arXiv:2301.00965v1 [cs.CV] 3 Jan 2023

OccluMix：通过语义引导的 Mixup 实现去遮挡虚拟试穿

Zhijing Yang†、Junyang Chen†、Yukai Shi*、Hao Li、Tianshui Chen、Liang Lin

摘要——**图像虚拟试穿旨在将个人图像中的布料替换为服装图像（店内服装），这一技术日益受到多媒体和计算机视觉领域的关注。以往方法成功保留了服装图像的特征，但遮挡仍是影响虚拟试穿真实感的顽固问题。在这项工作中，我们首先对遮挡进行全面分析，并将其归纳为两类：i）固有遮挡：原有服装的残影仍存在于试穿图像中；ii）获得性遮挡：目标服装扭曲到不合理的身体部位。基于深入分析，我们发现，遮挡可以通过一种新颖的语义引导 Mixup 模块进行模拟。该模块能够生成特定语义的遮挡图像，并将其与试穿图像配合使用，以促进去遮挡试穿（DOC-VTON）框架的训练。具体而言，DOC-VTON 首先对试穿人物进行精细化语义解析。在语义引导和姿态先验的辅助下，以复制粘贴的方式，将各种复杂纹理选择性地融合到人体部位上。随后，利用生成模块（GM）负责合成最终的试穿图像，并联合学习去遮挡。与最先进的方法相比，DOC-VTON 通过减少遮挡影响，获得了更佳的感知质量。**

索引词—**深度学习、虚拟试穿、遮挡处理、数据增强。**

I. 引言

VIRTUAL 虚拟试穿是一种热门应用，通过将所需的店内服装转移到参考人物身上来实现。随着电子商务需求的增长，虚拟试穿受到了越来越多的关注。尽管近期试穿网络的发展已能够通过生成合身的形状和具有逼真视觉质量的试穿图像来实现真实的服装变形 [1]—[7],但在变形后的试穿图像中定位并解决遮挡效应仍然是一项巨大挑战。

为研究现有视觉试穿方法中的遮挡效应，我们遵循 CP-VTON+ [8] 的流程，探索两种代表性方法的结果 [2], [7]。在我们的实验中，ACGPN [2]代表基于解析器的方法，而 PF-AFN [7] 代表无解析器方法。
如图 1所示，在第一行和第二行中，试穿图像中的遮挡表现为先前服装的残影。我们将这种遮挡称为固有遮挡。通常，固有遮挡是由泛化能力较差和人体解析错误造成的。如图1第三行和第四行所示，遮挡效应是由新服装的形状变形错误造成的。我们将其称为获得性遮挡。因此，我们主要展示

† 前两位作者贡献相同。代码见:
https://github.com/JyChen9811/DOC-VTON

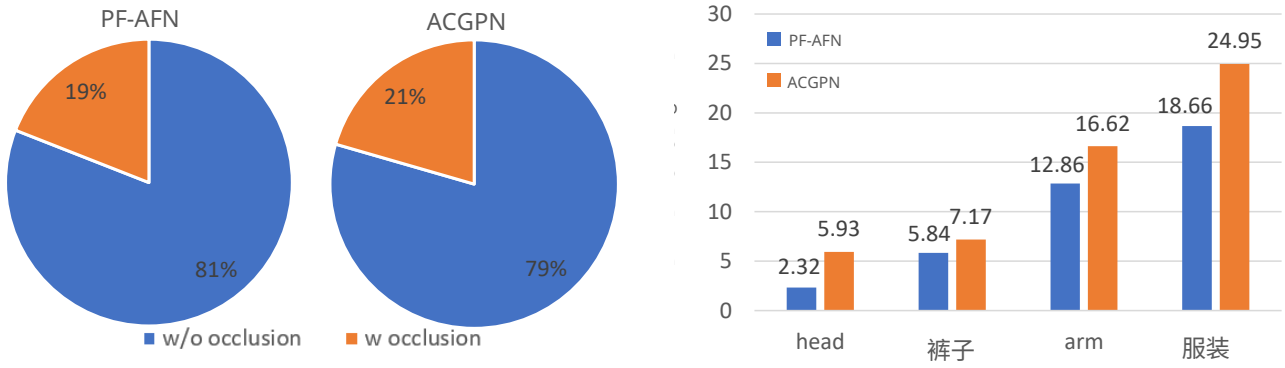


图1：在 VITON [1] 数据集上，当前最先进模型的试穿结果 [2], [7] 中，原有服装和新服装都出现了不理想的遮挡。为解决遮挡现象，我们尝试将大多数遮挡归为两类（即，固有遮挡和获得性遮挡）。在补充文件中，我们提供了大量遮挡样本，以验证固有遮挡和获得性遮挡这两种类型。

遮挡问题分为两种形式（即，固有遮挡与获得性遮挡）。试穿流程中的服装变形模块很容易受到空间变换的误导；当人体姿态出现较大变化时，遮挡效应会导致明显的质量下降。为此，一些先驱性的合成式模型 [1], [3], [9], [10] 受困于上述缺陷，生成的图像质量有限。

在图2中，我们通过人工标注研究了虚拟试穿方法中的遮挡效应。具体而言，4064张图像 [8]均由人工标注了遮挡是否存在及其位置。统计结果表明，遮挡仍是试穿任务中的主要挑战，出现率为 20%。我们进一步计算了结果中各身体部位的平均弗雷歇 Inception 距离（FID）值。手臂和服装部位的 FID 分数及空间方差较大，这促使我们聚焦这些具有挑战性的人体部位，以提升图像质量。

为解决上述问题，我们提出了一种鲁棒的虚拟试穿去遮挡框架（DOC-VTON），充分利用试穿图像的语义布局。DOC-VTON 自适应地为生成模块（GM）执行裁剪与粘贴操作 [11], [12]，以实现去遮挡。具体而言，
DOC-VTON 由三个



(a) (b

)图2：(a) 为了探究遮挡样本的比例，我们在人工标注的辅助下统计了 CP-VTON 数据集上 PF-AFN 和 ACGPN 结果中的遮挡示例+ [8] 。(b) PF-AFN 生成的不同身体部位的弗雷歇 Inception 距离（FID）分数 [7]以及 ACGPN [2]。

模块组成：i）服装变形模块（CWM），利用外观流将店内服装及其对应的掩码变形为合身形状；ii）遮挡混合模块（OccluMix），基于锐化解析网络（SPN）的语义布局，模拟不同的遮挡情况并生成 OccluMix 样本；iii）生成模块（GM），用于将 OccluMix 样本中的服装迁移到真实人物图像上，从而联合实现最终试穿图像的生成与去遮挡。

借助上述模块，DOC-VTON 在虚拟试穿中采用语义引导混合策略来学习去遮挡。总而言之，本文的主要贡献如下：

- 我们对当前视觉试穿算法中的遮挡效应进行了全面分析。这是首次尝试分析这一问题，有助于推动去遮挡虚拟试穿的进一步研究。
- 我们提出了一种简单但颇具挑战性的虚拟试穿遮挡混合策略（OccluMix），通过在语义和姿势的引导下混合复杂程度各异的纹理，生成具有挑战性的遮挡试穿人物。
- 与通用解析流程相比，我们研究了一种锐化解析网络（SPN），用于迭代解析试穿图像。SPN 不仅能精准处理不合理的变形部分，还能为 OccluMix 提供区域。
- 大量实验和评估表明，我们的方法在 VITON任务中，无论定性还是定量评估，均取得了最佳的最先进结果。

在本文余下部分中，第二节简要概述现有虚拟试穿方法中的遮挡现象及数据增强的衍生方法。第三节对最先进方法中的遮挡问题进行全面分析。第四节介绍 DOC-VTON 流程，并详细阐释 OccluMix。在第五节中，我们讨论 OccluMix 与其他数据增强方法的比较。在第六节中，我们通过与现有最先进方法进行比较，开展实验以验证 DOC-VTON 的有效性和效率。最后，我们在第七节中总结本文，并展望未来研究方向。

II. 相关工作

虚拟试穿。现有基于深度学习的虚拟试穿方法主要分为基于三维模型的方法 [13]–[16]和基于二维图像的方法 [1], [2], [7],[9], [17]–[19]。前者需要进行三维测量，会带来额外的计算开销。因此，基于二维图像的方法更适用于现实场景。例如，VITON [1], CP-VTON [9], ACGPN [2], ClothFlow [3], DCTON [20]和 RT-VTON [21] 使用人体表示作为输入来生成穿着服装的人物。此外，WUTON [10]和PF-AFN [7] 采用了无解析器方法。在服装变形方法中，VITON [1], CP-VTON [9],ACGPN [2], DCTON [20] 和 VITON-HD [22]使用薄板样条（TPS）[23] 变换，将目标服装变形为合身的形状。然而，TPS 变换的变形能力有限。为此，RT-VTON [21] 提出了半刚性变形方法，以使变形服装与预测语义对齐。ClothFlow [3] 和 PF-AFN [7] 使用外观流 [24], ，能够轻松地将目标服装平滑地变形并贴合到目标人物身上。

作为典型的试穿流程，VITON 和 CP-VTON 使用粗略形状和姿态图，以确保对任意服装的泛化能力。然而，当解析结果变得不准确时，基于解析器的方法 [1], [2], [9], [25]会生成质量较差的试穿图像。近期，PFAFN [7] 提出了一种开创性的无解析器知识蒸馏方法，摆脱了不准确分割带来的干扰。尽管如此，PFAFN 生成的图像仍会遇到固有遮挡和获得性遮挡。为此，我们提出了一种用于虚拟试穿的新型去遮挡方法（DOC-VTON）。DOC-VTON 能够处理目标服装与参考人物之间的错位，并减少由先前服装造成的重影效应。

去遮挡。去除目标物体上的部分遮挡是一项重要的计算机视觉任务 [26]–[31]。Sail [32] 提出了一种新颖的自监督框架，无需人工标注即可利用真实世界数据解决场景去遮挡问题



图3：固有遮挡分析。对于无解析器虚拟试穿结果，无论试穿何种服装，先前服装的残影都会残留。与无解析器方法相比，基于解析器的方法会因解析结果失败而生成扭曲图像。

。基于生成对抗网络的方法 [29] 被用于修复面部的遮挡区域。在虚拟试穿任务中，服装和人体被遮挡的问题同样存在。为避免遮挡现象，基于解析器的方法 [1], [2], [9], [25]使用人体解析器来理解身体部位的空间布局。然而，基于解析器的方法往往会生成质量较差且存在明显遮挡的试穿图像。最近，PFAFN [7]提出了一种开创性的无解析器方法，但图像质量仍受不合适的形状变形服装以及通用模型生成能力的限制。为解决上述问题，我们将 OccluMix 策略引入虚拟试穿任务。

数据增强。近年来，一些研究采用数据增强（DA）[33]–[39] 来增强泛化能力。Mixup [12], [40]通过在图像层面进行凸线性插值来实现数据增强。CutMix [11]提出从输入图像中裁剪一块区域并粘贴到其他图像中，以实现数据增强。尽管 MixUp 和 CutMix 展现出实际效果上的改进，但它们并未利用显著性、语义和光流等图像先验知识进行引导。随机裁剪 [39] 将图像裁剪为特定尺寸，从而生成合成数据。与 Mixup 和 Cutmix 相比，它能够灵活地保留先验知识，并解决不同尺寸试穿图像 [38]造成的失真问题。然而，在简单缩放图像上解决遮挡现象（即固有遮挡和获得性遮挡）仍是一项重大挑战。受 Supermix [36], 启发，我们研究了一种面向目标的数据增强方法，利用人体解析先验 [41] 来训练试穿生成模型。



图4：试穿图像继承了扭曲变形服装的扭曲图案。

三、遮挡分析

现有最先进的二维虚拟试穿方法可分为基于解析器的方法和无解析器方法。然而，这两种方法在试穿图像中仍然存在遮挡现象。为了研究这两种方法中的遮挡影响，我们遵循 CP- VTON 的流程+ ，探索 ACGPN 和 PF-AFN 中的遮挡现象。前者代表基于解析器的方法，PF-AFN 则代表无解析器方法。我们首先探究扭曲变形服装所导致的遮挡效应。如图 4所示，如果变形结果仍呈现扭曲形态，生成结果也会保留这种形态，并以遮挡的形式呈现。我们将这种遮挡称为获得性遮挡，

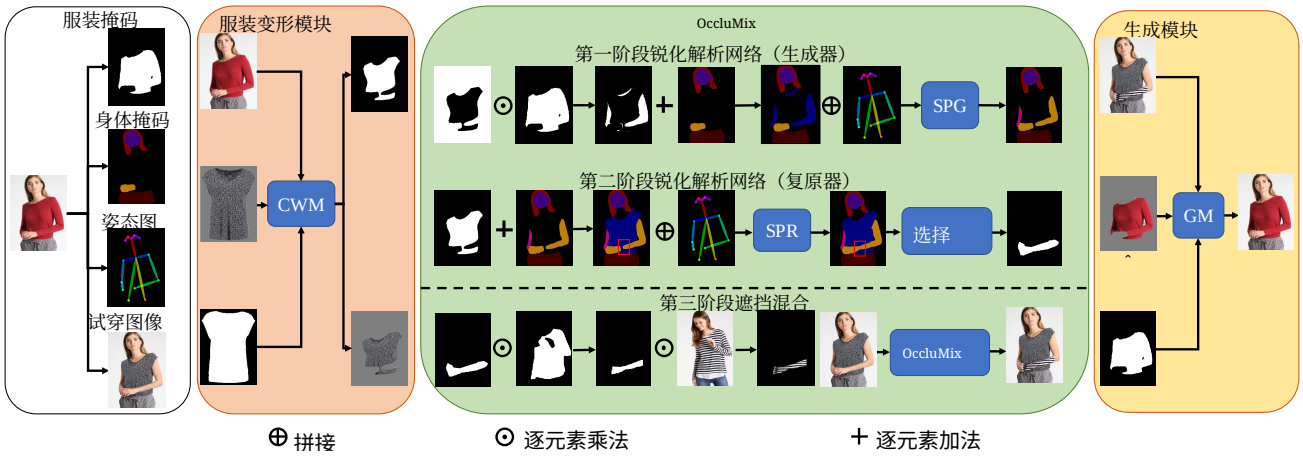


图 5：我们提出的方法示意图。OccluMix 通过将不同的服装纹理裁剪并粘贴到试穿图像的困难区域，生成增强图像。为了模拟试穿图像中具有挑战性的遮挡出现情况，我们需要识别试穿图像中的人体组成部分。我们首先获得基于解析器的输入图像 X_A，然后使用服装变形模块（CWM）预测变形服装 T_C^W和变形服装掩码 M_C^W；在步骤 I 中，锐化解析生成器（G）首先将服装掩码 M_C与 $\tilde{M}_C^W$相乘，得到异常区域 M_e，然后将其与身体部位 M_w 结合，以生成身体部位的潜在位置 M_w^p（包括头部、手臂和裤子）。随后，G 使用姿态图 M_p和身体部位的潜在位置 M_w^p生成身体部位的粗略掩码 M_w^g；在步骤 II 中，锐化解析恢复器（R）细化躯干的粗略掩码 M_t^o，得到完整的躯干掩码 M_t^c。在步骤 III 中，我们使用 DensePose [42]选择一个具有挑战性的片段 Y_A^r，与辅助服装 X_B 相乘以获得纹理遮挡 C_B。然后，我们将 C_B 与试穿图像 X_A^s 混合，得到 OccluMix 样本 X。最后，我们利用生成模块，通过变形服装信息 I_c、M_c 和 X生成试穿ˆ ˆ ˆ˜图像 X_A 。

它由扭曲的变形结果引起。看起来，只要变形结果恰当，生成器便能生成干净的试穿图像；然而，生成阶段同样会出现遮挡效应。如图 3所示，试穿结果会残留先前服装的残影。对于无解析器生成器，我们发现，无论试穿何种服装，残影都会残留。由于无解析器生成器需要对参考图像中的试穿区域进行分类，我们认为残影是泛化能力较差所致。与无解析器生成器相比，基于解析器的生成器会因解析结果失败而生成扭曲图像。我们将先前服装造成的这种遮挡效应称为固有遮挡。因此，我们主要展示两种形式的遮挡问题。（即固有遮挡和获得性遮挡）。

IV. 方法

如图 5 所示，所提出的 DOC-VTON 由三个模块组成。首先，服装变形模块旨在将目标服装图像变形到真实图像上。其次，OccluMix 模块通过语义信息逐步生成试穿图像中身体部位的掩码，并利用裁剪与粘贴策略在试穿图像上产生遮挡效应，从而生成 OccluMix 样本。最后，生成模块合成试穿图像。

A. 服装变形模块（*CWM*）

遵循 PF-AFN 的训练流程 [7], 我们使用二阶约束，以更好地保留布料特性

，该约束定义如下：

$$L_{sec} = \sum_{i=1}^{N} \sum_{p} \sum_{\pi \in N_p} Char(f_i^{p-\pi} + f_i^{p+\pi} - 2f_i^p), \quad (1)$$

其中，f_i^p和 $f_i^{p-\pi}$ 和 $f_i^{p+\pi}$ 分别表示第 i 个通道特征图上点 p 及其邻域点 $p-\pi$ 和 $p+\pi$ 的值。π 表示邻域的尺度。$Char$ 是广义 Charbonnier损失函数 [43]。N_p由第 p 个点周围的垂直、水平及两个对角邻域构成。

CWM 将目标服装变形为合身的形状，同时保留服装细节。当参考人物以简单姿势站立时，其效果良好。当参考人物呈现复杂姿势时，例如躯干扭转且双手挡在身体前方，变形服装可能会遮住部分人体。

B. OccluMix

1) 锐化解析网络（*SPN*）：提出锐化解析网络（SPN），用于改进变形服装，并生成人物的身体部位（例如手臂）。许多以往的工作忽视了一个事实：准确的解析结果可以纠正不合理的变形过程。为解决这一问题，我们采用锐化解析机制，在试穿变换过程中改进人体解析的细节失真。

具体而言，假设一个人试穿新服装，我们将原始图像中的服装掩码和身体部位掩码定义为 M_c 和 M_w（包括头部、手臂和裤子），将试穿图像中的服装掩码和身体部位掩码定义为 M_c^s和 M_w^s，并将 M_p 定义为参考人物的骨架信息。由于缺少 M_w^s，我们使用布料和躯干先验，结合

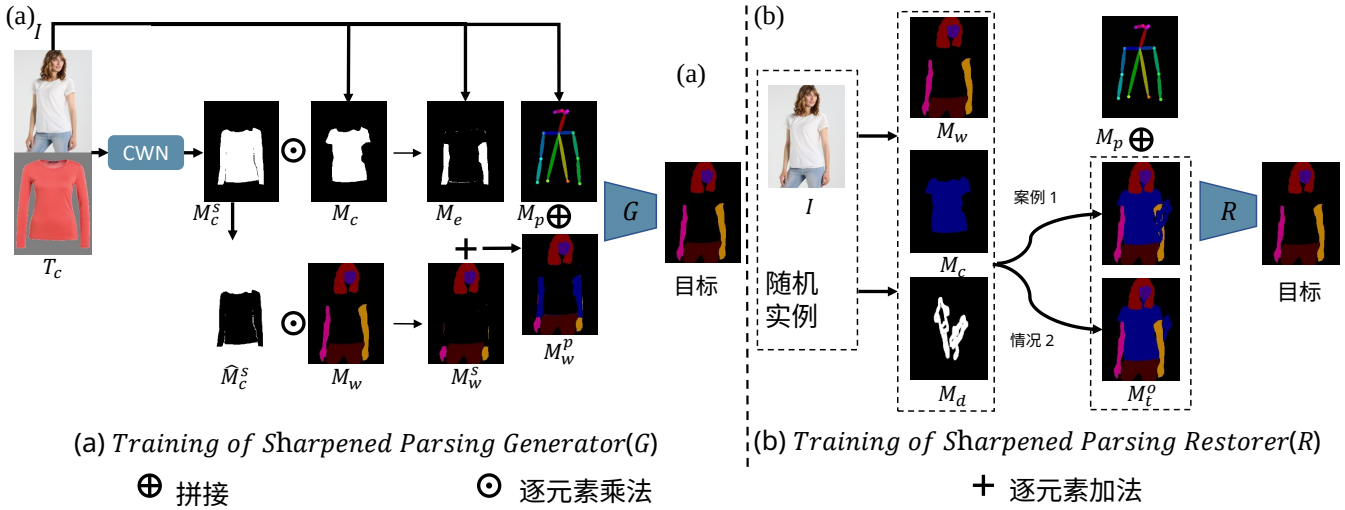


图6：(a) 向生成器输入姿态图 M_p 和潜在位置 M_w^p，以生成身体的粗略掩码。(b) 训练复原器以细化身体掩码。在此阶段，我们划分了两种训练情况。情况 1 鼓励对身体部位的掩码进行部分补全。情况 2 则防止复原器过度补全。

变形服装来获得它。M_w^s 的公式定义如下：

$$M_w^s = M_w \odot (1 - M_c^s). \tag{2}$$

其中 $\odot$ 表示逐元素乘法。M_e 是 M_c^s 与 M_c之间异常布料的掩码，表示生成躯干区域的潜在位置。M_e 的公式如下：

$$M_e = M_c^s \odot (1 - M_c). \tag{3}$$

然后，通过组合 M_e 和 M_w^s，得到试穿图像中身体部位的潜在位置 M_w^p，如下所示：

$$M_w^p = M_w^s + M_e. \tag{4}$$

如图 6 (a) 所示，由于已获得 M_w^p 和 M_w，我们可以使用 $(M_p,\ M_w^p) \xrightarrow{G} M_w$ 的过程来实现 G，根据潜在位置生成躯干的其余掩码M_w^p，并以标签 M_w为监督。

然而，当参考人物以扭曲姿势站立时，目标服装的严重变形可能会破坏 M_w中的身体细节。为解决这一问题，我们从 M_d 不规则掩码数据集 [44] 中引入 M_c，并将其与 M_t^o合并，以模拟身体部分细节丢失的失效情况。其公式定义如下：

$$M_t^o = M_w \odot (1 - M_d) + M_c \cup M_d. \tag{5}$$

然后，我们将其输入复原器 R，而 $(M_t^o,\ M_p) \xrightarrow{R} M_w$ 是细化身体部位细节的过程。请注意，复原器的训练细节见图 6 (b)。

如图7所示，M_o 身体部位的细节失真通过复原器得到增强。细化试穿图像的解析掩码后，我们可以利用它挖掘用于裁剪并粘贴纹理遮挡的区域，并对 OccluMix 数据进行建模。



图7：语义修复组件的效果。当参考人物处于扭曲姿势时，锐化部件生成器会为试穿图像提供错误的语义信息，而修复步骤可以解决这一问题。

2）遮挡混合：本节介绍专为试穿任务设计的数据增强（DA）策略 Occ- luMix。实用的试穿数据增强方法需要模拟具有挑战性的遮挡，并为试穿网络提供良好的正则化。顾名思义，OccluMix 会在试穿图像上叠加不同的服装，促使网络在试穿变换过程中复原逼真的细节。

图 1右上角展示了，当人物穿着纹理复杂的服装并想换上新布料时，个人图像中会残留复杂纹理的残影。为此，我们确保 OccluMix 中存在一定比例的复杂纹理。在实验中，我们使用灰度共生矩阵 [45] 来估计服装复杂度的熵。此外，我们根据纹理复杂度将服装分为两类



图 8：左侧服装包含圆点、条纹及各种其他纹理，因此较为复杂；右侧服装的纹理较少，因此更为简单。

(即 简单和复杂):

$$ENT = -\sum_{i=1}^{k}\sum_{j=1}^{k} G(i,j) log G(i,j), \tag{6}$$

其中， $G(i,\ j)$ 表示不同灰度值出现频率的归一化结果。

如图 8所示，左侧服装被归类为复杂纹理，右侧服装则被归类为简单纹理。随后，我们根据参考人物的纹理分类对训练对进行划分。

$$T = \begin{cases} 1, & ENT >= 2.5, \\ 0, & ENT < 2.5, \end{cases} \tag{7}$$

其中， $T = 1$ 表示服装属于复杂纹理类别，另一个则属于简单纹理类别。

令 $x \in \mathbb{R}^{W\times H\times C}$ 和 y 分别表示随机采样的图像及其对应标签。示例 A（即 $(x_A,\ y_A)$) 是训练样本。示例 B （即（$x_B,\ y_B$））则根据复杂系数 λ从复杂纹理集或简单纹理集中选取。

如图 9所示，试穿混合的目标是通过组合两个训练样本 $\tilde{x}$ ，生成一个新的训练样本$(x_A,\ y_A)$ 和 $(x_B,\ y_B)$。公式如下:

$$\begin{aligned} C &= y_A^r \odot y_B^c, \\ \tilde{x} &= C \odot x_B + (1 - C) \odot x_A, \end{aligned} \tag{8}$$

其中 y_A^r 是用于遮挡混合的选定区域。我们首先统计试穿图像中每个人体部位的遮挡分布。随后，使用 DensePose [42]传递锐化解析网络的结果，并根据 y_A^r 进行筛选，筛选依据是遮挡分布。此外， y_B^c 是人物的服装掩码， x_B 而 C 是纹理遮挡的掩码。如图5所示，我们将裁剪后的服装混入输入数据，以生成增强图像。最后，将增强图像输入生成器，从而获得在真实图像中穿上服装的试穿图像。

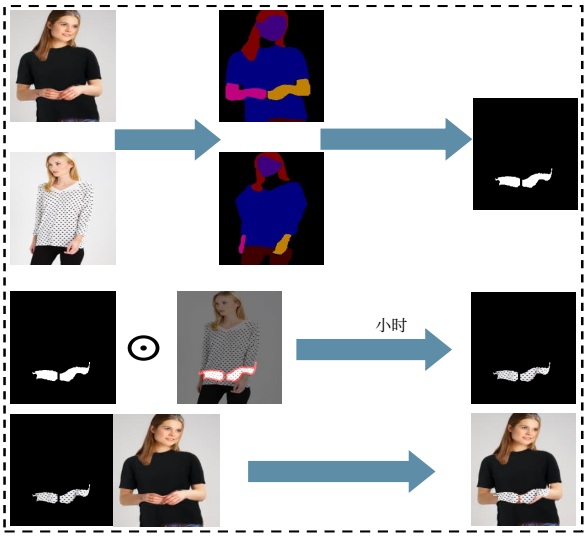


图 9：遮挡混合的过程。我们尝试对试穿图像中的遮挡进行建模，但很难将纹理遮挡限制在典型区域内，以模拟真实遮挡。为应对这一挑战，我们在 OccluMix 中使用锐化解析网络（SPN）。

C. 生成模块

在此模块中，我们采用 Res-UNet [3] 作为生成模块（GM）的骨干架构。它不仅能够保留变形服装的特征，还能保留人体各部位的细节。

在训练阶段，通过最小化 L_g来优化生成模块的参数，如下所示:

$$L_g = \alpha_l L_{euc} + \alpha_p L_{per}, \tag{9}$$

其中， L_{euc} 是逐像素 L1 损失， L_{per} 是感知损失 [46] ，用于提升试穿图像的视觉质量。其公式如下:

$$L_{euc} = \|I^G - I\|_1, \tag{10}$$

$$L_{per} = \sum_m \|\phi_m(I^G) - \phi_m(I)\|_1, \tag{11}$$

其中， I^G 和 I 分别表示生成图像和真实图像。 ϕ_m 表示在 ImageNet [48]上预训练的 VGG-19 [47]网络中的第 m个特征图。

V. 讨论

A. 与 *Mixup* 及其衍生方法的区别。

Mixup 及其衍生方法在随机图像块内混合图像内容，却没有充分考虑图像块的空间信息和语义信息。根据表I，我们发现这些随机图像块会导致 FID 性能下降。

与传统的 Mixup 流程相比，OccluMix 在语义引导下对伪图像和其他参考图像进行剪切粘贴，克服了随机图像块混合的缺陷。

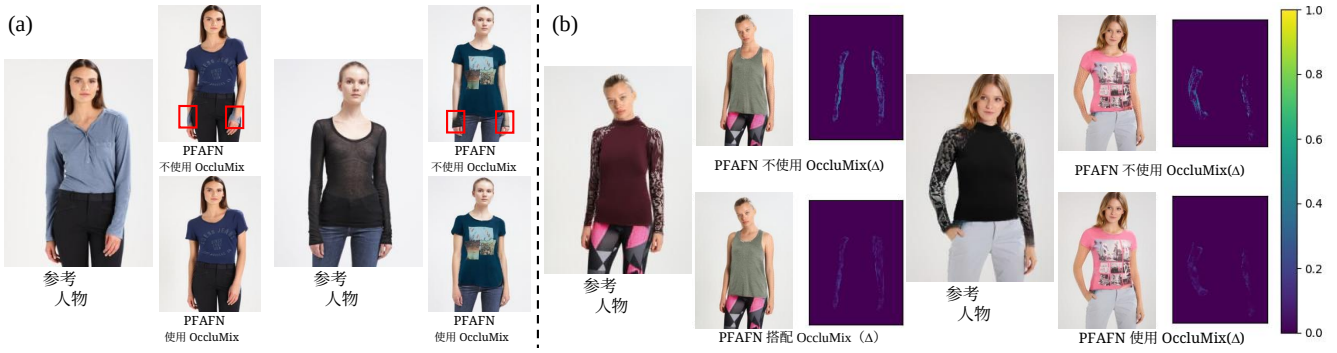


图10：OccluMix 在虚拟试穿推理过程中的定性效果。(a) OccluMix 能够成功生成干净的试穿图像，而基线生成的试穿图像仍带有前一件服装造成的遮挡。(b) △ 是试穿图像与真实值之间的绝对残差强度图。从对 △的比较可以看出，OccluMix 消除了服装纹理的残影。在(a)和(b)中，OccluMix 生成的试穿图像都具有更好的视觉连贯性。放大查看效果更佳。

表 I：与 Mixup 及其衍生方法的 FID 对比。我们报告在 VITON [7]数据集上训练的基线模型（PF-AFN [1] ）的性能。 δ 表示有无数据增强时的性能差距。

方法	FID ↓(δ)
PF-AFN	10.09 (+0.00)
Cutout [49]	10.36 (-0.27)
CutMix [11]	10.31 (-0.22)
Mixup [12]	10.17 (-0.08)
CutBlur [35]	9.94 (+0.15)
OccluMix	9.66 (+0.43)

B. 模型通过遮挡混合学到了什么？

与其他防止模型过度自信地进行预测的域适应方法类似，OccluMix 能够避免模型无法区分简单图像和失真图像，并帮助其减轻遮挡效应。这一点可通过图10加以说明。需要注意的是，OccluMix 的优势不仅体现在试穿图像的视觉连贯性上，也体现在残差强度图的减弱上。我们推测，这种提升源于伪图像分布更加均衡。现在，模型已经学会区分简单数据和失真数据，从而能够学习应当“在哪里”以及“如何”处理遮挡影响。

VI. 实验

在本节中，我们首先将 DOC-VTON 与最先进方法进行比较。然后，我们进行详细的消融研究，分析各个组成部分。最后，我们提供大量遮挡样本，以验证固有遮挡与获得性遮挡，并进一步展示遮挡对人类视觉系统的负面影响。此外，为了进一步验证所提出的 DOC-VTON 的性能，我们对遮挡图像（由 PF-AFN 生成）和干净图像（由我们的 DOC-VTON 生成）进行视觉比较。

A. 数据集

实验在 VITON数据集上进行 [1] 该数据集也用于 CP-VTON [9], CP-VTON+ [8]、ACGPN [2] 和 PF-AFN [7]。VITON 包含 14221 对图像的训练集和 2032 对图像的测试集，每对图像均由一张目标服装图像和一张女性照片组成，分辨率为 256 × 192。VITON-HD [22] 包含 11,647 组训练集样本和 2032 组测试集样本，每张图像的分辨率为 1024 × 768。所有评估和可视化均在测试集上进行。

B. 实现细节

1) 网络架构： DOC-VTON由CWM、OccluMix和生成模块组成。CWM的结构由双金字塔特征提取网络（PFEN） [50] 和渐进式外观流估计网络（AFEN） [7]组成。OccluMix中锐化解析网络的生成器采用与U-Net [51]相同的结构。生成模块的结构为Res-UNet [3]。在我们的实验中，VITON的图像分辨率为256× 192 ，VITON-HD数据集的图像分辨率为 1024 × 768 。

2) 训练： 我们训练锐化解析生成器（G）和复原器（R）200个周期。由于手臂与服装之间存在较大的空间变换，我们采用困难样本挖掘策略。在前100个训练周期中，我们有选择地使用具有挑战性的身体部位（例如手臂）训练模型。在最后100个周期中，我们使用整个身体部位（包括头部、手臂和裤子）共同训练模型。随后，我们使用OccluMix为生成模块生成样本。它们均采用0.0005的初始学习率，并使用Adam优化器优化网络，超参数为 $\beta_1 = 0.5$和$\beta_2 = 0.999$。

3）测试：测试过程与训练过程相同。将参考人物图像、目标服装、人体解析结果和人体姿态估计作为 DOC-VTON 的输入，以生成试穿图像。

C. 定性结果

为进一步验证所提出的 DOC-VTON 框架的性能，我们将其与其他方法进行视觉比较



图11：VITON数据集上的视觉比较。与四种最先进的试穿方法相比 [1], [2], [7], [9], 我们的方法生成了更加逼真的试穿图像。**借助提出的去遮挡策略，我们的方法不仅能处理变形服装中的不合理部分，还能清晰地避免原服装残影。**



图12：VITON-HD 数据集上的大量视觉比较。从多个角度来看，我们的方法都能生成高质量图像。

我们提出的方法与 CP-VTON [9], ClothFlow [3],ACGPN [2] 和 PF-AFN [7]。

1）VITON 上的结果：如图11的第一行和第二行所示，当人物摆出复杂姿势时（例如站立并抬起一只手臂至脸部附近），手臂会遮挡目标服装。在这种情况下，基线模型都无法处理变形过程，导致手臂图像变形或袖子破损。由于 TPS 的自由度有限 [23], 或外观流存在错位 [3], 基线模型的变形方法无法复原严重的非刚性变形。

在最后一行中，基线方法生成的图像含有明显的瑕疵，例如 CP-VTON [9], 、ClothFlow [3] 和 ACGPN [2] 存在纹理杂乱、边界模糊和颜色混杂等问题。为此，由于它们严重依赖解析结果来驱动图像生成，因此容易受到分割错误的影响。尽管 PF-AFN [7] 无需使用解析结果即可驱动图像生成，但它在生成伪造身体部位方面并未表现得更好。此外，当目标服装与原始服装之间存在较大差异时（例如，人物穿着长袖服装，而目标服装是短袖），PF-AFN [7] 无法在长袖袖口处生成手臂，因为模型未能充分理解服装的形状。为此，最先进模型在处理某些特殊情况（例如长袖袖口、复杂的服装纹理）时鲁棒性较差。

相比之下，所提出的 DOC-VTON 能够实现逼真的虚拟试穿，同时处理变形过程以避免有害的遮挡，并保留目标服装和人体部位的细节。得益于 OccluMix，我们的模型能够稳健地从原始服装生成未标注的身体部位（例如手臂、手和手指）。DOC-VTON 的所有定性结果都清楚地验证了其相较于 CP-VTON、ClothFlow、ACGPN 和 PF-AFN 的优越性。

*2）VITON-HD*上的结果：为验证我们方法的泛化能力，我们展示了在VITON-HD [22]数据集上的结果。由于更高分辨率下的伪影更容易被观察到，因此生成高度逼真的试穿结果更具挑战性。如图12所示，结果表明，我们的DOC-VTON能够在VITON-HD数据集上生成高质量图像。

D. 定量评估

对于虚拟试穿，试穿图像由目标服装和参考人物图像生成。由于缺少试穿图像的真实值，我们无法使用需要与标签进行计算的逐点指标（例如，SSIM、PSNR和学习感知图像块相似度）。本文采用弗雷歇 Inception 距离（FID） [52] 来衡量试穿图像的多样性。FID分数越低，结果质量越高。此外，Inception 分数（IS） [53]将不会采用，因为Rosca等人 [54]指出，在ImageNet以外的数据集上训练的模型使用IS进行评估会产生误导性结果。

表 II列出了 CP-VTON 试穿结果的 FID 分数 [9], CP-VTON + [8]、ClothFlow [3],ACGPN [2],

表 II：VITON数据集上不同方法的 FID 分数。

方法	手臂区域	变形服装	试穿结果
	FID ↓		
CP-VTON	24.42	33.17	24.43
CP-VTON+	22.36	30.21	21.08
ACGPN	16.62	24.95	15.67
ClothFlow	18.42	22.50	14.43
DCTON	13.27	30.18	14.82
RT-VTON	12.70	20.59	11.66
PF-AFN	12.86	18.67	10.09
Ours	**11.14**	**18.18**	**9.54**

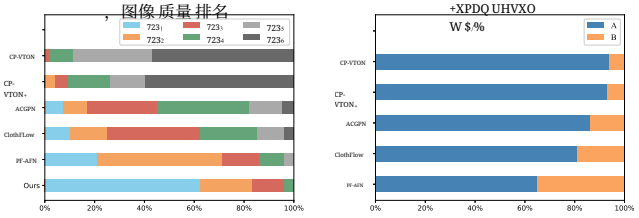


图13：我们招募了25名志愿者完成两项任务，对用户研究进行深入的质量评估。左图为各方法的质量排名。右图通过A/B测试比较DOC-VTON与其他方法。在右图中，标签A表示DOC-VTON被认为优于对比方法的比例，标签b表示对比方法被认为优于DOC-VTON的比例。

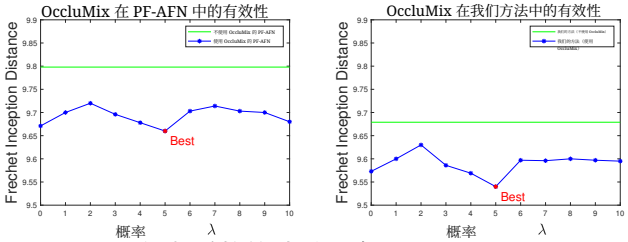


图14：复杂系数的消融研究 λ。

如图 13所示，我们的 DOC-VTON 在两项任务中都获得了最高的投票率。这验证了 DOC-VTON 相较于其他方法的显著优势。人类感知研究表明，所提出的方法能够有效去除有害遮挡，并在试穿任务中充分利用人类视觉系统。

DCTON [20],RT-VTON [21], PF-AFN [7]以及我们提出的 DOC-VTON 在 VITON 数据集上的表现。我们提出的 DOC-VTON 优于其他方法，这表明 DOC-VTON 能够提升试穿图像的感知质量，处理服装与人物之间的大幅错位，并合成逼真的试穿结果。

E. 人类感知研究

试穿图像上的细微变化（例如去除部分遮挡或残影）对人类视觉连贯性起着重要作用。由于 FID 对合成图像中的细微变化并不敏感，因此无法体现我们方法的有效性。我们还招募了 25 名志愿者，开展了两项用户研究。第一项任务旨在验证各个方法的优越性。具体而言，CP-VTON [9], ClothFlow [3], ACGPN [2], PF-AFN [7], 和 DOC-VTON 分别根据 300 张指定的参考图像生成试穿图像。我们要求每位志愿者对每组图像中的方法进行排名。图 13左侧图像清楚地展示了我们方法的优势。我们的方法排名第一，占比为 61.83 %。第二项任务是将我们的方法与其他方法逐一比较。我们将前述合成图像分成五组 a / b（即 a 为我们的方法，b 为其他方法）。我们要求每位志愿者选出视觉质量更好的图像。如图13右侧所示，我们的 DOC-VTON 始终比其他方法获得更高评价。

F. 消融实验

在本节中，我们分析 OccluMix 中促使模型具备鲁棒性的各个组件。

复杂系数 λ。

如上文所述，我们在训练阶段引入了 OccluMix方案。为验证 OccluMix 的有效性，我们设计了一种基线方法（即 ‘无 OccluMix’），该方法在原始模型上进行微调。如图 14所示，移除 OccluMix方案后，‘无 OccluMix’的 FID 降低了 0.08。为了进一步分析 OccluMix方案，我们将 λ 设置为 0 到 1，间隔为 0.1。对 λ 进行连续采样，表示从简单到复杂的混合范式。如图 14所示，我们发现，当使用最简单的混合模式时，模型表现出平凡的结果（$\lambda = 0.0$ 表示 DOC-VTON 仅使用简单纹理进行 OccluMix）。我们还对 $\lambda = 0.5$进行均衡分布，这表示 OccluMix 使用了两种比例均衡的纹理复杂度。如图 14 所示，均衡分布的性能最佳。由于不平衡的混合分布会导致性能明显下降，因此我们在 OccluMix 中采用平衡复杂系数。

OccluMix 的消融实验。

我们展示了在训练阶段使用 OccluMix方案的效果消融研究。同时，“不使用 OccluMix”的网络架构与“使用 OccluMix”相同，只是在训练过程中不采用 OccluMix

表III：DOC-VTON 的消融研究。FID 越低，结果越好。

方法	FID ↓
不使用 SPN	9.80
使用 SPN	9.68
不使用 OccluMix	9.80
使用 OccluMix	9.66
Ours	9.54

策略和 SPN。如图 10所示，我们注意到，使用 OccluMix方案训练的模型生成的试穿图像优于普通模型。借助 OccluMix，模型能够学会如何生成逼真的身体部位。如表 III所示，“使用 OccluMix”相比“不使用 OccluMix”，FID 提升了 0.14。

锐化解析网络。

我们展示了针对“锐化解析网络（SPN）”效果的消融研究。由于 SPN 的结果包含试穿图像中人体各身体部位的语义信息，因此，当目标服装在人体现出不合理变形（例如手臂、手部）时，我们可以利用这些信息引导变形过程。如图15所示，在 SPN 的引导下，变形服装能够克服复杂的空间交互，以及粗略变形服装与身体部位之间的失真。否则，变形服装的粗略形状会导致身体变形。如表 III 所示，与 “ 不使用 SPN” 相比，“ 使用 SPN” 的 FID 提升了 0.12。



图15：锐化解析网络效果的消融研究。

VII. 结论与局限性

本文介绍了遮挡混合（Occlu-Mix），这是一种新的数据增强方法和策略，旨在训练更强大的试穿模型。我们提出了一种通过数据增强实现的全新去遮挡方法，使模型能够去除部分遮挡，从而生成逼真的试穿图像。大量评估清楚地验证了其明显优势 DOC-VTON 在遮挡效应较小的情况下优于最先进方法。

尽管 DOC-VTON 解决了特定试穿数据集中的遮挡问题，但在分布外（OOD）图像上的表现仍然有限。因此，DOC-VTON 可能会受到光照、姿势和背景条件的限制。在后续工作中，我们将继续在 DOC-VTON 中应用无监督二维到三维变换，以开发 ODD 虚拟试穿框架。此外，由于我们的方法不仅可以去除遮挡，还可以暴露身体部位，因此我们已在开源代码中声明采用 BSD许可证，以避免这一潜在的社会影响。

参考文献

[1] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, “Viton：基于图像的虚拟试穿网络”，载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2018年，第7543–7552页。[2] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, and P. Luo, “通过自适应生成与保留图像内容实现逼真虚拟试穿”，载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2020年，第7850–7859页。[3] X. Han, X. Hu, W. Huang, and M. R. Scott, “Clothflow：用于着装人物生成的流模型”，载于*IEEE*国际计算机视觉会议（*ICCV*）论文集，2019年，第10 471–10 480页。[4] S. Choi, S. Park, M. Lee, and J. Choo, “Viton-hd：通过错位感知归一化实现高分辨率虚拟试穿”，载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2021年，第14 131–14 140页。[5] F. Yang and G. Lin, “Ct-net：用于任意几何变化服装迁移的互补迁移网络”，载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2021年，第9899–9908页。[6] C. Ge, Y. Song, Y. Ge, H. Yang, W. Liu, and P. Luo, “实现高度逼真虚拟试穿的解耦循环一致性”，载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2021年，第16 928–16 937页。[7] Y. Ge, Y. Song, R. Zhang, C. Ge, W. Liu, and P. Luo, “通过蒸馏外观流实现无需解析器的虚拟试穿”，载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2021年，第8485–8493页。[8] M. R. Minar, T. T. Tuan, H. Ahn, P. Rosin, and Y.-K. Lai, “Cp-vton+：保留服装形状和纹理的基于图像的虚拟试穿”，载于*IEEE/CVF*计算机视觉与模式识别会议研讨会（*CVPRW*）论文集，2020年。[9] B. Wang, H. Zheng, X. Liang, Y. Chen, and L. Lin, “迈向保留特征的基于图像的虚拟试穿网络”，载于欧洲计算机视觉会议（*ECCV*）论文集，2018年，第589–604页。[10] T. Issenhuth, J. Mary, and C. Calauzenes, “不需要掩码的内容就不要掩码：一种无需解析器的虚拟试穿方法”，载于欧洲计算机视觉会议（*ECCV*）论文集，2020年，第619–635页。[11] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, “Cutmix：利用可定位特征训练强分类器的正则化策略”，载于*IEEE* 国际计算机视觉会议（*ICCV*）论文集，2019年，第6023–6032页。[12] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup：超越经验风险最小化”，国际学习表征会议，2017年。[13] I. Santesteban, M. A. Otaduy, and D. Casas, “基于学习的虚拟试穿服装动画”，*Computer Graphics Forum*，第38卷，第2期，第355–366页，2019年。[14] D. Rohmer, T. Popa, M.-P. Cani, S. Hahmann, and A. Sheffer, “动画褶皱：通过增强粗略布料模拟生成逼真的褶皱”，*ACM* 图形学汇刊（*TOG*），第29卷，第6期，第1–8页，2010年。[15] I. Santesteban, M. A. Otaduy, and D. Casas, “基于学习的虚拟试穿服装动画”，*Computer Graphics Forum*，第38卷，第2期，第355–366页，2019年。

[16] R. Brouet, A. Sheffer, L. Boissieux, and M.-P. Cani, “保留设计的服装迁移”，ACM 图形学汇刊（TOG），第31卷，第4期，第1–11页，2012年。[17] H. Dong, X. Liang, X. Shen, B. Wang, H. Lai, J. Zhu, Z. Hu, and J. Yin, “迈向多姿态引导虚拟试穿网络”，载于 IEEE 国际计算机视觉会议（ICCV）论文集，2019年，第9026–9035页。[18] B. Hu, P. Liu, Z. Zheng, and M. Ren, “Spg-vton：用于多姿态虚拟试穿的语义预测引导”，IEEE 多媒体汇刊，第24卷，第1233–1246页，2022年。[19] C. Du, F. Yu, M. Jiang, A. Hua, X. Wei, T. Peng, and X. Hu, “Vton-scfa：基于语义约束和流对齐的虚拟试穿网络”，IEEE 多媒体汇刊，2022年。[20] C. Ge, Y. Song, Y. Ge, H. Yang, W. Liu, and P. Luo, “实现高度逼真虚拟试穿的解耦循环一致性”，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2021年，第16 928–16 937页。[21] H. Yang, X. Yu, and Z. Liu, “通过循环三层变换实现全范围虚拟试穿”，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2022年，第3460–3469页。[22] S. Choi, S. Park, M. Lee, and J. Choo, “Viton-hd：通过错位感知归一化实现高分辨率虚拟试穿”，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2021年，第14 131–14 140页。[23] J. Duchon, “索伯列夫空间中使旋转不变半范数最小化的样条”，载于《多变量函数构造理论》。Springer，1977年，第85–100页。[24] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros, “通过外观流进行视图合成”，载于欧洲计算机视觉会议（ECCV）论文集，2016年，第286–301页。[25] R. Yu, X. Wang, and X. Xie, “Vtnfp：兼顾人体与服装特征保留的基于图像的虚拟试穿网络”，载于 IEEE 国际计算机视觉会议（ICCV）论文集，2019年，第10 511–10 520页。[26] X. Zhan, X. Pan, B. Dai, Z. Liu, D. Lin, and C. C. Loy, “自监督场景去遮挡”，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2020年，第3784–3792页。[27] J. Hur and S. Roth, “联合光流与遮挡估计的迭代残差细化”，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2019年，第5754–5763页。[28] L. Qi, L. Jiang, S. Liu, X. Shen, and J. Jia, “基于 KINS 数据集的模态无关实例分割”，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2019年，第3014–3023页。[29] N. Zhang, N. Liu, J. Han, K. Wan, and L. Shao, “通过深度级联引导学习实现人脸去遮挡”，IEEE 多媒体汇刊，2022年。[30] X. Dong, J. Shen, D. Yu, W. Wang, J. Liu, and H. Huang, “具备遮挡感知能力的实时目标跟踪”，IEEE 多媒体汇刊，第19卷，第4期，第763–771页，2016年。[31] F. Angelini, Z. Fu, Y. Long, L. Shao, and S. M. Naqvi, “基于二维姿态且能处理遮挡的实时人体动作识别”，IEEE 多媒体汇刊，第22卷，第6期，第1433–1446页，2020年。[32] Y.-T. Hu, H.-S. Chen, K. Hui, J.-B. Huang, and A. G. Schwing, “Sail-vos：语义模态无关实例级视频目标分割——合成数据集与基线方法”，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2019年，第3105–3115页。[33] H. Guo, Y. Mao, and R. Zhang, “Mixup 作为局部线性流形外正则化”，载于 AAAI 人工智能会议（AAAI）论文集，2019年，第3714–3722页。[34] M. Hong, J. Choi, and G. Kim, “Stylemix：分离内容与风格以增强数据增强”，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2021年，第14 862–14 870页。[35] J. Yoo, N. Ahn, and K.-A. Sohn, “重新思考图像超分辨率的数据增强：全面分析与新策略”，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2020年，第8375–8384页。[36] A. Dabouei, S. Soleymani, F. Taherkhani, and N. M. Nasrabadi, “Supermix：监督混合式数据增强”，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2021年，第13 794–13 803页。[37] Q. Wang、W. Min、Q. Han、Q. Liu、C. Zha、H. Zhao 和 Z. Wei，“用于车辆重新识别数据增强的跨域适应标签”，*IEEE* 多媒体汇刊，第 24 卷，第 1031–1041 页，2022年。[38] T. Kang、S. Park、S. Choi 和 J. Choo，“通过随机图像裁剪进行数据增强，实现高分辨率虚拟试穿（VITON-Crop）”，*arXiv* 预印本 *arXiv:2111.08270*，2021年。[39] R. Takahashi、T. Matsubara 和 K. Uehara，“通过随机图像裁剪和拼接增强深度卷积神经网络的数据”，*IEEE* 视频技术电路与系统汇刊，第 30 卷，第 9 期，第 2917–2931 页，2019年。[40] Y. Tokozume、Y. Ushiku 和 T. Harada，“用于图像分类的类间学习”，载于*IEEE/CVF* 计算机视觉与模式识别会议（*CVPR*）论文集，2018年，第 5486–5494 页。[41] P. Li、Y. Xu、Y. Wei 和 Y. Yang，“人体解析的自我校正”，*IEEE*模式分析与机器智能汇刊，2020年。[42] R. A. G¨uler、N. Neverova 和 I. Kokkinos，“DensePose：野外密集人体姿态估计”，载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2018年，第 7297–7306 页。[43] D. Sun、S. Roth 和 M. J. Black，“当前光流估计实践及其背后原理的定量分析”，国际计算机视觉期刊，第 106 卷，137，2014年 2，第 115 页。[44] G. Liu、F. A. Reda、K. J. Shih、T.-C. Wang、A. Tao 和 B. Catanzaro，“使用部分卷积修复不规则孔洞”，载于欧洲计算机视觉会议（*ECCV*）论文集，2018年，第 85–100 页。[45] M. Partio、B. Cramariuc、M. Gabbouj 和 A. Visa，“使用灰度共生矩阵进行岩石纹理检索”，载于第 *5* 届北欧信号处理研讨会论文集，第 75 卷，2002年。[46] J. Johnson、A. Alahi 和 L. Fei-Fei，“用于实时风格迁移和超分辨率的感知损失”，载于欧洲计算机视觉会议（*ECCV*）论文集，2016年，第 694–711 页。[47] K. Simonyan 和 A. Zisserman，“用于大规模图像识别的超深卷积网络”，*arXiv* 预印本 *arXiv:1409.1556*，2014年。[48] A. Krizhevsky、I. Sutskever 和 G. E. Hinton，“使用深度卷积神经网络进行 ImageNet 分类”，载于神经信息处理系统进展，第 25 卷，2012 年。[49] T. DeVries 和 G. W. Taylor，“ 使用 Cutout 改进卷积神经网络的正则化”，*arXiv* 预印本 *arXiv:1708.04552*，2017年。[50] T.-Y. Lin、P. Doll´ar、R. Girshick、K. He、B. Hariharan 和 S. Belongie，“用于目标检测的特征金字塔网络”，载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2017年，第 2117–2125 页。[51] O. Ronneberger、P. Fischer 和 T. Brox，“U-Net：用于生物医学图像分割的卷积网络”，载于医学图像计算与计算机辅助干预国际会议（*MIC-CAI*），2015年，第 234–241 页。[52] M. Heusel、H. Ramsauer、T. Unterthiner、B. Nessler 和 S. Hochreiter，“通过双时间尺度更新规则训练的生成对抗网络收敛于局部纳什均衡”，载于神经信息处理系统进展，第 30 卷，2017年。[53] T. Salimans、I. Goodfellow、W. Zaremba、V. Cheung、A. Radford 和 X. Chen，“改进生成对抗网络训练的技术”，载于神经信息处理系统进展，第 29 卷，2016年。[54] M. Rosca、B. Lakshminarayanan、D. Warde-Farley 和 S. Mohamed，“生成对抗网络自动编码的变分方法”，*arXiv* 预印本 *arXiv:1706.04987*，2017年。