摘　要

随着消费者在线购物需求的不断增加，电商平台持续改进和优化购物体验，虚拟试衣技术解决网络购物中的服装试穿难题。传统的虚拟试衣技术需要高昂的设备进行 3D 建模，消耗大量的计算资源和时间。而基于图像的虚拟试衣技术则可以直接利用用户提供的图像进行试衣模拟，无需过多的资源投入，更适用于线上购物，试穿结果可以方便、快速地在电商平台上呈现和分享，用户能够及时获得反馈和意见。但是目前深度学习技术仍然存在一些局限性：现有的图像虚拟试衣模型在生成的试穿效果图像时缺少服装细节、目标服装不能精确的对齐人体姿态、手臂遮挡服装时边界模糊不自然并且伴随着服装颜色失真。基于多级注意力机制的虚拟试衣模型提供的一种新的解决方案，开展了以下工作：

（1）提出了一种基于多级注意力机制的虚拟试衣模型，围绕 U-Net 网络中的跳跃连接提出结合多级注意力机制的改进。虚拟试衣技术对生成的服装掩码要求较高，传统的 U-Net 网络提取到的特征包含大量冗余信息，生成的掩码比较粗糙，从而导致手臂遮挡变形。通过在 U-Net 网络每个尺度的跳跃连接层上引入混合注意力机制，即通道与空间注意力串行结构。分别在通道和空间维度上学习重要特征的位置和强度。充分保留纹理细节，缓解边界模糊不自然情况，从而提升模型性能。

（2）提出了一种联合损失函数，引入马尔可夫判别器损失和感知损失，用于训练模型。利用有监督的学习策略，指导掩码、调整服装、试穿结果的生成，缓解服装变形不精准，目标服装无法对齐人体姿态。从而提升试穿效果的真实度，使服装与手臂交界处更加平滑。

（3）在 VITON 数据集上进行定量和定性的实验结果和分析，设置对模型的注意力模块和损失函数的消融实验，与现有的虚拟试衣模型做对比试验，展示各个模块对试穿结果的影响，验证各个模块的重要性。最后展示在现实生活中的实际试穿效果。

关键词：虚拟试衣；多级注意力机制；联合损失；图像生成

Abstract

With the increasing demand of consumers for online shopping, e-commerce platforms continue to improve and optimize the shopping experience. Virtual fitting technology solves the problem of clothing fitting in online shopping. Traditional virtual fitting technology requires expensive equipment for 3D modeling, which consumes a lot of computing resources and time. Image-based virtual fitting technology can directly use the images provided by users to simulate the fitting, which is more suitable for online shopping. The fitting results can be easily and quickly presented and shared on the e-commerce platform, and users can obtain feedback and opinions in time. However, there are still some limitations in the current deep learning technology : the existing image virtual try-on model lacks clothing details when generating the try-on effect image, the target clothing cannot accurately align the human body posture, the boundary is blurred and unnatural when the arm blocks the clothing and is accompanied by clothing color distortion. The virtual fitting model based on multi-level attention mechanism provides a new solution, and the following work has been carried out :

（1） A virtual fitting model based on multi-level attention mechanism is proposed. The virtual fitting technology has higher requirements for the generated clothing mask. The features extracted by the traditional U-Net network contain a large amount of redundant information, and the generated mask is rough, resulting in arm occlusion deformation. By introducing a hybrid attention mechanism on the skip connection layer at each scale of the U-Net network, that is, the channel and spatial attention serial structure. The position and intensity of important features are learned on the channel and spatial dimensions respectively. Retain texture details and alleviate the unnatural situation of boundary blurring, thereby improving model performance.

（2） A hybrid loss function is proposed, which introduces Markov discriminator loss and perceptual loss for training model. Using supervised learning strategies to guide masks, adjust clothing, and try on the generation of results to alleviate inaccurate clothing deformation, the target clothing cannot align the human body posture. So as to improve the authenticity of the try-on effect and make the junction of clothing and arm smoother.

（3）Quantitative and qualitative experimental results and analysis are carried out on the VITON dataset. The ablation experiments of the attention module and loss function of the model are set up. Compared with the existing virtual fitting model, the influence of each module on the fitting results is demonstrated, and the importance of each module is verified. Finally, the actual fitting effect in real life is shown.

KeyWord: Virtual Try-on；Attention Mechanism；Generative Adversarial Network； Image Generation

目 录

1 绪 论

1.1 研究背景与意义

互联网的快速发展为人们提供了新的生活、工作和学习空间。以前，购物需要携带银行卡和现金出门，而现在只需一部小小的手机，就可以足不出户地轻松购买所需商品。科学技术的不断进步加速了社会的全面振兴，经济迅猛发展，人民生活水平得到了提高，人均可支配收入不断增加，这些都促使国民总体消费水平不断上升。根据国家统计局最新公布的数据，2022 年全国网上零售额中，穿着类商品同比增长 3.5%。如图 1.1 所示，2021 年中国网上零售额总额为 13.09 万亿元，还保持着 14.10%的增速，每年网上零售额不断攀升，服装行业也凭借网络购物高效、便捷的特点而蓬勃发展。网络购物为消费者提供了更多选择和便利，这也为服装行业带来了更多机遇和挑战。随着消费者对于在线购物需求的不断增加，电商平台需要持续改进和优化购物体验，并提高消费者对品牌的认知和信任度，从而在市场竞争中占据优势。

此外，随着审美意识的觉醒，消费者对美的需求越来越高，更加注重服装的品质和个性化需求的满足。为了满足消费者的需求，服装行业也在不断创新和探索，引入新兴技术如智能穿戴技术、可穿戴设备，虚拟试衣间等，将其应用于服装设计和销售中，提供更加智能化、便捷和舒适的试穿新体验。这些技术的应用，为消费者提供了更多个性化和多样化的选择，推动了服装行业的迅速发展。

自 2019 年 12 月武汉通报第一例不明原因的肺炎以来，新冠病毒开始在全球范围内肆虐，由于患者在潜伏期内也会具有传染性，这给人们的生命安全带来了极大的威胁。为了应对疫情，全国各地开始了长达三年的抗击疫情和居家隔离。传统的线下购物方式可以让消费者自行挑选不同的服装进行试穿，选择合适的尺码和中意的颜色。然而，在实体店铺挑选到心仪的衣服，消费者常常需要货比三家，耗费大量时间和精力。此外，由于新冠病毒具有传染性，前往实体店铺进行购物可能会增加被感染的风险。

相比传统的线下购物方式，网上购物成为了人们购物的主流方式。网上购物为商家开拓市场渠道、斩获订单、降低成本，不再需要支付昂贵的店面租金，从而节省人力成本，商家可以通过网上购物以更低的价格提供服装，消费者也可以以更低的价格获得更多优质的服装。此外，网上购物可以让消费者在不受

时间和地点限制的情况下轻松货比多家，更加方便快捷。由于网上购物是一种零接触式的购物方式，消费者不必前往实体店铺，因此可以减少被感染的风险，消费者也更加青睐于网上购物，并愿意通过这种方式购买服装。

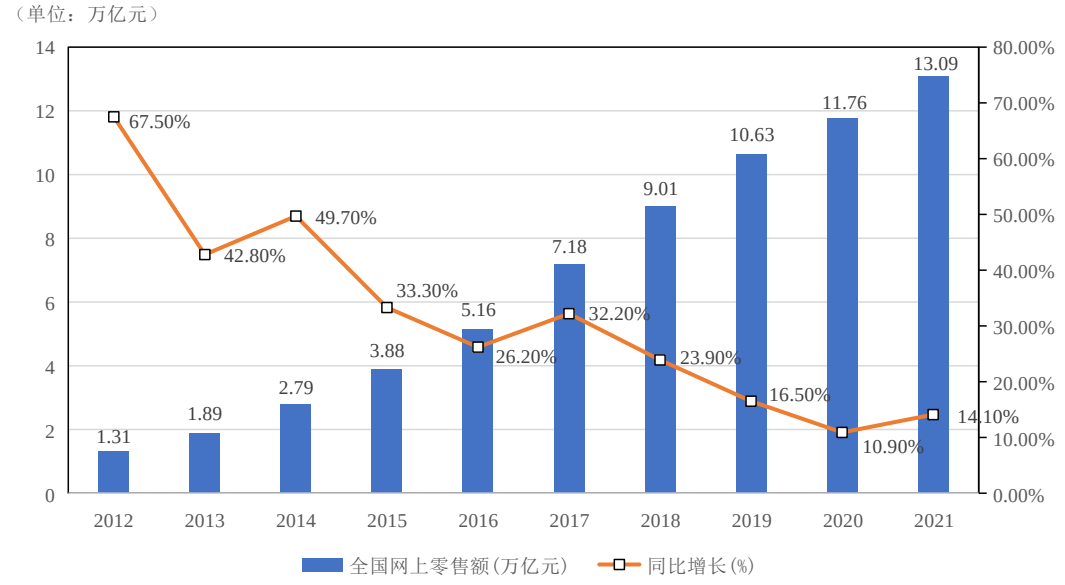

图 1.1 2012-2021 年中国网上零售额及增长率

网络电商平台的服装信息的传播和获取存在一定的局限性，虽然网络平台可以提供各种形式的商品信息，但消费者只能通过有限的信息形式来了解商品的实际情况，例如文字和图片等。这些信息可能无法全面地反映商品的实际情况，导致消费者难以做出准确的购买决策。同时，网络平台也难以提供真实场景下的试用和感受。由于衣服的款式、尺码、面料等因素，购买的服装可能不适合自己的身材或穿搭风格，从而导致消费者在收到商品后发现与自己的期望不符，进而产生退货等问题。因此，如何利用现有的技术手段来提高消费者在网络购物过程中的购买体验和满意度，解决网络购物的服装试穿难题成为了一个研究热点。一些电商平台已经推出了虚拟试衣间[1]等技术，允许消费者上传自己的照片，在虚拟场景中试穿衣服，以获得更为真实的试穿效果。

虚拟试衣技术的提出与应用带来多方面的效益，例如为商家节省经营成本，顾客也不用花费大量的时间和精力在不同店铺之间挑选喜欢的衣服，进而提高店铺的成交量。通过虚拟试穿技术，顾客可以更准确地选择适合自己的服装，了解衣服的上身效果和搭配情况，从而降低网购过程中退换货的情况。同时，还可以提高沟通效率，激发顾客的购买欲望，产生新的购买力，减少人员接触，降低被感染的风险。

虚拟试衣技术根据输入输出类型和算法模型的不同分为基于图像[2]和基于三维建模[3]两类方法。三维虚拟试衣模型是一种利用数字仿真技术和计算机图

形学原理，通过扫描消费者的三维身体信息，建立出完整的服装模型，让消费者可以根据自己的喜好和风格来搭配服装，从而获得最佳的试穿效果。这种方法的准确性和可靠性极高，因为它能够精确地模拟衣物与人体之间的几何变形，从而获得更加完美的试穿体验。但是，该方法需要使用昂贵的三维扫描设备和大量精确的数据进行建模，所需成本较高，实现起来更为复杂，对于当前的网购服装等场景，其适用性受到了限制。相对而言，基于二维图像的虚拟试衣技术虽然精度不及三维技术，展示效果略逊一筹，但其成本更低，设备要求不是特别高，建模工作也较为简单。此外，该技术还能够为消费者提供多样化的试穿选择，有效提高试穿效率，满足消费者的需求。因此，基于二维图像的虚拟试衣技术在线上购物场景中有着广泛的应用前景和市场潜力。

基于二维图像的虚拟试衣技术，可以看成是图像生成问题。该技术将消费者的人体图像和选定的目标衣服图像作为输入，首先对输入的人体图像进行预处理获取人体表征，然后生成与人体姿势相匹配的扭曲服装，两者融合得到消费者穿着目标服装的试穿图像。

本文旨在以二维图像为基础，建立一种虚拟试衣模型，通过输入消费者的图像和选择的试穿服装进行处理，尽可能地“还原”真实试穿场景，实现在消费者身上试穿指定衣物的效果，并减少因目标服装与人体姿势不匹配而引起的视觉误差，避免丢失目标服装图像特征和人体手臂遮挡边缘模糊的问题。在保证基本试穿效果的同时，生成高质量的虚拟试穿图像。相比于三维方法，本文所提出的虚拟试衣技术在试穿真实度方面存在一定的差距，但能够在一定程度上帮助消费者选择合适的服装，为消费者提供试穿参考，更加方便、快捷、实用应用到网络购物当中，提升购物体验。

1.2 国内外研究现状

虚拟试衣技术根据输入输出类型和算法模型的不同可分为两类：一类是基于二维图像的虚拟试衣技术，主要通过利用深度学习的方法生成二维的试穿图像；还有一类是基于三维建模的虚拟试衣方法，该方法主要通过利用三维扫描设备对消费者人体体型进行三维人体建模，再对不同的衣服进行服装建模，最终得到三维建模试穿效果。

1.2.1 基于三维建模的虚拟试衣方法

基于三维建模的虚拟试穿过程中，服装和人体模型之间的匹配仍然是研究人员面临的一个问题。Li[4]等人提出了一种基于特征点匹配的三维模型试穿方法。首先利用双二次贝塞尔曲面来估计两个网格模型上每个顶点的几何属性，

计算高斯曲率与平均曲率。然后从两个模型中提取特征点，并构造一个新的匹配函数根据曲率和约束关系进行匹配。Meng[5]等人提出了一种用于虚拟试穿、试穿评估和款式编辑的新型计算机辅助设计 (CAD) 解决方案，以加快服装设计过程。采用交叉参数化技术在模型表面绘制服装图案，提出了一种称为 Hybrid 弹出式的新变形方法来近似模拟虚拟试穿形状。况鹰[6]构建了一个基于 Kinect 动态捕捉的虚拟试衣技术，人机交互体验和实时模拟效果还原了虚拟试衣的真实感。Drape[7]是一种由 Guan 等人开发的技术，任何姿态和体型的人体上都可以自动生成逼真的服装动画。Sekine 等人[8]提出了一种新的方法，旨在通过利用单镜头深度图像来估计消费者的三维体形，并实现无缝调整试穿服装。

Ma 等人[9]提出的三维人体模型被广泛用于人体姿势和运动的分析。然而，现有模型是从穿着最少的人身上学习的，无法捕捉到普通图像和视频中人物穿搭的复杂性。为了解决这个问题，提出训练一个有条件的 Mesh-VAE-GAN[10]，从 SMPL 身体模型中学习服装变形，模型被命名为 CAPE，表示全局形状和精细的局部结构，有效地将 SMPL 身体模型扩展到服装上。Zhao 等人[11]提出了一种多姿态图像虚拟试衣方法，兼顾了二维与三维方法的优点。具体而言，该方法利用所提出的MPM模块，从图像中估计目标人体的分割图和深度图，同时获取2D与3D的辅助信息。与一般的方法不同，MPM首先利用一种新的自适应仿射变换，将店内服装变换到合适的尺寸和位置，然后进行非刚性 TPS 变形。为了添加高频深度细节，作者还提出了深度细化模块（DRM），进一步利用扭曲的服装和保留的人物部分中的亮度变化，来细化初始深度图，从而指导最终的合成过程。Peotopsaltou 等人[12]较早提出虚拟试衣间这一概念，后面 Cordier [13]等人构建的线上服装商店个性化地展示虚拟试衣效果，使用获取到的标准尺寸参数创建虚拟人物，消费者可以进行实时交互，以及动态调整不同的姿态，选择合适的服装。

此外，三维虚拟试衣软件在国内外发展趋于成熟，例如优衣库的虚拟试衣间；新加坡的 Browzwear：一款以色彩和渲染著称的三维虚拟试衣软件，可用于设计和生产服装；韩国的 3D-CLO：一款功能强大的三维虚拟试衣软件，可以进行服装设计、模拟和展示；天猫 AR Door 虚拟试衣间等。随着技术的不断发展，三维虚拟试衣技术在未来将继续得到改进和完善，使得虚拟试衣体验更加真实、方便和实用，为消费者提供更好的购物体验。

虽然三维虚拟试衣可以等多方面展示效果，但是需要占用大量的计算资源来进行模拟和渲染，建模工作需要娴熟技术和时间成本。在当前的网购服装等场景中的应用存在一定的局限性。

1.2.2 基于深度学习的虚拟试衣方法

基于深度学习的图像虚拟试衣技术根据其模型结构的不同，大致可以分为两类：一类是采用编码器-解码器结构，另一类则是采用生成式对抗网络[14]（Generative Adversarial Networks, GANs）。VITON[15]模型将试穿过程分解成多个步骤，首先，通过解码网络解析出人体和服装的特征，模型逐渐细化得到最终的试穿效果，同时，该模型的优化过程主要集中在改善服装的形变和融合效果。第二种是利用生成式对抗网络，生成以假乱真的试穿图。然而，生成过程缺乏精细的人体表征，因此主要的优化方向是解决人体姿势和服装扭曲变化的问题，生成对抗网络在处理具有较大空间变形的问题时，往往存在输出模糊、保留目标衣服细节少等问题。

人体姿态估计技和服装解析技术的发展，为图像虚拟试衣技术的实现提供了强有力的支持，对虚拟试衣技术的实现和应用具有重要的意义。人体姿态估计是计算机视觉中一个基础而重要的问题，为许多高层次的语义任务和下游应用场景提供了基础支持。指的是估计人体关键点（例如头部、左手、右脚等）在空间中的位置，以推断出人体的姿态。基于卷积神经网络的姿态估计根据网络结构可分为基于回归和基于检测两种方法。

基于回归的方法将人体姿态估计问题看作一个从输入图像到人体连接坐标和姿态模型参数的直接映射问题。Toshev 等人[16]提出了一个直接从输入图像中预测关键点坐标的模型，而不用通过中间步骤进行特征提取和分类—Deep Pose，Deep Pose 采用级联多层经 DNN 来预测人体各个关键点的相对坐标。每一个阶段都拿上一阶段的输出坐标作为输入，并进一步预测更为准确的坐标位置。最终，将预测得到的归一化的相对坐标转换为绝对坐标。Nibali 等人[17]提出了一种直接从图像中学习坐标的新思路。基于热图的方法在预测每一个位置得分情况，来表征该位置属于关键点的置信度。然而，从单一的图像中准确预测人体关键点坐标仍然是一个具有挑战性的问题。Hahn 等人[18]提出了一种新的回归方法，基于骨骼结构感知，并采用组合损失函数来限制姿态之间的长距离关联，从而大大提升了姿态估计的准确性。

基于检测的方法是将身体部位作为检测目标。Papandreou 等人[19]首先利用 Faster RCNN 模型[20]先检测出人，然后再做单人关键点检测。Open Pose 是一种基于自底向上的人体姿态估计算法，由 Cao 等人[21]提出。这种方法使用卷积神经网络从图像中提取特征，并输出 18 个预测置信度图来检测出图像中的所有关键点，任意一个图均表示人体骨架中的某一个坐标点，即人体关键点，如图 1.2 所示。还有一个阶段是预测一个 38 个亲和力场的集合，如图 1.3 所示。使用一

种启发式算法来将关键点进行匹配和连接，从而形成肢体。该算法会首先根据置信度图检测出所有的关键点，然后在所有关键点之间构建一个图，其中每个关键点表示一个节点，边表示两个关键点之间的连接。接下来，Open Pose 会利用亲和力场来度量关键点之间的连接程度，然后使用一种迭代算法来确定每个肢体的最优连接。最终，通过将连接的关键点进行分组，Open Pose 能够将图像中的多个人体姿态分离出来。当前应用最广泛的人体姿态估计算法之一，已经被广泛应用于虚拟试衣等领域。在本文所使用的数据集中，同样利用 Open Pose 进行预处理操作。

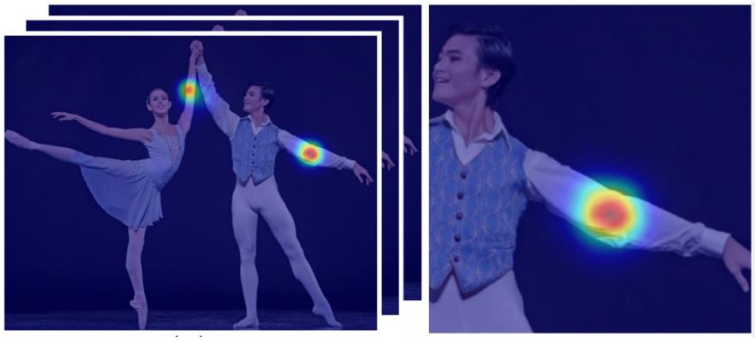

图 1.2 身体部位置信度图

图 1.3 亲和力场

服装解析技术是一种复杂的语义分割任务，旨在将带有服装类别的人体分割成多个像素级的语义部分，并给出相应的标签，如头发、头部、脖子等多个类别。由于服装更新迭代快，种类、款式、穿衣风格复杂多变，因此这个任务具有很大的挑战性。

Liu等人[22]提出了一种新的模型，旨在将姿态估计和像素级分类相结合，实现自动解析服装图像的弱监督方法，并且还能够通过颜色类别标签来替换像素级标签。Liang 等人[23]将人体图像分解为语义时尚/身体区域，转化为活动模板回归(Active Template Regression, ATR)问题，深度卷积神经网络用于在输入人体图像和人体解析结构输出之间建立端到端的关系。Liang等人[24]在研究中提出了结合上下文的卷积神经网络，主要任务是解决人体解析问题。目的是从低层加入情境化的信息，这个框架是将交叉层内容、全局图像层内容和局部超像素内

容整合成一个统一的网络。Gong 等人[25]提出的自监督学习方法“Look Into Person”(LIP)能够将人体姿态结构融入解析结果中，而且无需额外监督人体关节信息。在本文中，同样使用该方法对数据集进行处理，成功对人体图像进行解析并获取相应的标签。

目前最常用的虚拟试衣方法只需要提供包含目标服装信息的图像[26, 27]和人物特征图[28]。2018 年，Han 等人提出的 VITON[15]方法和 Wang 等人[29]基于 VITON 改进所提出的 CP-VTON 方法成为了后续研究的基础。Clo-VTON 研究[30]提出整合 2D 图像的高质量图像生成和 3D 模型的目标人体姿态变形的方法。ACGPN[31]使用语义分割代替原有的与服装无关的人体表示，引入图层的概念，将手和衣服分为两个独立的图层进行处理，自适应地判断所有不发生改变的区域，并减少它们之间的相互影响。由于缺少人体解析结构，导致部分裤子、手臂等部分出现像素缺失的情况。Ge 等人[32]首次将知识蒸馏应用到虚拟试衣网络，与以往的将人体分割作为模型输入不同，通过避免微小误差的分割结果而导致生成不真实的图像的问题，降低对人体解析的依赖程度。Kedan 等人[33]通过捕捉重要细节（例如纽扣、纹理、逼真的裙摆和服装之间的交互）生成高质量试穿图像。Choi 等人[34]利用分割图来指导模型生成，穿着目标服装的效果比较粗糙， 提出 ALIAS 生成器来解决未对齐的区域同时具有较高分辨率。Dong 等人[35]提出一个多姿态引导的视觉试穿网络（MG-VTON），根据不同的人体姿势设计的虚拟换衣系统。Hsieh 等人[36]提出一种新颖的 Fashion On 网络，可以合成以任意姿势穿着不同衣服的消费者图像。Dong 提出了流指导变换的生成对抗网络（FW-GAN）[37]，旨在操纵姿势和衣服的同时生成连贯自然的视频。Wu 等人[38]提出 M2E-Try On 网络，将模特衣服的姿势和纹理转移到所需的外观。LGVTON[39]提出了一个掩码生成器模块，试图预测人体模型的真实分割掩码，这反过来又指导图像生成器模块解决变形问题。Liu 等人[40]旨在设计一个统一的框架，解决人体图像合成的问题，包括人类运动模仿、外观迁移和新的图像合成。这意味着该模型一旦被训练，就可以用来处理上述所有的任务。

1.3 目前主要存在的问题

随着技术的发展和数据量的增加，深度学习在虚拟试衣领域中的应用得到了迅速的发展，取得了卓越的成就。但是目前服装的种类和款式不断增多，更新速度也越来越快。虚拟试衣模型需要能够快速适应新的服装款式和变化，同时保证试穿效果的准确性。现有的虚拟试衣技术与用户的期望仍存在一定差距，主要困难如下：

（1）试穿效果不佳。目前的虚拟试衣技术虽然在理想状态下能够很好地提取人体和目标服装的特征，但在现实生活中却面临着许多挑战和限制。模型的稳定性在某些复杂场景下会受到干扰，例如手臂遮挡、人物肤色、背景光照、服装材质等因素都会影响试穿效果的准确性。因此，如何提高虚拟试衣模型的鲁棒性和多样性，以实现更加逼真的试穿效果，是当前虚拟试衣技术需要解决的一个重要难题。

（2）训练样本不足。虚拟试衣技术需要海量的数据支撑，以便进行精准的建模和训练，但是这些数据的获取和处理成本非常高，需要耗费大量的时间和资源。在实际应用中，很难涵盖所有的人体形态和服装款式，虚拟试衣模型可能无法准确地捕捉到复杂的形态和特征，从而限制了其应用范围。因此，如何利用有限的训练样本来提高虚拟试衣模型的泛化性能，也是当下亟待解决的问题。

1.4 本文主要工作

本文的研究内容是基于图像的虚拟试衣模型。相较基于三维建模的虚拟试衣技术需要高昂的设备和大量的精确的数据进行实时的物理渲染，基于图像的虚拟试衣技术对设备的要求和计算机资源相对较小，更适用于线上购物，试穿结果可以方便、快速地在电商平台上呈现和分享，用户能够及时获得反馈和意见。由于目前深度学习技术仍然存在一些局限性：现有的图像虚拟试衣模型在生成的试穿效果图像时缺少服装细节、目标服装不能很好的对齐人体姿态、边界模糊不够丝滑、手臂遮挡变形，服装颜色失真等。针对以上列出的问题，本文工作内容如下：

（1）提出了一种基于多级注意力机制的虚拟试衣模型。虚拟试衣技术对生成的服装掩码要求较高，传统的 U-Net 网络提取到的特征包含大量冗余信息，生成的掩码比较粗糙，从而导致手臂遮挡变形。通过在 U-Net 网络每个尺度的跳跃连接层上引入混合注意力机制，即通道与空间注意力串行结构。分别在通道和空间维度上学习重要特征的位置和强度。保留纹理细节，缓解边界模糊不自然情况，从而提升模型性能。

（2）提出了一种联合损失函数，用于训练模型。利用有监督的学习策略，指导掩码、调整服装、试穿结果的生成，缓解服装变形不精准，目标服装无法对齐人体姿态。从而提升试穿效果的真实度，使服装与手臂交界处更加平滑。

（3）在 VITON 数据集上进行定量和定性的实验结果和分析，设置对模型的注意力模块和损失函数的消融实验，与现有的虚拟试衣模型做对比试验，展

示各个模块对试穿结果的影响，验证各个模块的重要性。最后展示在现实生活中的实际试穿效果。

1.5 本文组织结构

本文分为五个章节进行阐述，详细内容如下：

第一章：绪论。主要介绍了虚拟试衣模型的研究背景与意义，国内外的研究现状和目前存在的一些问题。

第二章：相关理论基础。详细阐述了生成对抗网络的基本原理，以及注意力机制的分类和实现方式。接着讲述了两种虚拟试衣模型，VITON 和 CP-VTON。

第三章：基于多级注意力机制的虚拟试衣模型。首先介绍了基础的 U-Net 网络，引入本文提出的 MA-VTON 模型，其中包括服装变形模块、改进的试穿模块以及训练模型采用的损失函数。

第四章：实验结果与分析。介绍了实验数据集以及预处理操作，网络实现的具体细节。通过与现有的模型做定量和定性的对比实验，和模型的注意力模块和损失函数的消融实验，对实验结果进行分析，验证 MA-VTON 模型的有效性。

第五章：总结与展望。基于图像的虚拟试衣模型，分析本文主要工作以及解决的问题，并对未来工作的一个展望。

2 相关理论基础

本质上，基于图像的虚拟试衣技术是一个图像生成问题。在本章中，首先介绍生成对抗网络。接着，根据注意力机制的计算方式进行分类，对其内容进行了详细的阐述。最后，介绍了两种虚拟试衣模型，VITON 和 CP-VTON。

2.1 生成对抗网络

生成对抗网络[14]的核心思想是：利用两个神经网络模型互相对抗学习，实现高质量数据样本的生成。两个神经网络模型分别是生成器和判别器，生成器的目标是生成与真实数据样本相似的数据样本，欺骗判别器将虚假数据样本误判为真实数据样本。而判别器的目标则是识别生成器生成的虚假数据样本，并对其进行准确的判别，提高自身的判别能力。

两种方式的训练都是交替完成的,生成器在生成虚假数据后将其传送给判别器,而判别器在确认该信息的准确性后将其反馈给生成器。不断调整各自的参数，这种过程将持续迭代下去,直至生成器生成的数据足够真实,使得判断器无法辨别真实信息与还是伪造的数据，如图 2.1 所示。

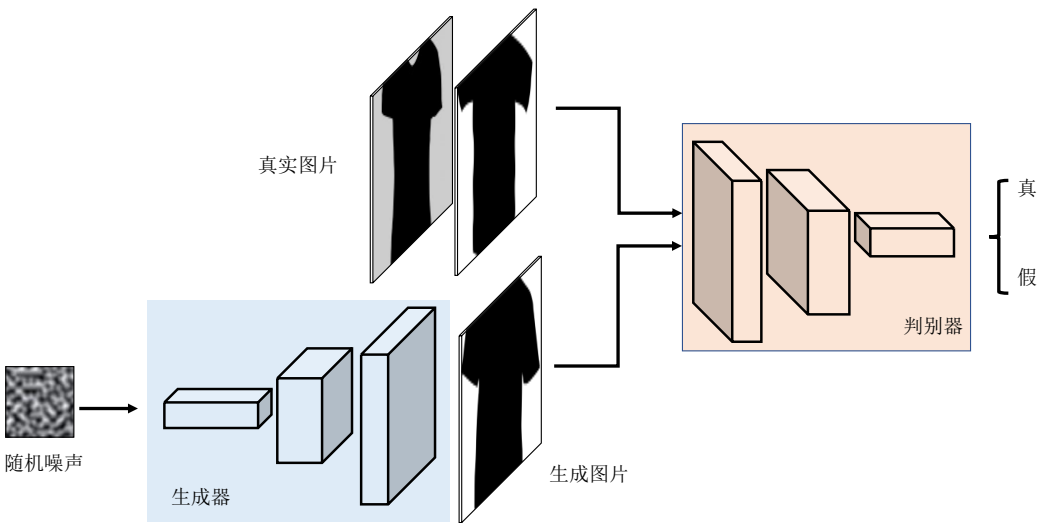

图 2.1 生成对抗网络结构

将上述过程公式化，可以得到公式(2-1)。

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

其中， z 为随机噪声， G 表示生成模型， D 代表判别模型； $x\sim p_{data}(x)$ 表

示 x 取自真实数据分布，$z \sim p_z(z)$ 表示 z 取自生成数据分布。

为了解决生成对抗网络中的最值优化问题，可以用公式来表达。具体来说，通过使用随机梯度下降等优化模型对判别器和生成器的参数进行迭代更新，从而逐步降低损失函数，提升模型的性能。最终，当生成器能够生成与真实数据分布相似的数据样本，且判别器无法区分样本真伪时，模型训练可终止。

(1) 优化判别器 D：

$$\max_{D} V(G,D) = \mathop{\mathrm{E}}_{x\sim P_r}\left[\log D(x)\right] + \mathop{\mathrm{E}}_{x\sim P_g}\left[\log\left(1-D(x)\right)\right] \tag{2-2}$$

(2) 优化生成器 G：

$$\min_{G} V(G,D) = \mathop{\mathrm{E}}_{x\sim P_g}\left[\log\left(1-D(x)\right)\right] \tag{2-3}$$

训练 GAN 时，首先固定生成器，通过优化判别器提高其区分真实数据和生成数据的能力，直至判别器输出 0.5，表示达到最优状态，无法准确区分两种数据。然后，固定判别器，优化生成器，使生成的数据能够更好地欺骗判别器。两个网络交替进行，直到模型收敛为止。

GAN 的发展可以被划分为两类：一类专注于提升模型的性能，比如图像风格转换[41, 42]、图像生成[43, 44]、风格迁移[45, 46]等；另一类则专注于解决 GAN 框架中存在的挑战，比如训练不够稳定[46]、模型崩溃[47]、样本缺乏多样性等。为了解决这些问题，研究者们提出了各种形式的正则化方法、优化算法、损失函数设计等技术手段，使得 GAN 能够应用于更广泛的领域，并取得更好的效果。如图 2.2 所示。

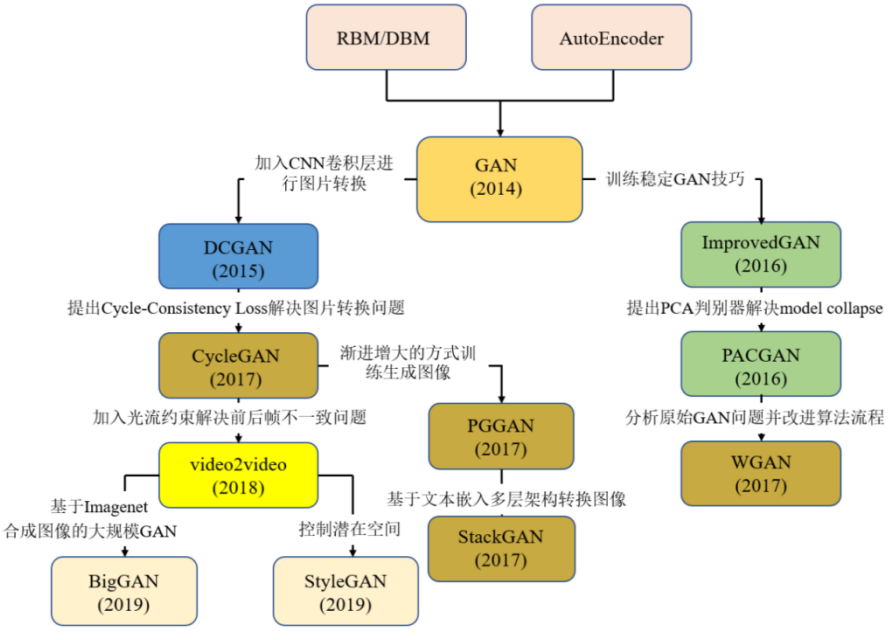

图 2.2 生成对抗网络的发展

2.2 注意力机制

注意力机制与人类眼睛看见事物的过程相似，通过模仿人类看的能力，对感兴趣的事物投入更多的关注，而忽视那些无关紧要的信息。旨在让系统像人类一样，能够在众多信息中筛选出关键信息，从而更加高效地完成任务。对事物的关注程度就是权重。注意力机制可分为强注意力、软注意力以及自注意力。这些实现方式的不同之处在于权重的分配方式以及计算方法等方面。通过了解这些注意力机制的基本特点，可以更好地理解注意力机制在深度神经网络中的应用。

2.2.1 强注意力机制

强注意力机制对输入的图像处理方式不同，只会留下有用的区域。与软注意力机制相比，强注意力机制不考虑输入信息的权重，因为是一种二值化的过程，随机预测输入是被关注还是被忽略。

在图像裁剪中，强注意力机制可以帮助模型定位并关注图片中最重要的区域，以便将其裁剪并保留下来。这种方法可以节省一定的时间和计算成本，并且可以将图像集中在最相关和最感兴趣的部分。然而，这种方法可能会丢失图像中的某些信息，因为没有被关注的部分将被裁剪掉。由于强注意力是不可微分的，因此通常需要采用强化学习的方法来训练网络。

2.2.2 软注意力机制

软注意力机制更加强调通道或空间特征的关系，被视为对每个输入特征的权重分配，权重范围从 0 到 1。权重越高，对应的输入信息在计算中的贡献越大。与强注意力相反，软注意力是可微的。由于关注了每一个输入，计算得到数据之间的关联性，所以较强注意力机制数据量更大。软注意力根据权重应用方式不同可分为空间注意力[48]和通道注意力[48]。

(1) 空间注意力：

空间注意力可以帮助神经网络更好地关注输入图片中的重要区域，即让神经网络“看哪里”，其重点在于强调哪些区域的特征对于目标任务的重要性更高。具有代表性的是谷歌 Deep Mind 团队提出的 STN(spatial transformer network)网络[49]。特征图之间存在一定的空间关系，例如，对于一张图像，经过第一层卷积的特征图通常会捕捉到一些低层次的边缘和纹理信息，而经过第二层卷积的特征图则会在第一层的基础上进一步提取更高层次的语义内容，这些特征图之间的空间位置和特征提取的语义信息存在一定的联系。通过学习通道间的像

素权重，得到权重矩阵，然后将其作用于原始特征图上。空间注意力的结构如图 2.3 所示。

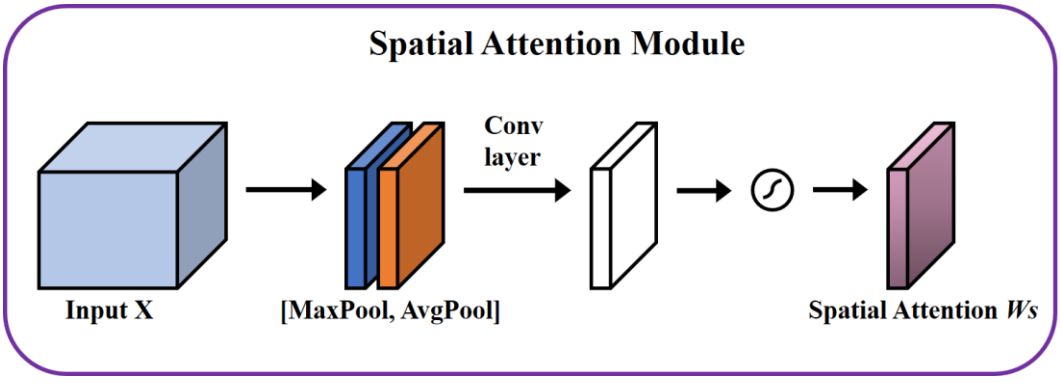

图 2.3 空间注意力模块

对于给定的输入特征图 X，进行按通道逐特征点最大池化和平均池化操作，获得两个跟原特征图尺寸相同的权重向量，将这两张特征图在通道维度上拼接，然后通过一个通道数为 1 卷积处理。最后，将卷积层的输出通过 Sigmoid 函数进行激活，赋予权重给特征图上的每一个特征点。具体地，该权重矩阵的计算公式(2-4)如下所示。

$$\begin{aligned} W_s(X) &= \sigma\left(conv^{7\times7}\left(\left[AvgPool(X);MaxPool(X)\right]\right)\right) \\ &= \sigma\left(conv^{7\times7}\left(\left[X^s_{avg};X^s_{max}\right]\right)\right) \end{aligned} \tag{2-4}$$

(2) 通道注意力：

通道注意力可以被理解为神经网络对于不同特征通道的关注程度，或者说是神经网络"看"哪些特征通道更加"关心"。最具代表的是 SENet[50]。通道注意力模块是一种自适应机制，用于提取重要特征通道。实现了对不同特征通道的加权调节，通过自适应地分配不同特征通道的权重，使网络能够更加聚焦于最具区分度和表征能力的特征通道，从而提高网络的泛化能力和鲁棒性。通道注意力结构如图 2.4 所示。

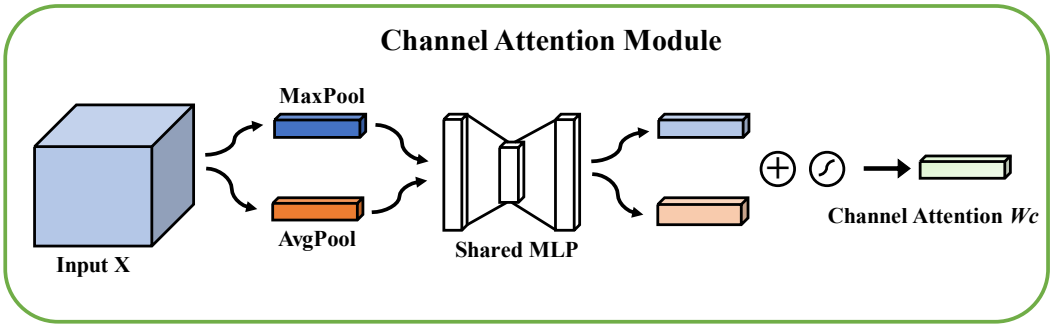

图 2.4 通道注意力模块

通道注意力机制的由两个阶段构成。给定一张输入特征图 X，经过平均池

化以及最大池化处理之后，得到两个权重向量。接下来分别输入到参数共享的感知机当中，并将得到的权重相加。将经过处理后的结果输入到 Sigmoid 函数中，可以得到每个特征通道的权重，权重的取值范围在 0 到 1 之间。接下来，将这些权重与原始特征图相乘，可以得到加权后的特征图。其计算公式(2-5)如下所示。

$$\begin{aligned} W_c(X) &= \varphi\left(MLP\left(AvgPool(X)\right)+MLP\left(MaxPool(X)\right)\right) \\ &= \varphi\left(H_1\left(H_0\left(X_{avg}^c\right)\right)+H_1\left(H_0\left(X_{max}^c\right)\right)\right) \end{aligned} \tag{2-5}$$

其中 W_c 为通道权重参数， X 为输入特征图， H 为共享感知层的权重参数，φ 为激活函数。

(3) 混合注意力[48]：

根据通道注意力的连接方式不同，可分为基于全连接的 SENet 和基于卷积的 ECANet[51]。CBAM[48]与 SENet 相比引入了空间注意力模块，效果得到大幅提升。因为不再是以往单一池化操作，而是通过相加或堆叠最大池化和平均池化。通道注意力模块使用相加的方式，空间注意力模块使用堆叠的方式。

混合注意力机制结合了空间和通道注意力的优势，不仅考虑了每个像素在空间上的重要性，也考虑了通道的重要性。与空间域和通道域注意力相比，混合域注意力不仅具有可解释性强的优点，而且在模型性能方面也有着显著的提升。混合注意力模块的结构如图 2.5 所示。

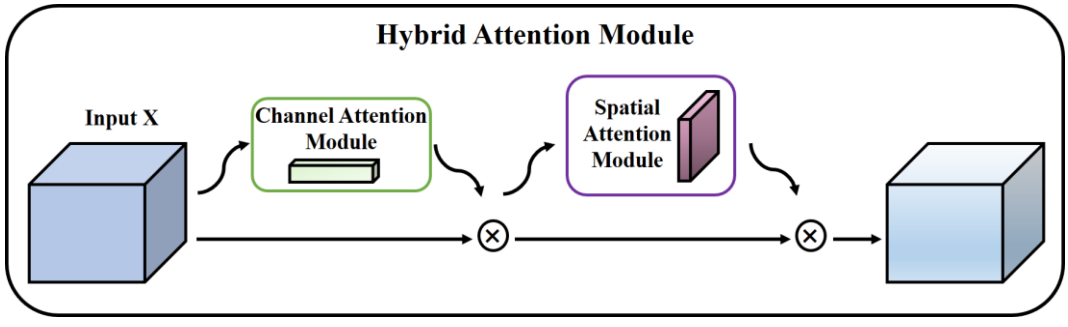

图 2.5 混合注意模块

将特征图 X 输入到通道注意力模块中进行处理，得到通道注意力权重。然后将通道注意力权重处与原特征图做点乘运算，得到新的特征图。接下来，采用空间注意力模块对新特征图进行处理，得到空间注意力权重。将空间注意力权重与新特征图进行点乘运算，可以得到混合注意力的最终输出。注意力机制不仅指导关注哪些重要的特征，还提高这些特征的表现力，并抑制那些不必要的特征。采用了通道注意力模块和空间注意力模块，分别在通道和空间维度上学习重要特征的位置和强度。

2.2.3 自注意力机制

自注意力能够根据输入数据之间的内部关系，为每个输入项赋予相应的权重。这种权重的计算是基于输入项之间的相互作用，从而决定每个输入项的重要性。自注意力拥有并行计算的优势，是上文提到的注意力机制所不能相媲美的。自注意力机制最初被应用于 NLP 中，但伴随技术发展，应用在计算机视觉方面也越来越多，例如谷歌提出的 VIT[52]模型已经在多种图像分类任务中取得了令人瞩目的成果。在 VIT 中，注意力机制分为两个主要组成部分：自注意力和跨注意力。自注意力用于对序列中的每个元素之间的关系进行交互和信息传递，而跨注意力用于在不同的序列之间进行交互和信息传递。自注意力结构如图 2.6 所示。

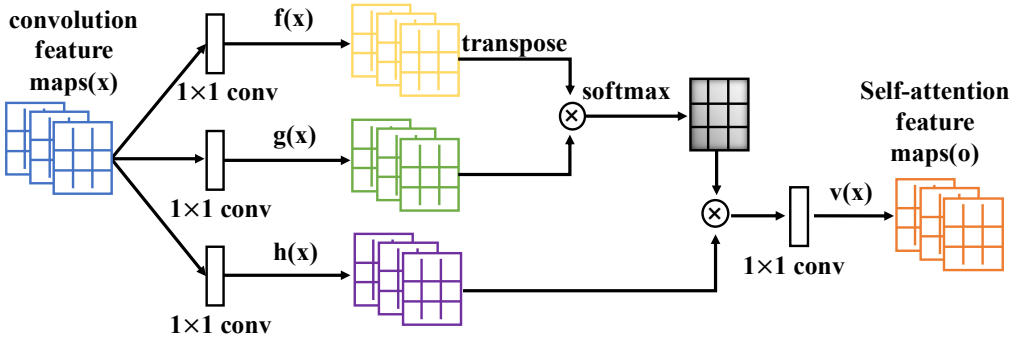

图 2.6 自注意力机制

在自注意力机制中，输入的特征图通常是由卷积网络（如 VGG[53]、ResNet[54]、Xception [55]等）提取的图像特征，通常会去掉最后的一些下采样层，从而得到与原始输入图像大小比较接近的特征图。通过自注意力计算得到特征图中不同位置之间的权重关系，从而实现对不同位置的特征信息加权处理。计算过程通常分为三个步骤：

1．第一步是将查询向量 q、键向量 k 和值向量 v 分别乘以对应的权重矩阵，得到对应的查询矩阵 Q、键矩阵 K 和值矩阵 V。这里的权重矩阵是需要通过训练学习得到的。

2．第二步是对查询矩阵 Q 和键矩阵 K 进行点积运算，以得到相似度矩阵，并在此基础上进行缩放操作（除以向量长度的平方根）。再使用 softmax 函数将相似度矩阵转化为权重矩阵，得到每个键对应的权重值。

3．加权求和：将每个值矩阵 V 乘以对应的权重值，并将结果加权求和，得到注意力输出向量。

自注意力机制由三个分支组成，分别是 Query，Key 和 Value。将输入的序列数据（如文本序列或图像特征图的序列）经过线性变换，如下公式(2-6)所示。

$$
\begin{aligned}
Q &= W_Q X \in \mathbb{R}^{d_3 \times N} \\
K &= W_K X \in \mathbb{R}^{d_3 \times N} \\
V &= W_V X \in \mathbb{R}^{d_3 \times N}
\end{aligned} \tag{2-6}
$$

其中， X 为输入， W_Q, W_K, W_V 为权重矩阵， Q 为查询矩阵， K 为键矩阵， V 为值矩阵。然后得到注意力输出向量 h_i，如下公式(2-7)所示。

$$
\begin{aligned}
h_i &= att\left((K,V), q_m\right) \\
&= \sum_{n=1}^{Y} \alpha_{mn} v_j \\
&= \sum_{n=1}^{Y} softmax\left(s\left(k_n, q_m\right)\right) v_n
\end{aligned} \tag{2-7}
$$

式中， $n, m \in [1, Y]$ 为向量序列的位置， α_{mn} 为连接权重 。

2.3 VITON

VITON 的主要目标是对给定人物的照片和目标服装的图片处理，并生成一张虚拟的试穿图，呈现出试穿者穿着目标服装的效果。首先，通过一个粗略的合成图像网络 G_C，生成一个变形后的目标服装，使其与试穿者的身体姿态相匹配。接着，使用一个细化网络 G_R，利用扭曲服装中包含的细节信息对模糊区域进行渲染，以达到更好的效果。VITON 网络的结构如图 2.7 所示。

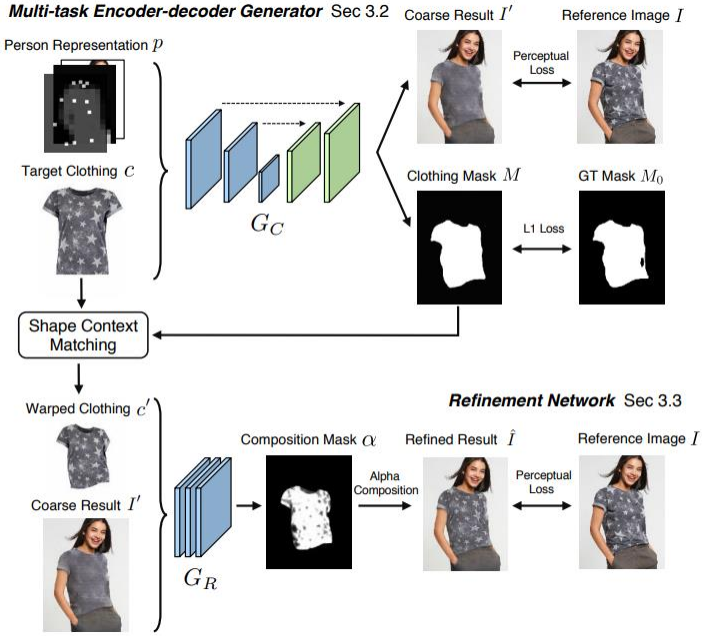

图 2.7 VITON 网络结构

VITON 模型并不直接将穿着衣服的人物照片输入到网络中进行处理。相反，该模型采用一系列预处理步骤，以摒弃输入图像中服装信息。该预处理阶段分

为三个小任务，通过将最终的结果整合起来，得到一种与服装无关的人物表示。使用这种人物表示，模型可以更好地控制生成的服装图像，使其更符合人体的姿势和其他身份特征，从而实现更准确和真实的试穿效果。下图 2.8 展示了这一过程的具体流程。

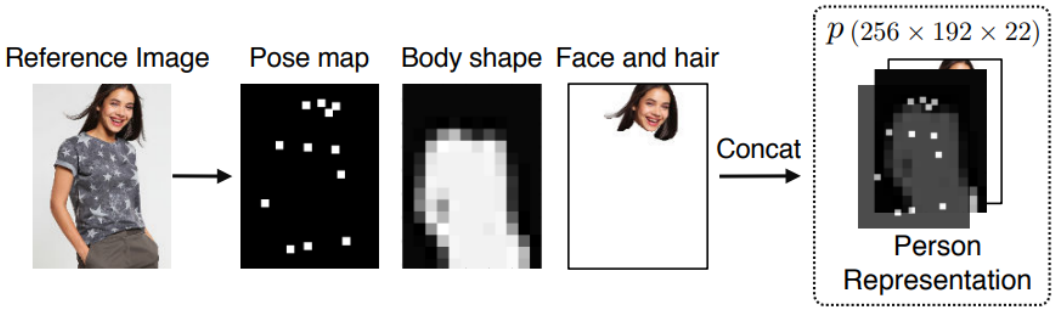

图 2.8 VITON 人物表示

人物表示分成三个部分：首先是由 18 个通道组成的人体姿势热图，用于表示人体的姿势信息。其次是 1 通道的边界模糊人体分割图，用于表示人物轮廓的体型信息，最后是由 3 个通道组成的面部和头发的 RGB 图像，更多的被用于描述人物的身份特征，这三个部分共同组成人物表示 p 。

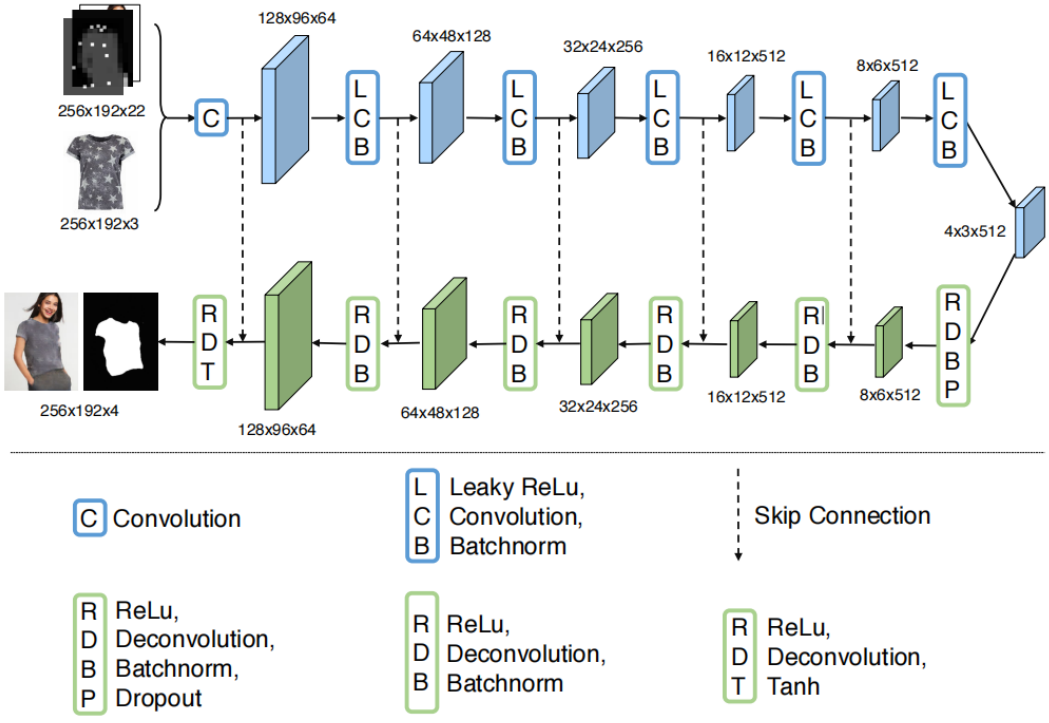

图 2.9 编解码生成器

如图 2.9 所示。编解码生成器：采用了一个多任务 U-Net 网络编解码生成器，通过跳跃连接来传递信息，生成粗糙试穿图像 I' 的同时，还有服装掩码 M 。生成的服装掩码不仅可以指导网络关注服装区域，还可以进一步用于完善生成的结果。

编解码生成器结构使用了感知损失 L_P 和 L_1 损失的组合损失函数。如式(2-8)所示。

$$L_{Gc}=\sum_{i=0}^{5}\lambda_i\left\|\phi_i\left(I'\right)-\phi_i\left(I\right)\right\|_1+\left\|M-M_0\right\|_1 \tag{2-8}$$

式中，$\sum_{i=0}^{5}\lambda_i\left\|\phi_i\left(I'\right)-\phi_i\left(I\right)\right\|_1$ 为感知损失 L_P，$\left\|M-M_0\right\|_1$ 为 L_1 损失，$\phi_i(\cdot)$ 指的是第 i 层卷积输出特征，λ_i 表示第 i 层的权重参数。

细化网络：虽然试穿图像的姿势，服装区域基本一致，但是目标服装的纹理细节不够完善。生成的图片虽然看起来比较真实，但并不知道在何处生成什么样的细节。引入细化网络进行图像渲染，提高生成效果。将扭曲服装 c' 和粗糙试穿图像 I' 输入到细化网络当中得到合成掩码 α，根据 α 合成最终的试穿图像 $\hat{I}$。其中 $\odot$ 表示矩阵点乘，公式(2-9)如下。

$$\hat{I}=\alpha\odot c'+(1-\alpha)\odot I' \tag{2-9}$$

多任务编解码生成器在生成粗糙结果 I'、扭曲服装掩码 M 的同时还要利用形状匹配技术，生成符合人体姿态的扭曲服装 c'，因此模型很难收敛到最优的状态，虽然可以生成试穿图像，但最终试穿效果却不尽如人意。

2.4 CP-VTON

CP-VTON 是在 VITON 模型的基础上进行改进和优化的。主要针对 VITON 模型中存在的问题进行了优化，更多是衣服和身体的贴合程度。具体来说，CP-VTON 提出了一个几何匹配模块 GMM，能够更好地处理衣服和身体之间的对齐问题，再经过提出的试穿模块 TOM，将已经对齐的扭曲服装转移到目标人物身上。

CP-VTON 跟 VITON 获取与服装无关的人物表示 p 方式相同，图 2.8 所示。GMM 模块将目标服装 c 根据人体姿势匹配变形得到扭曲服装 $\hat{c}$，如图 2.10 所示。首先，使用卷积神经网络分别提取 p 和 c 的高级特征，使用一个关联层将提取到的两个特征进合并，然后输入到回归网络中得到转换参数 θ。根据 θ 对目标服装做 TPS 转换得到扭曲服装 $\hat{c}$。GMM 模块通过三元组数据 (p,c,c_t) 进行端到端训练，其损失函数如式(2-10)所示。

$$L_{GMM}(\theta)=\left\|\hat{c}-c_t\right\|_1=\left\|T_\theta(c)-c_t\right\|_1 \tag{2-10}$$

其中，c_t 为直接从穿着目标服装的真实图片处理得到的。

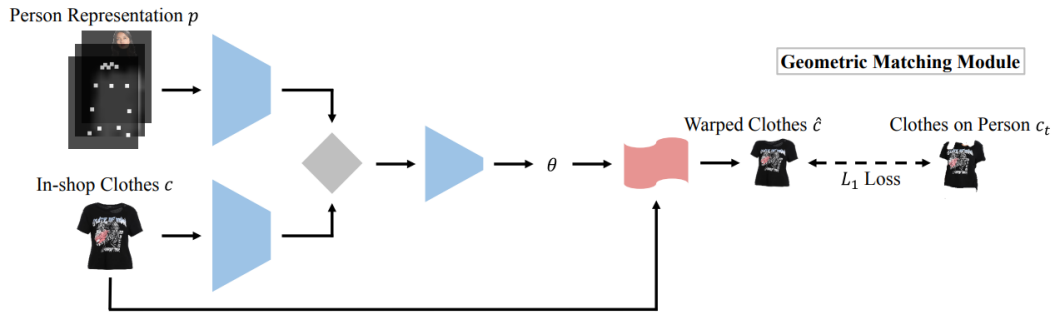

图 2.10 几何匹配模块

将扭曲服装 $\hat{c}$ 直接复制到人物身上，尽管拥有较多的服装细节，但是会出现边缘模糊，手臂消失等情况。还有一种方法是使用 U-Net 结构进行转换，这样可以缓解上述问题，但衣服与人体对齐存在不稳定性。TOM 试衣模块将两者进行整合，如图 2.11 所示。

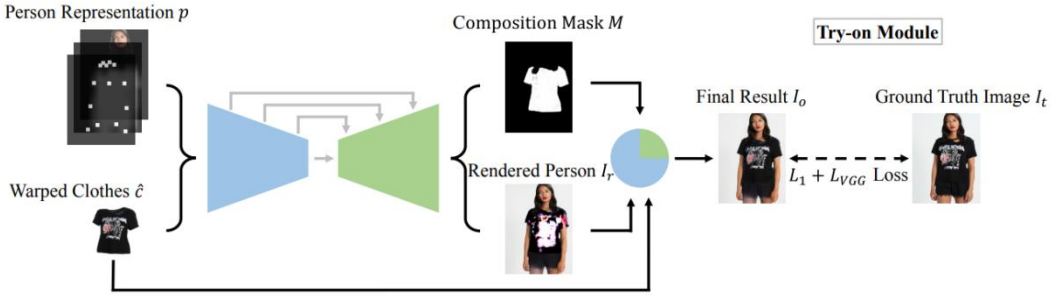

图 2.11 试衣模块

利用 U-Net 网络将输入的人物表示 p 和扭曲服装 $\hat{c}$ 生成粗略试穿图像 I_r 和掩码 M 。使用 M 将 I_r 与 $\hat{c}$ 融合生成最终输出 I_o ，公式如下(2-11)。

$$I_o = M \odot \hat{c} + (1-M) \odot I_r \tag{2-11}$$

TOM 模块采取三元组数据 (p, c, I_t) 进行端到端训练，损失函数同样为感知损失和 L_1 损失的组合损失函数，试穿模块总体损失为：

$$L_{\text{TOM}} = \lambda_{L1} \left\| I_o - I_t \right\|_1 + \lambda_{vgg} L_{\text{VGG}}(\hat{I}, I) + \lambda_{mask} \left\| 1-M \right\|_1 \tag{2-12}$$

其中，$\left\| 1-M \right\|_1$ 表示生成掩码 M 使用 L_1 正则化，保留更多生成细节。

2.5 本章小结

在本章中，介绍了生成对抗网络。随后，详细探讨了注意力机制的几种分类，并对其内容进行了详细阐述。最后引入了经典虚拟试衣网络 VITON 和 CP-VTON，介绍两者的网络结构和工作流程。

3 基于多级注意力机制的虚拟试衣模型

基于深度学习的虚拟试衣技术是近年来虚拟试衣领域的研究热点之一。基于三维建模的虚拟试衣技术需要使用昂贵的设备和大量精确的数据来进行实时的物理渲染。相比之下，基于图像的虚拟试衣技术所需的设备和计算机资源要求相对较小，更适合用于线上购物。试穿结果可以方便地在电商平台上展示和分享，用户可以迅速获得反馈和建议。CP-VTON 相比于 VITON 在生成图像质量上有一定的提升，但是在处理遮挡问题和服装细节方面仍然存在不足。为了提高试穿效果、解决边界模糊，遮挡变形等问题，本文提出一种基于多级注意力机制的虚拟试衣模型(Multi-level Attention Mechanism Virtual Try-On,MA-VTON)。本章主要介绍 MA-VTON 模型的研究基础、网络结构以及损失函数等内容。

3.1 研究基础

U-Net[56]是一种用于图像分割的深度学习网络，最初是为了解决医学领域的复杂问题而提出的，可以有效地分割细胞、组织、器官等复杂的图像，并且在计算机视觉领域取得了巨大的成功。U-Net的构建技术为深度学习模型的设计提供了重要的灵感，许多出色的深度学习模型都借助于它的强大功能，得到了显著的改进和优化。因为 U-Net 的网络结构呈现出“U”字形，所以被称为 U-Net，具体的网络结构如下图 3.1 所示。

U-Net的网络结构由三个部分组成：左侧为特征提取，中间为特征拼接，右侧为特征融合。其中，左侧特征提取是关键步骤之一，通过四次下采样逐渐减小输入图像尺寸，提取更多特征，实现更高效的网络。

网络中共有四个拼接操作，通常被称为“跳跃连接”。主要目的是将不同尺度深层和浅层的信息融合起来，增强特征的表达能力。在进行拼接操作时，需要注意的是，不仅图片的尺寸需要保持一致，而且特征的通道数也必须相同才能进行拼接。

通过上采样，可以将图片的尺寸扩大，并提取出更多的尺度信息。该部分包括四个操作，每个操作都会减少特征图的通道数。在上采样过程中，左侧特征提取网络将特征图传递给右侧网络，即左侧特征和右侧特征进行拼接融合。

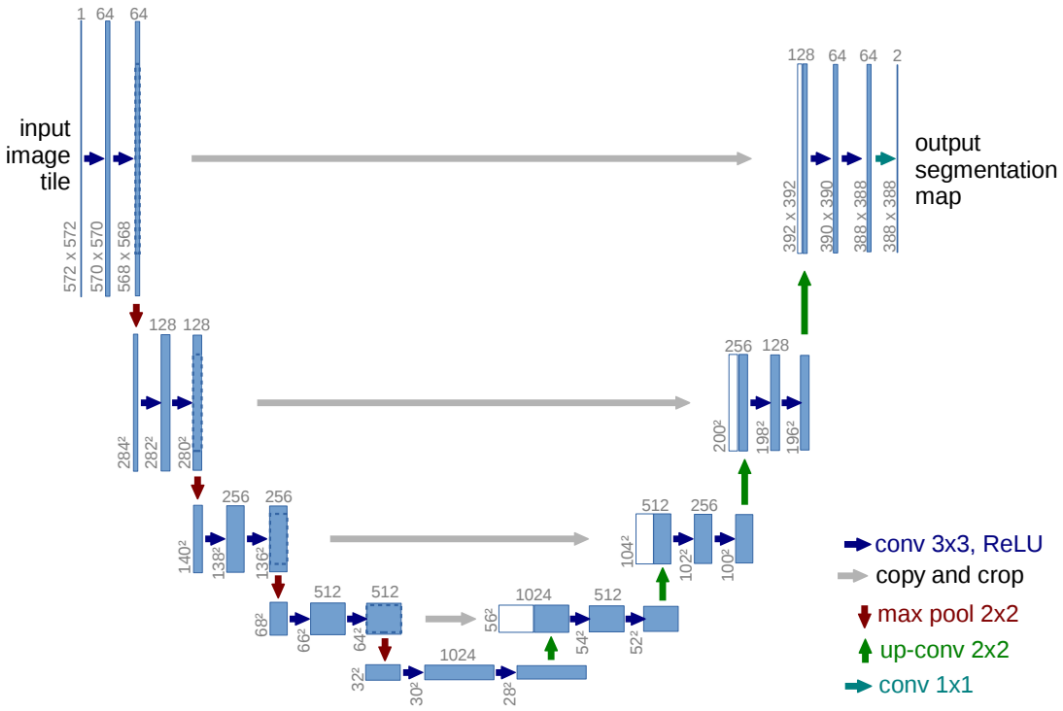

图 3.1 U-Net 网络结构

U-Net的编解码部分由多个卷积层和池化层组成，这些局部运算关注并提取局部信息。然而，为了提取长距离信息，需要增加网络深度。这种方式存在一定缺陷，即参数量过大，且过多的下采样会导致空间信息的丢失更多。在图像分割任务中，需要精确预测每个像素，因此空间信息的丢失会导致分割结果不准确。

3.2 MA-VTON 模型

深度学习模型中，底层特征提供了图像局部内容和纹理细节，而高层特征包含了更丰富的全局内容和语义信息。为了综合利用图像的局部和全局信息，同时捕捉图像的细节和语义信息，将底层和高层特征进行融合，这样可以帮助高层网络获得更多的局部信息。此外，深度模型提取的特征通常会包含大量冗余信息，这些冗余特征会对检测结果产生负面影响。因此，引入多级自注意力机制，可以突出图像中的重要特征，从而提高模型的检测精度。本文基于 CP-VTON 网络结构，提出了一种基于多级自注意力机制的虚拟试衣模型。模型由几何匹配模块和试穿模块两部分组成。本文将上述提出的模型命名为 MA-VTON(Multi-level Attention Mechanism Virtual Try-On)模型。

3.2.1 服装变形模块

服装变形模块是用于将目标服装按对应人体姿势和区域进行变形。该过程