语义多模态图像合成

Zhen Zhu[1] , Zhiliang Xu[1*] , Ansheng You[2], Xiang Bai[1†]
[1]Huazhong University of Science and Technology, [2]Peking University
fzzhu, zhiliangxu1, xbaig@hust.edu.cn, youansheng@pku.edu.cn

摘 要

在本文中，我们专注于语义多模态图像合成（SMIS）任务，即在语义级别生成多模态图像。先前的工作试图使用多个特定类别的生成器，以限制其在少数类的数据集中的使用。相反，我们提出了一种新颖的组减少网络（GroupDNet），它利用生成器中的组卷积并逐步减少解码器中卷积的组数。因此，GroupDNet 在将语义标签转换为自然图像方面具有更大的可控性，并且对于具有多个类的数据集具有合理的高质量效益。在几个具有挑战性的数据集上进行的实验证明了 GroupDNet 在执行 SMIS 任务方面的优越性。我们还表明 GroupDNet 能够执行各种有趣的合成应用程序。代码和模型被分享在以下地址：

https://github.com/Seanseattle/SMIS。

1. 介绍

语义图像合成，即将语义标签转换为自然图像，具有许多实际应用，并引起了社区的广泛关注。它本质上是一对多映射问题，即无数可能的自然图像对应于一个语义标签。先前的工作针对该任务采用了不同的策略：采用变分自动编码器[38、58、2、13]，在训练时引入噪声[21]，构建多个子网络[12]和包括实例级特征编码[43]等。虽然这些方法在改善图像质量和扩展更多应用方面取得了非凡的成就，但我们还是采取了进一步的措施，特别专注于特定的多模态图像合成任务，该任务增加了更大的灵活性来控制生成的结果。

想象一下从人类构建草图创建内容的场景。借助语义到图像的翻译模型，可以将草图转换为真实的

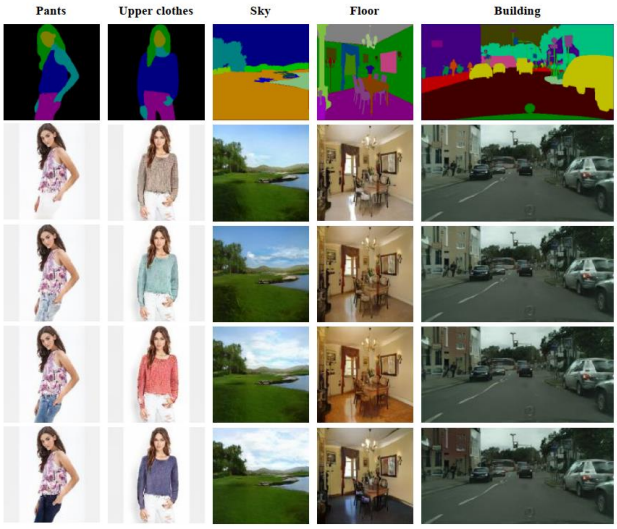


图 1：演示语义多模态图像合成（SMIS）任务。图像上方每列的文本表示整个列中正在变化的语义区域。第一行代表输入标签，其余行是通过我们的方法生成的图像。

人的形象。总体上看起来不错，但是上衣不适合您的口味怎么办？这就是随之而来的问题-这些模型不支持多模态合成，或者当这些模特更换上衣时，其他部分也会相应变化。这些都不符合您的期望。总而言之，我们可以将这种用户可控制的内容创建场景理解为执行一项任务，该任务在语义级别上产生多模态结果，而其他语义部分则保持不变。我们将此任务总结为：语义多模态图像合成（SMIS）。如图 1 所示，对于每种语义，我们都有其特定的控制器。通过调整特定类别的控制器，仅相应的区域会发生相应地更改。

该任务的直观解决方案是为不同的语义构建不同的生成网络，然后通过融合不同网络的输出来生成最终图像。它与[12]的总体方案非常相似，后者专注于肖像编辑。但是，在越来越多的类别下，这类方法很快会遇到性能下降，且训练时间线性增加和计算资源消耗大的问题。

为了使网络更加优雅，我们将所有生成过程统一为

* 同等的贡献
† 通讯作者

一种模型，采用组卷积创造性地替换生成器中的所有常规卷积。当卷积的组数等于类的组数时，我们的策略在数学和函数上均等效于[12]。但是，我们采用的另一策略则以不同的方式设置路径-在前向传播过程中，我们减少了解码器卷积中的组数。我们观察到不同的类别彼此之间具有内部相关性，例如，草和树叶的颜色应该高度相似。在这种情况下，逐步地合并组将为模型提供足够的能力，以建立不同类别之间的相互关系，从而提高整体图像质量。此外，当数据集的类别数量很大时（例如 ADE20K [55]），该策略还可以大大缓解计算消耗问题。我们将配备了这两种策略的生成器称为组减少网络（GroupDNet）。为了评估 GroupDNet 在 SMIS 任务上的绩效，我们提出了两个新的指标，分别称为平均特定类多样性（mCSD）和平均其他类别多样性（mOCD）。当某些语义部分发生巨大变化而其他部分保持不变时，mCSD 指标往往会保持较高的值，而 mOCD 往往会较低。

我们在几个具有挑战性的数据集上进行了实验：DeepFashion [32]，Cityscapes [5]和 ADE20K [55]。结果表明，我们的 GroupDNet 在生成过程中引入了更多的可控制性，从而生成了语义上的多模态图像。此外，GroupDNet 在图像质量方面一直保持与以前最先进的方法的竞争力，展现了 GroupDNet 的优越性。此外，GroupDNet 在生成过程中引入了很多可控性，并且具有各种有趣的应用程序，例如外观混合，语义操纵和样式变形。

2. 相关工作

生成模型。 生成对抗网络（GAN）[11]由生成器和判别器组成，即使对于非常具有挑战性的数据集，也具有惊人的生成清晰图像的能力[11、23、3、24]。变分自编码器（VAE）[26]包含一个编码器和一个解码器，并要求编码器的潜码产生量符合高斯分布。其合成结果通常表现出很大的多样性。一些方法[2]在其模型中结合了 VAE 和 GAN，可生成逼真而多样的图像。

条件图像合成。 条件生成对抗网络[36]激发了广泛的条件图像合成应用，例如图像到图像的转换[21，19，43，57，29，31，17]，超分辨率[28， 22]，域适应[16、54]，单一模型图像合成[56、39、35、40]，样式转换[18、7、22]，人物图像生成[33，59，13]和从文本进行图像合成[50，51]等。我们专注于将条件语义标签转换为自然图像，同时在语义级别上为该任务增加更多多样性和可控性。

多模态标签到图像合成。 在多模态标签到图像合成任务中，已有许多工作[1、30、42]。陈等[4]避免使用 GAN 同时利用级联细化网络生成高分辨率图像。王等[43]在编码器的输出中添加了附加的实例级特征通道，从而可以对生成的结果进行对象级控制。王等[42]使用图像的另一个来源作为时尚的例子来指导生成过程。Park 等[38]将 VAE 纳入其网络，该网络使生成器能够产生多模态图像。李等[30]采用隐式最大似然估计框架来缓解 GAN 的模式崩溃问题，从而鼓励多样化的产出。Bansal 等[1]使用经典的工具以指数方式将画廊的形状，环境和零件与语义标签输入进行匹配，从而产生不同的结果。与这些作品不同，我们专注于语义多模态图像合成，它要求在语义级别而不是全局级别具有细粒度的可控制性。顾等[12]为每个面部组件构建了几个自动编码器，以提取不同的组件表示，然后将它们合并到人像编辑任务中的下一个前景生成器中。我们的工作与这项工作高度相关，因为这两种方法都通过处理具有不同参数的不同类来应对 SMIS 任务。但是，我们在解码器中逐渐减少组号的独特设计使我们的网络能够处理可能无法使用其方法的许多类的数据集。

组卷积。先前的工作[27、46、53、34、45]指出组卷积对于降低计算复杂度和模型参数是有利的，因此它们已被广泛用于轻量级网络。马等[34]提到过度使用组卷积会导致较大的内存访问成本（MAC）。尽管在网络中使用具有小的组卷积甚至没有组卷积也是理想的，但我们在实验中表明，完全避免解码器中的组卷积对于 SMIS 任务的性能存在问题。此外，我们的减少组数策略充分缓解了巨大的 MAC 问题，使其适用于实际应用。

3. 语义多模态图像合成

3.1. 问题定义

令 M 表示语义分割掩码。假设数据集中有 C 个语义类。H 和 W 分别代表图像的高度和宽度。作为一个非常

直接的标签到图像的转换的方式，即生成器 G 需要 M 作为条件输入来生成图像。但是，为了支持多模态生成，我们需要另一个输入源来控制生成多样性。通常，受 VAE 的启发，我们采用编码器提取潜在编码 Z 作为控制器[26]。接收到这两个输入后，可以通过 O = G（Z，M）来产生图像输出 O。但是，在语义多模态图像合成（SMIS）任务中，我们的目标是通过干扰特定类的潜在编码来产生语义上多样化的图像，该类潜在编码独立地控制了其相应类的多样性。

3.2. 挑战

对于 SMIS 任务，关键是将潜在编码划分为一系列特定于类别的潜在编码，每个潜在编码仅控制特定语义类的生成。传统的卷积编码器不是最佳选择，因为所有类的特征表示都在潜在编码内部纠缠在一起。即使我们有特定于类的潜在编码，在如何使用编码方面仍然存在问题。正如我们将在实验部分中说明的那样，用特定于类的代码简单地替换 SPADE [38]中的原始潜在编码具有有限的处理 SMIS 任务的能力。这种现象激励我们，我们需要在编码器和解码器中都进行一些体系架构修改，以更有效地完成任务。

3.3. GroupDNet

基于上面的分析，我们现在提供有关此任务的解决方案-组减少网络（GroupDNet）的更多详细信息。考虑到其在标签到图像生成任务中的卓越性能，GroupDNet 的主要体系架构从 SPADE [38]中汲取了设计灵感。 GroupDNet 的主要修改是将典型的卷积替换为组卷积[27]，以实现特定于类别的可控性。在下面，我们将首先简要概述我们的网络体系架构，然后描述我们对网络的不同组件所做的修改。

概述。从图 2 可以看出，GroupDNet 包含一个编码器和一个解码器。受到 VAE [26]和 SPADE [38]的启发，编码器 E 产生潜码 Z，该潜码 Z 在训练过程中应遵循高斯分布 N（0，1）。在测试期间，编码器 E 被丢弃。来自高斯分布的随机采样编码替代了 Z。为了实现这一点，我们使用重参数化技巧[26]在训练过程中启用了可微分的损失函数。具体而言，编码器通过两个完全连接层预测平均矢量和方差矢量，以表示编码后的分布。编码分布与高斯分布之间的差距可以通过施加 KL 散度损失来最小化：

$$\mathcal{L}_{\mathrm{KL}} = \mathcal{D}_{\mathrm{KL}}(E(I)||\mathcal{N}(0,1)), \tag{1}$$

其中$\mathcal{D}_{\mathrm{KL}}$代表 KL 散度。

编码器。 令M_c表示 c 类的二进制掩码，$X \in \mathbb{R}^{H\times W}$为输入图像。 通过将 X 拆分为不同语义类别的不同图像，我们有：

$$X_c = M_c \cdot X. \tag{2}$$

此操作减少了对 E 进行特征解缠的依赖性，从而节省了更多函数来对特征进行精确编码。编码器的输入是这些图像的串联：$S = \underset{c}{\mathrm{cat}} X_c$。 E 中的所有卷积具有相同的组数，即类的总数 C。从输入端和体系架构侧，我们将不同的类解耦以彼此独立。结果，被编码的潜码 Z 由所有特定于类的潜码Z_c（Z 的离散部分）组成。Z_c在即将到来的解码阶段中充当 c 类的控制器。与产生两个向量作为高斯分布的均值和方差预测的一般方案不同，我们的编码器通过卷积层产生均值图和方差图，以将结构信息大量保留在潜码 Z 中。

解码器。 接收到潜在编码 Z 后，解码器会在语义标签的引导下将其转换为自然图像。问题是如何利用语义标签正确地指导解码阶段。有几种方法可以实现此目的，例如将语义标签连接到输入或在解码器的每个阶段进行调节。前一种不适合我们的情况，因为解码器输入的空间大小非常有限，这将严重丢失语义标签的许多结构信息。我们选择后者，并选择一种典型的高级模型-SPADE 生成器[38]作为我们网络的基础。如[38]所述，SPADE 是某些条件归一化层的更通用形式[6，18]，并且在语义图像合成中显示出产生像素级导航的卓越能力。遵循在生成器中使用所有组卷积的一般思想，我们将 SPADE 模块中的卷积层替换为组卷积，并将此新的条件模块称为条件组归一化（CG-Norm），如图 2 所示。然后，通过动态合并 CG 范数和组卷积来组成称为条件组块（CG-Block）的网络块。CG-Block 的体系架构也显示在图 2 中。

同样，令$\mathbf{F}^i \in \mathbb{R}^{H^i\times W^i}$表示解码器网络的第 i 层的特征图，$\mathcal{G}^i$表示第 i 层的组数。 而且，N，$D^i$，$H^i$

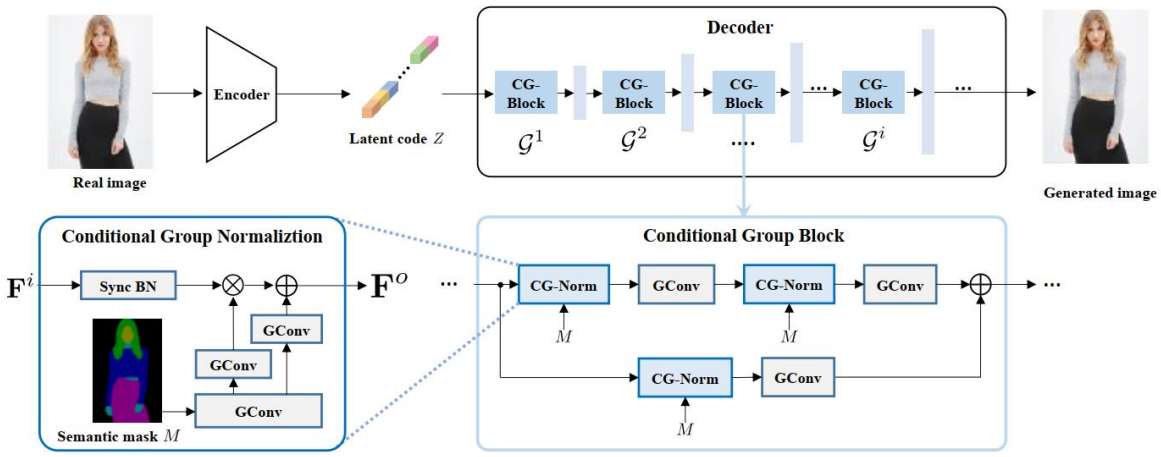


图 2：我们的生成器（GroupDNet）的体系架构。“GConv”表示组卷积，“Sync BN”表示同步批归一化。$\mathcal{G}^i$是第 i 层的组数。注意对于 GroupDNet 的 i ≥ 1，通常$\mathcal{G}^i \geq \mathcal{G}^{i+1}$。

和W^i分别是批处理大小，通道数，特征图的高度和宽度。如图 2 所示，CG-Norm 内部的组卷积会将语义标签输入转换为像素级调制参数$\gamma \in \mathbb{R}^{D^i \times H^i \times W^i}$和$\beta \in \mathbb{R}^{D^i \times H^i \times W^i}$。特征输入$\mathbf{F}^i$首先将通过批量归一化层[20]来归一化$\mathbf{F}^i$：

$$\mathrm{BN}(\mathbf{F}^i) = \gamma_{\mathrm{BN}} \left(\frac{\mathbf{F}^i - \mu(\mathbf{F}^i)}{\sigma(\mathbf{F}^i)} \right) + \beta_{\mathrm{BN}}, \tag{3}$$

在这里$\gamma_{\mathrm{BN}}, \beta_{\mathrm{BN}} \in \mathbb{R}^D$是从数据中学到的仿射参数。$\mu_d$和$\sigma_d$是通过每个特征通道的批处理大小和空间尺寸计算的：

$$\begin{aligned} \mu_d(\mathbf{F}^i) &= \frac{1}{NH^iW^i} \sum_{n=1}^{N} \sum_{h=1}^{H^i} \sum_{w=1}^{W^i} \mathbf{F}^i_{ndhw} \\ \sigma_d(\mathbf{F}^i) &= \sqrt{\frac{1}{NH^iW^i} \sum_{n=1}^{N} \sum_{h=1}^{H^i} \sum_{w=1}^{W^i} (\mathbf{F}^i_{ndhw})^2 - (\mu_d(\mathbf{F}^i))^2} \end{aligned} \tag{4}$$

之后，输出$\mathrm{BN}(\mathbf{F}^i)$与先前预测的像素级 γ和 β进行交互，并产生一个插入了语义信息的新特征图$\mathbf{F}^o$。将公式 3 考虑进来，

$$\begin{aligned} \mathbf{F}^o &= \gamma \cdot \mathrm{BN}(\mathbf{F}^i) + \beta \\ &= \gamma \cdot \gamma_{\mathrm{BN}} \left(\frac{\mathbf{F}^i - \mu(\mathbf{F}^i)}{\sigma(\mathbf{F}^i)} \right) + (\gamma \cdot \beta_{\mathrm{BN}} + \beta). \end{aligned} \tag{5}$$

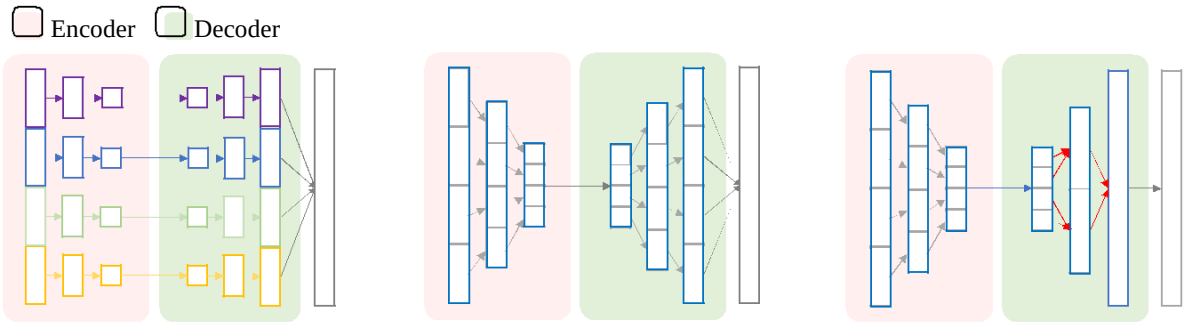
当 i 变大时，组号最终减少为 1。在进行常规卷积后，特征映射到三通道 RGB 图像 O。

3.4. 其他解决方案

除了 GroupDNet 之外，执行 SMIS 任务的简单解决方案是构建一组编码器和解码器，每个重点放在特定的语义类上，如图 3（a）所示。基本思想是独立对待每个类别，然后融合不同子网的结果。为简单起见，我们将此类网络称为“多个网络（MulNet）”。具有类似想法的另一种替代方法是在整个网络中使用组卷积[27]。如图 3（b）所示，用组卷积[27]替换编码器和解码器中的所有卷积，并将组号设置为等于类号，这就是组网络（GroupNet）。如果每个组中的通道号等于 MulNet 单个网络中相应层的通道号，则在理论上与 MulNet 等效。图 3（c）说明了我们的 GroupDNet。GroupDNet 和 GroupNet 之间的主要区别是解码器中组的数量单调减少。尽管此修改看起来很简单，但它带来了一些明显的好处，主要是在以下三个方面：

类平衡。值得注意的是，不同的类具有不同数量的实例[32、5、55]，并且需要不同的网络容量来对这些类进行建模。MulNet 和 GroupNet 很难找到合适的网络设计来平衡所有类别。更重要的是，并非所有的类都出现在一张图像中。在这种情况下，MulNet 和 GroupNet 不可避免地会浪费大量的计算资源，因为它们必须在训练或测试期间激活所有类别的所有子网或子组。但是，在 GroupDNet 中，不平衡类与其相邻类共享参数，从而极大地减轻了类不平衡问题。

类相关。在自然世界中，语义类别通常与其他类别具有关联，例如，草的颜色和树叶的颜色相似，并且建筑物影响附近道路的日照等。为产生合理的结果，MulNet 和 GroupNet 都有一个融合模块（我们中的几个常规卷积



(a) MulNet (b) GroupNet (c) GroupDNet

图 3：MulNet（a），GroupNet（b）和 GroupDNet（c）的示意图。 请注意，MulNet 和 GroupNet 的最后一层是融合模块，该模块由几个正常的卷积层组成，以融合不同类的结果。

情况）在解码器的末尾将不同类的特征合并到一个图像输出中。通常，融合模块大致考虑不同类别的相关性。但是，我们认为这是不够的，因为不同类的相关性太复杂，以至于无法通过使用具有受限制的感受野的简单组件来全面探索。一种替代方法是使用一些网络模块（如自注意力模块）来捕获图像的远程依赖性，但其过高的计算量会阻碍其在此类情况下的使用[49]。但是，GroupDNet 将这些关系贯穿整个解码器。因此，它可以更准确，更彻底地利用相关性。结果，与其他两种方法生成的图像相比，GroupDNet 生成的图像更好，更真实。

GPU 内存。 为了确保 MulNet 的每个网络或 GroupNet 中每个类的分组参数具有足够的容量，总通道数将随着类数的增加而显著增加。达到极限后，图形卡的最大 GPU 内存将不再能够容纳一个样本。正如我们在 ADE20K 数据集上粗略估计的那样[55]，即使将批大小设置为 1，一张 Tesla V100 图形卡也无法容纳足够容量的模型。但是，由于不同类共享参数，因此在 GroupDNet 中问题不那么严重。不必为每个类别设置如此多的通道。

3.5. 损失函数

我们采用与 SPADE [38]相同的损失函数：

$$\mathcal{L}_{\text{full}} = \arg\min_{G}\max_{D} \mathcal{L}_{\text{GAN}} + \lambda_1 \mathcal{L}_{\text{FM}} + \lambda_2 \mathcal{L}_{\text{P}} + \lambda_3 \mathcal{L}_{\text{KL}}. \tag{6}$$

$\mathcal{L}_{\text{GAN}}$是 GAN 损失的铰链版本，$\mathcal{L}_{\text{FM}}$是真实和合成图像之间的特征匹配损失。具体来说，我们使用多层判别器从真实和合成图像中提取特征，然后计算这些成对特征之间的 L1 距离。 同样，$\mathcal{L}_{\text{P}}$是为样式转换建议的感知损失[22]。使用预训练的 VGG 网络[41]以获得配对的中间特征图，然后计算这些配对图之间的 L1 距离。$\mathcal{L}_{\text{KL}}$是 KL 散度损失项，如等式 1。我们设置 λ1 = 10; λ2 = 10; λ3 = 0.05，与 SPADE [38]相同。

4. 实验

4.1. 实现细节

我们将谱归一化[37]应用于生成器和判别器中的所有层。遵循两个时间尺度更新规则（TTUR）[14]，将生成器和判别器的学习率分别设置为 0.0001 和 0.0004。我们使用 Adam 优化器[25]并设置 β1 = 0; β2 = 0.9。 所有实验均在至少 4 个 P40 GPU 上进行。 此外，我们使用同步批处理归一化来跨多个 GPU 同步均值和方差统计信息。补充材料中提供了更多详细信息，例如详细的网络设计和更多的超参数。

4.2. 数据集

我们对三个非常具有挑战性的数据集进行了实验，包括 DeepFashion [32]，Cityscapes [5]和 ADE20K [55]。之所以选择 DeepFashion，是因为该数据集显示出所有语义类别之间的差异，这自然适合于评估模型进行多模态合成的能力。因此，我们在此数据集上比较了几个基线模型，以评估我们的模型在 SMIS 任务上的强大功能。 Cityscapes 中图像的尺寸非常大，因此测试模型在此数据集上生成高分辨率图像的能力是适当的。 ADE20K 因其种类繁多而极具挑战性，我们发现很难使用有限的 GPU 在 ADE20K 上训练 MulNet 和 GroupNet。 更多细节可以在补充材料中找到。

4.3. 指标

平均 SMIS 多样性。 为了评估为 SMIS 任务设计的模型的性能，我们介绍

两个新指标名为：特定于类别的多样性（mCSD）和平均其他类别的多样性（mOCD）。我们基于 LPIPS 指标[52]设计新的指标，该指标用于通过计算图像对深层特征之间的加权 L2 距离来评估模型的生成多样性。对于相同的语义标签输入，我们仅通过对语义类 c 的潜在编码Z_c进行调制，就可以为每个语义类生成 n 个图像。 因此，我们有一组图像 $S=\{I_1^1,...,I_1^n,...,I_{\mathcal{C}}^1,...,I_{\mathcal{C}}^n\}$。最后，mCSD 和 mOCD 的计算公式为：

$$\mathrm{mCSD}=\frac{1}{\mathcal{C}}\sum_{c=1}^{\mathcal{C}}L_c,\ \mathrm{mOCD}=\frac{1}{\mathcal{C}}\sum_{c=1}^{\mathcal{C}}L_{\neq c}. \qquad (7)$$

其中L_c是采样的 m 对之间的 c 类语义区域的平均 LPIPS 距离[52]，而$L_{\neq c}$表示相同对之间所有其他类别的区域中的平均 LPIPS 距离[52]。在我们的设置中，我们设置 n=100，m=19，根据[58，19]。ImageNet 经过预训练的 AlexNet [27]被用作深度特征提取器。SMIS 任务的更高性能要求特定语义区域的高度多样性（较高的 mCSD）以及所有其他区域的较低多样性（较低的 mOCD）。此外，我们还通过针对相同语义标签生成全局多样的结果来报告总体 LPIPS 距离。

人工评估指标。 我们进一步介绍了人工评估，以评估生成模型是否在 SMIS 任务中表现良好。我们招募了 20 位在生成任务方面具有研究经验的志愿者。我们向他们展示了一个输入蒙版以及仅从一个模型生成的两个图像。这两个图像是多模态结果，仅在一个随机语义类别的区域中有所不同。志愿者判断给定的两个图像是否仅在一个语义类别中有所不同。仅在一个语义类中被判断为在语义上不同的对的百分比表示对模型在 SMIS 任务上的性能的人工评估。我们将此度量缩写为 SHE（SMIS 人类评估）。在每个阶段，为志愿者提供 50 个无限制回答时间的问题。

Frechet´初始距离。 我们使用 Frechet 初始距离（FID）[15]计算合成结果的分布与实际图像的分布之间的距离。较低的 FID 通常暗示所生成图像的保真度更高。

细分效果。 如果生成的图像的预测标签看起来很逼真，则它们与原始图像的标签高度相似是合理的。因此，我们采用先前工作[4、43、38]中的评估协议来测量生成图像的分割精度。我们在不考虑可以忽略的类别的情况下，报告了平均交并比（mIoU）和像素精度（Acc）指标的结果。图片

图 4：GroupDNet 与其他基准模型之间的定性比较。 前两行表示通过更改其上衣潜在编码的不同模型的结果，而后两行则表示其在更改裤子潜在编码时的结果。请注意，对于那些没有特定于类的控制器（例如 VSPADE）的模型，我们会更改其整体潜在编码以生成不同的图像。

使用训练有素的分割模型 Uper-Net101[44]评估 ADE20K，使用 DRN-D-105 [48]评估 Cityscapes，使用现成的人体解析器 CIHP [9]评估 DeepFashion。

4.4. 结果

除了以下各节，我们还将在补充材料中更充分地说明模型设计的合理性，以供感兴趣的读者参考。

4.4.1 SMIS 上的比较

对可能为 SMIS 任务修改的模型的基本要求是，它们应具有进行多模态图像合成的能力。我们比较了几种支持多模态图像合成的方法，以证明 GroupDNet 的优越性：

- **变异 SPADE [38]（VSPADE）**具有将实际图像处理为均值和变异矢量的图像编码器，其中应用了 KL 散度损失来支持多模态图像合成。详细的描述可以在他们的论文中找到。
- **BicycleGAN [58]**将给定的图像输入映射为潜在编码，然后将其与标签输入组合以产生输出。由于潜在编码受 KL 散度损失的约束，因此可以用高斯分布中的随机样本代替它。
- **DSCGAN [47]**通过在生成器上引入显式正则化来补充 BicycleGAN，以减轻先前模型的模式坍塌问题。

除了第二节中介绍的 MulNet 和 GroupNet 之外。 在图 3.4 中，我们还通过替换 VSPADE 模型的编码器/解码器中的卷积，对卷积进行分组，分别将卷积分组为等于数据集类编号的组编号，分别表示为 GroupEnc / GroupDec。 注意 MulNet，GroupNet，GroupEnc，GroupDec

和 VSPADE 使用与 GroupDNet 相同类型的多尺度判别器[43]和训练设置进行训练。为了公平地比较性能，我们平衡了这些模型的参数数量，以减轻对使用更多参数带来性能改进的怀疑。对于 BicycleGAN 和 DSCGAN，我们采用其原始的训练和测试协议。

表中给出了定量和定性的结果。分别参见图 1 和图 11。定量结果证明了 GroupDNet 的整体优势。通常，GroupDNet 具有最佳的图像质量（最低的 FID）和整体的多样性（最高的 LPIPS）。就 SMIS 任务的性能而言，MulNet 和 GroupNet 比 GroupDNet 稍好，因为有证据表明它们具有更大的 mCSD 或更低的 mOCD。但是，MulNet 和 GroupNet 的图像质量不令人满意（FID 高），MulNet 的 FPS 比 GroupDNet 低得多。在 SHE 指标方面，GroupDNet 也比 MulNet 和 GroupNet 更具竞争力。尽管 VSPADE 具有相当大的 mCSD，但它的 mOCD 也非常大，表明它在 SMIS 任务上的表现不尽人意。在 BicycleGAN 和 DSCGAN 中也观察到相同的现象，其 FID 相对高于 VSPADE，这显示出 SPADE 体系架构的优势。从 VSPADE 和 GroupDec 的高 mOCD 值（它们的编码器由常规卷积组成），我们得出结论，组编码器是实现 SMIS 任务高性能的关键。但是，GroupDNet 的出色性能表明，与 GroupEnc 相比，解码器中的组递减修改也是有效的，并带来了进一步的性能提升。考虑到速度，视觉质量和 SMIS 任务的性能，GroupDNet 收集了这些信息，是一个很好的权衡模型。

根据定性结果，很明显，MulNet，GroupNet，GroupEnc 和 GroupDNet 能够生成语义上的多模态图像，而其他则不能。但是，MulNet，GroupNet，BicycleGAN 和 DSCGAN 的图像质量远不能令人满意，因为它们的图像在视觉上难以令人信以为真。GroupEnc 的图像质量更好，但在 SMIS 任务中却下降了。从图 11 的前两行可以看出，当将上装更改为另一种样式时，GroupEnc 也会略微改变牛仔裤短裤的颜色。

4.4.2 在标签到图像翻译上的比较

在本节中，我们主要通过与 FID，mIoU 和 Accuracy 度量标准上的一些标签到图像方法进行比较，来评估我们方法生成的图像质量。作为比较方法，我们选择了四个最新的方法：BicycleGAN [58]，DSCGAN [47]，pix2pixHD [43] 和 SPADE [38]。比较在

Models	FID↓	mCSD↑	mOCD↓	LPIPS↑	SHE↑	Speed↑	# Param↓
MulNet	12.07	0.0244	0.0019	0.202	79.2	6.3	105.1
GroupNet	12.58	0.0276	**0.0017**	0.203	**83.7**	8.2	97.7
Group Enc	10.83	0.0232	0.0065	0.217	69.3	19.6	105.5
Group Dec	9.84	0.0003	0.0257	0.206	26.4	12.1	111.3
VSPADE [38]	10.02	0.0304	0.1843	0.207	23.6	20.4	106.8
BicycleGAN [58]	40.07	**0.0316**	0.2147	**0.228**	24.8	66.9	**58.4**
DSCGAN [47]	38.40	0.0245	0.1560	0.163	27.6	**67.2**	**58.4**
GroupDNet	**9.50**	0.0264	0.0033	**0.228**	81.2	12.2	109.1

表 1：与基准模型的定量比较结果。 “SHE”是指对模型在 SMIS 任务上的绩效进行人工评估。 我们使用每秒帧数（FPS）表示模型的“Speed”。 “# Param”表示参数的数量，单位为“M”，表示百万。 对于 mCSD，越高越好。 对于 mOCD，越低越好。

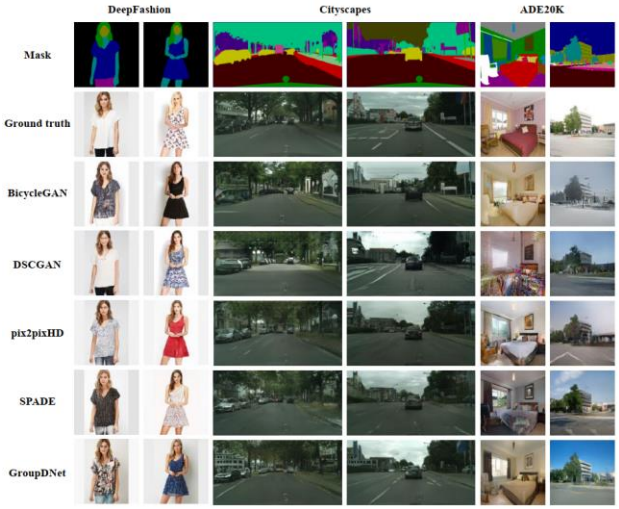


图 5：与 SOTA 标签到图像方法的定性比较。 图像从上至下分别代表 DeepFashion，Cityscapes 和 ADE20K 上的实验。

跨 DeepFashion，Cityscapes 和 ADE20K 数据集上执行。如果有的话，我们会评估从他们的官方 GitHub 存储库下载的已训练的模型的性能。对于原始论文中未包含的那些实验，我们遵循他们的代码，并使用与 GroupDNet 相似的设置运行实验。定量结果显示在表格 2 中。通常，由于我们的网络是基于 SPADE 构建的，因此在 DeepFashion 和 Cityscapes 数据集上，其性能几乎与 SPADE 相同。虽然在 ADE20K 数据集上，我们的方法不如 SPADE，但仍优于其他方法。这种现象一方面显示了 SPADE 体系架构的优越性，另一方面也暴露了 GroupDNet 仍在努力处理具有大量语义类的数据集。

图 5 显示了 DeepFashion，Cityscapes 和 ADE20K 的定性比较。通常，GroupDNet 生成的图像比其他图像更真实，更合理。这些视觉结果始终显示 GroupDNet 生成图像的高质量图像，从而证明了其在各种数据集上的有效性。

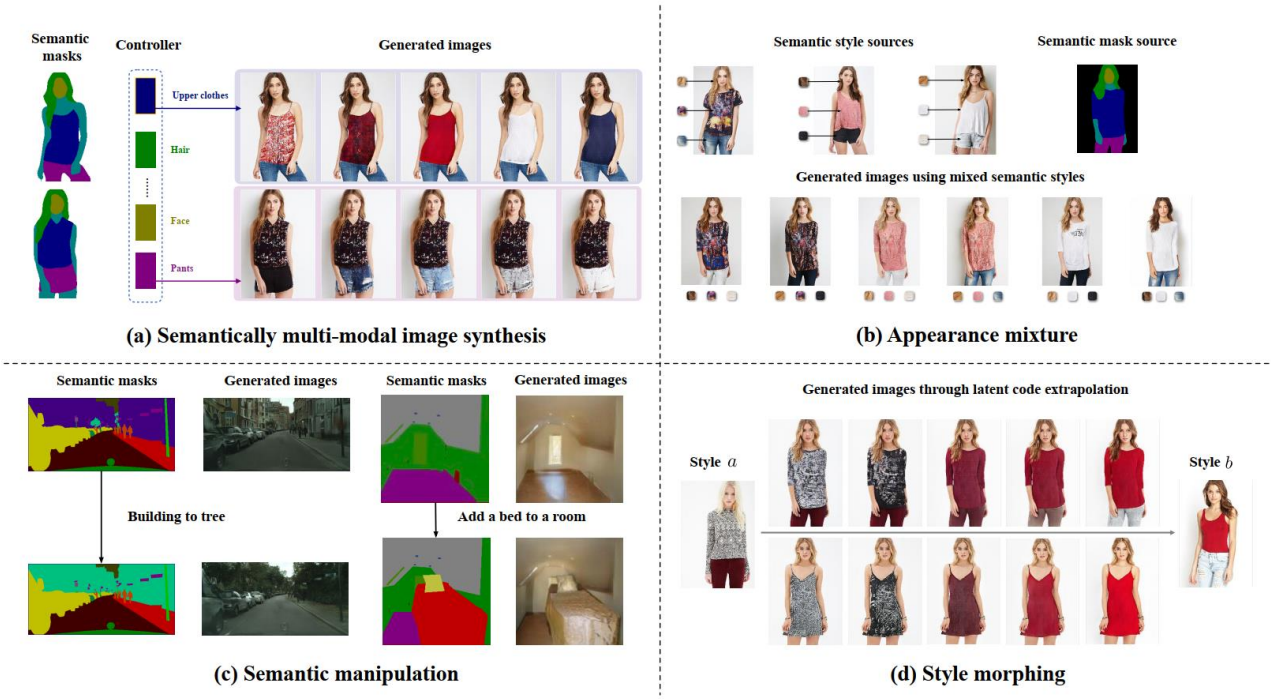


图 6：所提出方法的示例应用。（a）演示语义多模态图像合成（SMIS）任务。（b）我们的 SMIS 模型在外观混合中的应用。我们的模型从不同的来源提取不同语义类别的样式，并通过将这些语义样式与给定的语义蒙版组合来生成混合图像。（c）我们的 SMIS 模型在语义处理中的应用。（d）我们的 SMIS 模型在图像插值中的应用。 放大以查看更好的细节。

Method	DeepFashion			Cityscapes			ADE20K		
	mIoU↑	Acc↑	FID↓	mIoU↑	Acc↑	FID↓	mIoU↑	Acc↑	FID↓
BicycleGAN [58]	76.8	97.8	40.07	23.3	75.4	87.74	4.78	29.6	87.85
DSCGAN [47]	81.0	98.3	38.40	37.8	86.7	67.77	10.2	58.8	83.98
pix2pixHD [43]	85.2	98.8	17.76	58.3	92.5	78.24	27.6	75.7	55.9
SPADE [38]	87.1	**98.9**	10.02	**62.3**	93.5	58.10	**42.0**	**81.4**	**33.49**
GroupDNet	**87.3**	**98.9**	**9.50**	**62.3**	**93.7**	**49.81**	30.4	77.1	42.17

表 2：与标签到图像模型的定量比较。 pix2pixHD 和 SPADE 的数值是通过在我们的机器上运行评估而不是在其论文上收集的。

4.4.3 应用

由于 GroupDNet 为生成过程贡献了更多的用户可控性，因此，除了 SMIS 任务外，它还可以应用于许多激动人心的应用程序，具体说明如下。补充材料中提供了更多结果。

外观混合。 通过在推理过程中利用 GroupDNet 中的编码器，我们可以收集人的不同身体部位的独特风格。给定人体解析蒙版，这些样式的每种组合都会呈现出独特的人像。通过这种方式，我们可以在给定人物图像库的情况下创建数千个多样且逼真的人物图像。此应用程序在图 6（b）中进行了演示。可以在我们的代码库中找到演示视频。

语义操纵。 与大多数标签到图像方法[43、38、30]相似，我们的网络也支持语义操纵。如图 6（c）所示，我们可以在房间里插入一张床或用树木等代替建筑物。

样式变换。将两个真实图像馈送到编码器会生成这些图像的两个样式编码。通过在这两个编码之间进行渐变，我们可以生成一系列图像，这些图像从图像 a 到图像 b 逐渐变化，在图 6（d）中示出。

5. 结论和未来工作

在本文中，我们提出了一种用于语义多模态合成任务的新型网络，称为 GroupDNet。我们的网络非常规地采用所有都是组卷积，并修改卷积的组数以在解码器中减少，从而大大提高了与其他可能的解决方案（例如，多个生成器）相比的训练效率。

尽管 GroupDNet 在语义多模态合成任务上表现良好，并且生成的质量相对较高，但仍有一些问题有待解决。首先，与 pix2pixHD 和 SPADE 相比，它需要更多的计算资源进行训练，尽管它比多个生成器网络快近 2 倍。其次，GroupDNet 仍然难以为多样性有限的数据集建模特定语义类的不同布局，尽管它演示了一些低级变化，例如照明，颜色和纹理等。

致谢

感谢 Taesung Park 对这个项目的慷慨帮助。这项工作得到了国家杰出青年支持计划和 HUST 学术前沿青年团队计划的支持，获得了 NSFC 61573160 的支持，谢谢白翔博士的支持。

参考文献

[1] Aayush Bansal, Yaser Sheikh, and Deva Ramanan. Shapes and context: In-the-wild image synthesis & manipulation. In Proc. CVPR, pages 2317–2326, 2019.

[2] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. CVAE-GAN: fine-grained image generation through asymmetric training. In Proc. ICCV, pages 2764–2773, 2017.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In Proc. ICLR, 2019.

[4] Qifeng Chen and Vladlen Koltun. Photographic image syn-thesis with cascaded refinement networks. In Proc. ICCV, pages 1520–1529, 2017.

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In Proc. CVPR, pages 3213–3223, 2016.

[6] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. In Proc. ICLR, 2017.

[7] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In Proc. CVPR, pages 2414–2423, June 2016.

[8] Xavier Glorot and Yoshua Bengio. Understanding the diffi-culty of training deep feedforward neural networks. In Proc. AISTATS, pages 249–256, 2010.

[9] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In Proc. ECCV, pages 805–822, 2018.

[10] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In Proc. CVPR, pages 6757–6765, 2017.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Proc. NeurIPS, pages 2672–2680, 2014.

[12] Shuyang Gu, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, and Lu Yuan. Mask-guided portrait editing with con-ditional gans. In Proc. CVPR, pages 3436–3445, 2019.

[13] Xintong Han, Zuxuan Wu, Weilin Huang, Matthew R. Scott, and Larry S. Davis. Compatible and diverse fashion image inpainting. In Proc. ICCV, 2019.

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilib-rium. In Proc. NeurIPS, pages 6626–6637, 2017.

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilib-rium. In Proc. NeurIPS, pages 6626–6637, 2017.

[16] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Dar-rell. Cycada: Cycle-consistent adversarial domain adapta-tion. In Proc., pages 1994–2003, 2018.

[17] Seunghoon Hong, Xinchen Yan, Thomas S. Huang, and Honglak Lee. Learning hierarchical semantic image manipu-lation through structured representations. In Proc. NeurIPS, pages 2713–2723, 2018.

[18] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In Proc. ICCV, pages 1510–1519, 2017.

[19] Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In Proc. ECCV, pages 179–196, 2018.

[20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal co-variate shift. In Proc. ICML, pages 448–456, 2015.

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adver-sarial networks. In Proc. CVPR, pages 5967–5976, 2017.

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In Proc. ECCV, pages 694–711, 2016.

[23] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In Proc. ICLR, 2018.

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In Proc. CVPR, pages 4401–4410, 2019.

[25] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Proc. ICLR, 2015.

[26] Diederik P. Kingma and Max Welling. Auto-encoding vari-ational bayes. In Proc. ICLR, 2014.

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural net-works. In Proc. NeurIPS, pages 1106–1114, 2012.

[28] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In Proc. CVPR, pages 105– 114, 2017.

[29] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In Proc. ECCV, pages 36–52, 2018.

[30] Ke Li, Tianhao Zhang, and Jitendra Malik. Diverse im-age synthesis from semantic layouts via conditional IMLE. CoRR, abs/1811.12373, 2018.

[31] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In Proc. NeurIPS, pages 700–708, 2017.

[32] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In Proc. CVPR, 2016.

[33] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuyte-laars, and Luc Van Gool. Pose guided person image genera-tion. In Proc. NIPS, pages 405–415, 2017.

[34] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet V2: practical guidelines for efficient CNN archi-tecture design. In Proc. ECCV, pages 122–138, 2018.

[35] Jiayuan Mao, Xiuming Zhang, Yikai Li, William T. Freeman, Joshua B. Tenenbaum, and Jiajun Wu. Program-guided image manipulators. In Proc. ICCV, 2019.

[36] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784, 2014.

[37] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In Proc. ICLR, 2018.

[38] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In Proc. CVPR, pages 2337–2346, 2019.

[39] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Sin-gan: Learning a generative model from a single natural im-age. In Proc. ICCV, 2019.

[40] Assaf Shocher, Shai Bagon, Phillip Isola, and Michal Irani. Ingan: Capturing and retargeting the ”dna” of a natural im-age. In Proc. ICCV, 2019.

[41] Karen Simonyan and Andrew Zisserman. Very deep convo-lutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014.

[42] Miao Wang, Guo-Ye Yang, Ruilong Li, Runze Liang, Song-Hai Zhang, Peter M. Hall, and Shi-Min Hu. Example-guided style-consistent image synthesis from semantic labeling. In Proc. CVPR, pages 1495–1504, 2019.

[43] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In Proc. CVPR, pages 8798–8807, 2018.

[44] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understand-ing. In Proc. ECCV, pages 432–448, 2018.

[45] Guotian Xie, Jingdong Wang, Ting Zhang, Jianhuang Lai, Richang Hong, and Guo-Jun Qi. Interleaved structured sparse convolutional neural networks. In Proc. CVPR, pages 8847–8856, 2018.

[46] Saining Xie, Ross B. Girshick, Piotr Dollar,´ Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In Proc. CVPR, pages 5987–5995, 2017.

[47] Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee. Diversity-sensitive conditional generative adversarial networks. In Proc. ICLR, 2019.

[48] Fisher Yu, Vladlen Koltun, and Thomas A. Funkhouser. Di-lated residual networks. In Proc. CVPR, pages 636–644, 2017.

[49] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial net-works. In Proc. ICML, pages 7354–7363, 2019.

[50] Han Zhang, Tao Xu, and Hongsheng Li. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proc. ICCV, pages 5908–5916, 2017.

[51] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiao-gang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stack-gan++: Realistic image synthesis with stacked generative ad-versarial networks. IEEE Trans. Pattern Anal. Mach. Intell., 41(8):1947–1962, 2019.

[52] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shecht-man, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In Proc. CVPR, pages 586–595, 2018.

[53] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural net-work for mobile devices. In Proc. CVPR, pages 6848–6856, 2018.

[54] Yang Zhang, Philip David, and Boqing Gong. Curricu-lum domain adaptation for semantic segmentation of urban scenes. In Proc. ICCV, pages 2039–2049, 2017.

[55] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In Proc. CVPR, pages 5122–5130, 2017.

[56] Yang Zhou, Zhen Zhu, Xiang Bai, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Non-stationary texture synthesis by adversarial expansion. ACM Trans. Graph., 37(4):49:1– 49:13, 2018.

[57] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proc. ICCV, pages 2242– 2251, 2017.

[58] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Dar-rell, Alexei A. Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In Proc. NeurIPS, pages 465–476, 2017.

[59] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In Proc. CVPR, pages 2347–2356, 2019.

附 录

A. 实现细节

网络体系架构。 在本节中，我们将为每个数据集提供详细的网络设计。我们在图 7 中演示了判别器的体系架构。请注意，对于不同的数据集，判别器的体系架构保持相同。在图 8 中，我们演示了不同数据集的编码器架构。图 9 描绘了 DeepFashion 和 Cityscapes 解码器的体系架构，而图 10 展示了 ADE20K 的解码器的体系架构。由于 ADE20K 的类别太多，我们降低每个组的通道数，以避免大量使用 GPU。在这种情况下，整体网络容量会减少，我们认为这对结果没有帮助。因此，我们增加了一些卷积层以扩大网络容量。因此，这使得 ADE20K 解码器的体系架构与其他两个数据集的体系架构不同。

训练细节。 我们在 DeepFashion 上训练了所有实验，共进行了 100 个迭代，其中在前 60 个迭代中，生成器和判别器的学习率保持不变，而在最后 40 个迭代中线性下降至 0。对于 Cityscapes 和 ADE20K 数据集，我们遵循 SPADE [38]的训练设置来训练 200 个时期，其中学习率从 100 到 200 个时期线性衰减至 0。图像大小为 256×256，但 Cityscapes 除外，为 512 × 256。DeepFashion 和 Cityscapes 的批处理大小为 32，而 ADE20K 的批处理大小为 16，这是因为大的通道数量可以满足 150 个类的足够容量的要求。在 SPADE [38]之后，使用 Glorot 初始化[8]初始化网络权重。

组号选择策略。 实际上，在 GPU 内存容量，批处理大小和参数数量等的限制下，很难设计出一种编程策略来决定数量的减少。但是，我们仍然遵循两个规则来设计组号： 1）在解码器的前几层中，数字急剧减少，从而大大降低了计算成本； 2）上一层的组号等于或是下一层的 2 倍。

B. 数据集

DeepFashion [32]。 DeepFashion（店内衣服检索基准）包含 52,712 人穿着时尚服装的图像。我们选择了大约 29,000 个训练图像和 2,500 个验证图像。之后，我们使用在 LIP 数据集[10]上预先训练的现成的人体解析器[9]来获得分割图。具体来说，给定输入图像，我们首先获取其分割图，然后将图重新组织为八类：头发，脸，皮肤（包括手和腿），上衣，下衣，袜子，鞋子和背景 。同时，我们会滤除具有一些稀有属性（例如手套等）的图像。我们之所以选择 DeepFashion，是因为该数据集显示出所有语义类的多样性，这自然适合评估模型进行多模态合成的能力。

Cityscapes [5]。 Cityscapes 数据集[5]从德国城市收集了 3,000 个训练图像和 500 个验证图像。 Cityscapes 中图像的尺寸很大，因此测试模型在此数据集上生成高分辨率图像的能力是适当的。

ADE20K [55]。 ADE20K 数据集[55]包含 20,210 个训练图像和 2000 个验证图像。该数据集包含多达 150 个语义类，因此对许多任务而言极具挑战性。ADE20K 因其种类繁多而极具挑战性，我们发现很难使用有限的 GPU 在 ADE20K 上训练 MulNet 和 GroupNet。

C. 补充结果

在图 11 中，我们在 DeepFashion 上显示了更多的消融定性结果。结论与我们在主要意见中提出的基本相同。要注意的一件事是，与 MulNet，GroupNet 和 GroupEnc 相比，我们的 GroupDNet 具有更好的颜色，样式和照明一致性，这是因为其设计考虑了不同类别之间的相关性。同样，GroupDec 和 VSAPDE 似乎也可以像 GroupDNet 一样考虑类关联，因为它们的解码器中的常规卷积有助于发现关联关系。但是，它们却失去了与 GroupDNet 不同的强大的 SMIS 可控性。这些结果坚定地证明了 GroupDNet 的有效性，并显示了其在 SMIS 可控制性和图像质量之间的平衡取舍。

在图 12，图 13 和图 14 中，我们在 pix2pixHD [43]和 SPADE [38]的 DeepFashion，Cityscapes 和 ADE20K 数据集上显示了所提出方法的其他比较结果。这些结果表明，GroupDNet 的图像质量比其他两种方法稍好，尤其是在 Cityscapes 数据集中保持对象结构有序且规则的方面（请参见这些图片中的建筑物和汽车）。

在随附于我们的代码库（见底部 1）的随附视频中，我们在所有数据集中展示了该模型的更多结果。此外，我们对示例应用程序进行了更直接的演示。该视频在主要提交内容中显示了更多示例性应用的结果和详细说明。从这些视频中，我们展示了 GroupDNet 在所有三个数据集上的 SMIS 性能以及为 SMIS 任务设计的模型的潜在应用。但是，我们在 Cityscapes 数据集上训练的模型似乎失去了语义可控性。例如，当

1 https://github.com/Seanseattle/SMIS

Models	FID↓	mCSD↑	mOCD↓	LPIPS↑	SHE↑	FPS ↑	# Para↓
GroupDNet	**9.50**	0.0264	**0.0033**	0.228	**81.2**	12.2	109.1
w/o map	11.01	0.0253	0.0036	0.217	79.5	11.5	109.3
w/o split	10.76	0.0054	0.0189	0.216	31.7	12.1	109.1
→GroupNorm	10.33	0.0256	0.0040	0.225	77.0	12.2	109.1
w/o SyncBN	9.76	0.0251	0.0037	0.216	79.3	12.3	109.1
w/o SpecNorm	10.42	**0.0290**	0.0153	**0.231**	46.3	**13.5**	109.0

表 3：在 DeepFashion 数据集上的消融实验的定量结果。

更改架构的潜在编码后，其他部分也会随之更改。我们不愿意将此现象视为重大缺陷。在某些情况下，我们确实希望整个图像可以随类特定的潜在编码的变化而变化，因为它可以增强生成的图像的保真度。例如，整体照明与某些物体上的照明之间的差异会使图像不真实且不自然。另一个问题是，在 Cityscapes 上生成的结果的多样性似乎非常有限。这是因为此数据集最初仅限于德国城市的场景。此外，数据集内的图像是在短时间间隔内拍摄的。因此，从白天到黑暗，它们都没有照明的多样性。看到这些结果，我们坚信语义多模态图像合成具有更多的应用和内在的科学价值，值得探索给定合适的数据集。将来，我们将对 GroupDNet 进行更多调查，并尝试改善其在 SMIS 中的性能。

D. 额外消融研究

为了支持 SMIS 任务并提高生成的图像的质量，我们对 GroupDNet 进行了一些较小的修改，包括：1）如原始文本中所述，将原始输入图像拆分为不同语义类别的不同图像；2）强制编码器生成均值图和方差图，而不是 SPADE 中使用的均值向量和方差向量。为了验证这些策略的效果，我们通过不在网络中使用消融组件来进行消融实验，因此我们得到了不分割原始图像（含/不含分割图像）的模型结果，并且不产生均值和方差的模型特征图（含/不含模型特征图）。对于后一种情况，请注意，我们在编码器的最后使用添加全局平均池化将映射压缩为向量，而不是使用常规的完全连接层来生成向量，这可能会削弱每个类的独立性。正如在表 3 中的结果所显示：分割输入图像或生成均值和方差图是 SMIS 任务的必要策略，否则模型将遭受 FID，mCSD 和 mOCD 性能下降的困扰。

此外，考虑到 GroupDNet 中组卷积的广泛使用，非常有趣的是，如果我们将组归一化作为模型中的主要非归一化层，将会产生什么影响，因为组归一化还可以在不同组的函数通道上单独运行。因此，我们通过更改原始模型中的所有归一化层以对归一化层进行分组并设置其组号等于其先前的卷积层（→GN）来进行另一项实验。我们还进行了实验，以通过丢弃模型中的几个归一化层的使用来调查它们的影响。基本上，我们有了模型，而没有使用同步批处理归一化（含/不含 SyncBN），也没有使用谱归一化[37]（含/不含 SpecNorm）。结果在表 3 中报告。更改归一化层以进行归一化或删除同步的批处理归一化已在大多数指标上稍微降低了性能。但是，删除 GroupDNet 中的谱归一化层将大大增加 mOCD，这表明谱归一化有助于 SMIS 任务。没有谱归一化的模型的 LPIPS 和 mCSD 度量甚至比 GroupDNet 高，这表明谱归一化可能会对模型生成的图像的多样性产生负面影响。

E. 讨论和未来工作

如前所述，GroupDNet 的局限性在于在 Cityscapes 数据集中捕获图像的类别多模态的能力有限。尽管数据集本身显示的对象外观变化很小，但我们认为仍然存在许多方法和体系架构上的修改，可以使 GroupDNet 或其他模型处理此类困难情况。

此外，我们还认为有必要设计一种干净有效的策略来决定如何为每个卷积或可能的归一化层设置组号。尽管我们没有给出明确的实验证据，但我们认为不同的组号配置可能会对效果产生一些影响。该结论很自然，因为不同的组号配置会确定不同的网络结构，并且不同的网络结构通常会影响网络性能。

此外，考虑到现在我们按照数据集提供者设置的顺序或我们随机设置的顺序将分割后的输入图像馈送到编码器，发现改变不同类别的输入顺序是否会对性能产生影响也很有趣。自然地讲，将相似的类放在一起可以使相应的区域和谐一致地变化，从而生成具有更高保真度的图像。

Models	DeepFashion	Cityscapes	ADE20K
Encoder	$C = 64, G = 8$	$C = 8, G = 35$	$C = 3, G = 151$
Decoder	$C = 160, G = \{8, 8, 4, 4, 2, 2, 1\}$	$C = 280, G = \{35, 35, 20, 14, 10, 4, 1\}$	$C = \{151, 64\}, G = \{151, 16, 16, 8, 4, 2, 1, 1\}$

表 4：我们的编码器和解码器的不同数据集的预定义超参数。 注意，在图 8，图 9 和图 10 中，“ C{i}”代表 C 括号内的第 i 个数字，“ G{i}”代表 G 括号内的第 i 个数字。

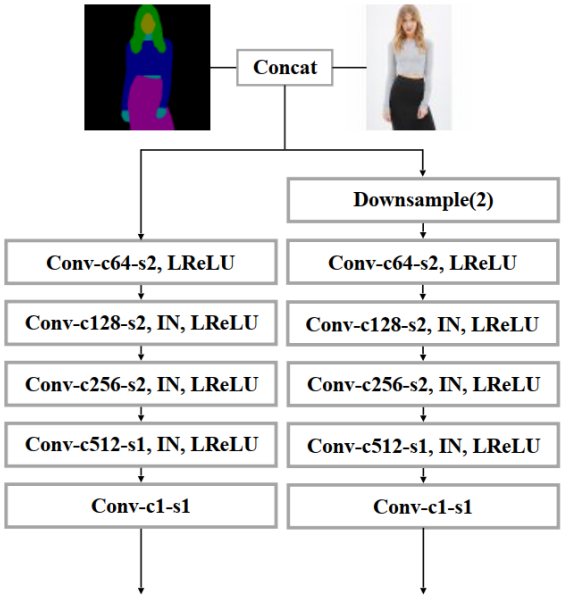


图 7：所有三个数据集的判别器架构。 注意“Conv”是指卷积层。“-c”，“-s”和“-g”之后的数字表示对应卷积的通道号，步幅和组号。 如果未指定，则卷积层的默认内核大小，步幅，填充和组号分别为 4、2、2、1。“IN”表示实例归一化层，“LReLU”表示 Leaky ReLU 层。 “Downsample（·）”表示一个平均池化层，其内核大小设置为括号内的数字。

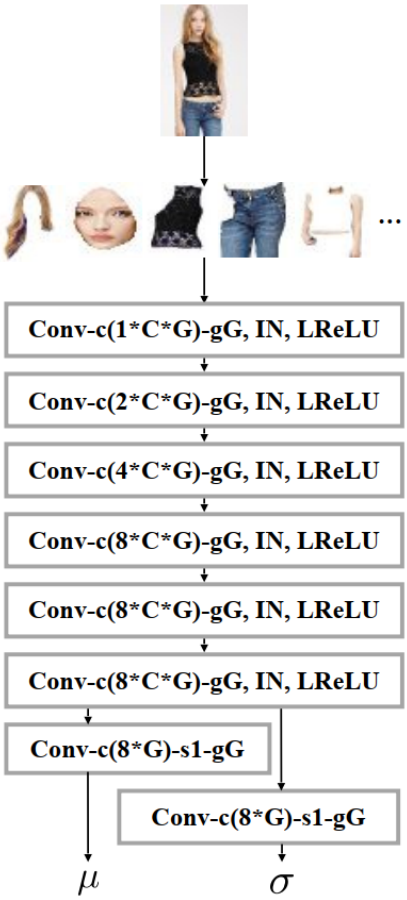


图 8：三个数据集的编码器架构。 注意“Conv”是指卷积层。 “-c”，“-s”和“-g”之后的数字代表相应卷积的通道号，步幅和组号。 如果未指定，则卷积层的默认内核大小，步幅，填充和组号分别为 3、2、1、1。“IN”表示实例归一化层，“LReLU”表示 Leaky ReLU 层。 在此，“C”和“G”是每个数据集的预定义数字，在表 4 中给出。

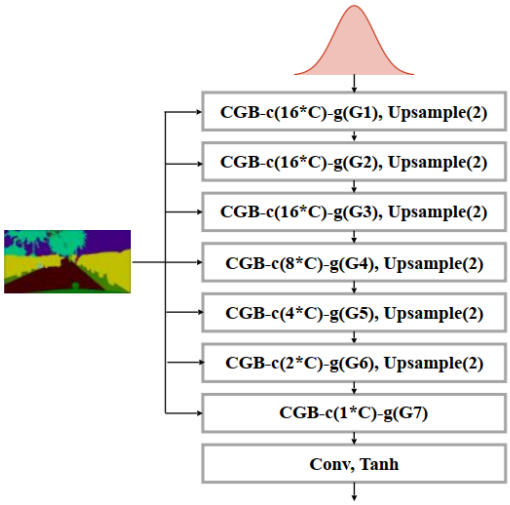


图 9：DeepFashion 和 Cityscapes 数据集的解码器架构。 注意“ CGB”是指我们的条件分组块（CG-Block）。 “-c”， “-s”和“-g”之后的数字表示相应卷积的通道号，步幅和组号。如果未指定，则卷积层的默认内核大小，步幅，填充和组号分别为 3、1、1、1。 在 CG 块中的每个 CG 归一化之后，都有一个 ReLu 层。 “ Upsample（·）” 表示内核大小设置为括号内数字的最近邻居的上采样层。 在此，“C”和“G”是每个数据集的预定义数字，在表 4 中给出。

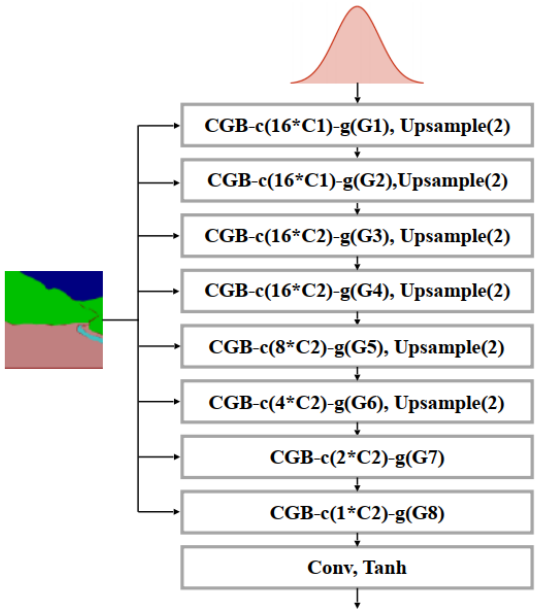


图 10：用于 ADE20K 数据集的解码器架构。 注意“ CGB”是指我们的条件分组块（CG-Block）。 “-c”， “-s”和“-g”之后的数字表示相应卷积的通道号，步幅和组号。 如果未指定，则卷积层的默认内核大小，步幅，填充和组号分别为 3、1、1、1。 在 CG 块中的每个 CG 归一化之后，都有一个 ReLu 层。 “ Upsample（·）” 表示内核大小设置为括号内数字的最近邻居向上采样层。 在此， “C”和“G”是每个数据集的预定义数字，在表 4 中给出。

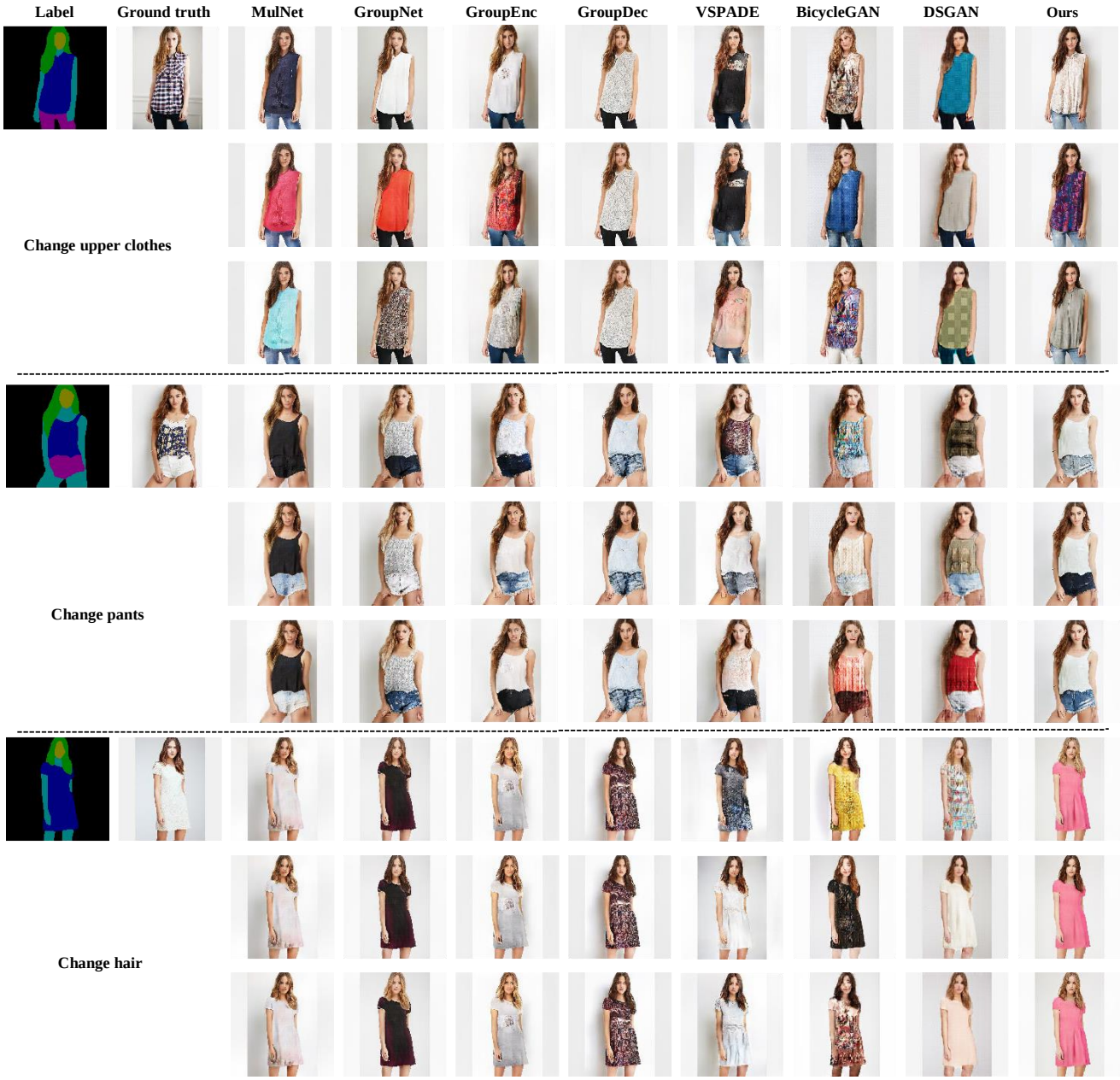


图 11：GroupDNet 与其他基准模型之间的定性比较。 前三行通过更改其上衣潜在编码表示不同模型的结果。 中间的三行表示通过更改其裤子潜在编码的不同模型的结果，而最后三行表示其更改头发潜在编码的结果。 请注意，对于那些没有特定于类别的潜在编码的模型（例如 VSPADE），我们会更改其整体潜在编码以生成不同的图像。

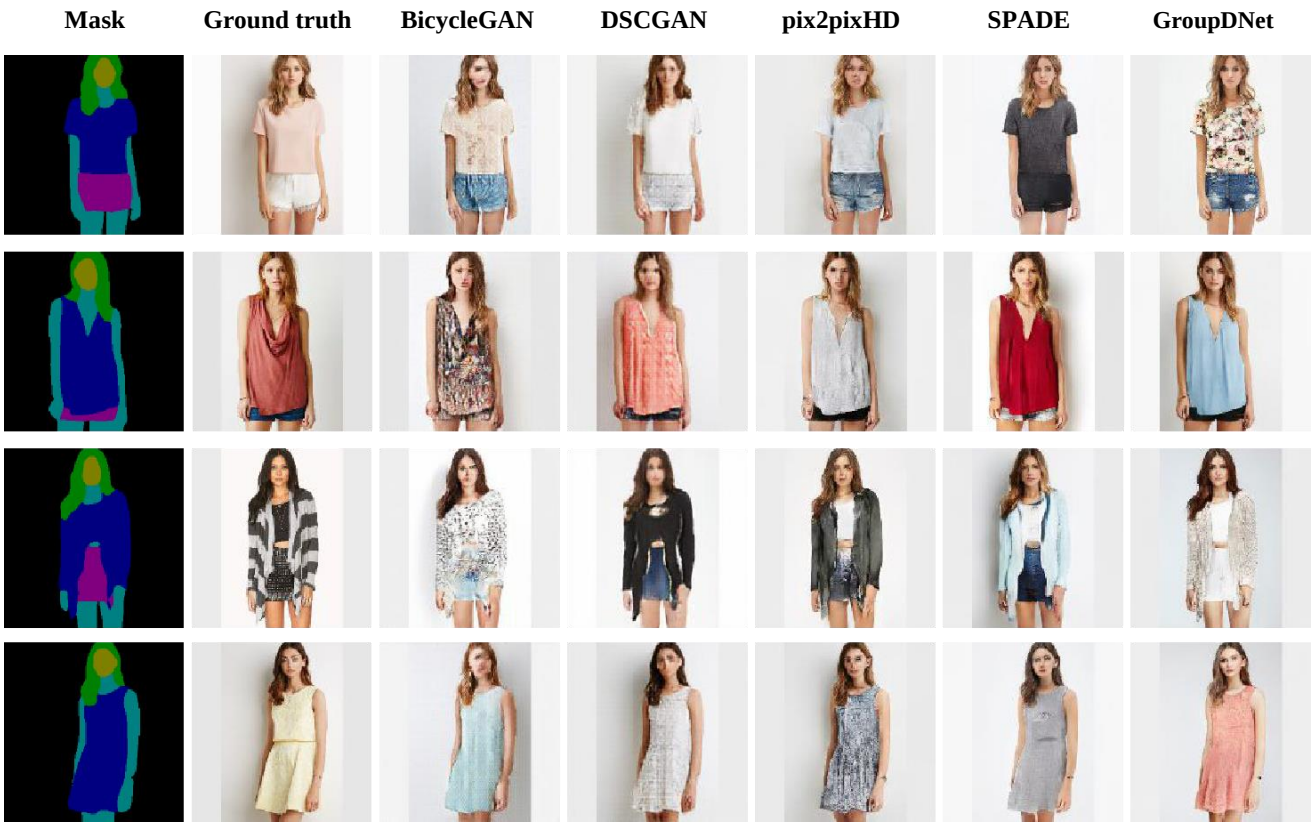


图 12：我们的模型与 DeepFashion 数据集上几种从标签到图像的方法的定性比较。

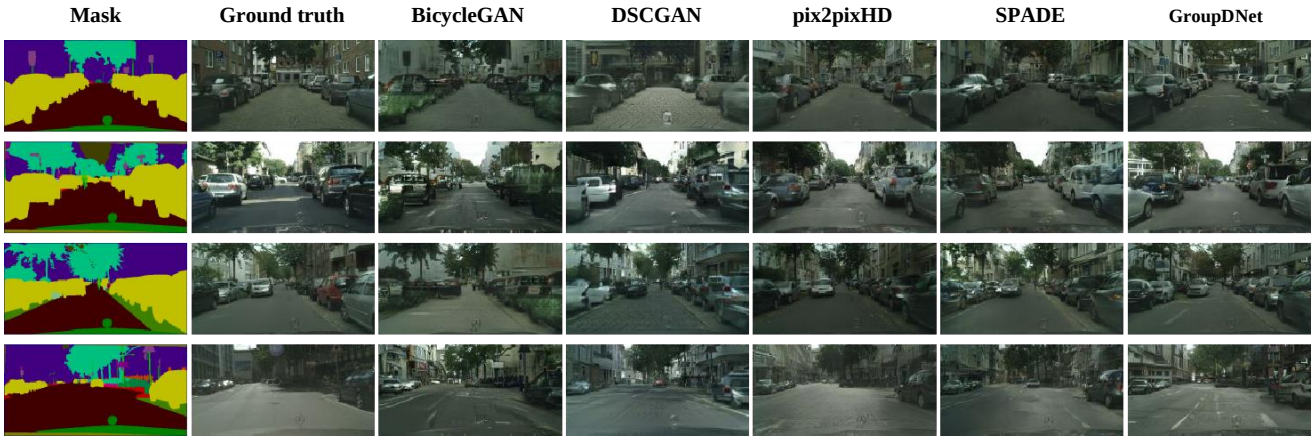


图 13：我们的模型与 Cityscapes 数据集上几种从标签到图像方法的定性比较。

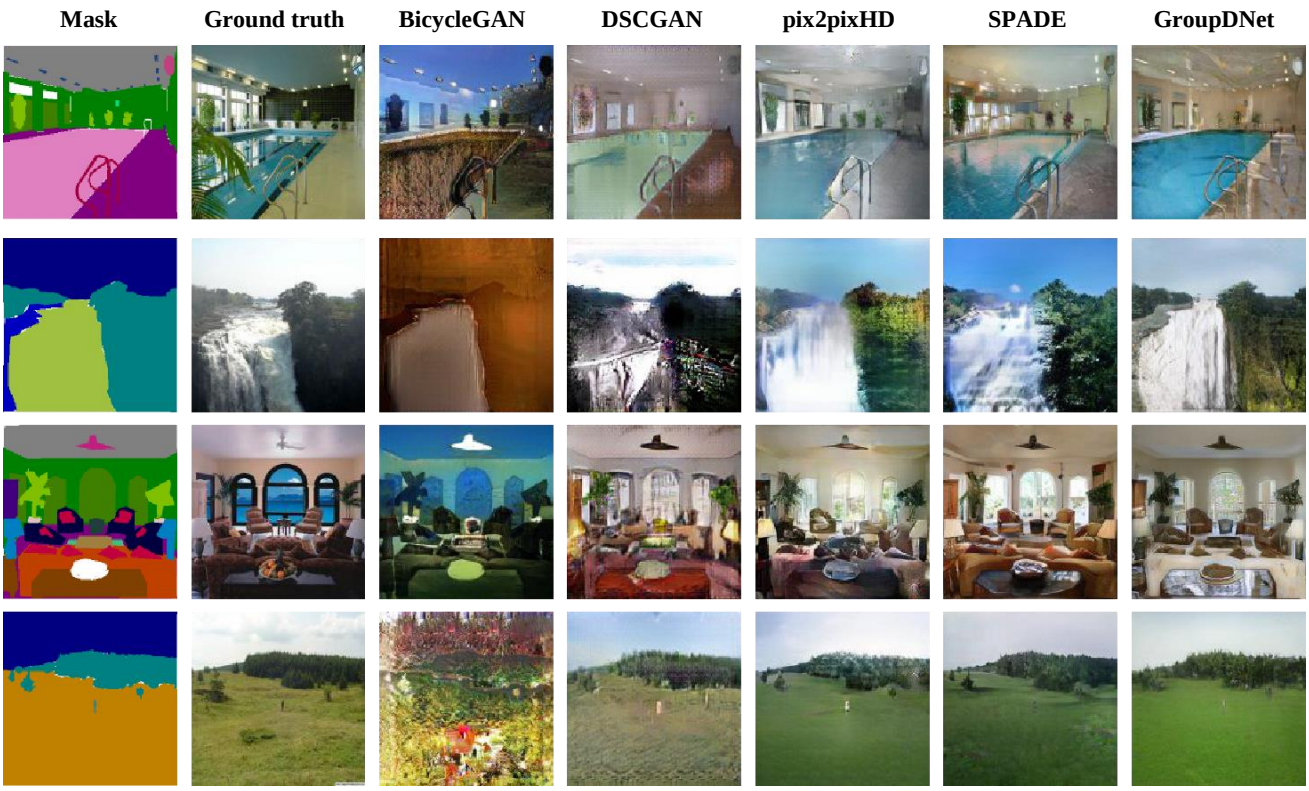


图 14：我们的模型与 ADE20K 数据集上几种从标签到图像方法的定性比较。