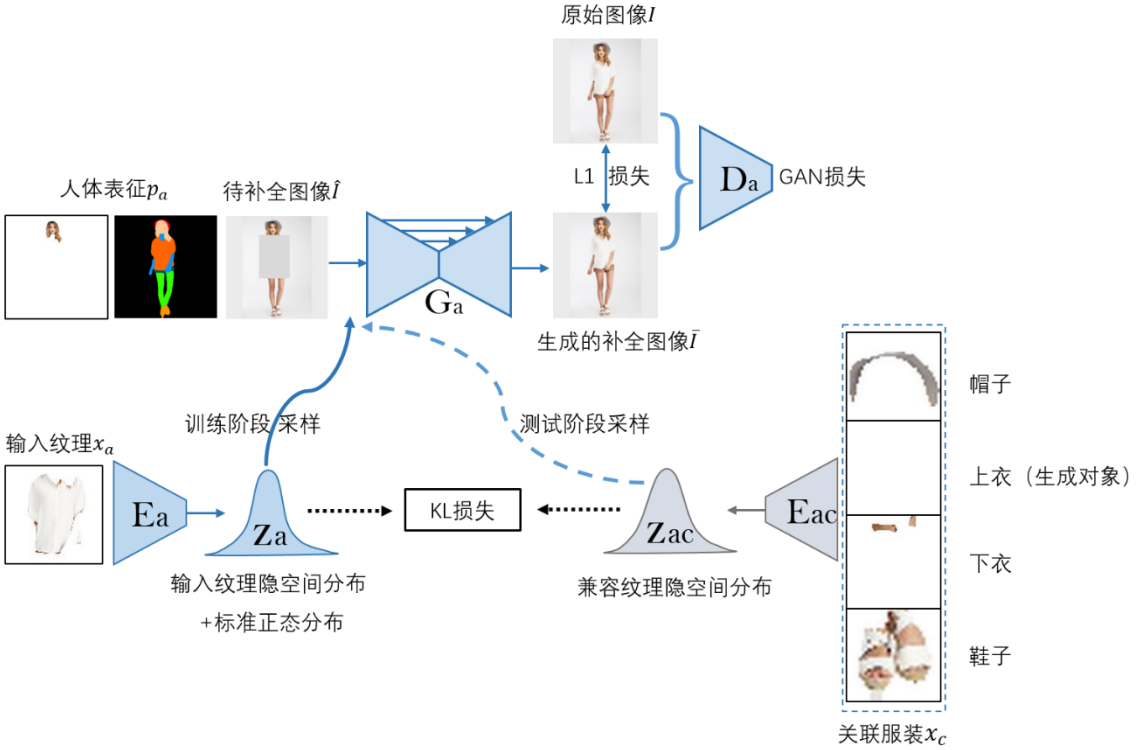

图 3-3 纹理合成网络整体架构[9]

纹理合成网络的损失函数相对于第一阶段的形状生成网络而言，更加复杂一些，大致可以分为三大类：KL 损失、外观匹配损失、对抗损失，以下将一一介绍每一类损失的构成情况。

1、KL 损失

类似于第一阶段，KL 损失可以表示为式(2-20)，即纹理编码器 E_a 和纹理兼容性编码器 E_{ac} 输出分布之间的差异。

$$L_{KL}=D_{KL}(E_s(x_s)\,||\,E_{sc}(x_c)) \tag{2-20}$$

2、外观匹配损失

外观匹配损失分为两部分：内容损失和风格损失。内容是一幅图像中不同物体的分布，因此让两个图像的上层特征尽量接近就能实现类似的内容。而风格相对抽象，用图像的纹理衡量更加确切一些，纹理信息反映在上层特征中就是特征图之间的相关性。

具体的，内容损失采用风格迁移任务中常常用到的感知损失，该损失的计算方式是，先将真实图像与生成图像分别输入到一个在 ImageNet 数据集上预训练好的 VGG-19 模型，然后保存标记为 conv1_2，conv2_2，conv3_2，conv4_2 和 conv5_2

的特征层输出，真实图像和生成图像分别对应的这五个特征层之间的距离，外加真实图像和生成图像本身在像素级别的距离，即为我们所需的内容感知损失，公式表示如(2-21)所示，其中，$\phi_0(I)$为真实图像，$\phi_0(\bar{I})$为第二阶段生成的图像，$\phi_l(I)$和$\phi_l(\bar{I})$表示第 i 个特征层输出，λ用来平衡不同特征层之间的权重。

$$L_{content}=\sum_{l=0}^{5}\lambda_l\left\|\phi_l(I)-\phi_l(\bar{I})\right\|_1 \tag{2-21}$$

相比只在像素级别做 L1 损失，特征层级别的差异更能反映图像结构上的差异，因为根据已有的卷积神经网络可视化研究[29]，我们可知，随着卷积网络的深入，网络学习到的特征逐渐从颜色、轮廓等相对底层的特征升级为纹理特征，直到学习到一些具有具体意义的实物个体。因此，通过最小化感知损失而非单一的图像 L1 损失更有利于生成真实的服装内容。

在计算风格损失时，我们则用到了格拉姆矩阵。格拉姆矩阵实际上就是特征之间的协方差矩阵，由概率论可知，协方差能够粗略反映两个变量之间的相关性（因为既没有减去均值也没有归一化），在我们的问题中就是不同特征层之间的相关性，因此，格拉姆矩阵能够反映一张图像的大致风格，要比较两张图像的风格差异，只需要计算它们的格拉姆矩阵的差。在计算格拉姆矩阵时也利用了计算感知损失时用到的 VGG-19 模型，在得到五个特征层的输出后，将每个特征图向量化后计算内积就得到了格拉姆矩阵$G_l\in\mathbb{R}^{C_l\times C_l}$，$C_l$为对应特征层的通道数，计算公式如下所示：

$$G_l(I)_{ij}=\sum_{k=1}^{H_lW_l}\phi_l(I)_{ik}\phi_l(I)_{jk} \tag{2-22}$$

得到格拉姆矩阵后，就可以将风格损失表示为下式，其中的γ用来平衡不同特征层之间的权重。

$$L_{style}=\sum_{l=1}^{5}\gamma_l\left\|G_l(I)-G_l(\bar{I})\right\|_1 \tag{2-23}$$

综合内容损失和风格损失，我们就得到了外观匹配损失：

$$\begin{aligned}L_{appearance}&=L_{content}+L_{style}\\&=\sum_{l=0}^{5}\lambda_l\left\|\phi_l(I)-\phi_l(\bar{I})\right\|_1+\sum_{l=1}^{5}\gamma_l\left\|G_l(I)-G_l(\bar{I})\right\|_1\end{aligned} \tag{2-24}$$

这里的参数λ和γ的选择分别参考了[6]和[30]。

3、对抗损失

不同于只从编码器输出的分布中进行采样得到生成器条件的形状生成网络，在训练纹理合成网络时，为了更好的提高判别网络D_a的鉴别能力和纹理生成器G_a

的输出图像质量，我们不仅从纹理编码器 E_a 输出的分布中采样，同时还从标准正态分布中进行采样，两个采样结果分别作为生成器 G_a 的条件得到输出 $\bar{I}$ 和 $\bar{I}_{\text{norm}}$。对于网络 D_a 来说，它的任务就是将尽可能的将原始图像 I 判为真、将生成器的两类输出 $\bar{I}$ 和 $\bar{I}_{\text{norm}}$ 判为假；反之，对于网络 G_a 和 E_a 来说，它们需要尽可能的蒙混网络 D_a，使得 $\bar{I}$ 和 $\bar{I}_{\text{norm}}$ 被判为真，只不过由于在生成 $\bar{I}_{\text{norm}}$ 过程中并没有 E_a 输出的指导，我们对它的生成效果要求更低。

对于不同的采样分布输入，生成器和编码器的损失函数表示不同：当条件输入是从纹理编码器 E_a 输出分布得到时，损失函数用平均特征匹配损失，如式(2-25)所示，

$$L_{adversarial}^{\bar{I}} = \left\| f_D(\bar{I}) - f_D(I) \right\|_2 \tag{2-25}$$

其中，$f_D(\bar{I})$ 和 $f_D(I)$ 表示判别网络 D_a 的最后一个特征层输出，该损失通过鼓励生成的图像 $\bar{I}$ 和真实图像 I 在判别网络中保持特征一致，加强了生成图像与真实图像的相似性。当条件输入是从标准正态分布中采样得到时，损失函数则参考了最小二乘对抗生成网络的损失函数形式，如式(2-26)所示，

$$L_{adversarial}^{\bar{I}_{\text{norm}}} = [f_D(\bar{I}_{\text{norm}}) - 1]^2 \tag{2-26}$$

因此，生成器和编码器的对抗损失可以表示为：

$$L_{adversarial}^{g/\text{e}} = L_{adversarial}^{\bar{I}} + L_{adversarial}^{\bar{I}_{\text{norm}}} \tag{2-27}$$

判别器的对抗损失则全部用最小二乘距离表示：

$$L_{adversarial}^{d} = [f_D(I) - 1]^2 + [f_D(\bar{I})]^2 + [f_D(\bar{I}_{\text{norm}})]^2 \tag{2-28}$$

综合以上三大类损失，我们可以得到生成器和编码器的损失函数、判别器的损失函数，如式(2-29)所示。通过优化 KL 损失、外观匹配损失、对抗损失，我们对纹理合成过程进行了约束，使其生成的图像做到真实性、视觉兼容性、多样性的统一。

$$\begin{cases} L_{g/e} = \lambda_{KL} L_{KL} + \lambda_{appearance} L_{appearance} + \lambda_{adversarial} L_{adversarial}^{g/\text{e}} \\ L_d = \lambda_{adversarial} L_{adversarial}^{d} \end{cases} \tag{2-29}$$

另外，在生成器和判别器中，我们还引入了注意力机制。当该机制被当做自注意力机制在解码器中使用时，它能有效利用远距离的空间信息；当跨越在编码层和解码层之间使用时，它能更好地捕获特征与特征之间的信息，因为该机制本身是用来引导图像生成时的关注重点的，当网络的编码层和解码层跳跃连接时，该机制能够获得更丰富的特征信息，网络能够根据实际情况，决定重点关注具有更加细粒度特征的编码层或是语义上对于生成更有力的解码层。

我们用到的注意力机制与原始的自注意力机制 24 相比有所不同，整体结构如图 3-4 所示。

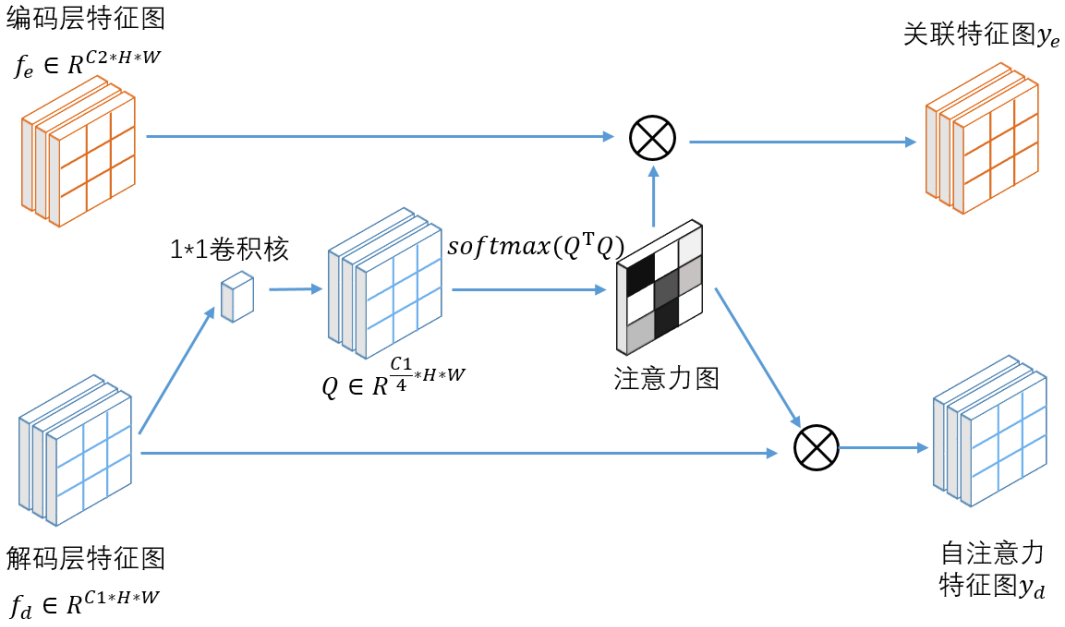

图 3-4 注意力机制整体架构

我们首先仿照自注意力机制，根据解码层特征图 f_d 得到注意力图，注意力权重计算方式为：

$$\beta_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^{N}\exp(s_{ij})}, \text{其中，} s_{ij} = Q(f_{di})^T Q(f_{dj}) \tag{2-30}$$

不同于自注意力机制，我们只用了一个 1*1 卷积，将解码层特征图变为 $Q(f_d)=W_q f_d$，将注意力图与解码层特征图做一系列运算得到自注意力特征图 y_d，表达式如下所示：

$$\begin{aligned} c_{dj} &= \sum_{i=1}^{N}\beta_{j,i} f_{di}, \\ y_d &= \gamma_d c_d + f_d \end{aligned} \tag{2-31}$$

计算完自注意力特征图后，我们再用编码层特征图和刚刚得到的注意力图来计算关联特征图 y_e，类似于式(2-31)，表达式可以写做式(2-32)，有些特别的是，这里的编码层特征图缺少了被遮挡部分的信息，它只对可见区域的的生成起指导作用，所以在计算关联特征图时使用了二值掩码 M 。

$$\begin{aligned} c_{ej} &= \sum_{i=1}^{N}\beta_{j,i} f_{ei}, \\ y_e &= \gamma_e(1-M)c_e + Mf_e \end{aligned} \tag{2-32}$$

3.3 本章小结

本章先概括的介绍了时尚图像补全网络的整体架构，后将整个网络分为形状生成网络和纹理合成网络两部分进行了详细论述，从模型的建立思路出发，将每一部分所涉及到的单个网络结构、每个网络的输入输出、所需数据的预处理、网络的优化目标函数及应用效果、网络的训练测试流程全面展开介绍。

第四章 模型的实验部分

4.1 实验设置

4.1.1 数据集

我们在 DeepFashion（In-shop Clothes Retrieval Benchmark）数据集[31]上进行了我们的实验，该原始数据集包括 52,712 张穿着时尚服装的人像图片。不同于以往那些需要成对数据的人像生成方法，例如，[6,7]等虚拟试装方法需要成对的穿着服装的人像和对应的服装产品图，[32,33]等基于姿势的人像生成方法需要成对的人身着同一服装但不同姿势的图像，我们的模型训练方法每次只需要单张人像，不过，因为我们的模型需要考虑服装间的兼容性，每张人像需要满足包含三件及以上服装单品的要求。依据以上要求，我们从原始数据集中筛选出了 9023 张符合标准的图像，随机选出 8023 张作为训练集、1000 张作为测试集并确保两者没有交集。

4.1.2 网络结构细节

我们的形状生成网络和纹理合成网络非常相似，所有网络的输入尺寸都是 256*256，所有的卷积操作后都加了谱归一化层[34]，网络的其他具体参数可以查看附录部分。形状编码器生成器 E_s、形状兼容性编码器 E_{sc}、纹理编码器 E_a、纹理兼容性编码器 E_{ac} 除了输入的通道数不同之外都一致，它们都输出一个正态分布的参数，从中采样得到隐变量 z，将该变量作空间扩展后，作为条件与生成器 G_s / G_a 的其他输入进行通道拼接，成为真正的网络输入。生成器 G_s 和 G_a 采用平均池化进行下采样，利用反卷积进行上采样，每一个采样层都采用了残差模块，下采样部分同编码器十分类似，仅仅移除了一个在最后用来生成分布参数的残差模块，上采样部分与下采样部分总体十分对称，只是在第二层后加了一个注意力层，该操作同时利用了上一层输出和下采样部分中对应的一层的输出。我们只在纹理合成网络加入了判别器 D_a，判别器可以看做是在生成器的下采样部分基础上，在第三层输出后加了一个自注意力层，在最后又加了一个残差模块、非线性操作和 3*3 的卷积操作。

4.1.3 模型优化细节

类似于以往基于编码器-解码器结构的生成网络的训练方式，我们采用 Adam 优化器，β_1=0.5，β_2=0.999，固定学习率 $lr = 0.0001$，形状生成网络共训练 20,000

次，纹理合成网络共训练 60,000 次，批大小都采用 16。不同损失函数的权重分别为：λ_{KL}=20，$\lambda_{appearance}$=20，$\lambda_{adversarial}$=1。

4.2 对比方法及评价指标

4.2.1 对比方法

为了验证我们的模型的有效性，我们将时尚图像补全网络（记做 N_{we}）与以下方法进行比较：

1、FiNet

FiNet[9]跟我们的任务一致，也通过二阶网络生成。不同点在于：我们的 N_{we} 整体网络规模相对于 FiNet 大大缩小，包括网络的深度、尺度、使用残差网络的个数以及相对应的网络可训练参数个数；我们的网络在第二阶段加入了一个判别器，利用对抗生成网络的思想进行图像生成；不同于 FiNet 只在补全部分计算损失，我们的损失函数覆盖了整个图像。

2、BicycleGAN

虽然 BicycleGAN [35]本身并不是做时尚图像合成的，但是由于它的原始任务——图像到图像的翻译也是图像生成任务，该模型也能实现我们的任务。为了做到比较的公平性，我们在筛选过后的 DeepFashion 数据集上重新训练了该模型。

3、没有使用二阶网络的 N_{we}（N_{we} w/o two-stage）

为了证实分两阶生成图像的确能够改善生成图像的质量，我们去掉生成人体语义分割图的步骤，利用第二阶段的模型一步生成 RGB 图像，唯一不同的是将生成器输入中的人体语义分割图替换成了人体关键点热度图。

4、没有使用对抗生成训练的 N_{we}（N_{we} w/o GAN）

为了证实对抗生成训练对提升图像质量的作用，我们在 N_{we} 基础上将第二阶段增加的判别器移除，不再使用生成器-判别器交替对抗的方式进行训练，仅仅使用外观匹配损失和 KL 损失优化网络。

5、无注意力机制的 N_{we}（N_{we} w/o Attention）

为了证实注意力机制能够使得合成图像具有更加清晰的结构，从而改善图像的视觉效果，我们在 N_{we} 基础上将所有网络（包括生成器、判别器）的注意力层、自注意力层移除，得到无注意力机制的 N_{we}。

6、无谱归一化的 N_{we}（N_{we} w/o SpecNorm）

为了证实谱归一化层的使用稳定了对抗生成网络的训练，我们在 N_{we} 基础上将第二阶段的所有网络（包括编码器、生成器、判别器）用到的谱归一化层移除，

得到无谱归一化的 N_{we}。

以上方法中，前两个属于其他研究的方法，用于检验模型的整体性能，后四个属于消融实验，即移除我们的模型的某类模块，用于证实该模块对改善网络性能的有效性和必要性。

4.2.2 评价指标

我们在评价以上方法时使用 Fréchet Inception Distance（FID）来衡量生成图像的质量，而没有利用对抗生成网络中常用的 Inception score（IS）。原因有以下几点：1、IS 基于的假设是：一个好的生成网络，它生成的每一张图像属于某个类别的概率应该远大于其他类别（说明生成图像界定清晰），同时，生成的图像在各个类别之间的分布应该尽可能的平均（说明不存在模式塌陷问题）。但是对于我们的生成网络而言，所有生成结果都是指定类别的服装图像，不可能做到类别的多样化，用 IS 衡量图像质量就不可行；2、不同于仅仅考虑生成样本之间关系的 IS 指标，FID 考虑了生成样本和真实数据之间的关系，它利用预训练好的 Inception V3 模型来提取顶层特征，在特征层面比较生成图像与真实图像的距离，是更加客观、更为全面的评价指标，更符合现实中人类所感知的图像真实性。

4.3 生成结果的定量分析

依据评价指标 FID，我们对各个方法生成的图像真实性进行定量评估，同时，为了评价的公平性，我们将各方法所使用到的各网络的参数个数标注出来，方便我们比较网络性能和网络规模，具体数据如表 4-1 所示，其中，‘/’表示该模型不存在对应网络。

对比表 4-1 的前三行，可以看到，我们采用的时尚图像合成网络 N_{we} 虽然在网络规模上比 FiNet、BicycleGAN 小很多，但是合成图像的质量比后两者反而更好，证实了我们的模型的有效性。N_{we} 和没有使用二阶生成网络的 N_{we} 区别不是特别明显，在一定程度上说明人体关键点热度图已经能够给网络带来不少人体信息，但是可视化结果显示，该网络不能对服装形状和纹理分开控制，这对于实际应用可能是一个问题。对比 N_{we} 和没有使用对抗生成训练的 N_{we}，即使网络规模略有缩小，但是生成的图像质量严重变差，这在定性分析中会有更直观的感受。对比 N_{we} 和无注意力机制的 N_{we}，虽然加入注意力机制使得生成器和判别器的可训练参数都有所上升，但是差异并不太大，且 FID 结果显示该机制的确有助于图像生成质量的提升。类似的，对比 N_{we} 和无谱归一化的 N_{we}，加入谱归一化层使得所有网络的可训练参数只有略微增加，但是图像生成质量却有大幅度提升，不仅如此，在下一小

节的定性分析中，我们还会看到谱归一化对于稳定对抗生成网络训练起到的作用。

表 4-1 不同生成模型的定量分析

生成方法	FID	G 网络参数个数	E 网络参数个数	Ec 网络参数个数	D 网络参数个数
N_{we}	**23.07**	3.0506M	0.9601M	0.9631M	1.5893M
FiNet	34.68	96.0798M	40.9167M	40.9219M	/
BicycleGAN	28.02	54.7950M	2.5900M	/	3.4590M
N_{we} w/o two-stage	24.28	3.0301M	0.9601M	0.9631M	1.5893M
N_{we} w/o GAN	35.90	3.0268M	0.9601M	0.9631M	/
N_{we} w/o Attention	27.32	2.5672M	0.9601M	0.9631M	1.1060M
N_{we} w/o SpecNorm	29.78	3.0208M	0.9402M	0.9521M	1.5782M

综上所述，我们的模型不仅在生成图像的性能上超越了网络规模更大的其他方法，消融实验还证明了使用二阶生成模型、对抗生成训练、注意力机制以及谱归一化的必要性。

4.4 生成结果的定性分析

4.4.1 真实性

在图 4-1 中，我们展示了六组用不同生成方法得到的图像，目标生成服装类型是上衣，每一行对应同样的输入，每一列对应一种生成方法，第一列是待补全图像 $\hat{I}$，第二列是真实的图像 I。

从横向来看，首先观察前两行，可以看到我们的方法和 BicycleGAN 在上衣与下衣的衔接部分做的最好，FiNet 不理想的原因可能是在计算损失时只考虑待补全部分的差异，网络在优化时遮挡部分没有受到可见区域的影响；没有使用二阶网络的 N_{we} 在视觉上也生成了不错的效果，但是由于其生成过程没有第一阶段输出的语义分割图的指导，无法分别控制形状和纹理的变化；无注意力机制的 N_{we} 产生的不自然的颜色、纹理衔接则证实了注意力机制的引入能够增强生成网络对于图像整体结构的把控能力。接着观察中间两行，注意到我们的模型在生成具有内外层

次结构的服装图像时得到了最好的结果，FiNet 几乎不能区分内外服装，BicycleGAN 生成的服装界限不分明且纹理杂乱，无注意力机制的 N_{we} 能够看到服装界限但是各部分纹理相对混乱，这也再次说明了注意力机制能够让网络输出更加具有意义的纹理。最后观察倒数两行，我们的模型生成图像的纹理更加细致逼真，而 FiNet 生成的更像是模糊的色块。

从纵向看，我们的方法从视觉效果来看更加真实，但是又有别于真实图像；FiNet 的输出相对模糊、无细致的纹理；BicycleGAN 因其本身服务于图像重构任务，生成的结果更像是在逼近真实图像；一阶生成网络生成的图像的质量看起来与二阶网络比较接近，但它的生成过程不受形状的约束，我们在实际使用时无法指定想要的服装形状；没有使用对抗生成模式训练的网络性能大幅下降，又因为对抗损失在总损失函数中占比很小，这就说明采用对抗生成网络的架构对网络的优化起到了很大的作用；没有使用注意力机制的网络生成的纹理有些违背认知，因而显得不真实；没有谱归一化的网络输出的图像存在对抗生成网络在训练初期或是训练失败时的“网格”效应，具体原因见 4.3.3 节。

图 4-1 相同输入不同生成方法比较

4.4.2 多样性

如图 4-2 所示，每一行分别对应一个待补全的人体语义分割图的不同生成结果，可以看到每个输出之间都存在或多或少的不同，这告诉我们通过控制隐空间变量 z 就能在一定程度上实现特定的输出效果，在图 4-2 中表现为服装中缝的拓宽。

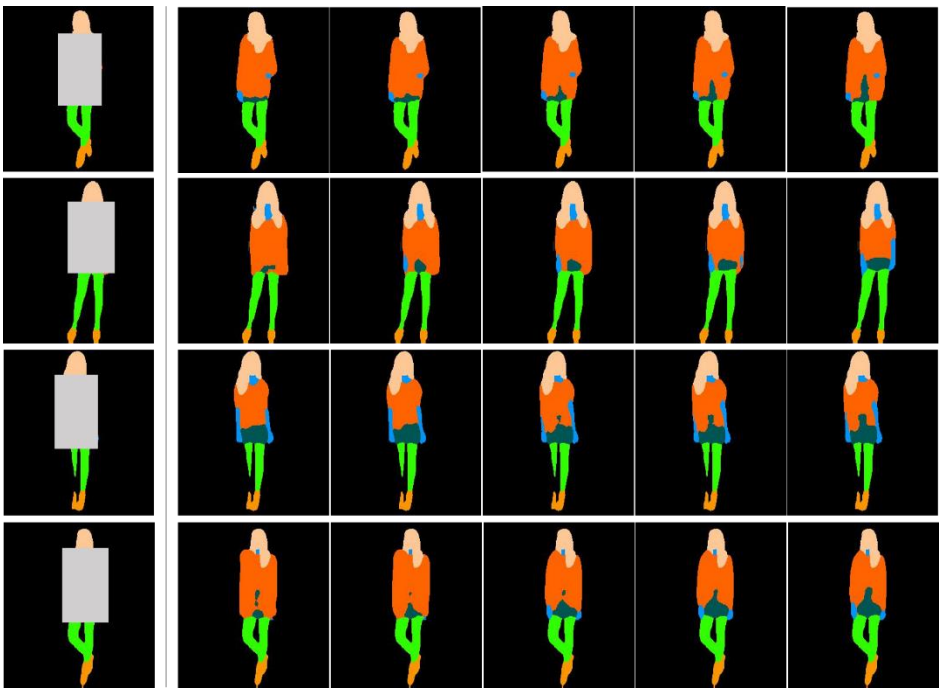

图 4-2 相同输入对应的不同输出（第一阶段）

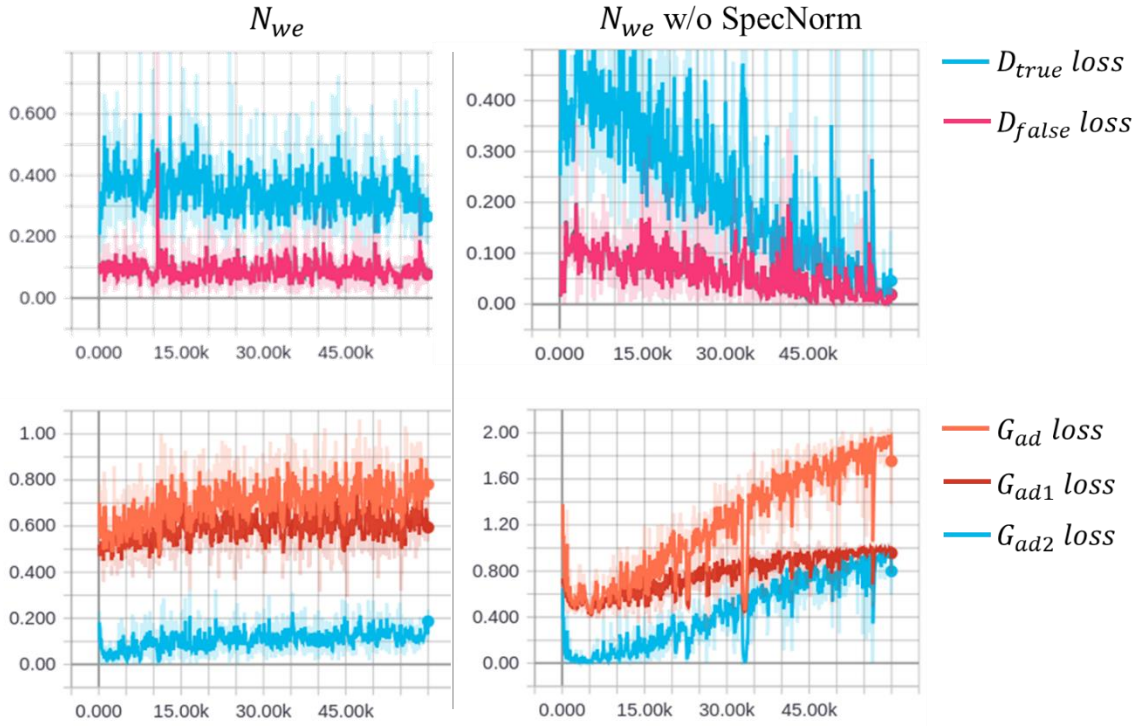

图 4-3 有/无谱归一化时损失函数随训练步数变化曲线

4.4.3 稳定性

如 4.3.1 节中所述，当我们的模型移除所有的谱归一化层后，生成图像会出现“网格”效应，该效应往往出现在对抗生成网络的训练初期，该现象说明网络可能训练失败，更说明了谱归一化层能够使得对抗生成网络的训练更加稳定。如图 4-3 所示，左/右侧分别是有/无谱归一化层的网络在训练时损失函数随训练步数的变化曲线，上/下分别是判别器/生成器的对抗损失，可以明显看到右侧的曲线一直在下降（判别器损失）或上升（生成器损失），而左侧的曲线只是在上下轻微波动，总体保持稳定。由对抗生成网络的优化原理可知，良好的双方博弈应该是双方实力的同时增强，各自的损失应该维持在一个稳定的值，因此谱归一化的确起到了维持网络训练的作用。

4.5 本章小结

本章首先介绍了一些模型实验的基本设置，如用到的数据集、模型参数和训练参数，接着列举并解释了我们用来对比模型性能的一些方法，包括和以往的研究工作、移除某些模块得到的退化模型。在定义完模型评价准则后，我们就展开了模型的定量、定性分析，不论是从客观的评价指标还是从实际的视觉效果看，我们的模型都存在一定的优势。

第五章 结束语

在总体思路上，本文将虚拟试装任务转化为时尚图像补全任务，使得模型训练从监督学习转化为自监督学习，大大减少对标注数据的依赖。为了实现虚拟试装的三大要求：真实性、多样性和视觉兼容性，我们采用了一个二阶时尚图像补全网络。该网络分为两部分：形状生成网络和纹理合成网络，两者均含有一个生成器和两个编码器，交互作用的编码器让网络学会利用图像的上下文信息（可见服装类别的图像），编码器将已知信息映射到隐空间，在隐空间分布中采样得到生成器的生成条件，生成器通过编码-解码的方式重构真实的原始图像，其中的采样操作又同时增加了生成结果的多样性。在纹理合成网络中另外加入一个判别器，利用对抗生成网络的思想实现更真实的图像合成。在网络结构设计中，我们在借鉴以往工作的基础上，大大缩小网络规模，并引入注意力机制、谱归一化等操作，在 DeepFashion 上做的相关测试显示，我们的模型相比以往方法性能有所提升，且网络使用的搭建技巧的确有其合理性。

由于我对生成模型所涉及的理论知识储备尚有欠缺，本文对于虚拟试装网络的研究还存在较大的探究空间。首先，目前只实现了第一阶段，即形状生成的多样性，且多样性不够显著，相关研究指出[36]，变分自编码所依据的概率模型、基于隐空间分布的假设决定了其被用作生成模型时，往往趋于输出固定的最优解，因此难以实现多模型学习。如何通过理论推导改进模型架构设置，进而实现一对多映射会是之后的研究主题。其次，目前仍然缺乏客观的兼容性的度量，如何建立类似于 FID score、Inception score 这样能够间接反映网络抽象属性的指标，也将会是下一阶段要探究的方向。因此，还有许多遗留问题等待我们去思考、解决。

致　谢

随着毕业的迫近，越发对成电校园、尤其是这个记录我过去两年成长足迹的教研室充满不舍，趁此撰写毕业论文之际，感谢那些在学术上、生活中帮助我、鼓励我的人们。

首先，谨向我的毕设导师胡洋老师致以诚挚的感谢，在过去近两年的时间里，胡老师不仅为我提供了良好的科研环境和实验条件，更重要的是，她严谨治学的态度深深地感染、激励了我，通过每周一次的组会，我从懵懂的无法理解学术词汇的小白成长到能够给大家讲解最新论文，通过每周例行的进度汇报，我学会了总结整理自己的工作，胡老师常常在与我讨论时帮助我找到更多的改进思路，可以说本文所得到的结果离不开她的悉心指导。

其次，我想感谢教研室的同学们，他们的存在不仅让我的科研之路充满欢乐，同时，同他们、特别是和我的师傅俞聪的交流，带给我许多实验的灵感，也让我能够避开前人走过的弯路，更加稳健的前进。即便未来就要离开这群可爱的伙伴，这段美好的、共同奋斗的记忆永不会消逝。

最后，我要特别感谢我的父母，感谢他们对我在成长道路上所做的许多关键性决定的理解与支持。我的父亲教会我永远保持对新事物的好奇，我的母亲教会我永远保持对生活的热爱。

附录

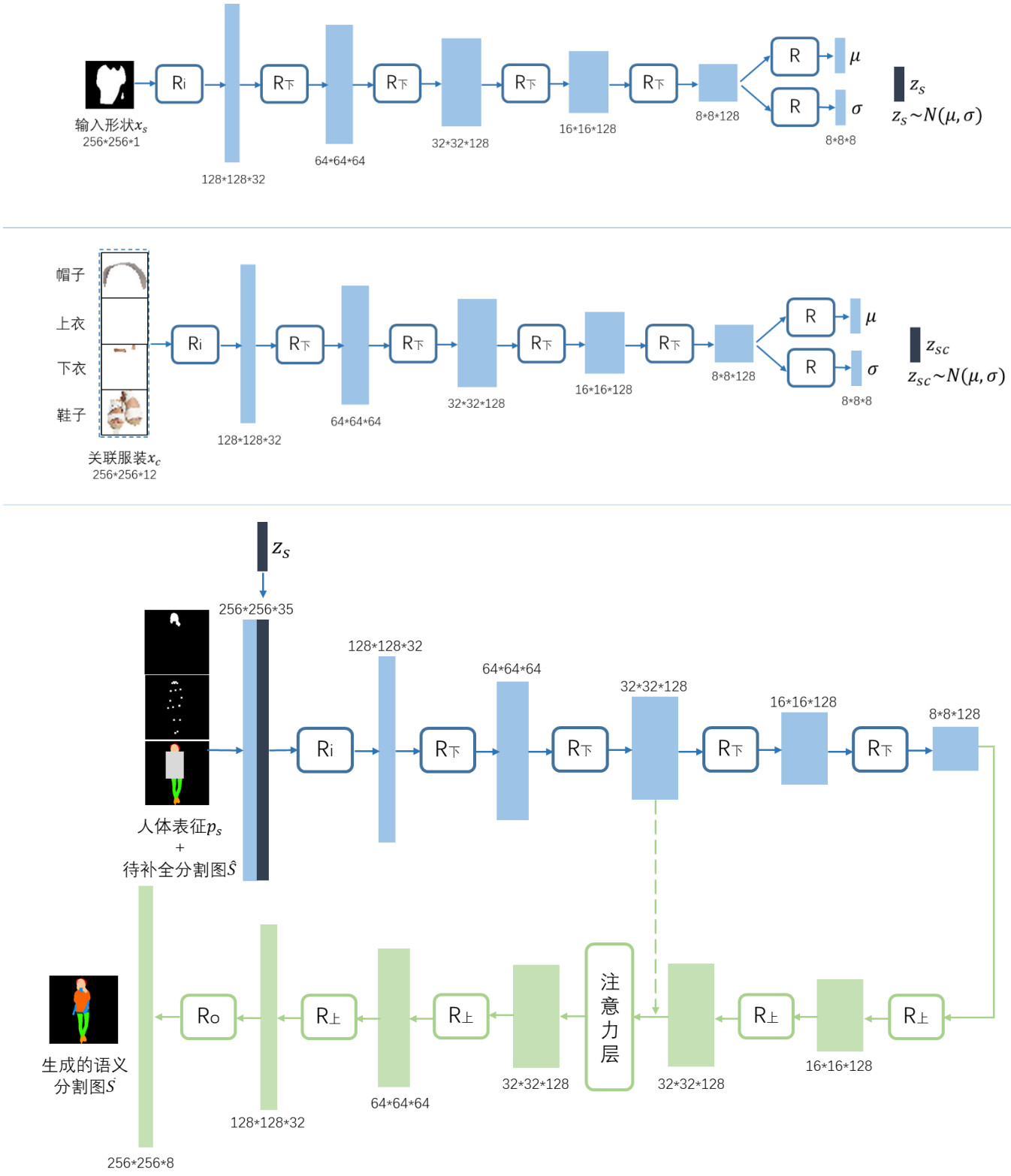

图 1 形状生成网络整体架构

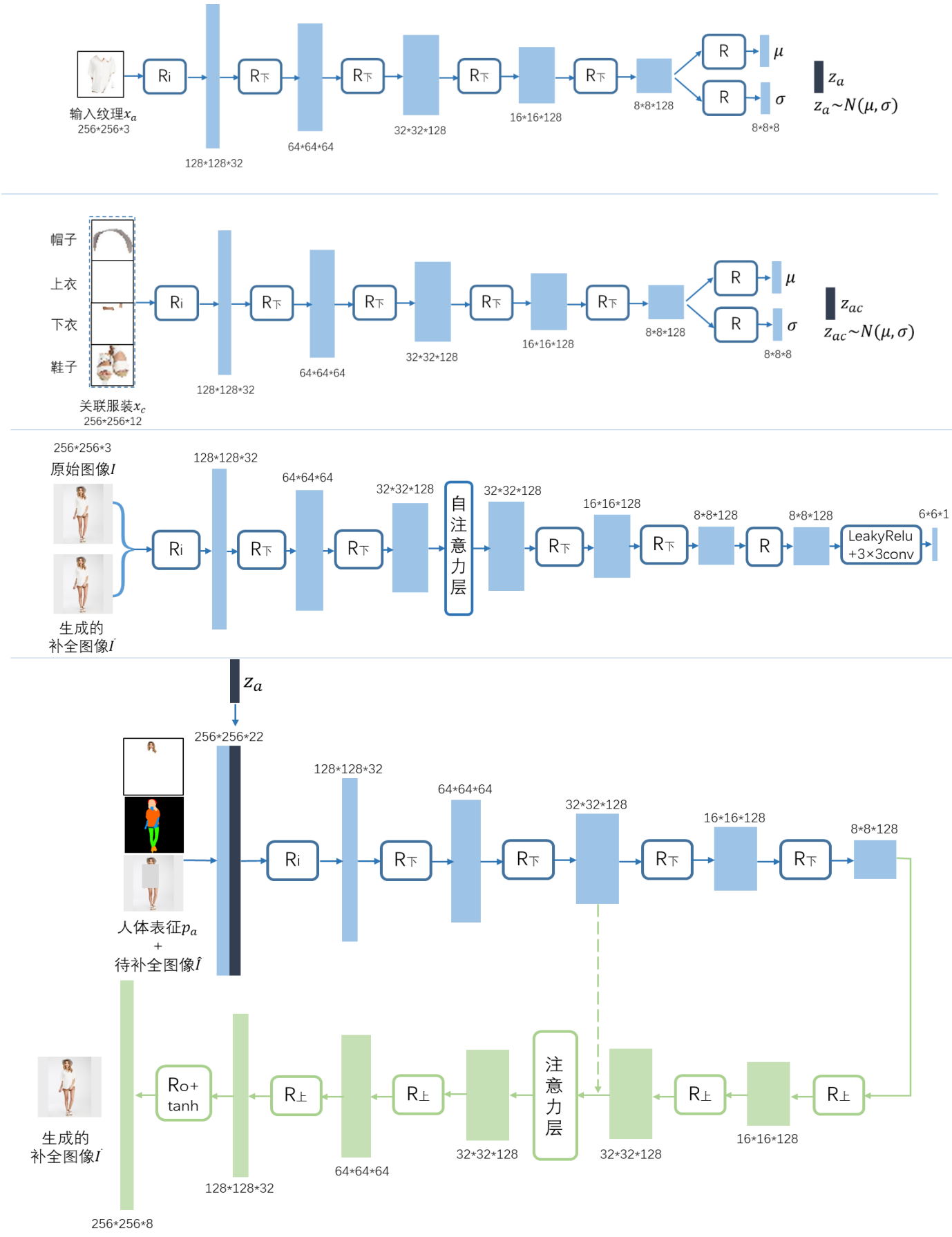

图 2 纹理合成网络整体架构

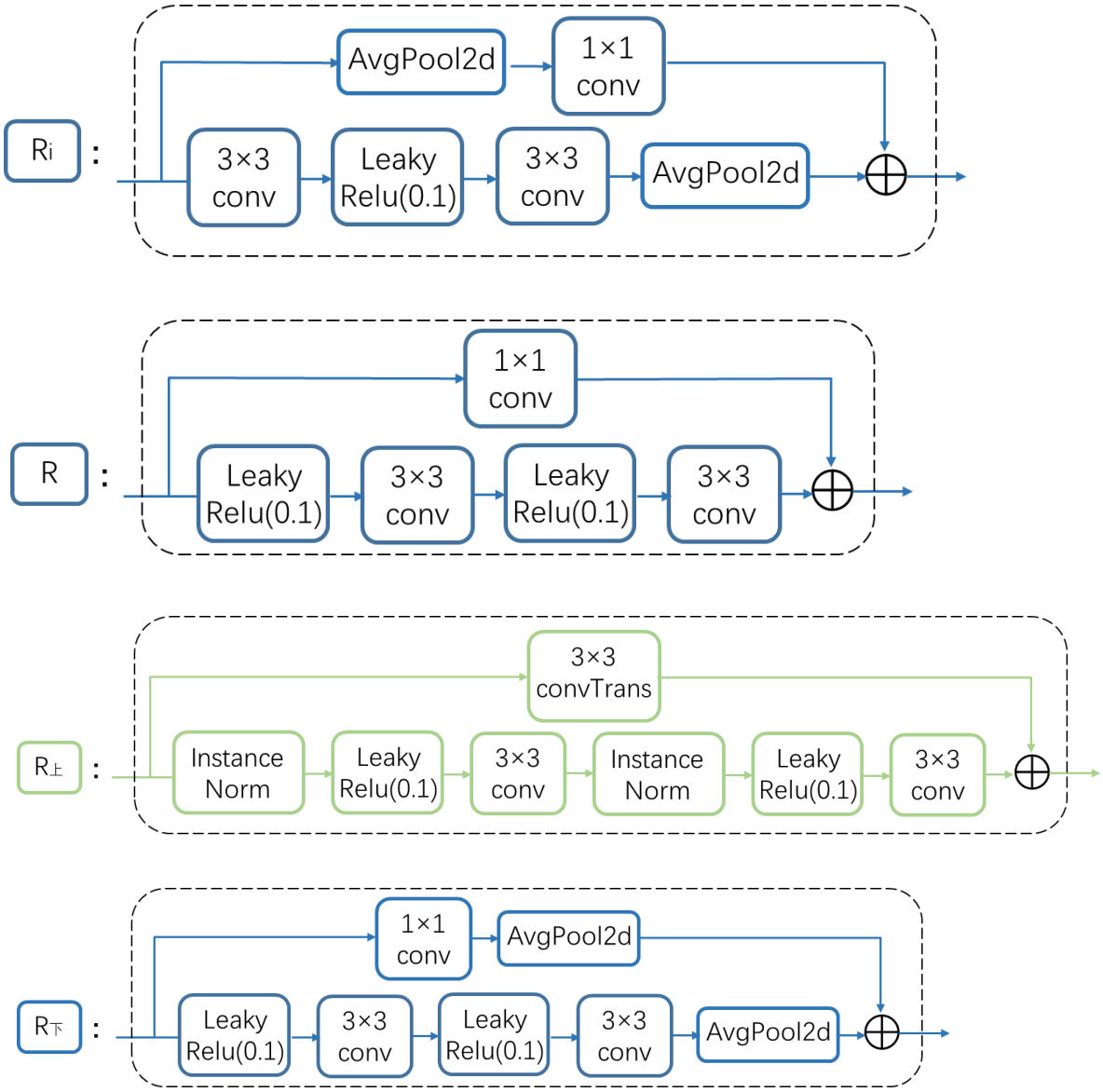

图 3 网络所涉及模块框图

参考文献

[1] Sekine M, Sugita K, Perbet F, et al. Virtual fitting by single-shot body shape estimation[C].Int. Conf. on 3D Body Scanning Technologies. Citeseer, 2014: 406-413.

[2] Chen W, Wang H, Li Y, et al. Synthesizing training images for boosting human 3d pose estimation[C].2016 Fourth International Conference on 3D Vision (3DV). IEEE, 2016: 479-488.

[3] Anguelov D, Srinivasan P, Koller D, et al. SCAPE: shape completion and animation of people[C].ACM transactions on graphics (TOG). ACM, 2005, 24(3): 408-416.

[4] Pons-Moll G, Pujades S, Hu S, et al. ClothCap: Seamless 4D clothing capture and retargeting[J]. ACM Transactions on Graphics (TOG), 2017, 36(4): 73.

[5] Jetchev N, Bergmann U. The conditional analogy gan: Swapping fashion articles on people images[C].Proceedings of the IEEE International Conference on Computer Vision. 2017: 2287-2292.

[6] Han X, Wu Z, Wu Z, et al. Viton: An image-based virtual try-on network[C].Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7543-7552.

[7] Wang B, Zheng H, Liang X, et al. Toward characteristic-preserving image-based virtual try-on network[C].Proceedings of the European Conference on Computer Vision (ECCV). 2018: 589-604.

[8] Chou C T, Lee C H, Zhang K, et al. PIVTONS: Pose Invariant Virtual Try-on Shoe with Conditional Image Completion[J].

[9] Han X, Wu Z, Huang W, et al. Compatible and Diverse Fashion Image Inpainting[J]. arXiv preprint arXiv:1902.01096, 2019.

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

[11] Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.

[12] Odena A, Olah C, Shlens J. Conditional image synthesis with auxiliary classifier gans[C].Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017: 2642-2651.

[13] Shen W, Liu R. Learning residual images for face attribute manipulation[C].Proceedings of the

IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4030-4038.

[14] Reed S, Akata Z, Yan X, et al. Generative adversarial text to image synthesis[J]. arXiv preprint arXiv:1605.05396, 2016.

[15] Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks[C].Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1125-1134.

[16] Bertalmio M, Sapiro G, Caselles V, et al. Image inpainting[C].Proceedings of the 27th annual conference on Computer graphics and interactive techniques. ACM Press/Addison-Wesley Publishing Co., 2000: 417-424.

[17] Bertalmio M, Vese L, Sapiro G, et al. Simultaneous structure and texture image inpainting[J]. IEEE transactions on image processing, 2003, 12(8): 882-889.

[18] Hays J, Efros A A. Scene completion using millions of photographs[J]. ACM Transactions on Graphics (TOG), 2007, 26(3): 4.

[19] Pathak D, Krahenbuhl P, Donahue J, et al. Context encoders: Feature learning by inpainting[C].Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2536-2544.

[20] Yang C, Lu X, Lin Z, et al. High-resolution image inpainting using multi-scale neural patch synthesis[C].Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6721-6729.

[21] Yan Z, Li X, Li M, et al. Shift-net: Image inpainting via deep feature rearrangement[C].Proceedings of the European Conference on Computer Vision (ECCV). 2018: 1-17.

[22] Song Y, Yang C, Lin Z, et al. Contextual-based image inpainting: Infer, match, and translate[C].Proceedings of the European Conference on Computer Vision (ECCV). 2018: 3-19.

[23] Yu J, Lin Z, Yang J, et al. Generative image inpainting with contextual attention[C].Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5505-5514.

[24] Zhang H, Goodfellow I, Metaxas D, et al. Self-attention generative adversarial networks[J]. arXiv preprint arXiv:1805.08318, 2018.

[25] Gong K, Liang X, Li Y, et al. Instance-level human parsing via part grouping network[C].Proceedings of the European Conference on Computer Vision (ECCV). 2018: 770-785.

[26] Chen Y, Wang Z, Peng Y, et al. Cascaded pyramid network for multi-person pose

estimation[C].Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7103-7112.

[27] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C].Advances in neural information processing systems. 2013: 3111-3119.

[28] Vasileva M I, Plummer B A, Dusad K, et al. Learning type-aware embeddings for fashion compatibility[C].Proceedings of the European Conference on ComputeVision (ECCV). 2018: 390-405.

[29] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C].European conference on computer vision. springer, Cham, 2014: 818-833

[30] Chen Q, Koltun V. Photographic image synthesis with cascaded refinement networks[C].Proceedings of the IEEE International Conference on Computer Vision. 2017: 1511-1520.

[31] Liu Z, Luo P, Qiu S, et al. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations[C].Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1096-1104.

[32] Ma L, Jia X, Sun Q, et al. Pose guided person image generation[C].Advances in Neural Information Processing Systems. 2017: 406-416.

[33] Siarohin A, Sangineto E, Lathuilière S, et al. Deformable gans for pose-based human image generation[C].Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3408-3416.

[34] Miyato T, Kataoka T, Koyama M, et al. Spectral normalization for generative adversarial networks[J]. arXiv preprint arXiv:1802.05957, 2018.

[35] Zhu J Y, Zhang R, Pathak D, et al. Toward multimodal image-to-image translation[C].Advances in Neural Information Processing Systems. 2017: 465-476.

[36] Zheng C, Cham T J, Cai J. Pluralistic Image Completion[J]. arXiv preprint arXiv:1903.04227, 2019.