带有空间自适应归一化的语义图像合成

Taesung Park[1] Ming-Yu Liu[2] Ting-Chun Wang[2] Jun-Yan Zhu[2,3]

[1]UC Berkeley [2]NVIDIA [3]MIT CSAIL

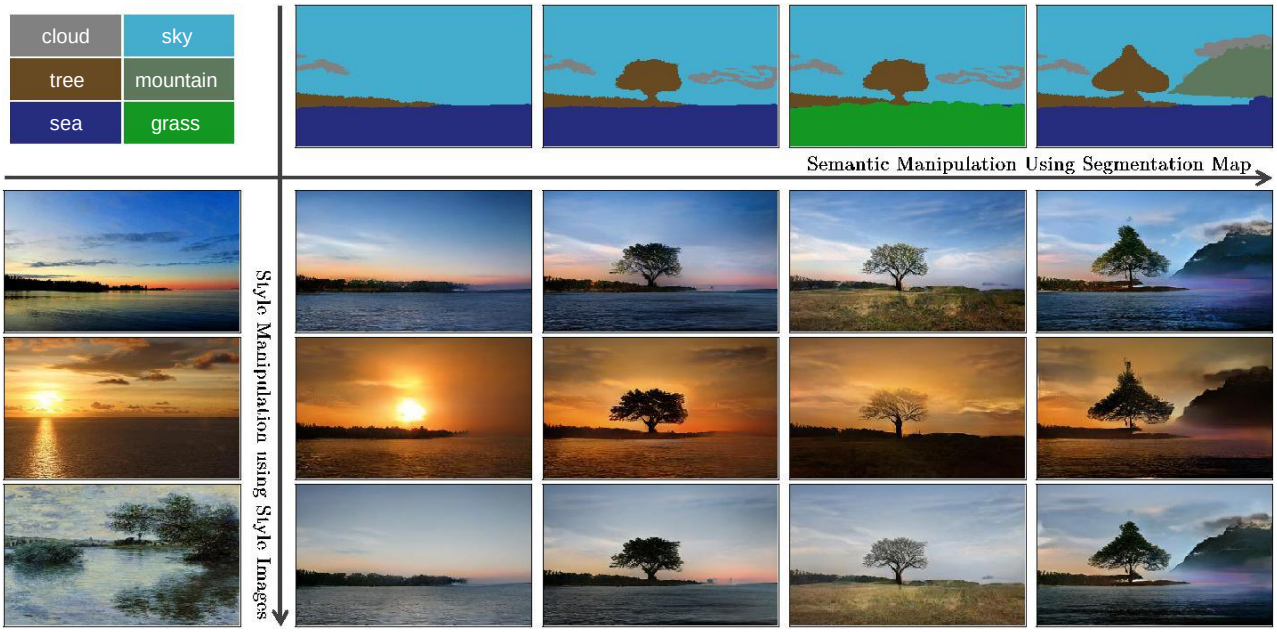

图 1：我们的模型允许用户在合成图像时控制语义和样式。 语义（例如，树的存在）通过标签图（在顶行中可视化）来控制，而样式通过参考样式图像（在最左列中可视化）来控制。 请访问我们的 website 查看交互式图像合成演示。

摘 要

我们提出了空间自适应归一化，这是一个简单但有效的层，用于在给定输入语义布局的情况下合成逼真的图像。 以前的方法直接将语义布局作为深层网络的输入，然后通过卷积的堆叠，归一化和非线性层进行处理。 我们证明这是次优的，因为归一化层倾向于“冲走”语义信息。 为了解决这个问题，我们建议使用输入布局通过空间自适应学习变换来调整归一化层中的激活函数。 几个具有挑战性的数据集的实验证明了我们提出的方法优于现有方法的优势在于既有视觉保真度又能输入布局对齐。 最后，我们的模型允许用户控制语义和样式作为合成图像。 代码将在 https://github.com/NVlabs/SPADE 开源。

1. 介绍

条件图像合成是指在一些输入数据上生成照片级真实感图像的任务。早期的方法是通过拼接图像数据库来计算输出图像[3,13]。重新分析方法使用神经网络直接学习映射[4,7,20,39,40,45,46,47]。 后一种方法通常更快并且不需要外部图像数据库。

我们感兴趣的是一种特定形式的条件图像合成，它将语义

Taesung Park 在他的 NVIDIA 实习期间为这项工作做出了贡献。

分割蒙版转换为逼真的图像。这种模式具有广泛的应用，如内容生成和图像编辑[7,20,40]。我们将这种形式称为语义图像合成。在本文中，我们表明，通过堆叠卷积，归一化和非线性层构建的传统网络架构[20,40]充其量是次优的，因为它们的归一化层倾向于“冲走” 输入语义掩码中的信息。

为了解决这个问题，我们提出了空间自适应归一化，这是一种条件归一化层，它通过空间自适应学习转换使用输入语义布局来调制激活函数，并且可以有效地在整个网络中提供语义信息。

我们对几个具有挑战性的数据集进行了实验，包括 COCO-Stuff [5,26]，ADE20K [48]和 Cityscapes [8]。我们表明，在我们的空间自适应归一化层的帮助下，与几种最先进的方法相比，紧凑的网络可以合成明显更好的结果。此外，一项广泛的研究表明，所提出的归一化层针对语义图像合成任务的几种变体都有有效性。最后，我们的方法支持多模态和样式引导的图像合成，实现可控制的多样化输出，如图 1 所示。

2. 相关工作

深度生成模型 可以学习合成随机采样的图像。最近的方法包括生成对抗网络（GAN）[12]和变分自动编码器（VAE）[22]。我们的工作建立在 GAN 之上，但目标是条件图像合成任务。 GAN 由生成器和判别器组成，其中生成器的目标是生成逼真的图像，使得判别器不能将合成图像与真实图像区分开来。

条件图像合成 以许多形式存在，这些形式在输入数据的类型中是不同的。例如，类条件模型[4,29,31,33]学习合成具有猫标签的图像。研究人员已经探索了基于文本生成图像的各种模型[16,36,43,46]。另一种广泛使用的形式是图像到图像的转换[18,20,23,27,49,50]，其中输入和输出都是图像。在这项工作中，我们专注于将分割蒙版转换为逼真的图像。我们假设训练数据集包含成对的分割掩模和图像。通过提出的空间自适应归一化，与领先方法相比，我们的紧凑型网络实现了更好的结果。

无条件规范化层 一直是现代深度网络中的一个重要组成部分，可以在各种分类器设计中找到，包括 AlexNet [24]中的本地响应归一化（LRN）和 Inception-v2 网络[19]中的批量归一化（BN）。其他流行的规范化层包括实例

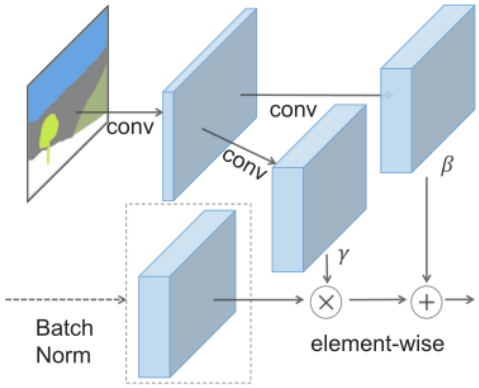

图 2：在 SPADE 中，首先将掩模投影到 embedding 空间，然后进行卷积以产生调制参数 γ 和 β 。 不同于先前的条件归一化方法，γ 和 β 并不是向量，而是具有空间维度的张量。 产生的 γ 和 β 会被相乘并加到归一化激活元素中。

归一化（IN）[38]，层归一化（LN）[2]，组归一化（GN）[41]和权重归一化（WN）[37]。我们将这些归一化层标记为不合理，因为与下面讨论的条件归一化层相比，它们不依赖于外部数据。

条件归一化层 包括条件批量归一化（Conditional BN）[10]和自适应实例归一化（AdaIN）[17]。两者都首先用于风格转移任务，后来被用于各种任务[9,18,29,31,34,45]。与早期的归一化技术不同，条件归一化层需要外部数据，并且通常如下操作。首先，将层激活归一化为零均值和单位方差。然后通过使用学习的仿射变换调制激活来对归一化激活进行非规范化，所述仿射变换的参数是从外部数据推断的。对于样式转移任务[10,17]，仿射参数用于控制输出的全局样式，因此在空间坐标上是均匀的。与先前的工作不同，我们提出的归一化层应用空间变化的仿射变换，使其适合于从空间变化的语义掩模进行图像合成。

3.语义图像合成

设 $\mathbf{m} \in \mathbb{L}^{H\times W}$ 是语义分割掩模，其中 $\mathbb{L}$ 是表示语义标签的整数集，H 和 W 是图像的高度和宽度。 m 中的每个条目表示像素的语义标签。我们的目标是学习可以将输入分割掩模 m 转换为逼真图像的映射函数。

空间自适应非规范化。 设 $\mathbf{h}^i$ 表示给定一批 N 个样本的深卷积网络的第 i 层激活层。设 C^i 是图层中通道的数量。设 H^i 和 W^i 是各图层中激活层的高度和宽度。我们提出一个新的

条件归一化方法称为 SPatially-Adaptive（DE）正则化(底部 1)（SPADE）。 与批量归一化[19]类似，激活以通道方式归一化，然后用学习的缩放和偏差进行调制。 图 2 显示了 SPADE 设计。 现场的激活值 $(n \in N, c \in C^i, y \in H^i, x \in W^i)$ 由下式给出：

$$\gamma^i_{c,y,x}(\mathbf{m})\frac{h^i_{n,c,y,x} - \mu^i_c}{\sigma^i_c} + \beta^i_{c,y,x}(\mathbf{m}) \qquad (1)$$

其中 $h^i_{n,c,y,x}$ 是标准化前的激活位置，μ^i_c 和 σ^i_c 是通道 c 中激活的平均值和标准偏差：

$$\mu^i_c = \frac{1}{NH^iW^i}\sum_{n,y,x} h^i_{n,c,y,x} \qquad (2)$$

$$\sigma^i_c = \sqrt{\frac{1}{NH^iW^i}\sum_{n,y,x}(h^i_{n,c,y,x})^2 - (\mu^i_c)^2}. \qquad (3)$$

（1）中的变量 $\gamma^i_{c,y,x}(\mathbf{m})$ 和 $\beta^i_{c,y,x}(\mathbf{m})$ 是归一化层的学习调制参数。与 BatchNorm [19]相反，它们依赖于输入分段掩码并且相对于位置（y，x）而变化。我们使用符号 $\gamma^i_{c,y,x}$ 和 $\beta^i_{c,y,x}$ 来表示将输入分段掩码 m 转换为第 i 个激活映射中的位置（c，y，x）处的缩放和偏置值的函数。我们使用简单的两层卷积网络实现函数 $\gamma^i_{c,y,x}$ 和 $\beta^i_{c,y,x}$，其详细设计可以在附录中找到。

事实上，SPADE 与几个现有的归一化层相关，并且是一个概括。首先，用图像类标签替换分割掩模 m，并使调制参数在空间上不变（即 $\gamma^i_{c,y_1,x_1} \equiv \gamma^i_{c,y_2,x_2}$ 和 $\beta^i_{c,y_1,x_1} \equiv \beta^i_{c,y_2,x_2}$ 对于任意 $y_1, y_2 \in \{1, 2, ..., H^i\}$ 和 $x_1, x_2 \in \{1, 2, ..., W^i\}$），我们得到条件批量归一化层[10]的形式。实际上，对于任何空间不变的条件数据，我们的方法简化为 Conditional BN。类似地，我们可以通过用另一个图像替换分割掩模来到达 AdaIN [17]，使调制参数在空间上不变并设置 N = 1。由于调制参数适应输入分割掩模，所以提出的 SPADE 是更适合语义图像合成。

SPADE 生成器。使用 SPADE，不需要将分割图提供给生成器的第一层，因为学习的调制参数已经编码了关于标签布局的足够信息。因此，我们丢弃了生成器的编码器部分，这是最近的架构中常用的[20,40]。这种简化可以在更轻量级的网络中实现。此外，类似于

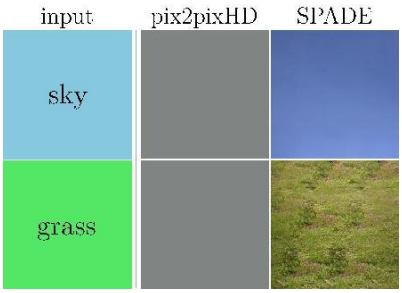

图 3：比较给定均匀分割图的结果：当 SPADE 生成器产生合理的纹理时，pix2pixHD [40]由于归一化层之后语义信息的丢失而产生相同的输出。

对于现有的类条件生成器[29,31,45]，新生成器可以将随机向量作为输入，为多模态合成实现简单而自然的方式[18,50]。

图 4 说明了我们的生成器架构，它采用了几个带有上采样层的 ResNet 块[14]。使用 SPADE 学习所有归一化层的调制参数。由于每个残差块以不同的比例操作，因此 SPADE 对语义掩码进行下采样以匹配空间分辨率。

我们使用 pix2pixHD 中使用的相同的多尺度判别器和损失函数来训练生成器，除了我们用铰链损失项[25,30,45]替换最小平方损失项[28]。我们测试了几个在最近的无条件 GAN 中使用的基于 ResNet 的判别器[1,29,31]，但是以更高的 GPU 存储器要求为代价观察到类似的结果。将 SPADE 添加到判别器也可以产生类似的性能。对于损失函数，我们观察到去除 pix2pixHD 损失函数中的任何损失项导致退步的生成结果。

为什么 SPADE 工作得更好？ 简短的回答是，它可以更好地保护语义信息不受常见归一化层的影响。具体来说，虽然 InstanceNorm [38]等归一化层是几乎所有最先进的条件图像合成模型[40]中必不可少的部分，但它们在应用于均匀或平滑语义分割遮罩时往往会冲掉语义信息。

让我们考虑一个简单的模块，首先将卷积应用于分割掩码，然后进行归一化。此外，让我们假设具有单个标签的分割掩模被给予作为模块的输入（例如，所有像素具有相同的标签，例如天空或草）。在此设置下，卷积输出再次均匀化，不同的标签具有不同的均匀值。现在，在我们将 InstanceNorm 应用于输出之后，无论输入的语义标签是什么，标准化激活都将变为全零。因此，语义信息完全丢失。此限制适用于各种生成器架构，包括 pix2pixHD 及其变体

1 条件归一化[10,17]使用外部数据对归一化激活进行非规范化；即，非规范化部分是有条件的.

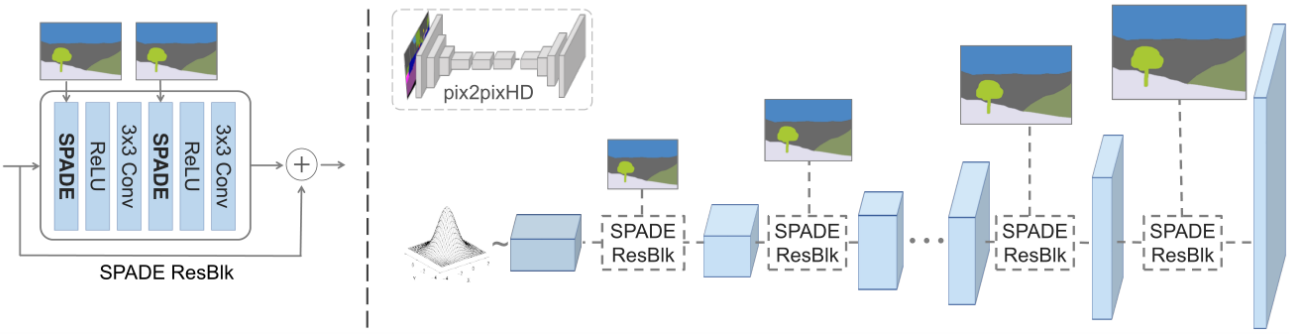

图 4：在 SPADE 生成器中，每个归一化层使用分段掩码来调制图层激活。（左）具有 SPADE 的一个残余块的结构。（右）生成器包含一系列具有上采样层的 SPADE 残差块。通过删除先前的图像到图像转换网络（pix2pixHD [40]）的下采样层，我们的架构通过较少的参数实现了更好的性能。

只要网络应用卷积然后归一化到语义掩码，就可以连接到所有中间层并使用语义掩码。在图 3 中，我们凭经验证明了这正是 pix2pixHD 的情况。由于分段掩码通常由几个统一区域组成，因此在应用归一化时会出现信息丢失问题。

相反，SPADE Generator 中的分割掩模通过空间自适应调制进行馈送而不进行归一化，仅前一层的激活被标准化。因此，SPADE 生成器可以更好地预先提供语义信息。它享受归一化的好处而不会丢失语义输入信息。

多模态合成。通过使用随机向量作为生成器的输入，我们的架构为多模态合成提供了一种简单的方法。即，可以将处理真实图像的编码器附加到随机矢量中，然后将其馈送到生成器。编码器和生成器形成变分自动编码器[22]，其中编码器试图捕获图像的样式，而生成器通过 SPADE 组合编码样式和分割掩模信息以重建原始图像。编码器还在测试时用作风格引导网络以捕获目标图像的样式，如图 1 所示。对于训练，我们添加 KL-Divergence 损失项[22]。

4. 实验

实施细节。 我们将 Spectral Norm [30]应用于生成器和判别器中的所有层。 生成器和判别器的学习率分别设置为 0.0001 和 0.0004 [15]。 我们使用 ADAM [21]并设置β1 = 0, β2 = 0.999。 所有实验均在配备 8 个 V100 GPU 的 NVIDIA DGX1 上进行。 我们使用同步均值和方差计算，即从所有 GPU 收集这些统计数据。

数据集。 我们对几个数据集进行了实验。

- *COCO-Stuff* [5]源自 COCO 数据集[26]。它拥有 118,000 张训练图像和 5,000 个从不同场景中捕获的验证图像。它有 182 个语义类。由于其多样性，现有的图像合成模型在该数据集上表现不佳。
- *ADE20K* [48]包含 20,210 张训练和 2,000 张验证图像。与 COCO 类似，数据集包含具有 150 个语义类的具有挑战性的场景。
- *ADE20K-outdoor* 是 ADE20K 数据集的一个子集，仅包含 Qi 等人使用的室外场景[35]。
- *Cityscapes* [8] 包含德国城市中的街景图像。训练和验证集大小分别为 3,000 和 500。最近的工作已经在 Cityscapes 数据集上实现了逼真的语义图像合成结果[35,39]。
- *Flickr Landscapes*。我们从 Flickr 收集 41,000 张照片，并使用 1,000 个样本作为验证集。我们使用预先训练的 DeepLabV2 模型[6]来计算输入分割掩模，而不是手动注释。

我们在同一训练集上训练竞争语义图像合成方法，并在每个数据集的相同验证集上报告它们的结果。

性能指标。 我们采用以前工作的评估协议[7,40]。 具体来说，我们在合成图像上运行语义分割模型，并比较预测分割掩模与输入对应的完全真实图像的匹配程度。 这是基于如下直觉：如果输出图像是真实的，那么训练有素的语义分割模型应该能够预测完全真实标签。 为了测量分割精度，我们使用平均交叉联合（mIoU）和像素精确（准）度量。 我们为每个数据集使用最先进的分割网络：针对 COCO-Stuff 的 DeepLabV2 [6,32]，针对 ADE20K 的 UperNet101 [42] 和针对 Cityscapes 的 DRN-D-105 [44]。 除了分割精度之外，我们还使用 Frechet'起始距离（FID）[15]进行测量

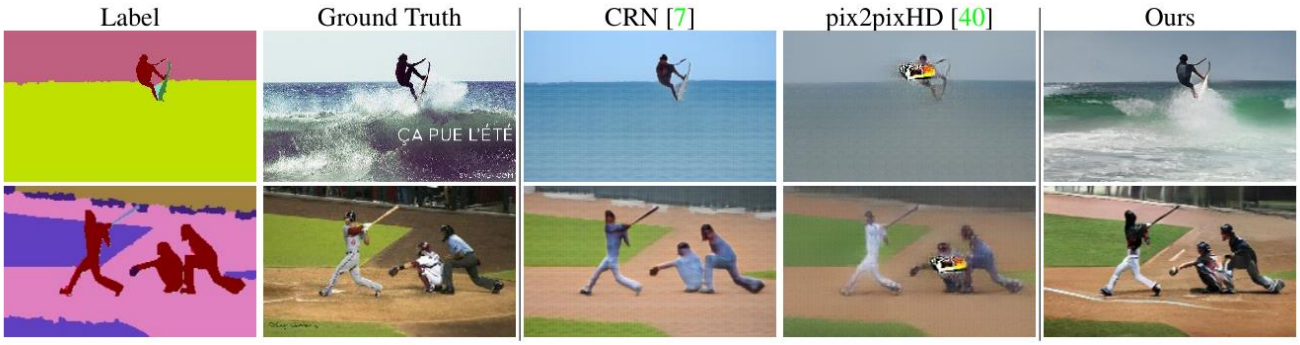

图 5：COCO-Stuff 数据集上语义图像合成结果的可视化比较。 我们的方法成功地从语义标签合成了真实的细节。

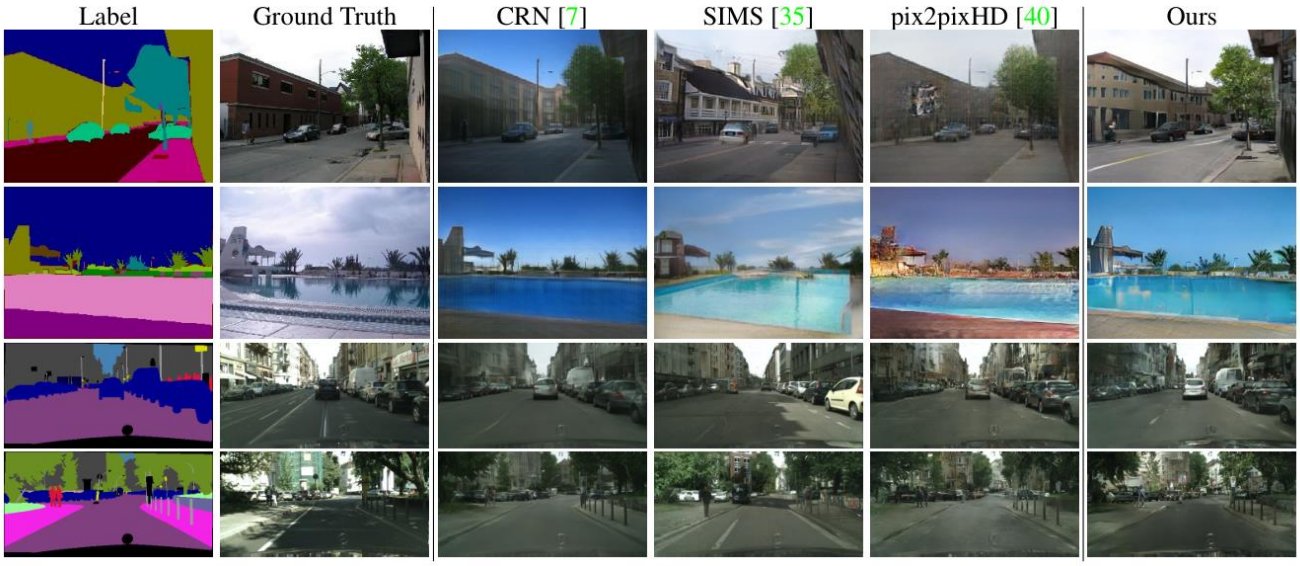

图 6：ADE20K 室外和 Cityscapes 数据集上语义图像合成结果的可视化比较。 我们的方法在尊重空间语义布局的同时产生逼真的图像。

	COCO-Stuff			ADE20K			ADE20K-outdoor			Cityscapes		
Method	mIoU	accu	FID	mIoU	accu	FID	mIoU	accu	FID	mIoU	accu	FID
CRN [7]	23.7	40.4	70.4	22.4	68.8	73.3	16.5	68.6	99.0	52.4	77.1	104.7
SIMS [35]	N/A	N/A	N/A	N/A	N/A	N/A	13.1	74.7	67.7	47.2	75.5	**49.7**
pix2pixHD [40]	14.6	45.8	111.5	20.3	69.2	81.8	17.4	71.6	97.8	58.3	81.4	95.0
Ours	**37.4**	**67.9**	**22.6**	**38.5**	**79.9**	**33.9**	**30.8**	**82.9**	**63.3**	**62.3**	**81.9**	71.8

表 1：我们的方法在所有基准数据集上的语义分段得分（平均 IoU 和整体像素精度）和 FID [15]方面优于当前领先方法。 对于 mIoU 和像素精度，越高越好。 对于 FID，越低越好。

合成结果的分布与实际图像的分布之间的距离。

基线。 我们将我们的方法与三种领先的语义图像合成模型进行比较：pix2pixHD 模型[40]，级联细化网络模型（CRN）[7]和半参数图像合成模型（SIMS）[35]。pix2pixHD 是目前最先进的基于 GAN 的连续图像合成框架。 CRN 使用深度网络重复精细化从低到高分辨率的输出，而 SIMS 采用半参数方法

它组合了训练集中的真实片段并细化了边界。 CRN 和 SIMS 都主要使用图像重建损失进行训练。 为了公平比较，我们使用作者提供的实现来训练 CRN 和 pix2pixHD 模型。由于使用 SIMS 合成图像需要对训练数据集进行许多查询，因此对于诸如 COCO-stuff 和完整 ADE20K 的大型数据集而言，它在计算上是超级艰巨的。 因此，我们尽可能使用作者提供的结果图像。

定量比较。 如表 1 所示，我们的

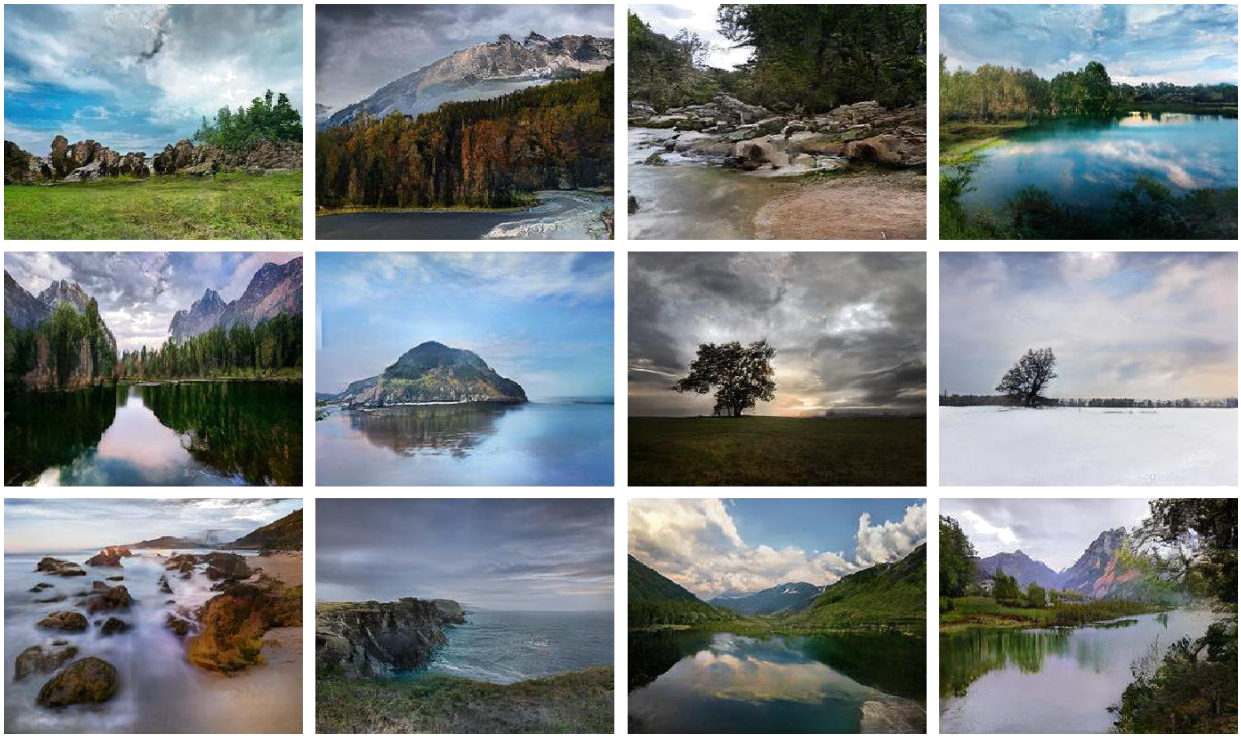

图 7：Flickr Landscapes 数据集上的语义图像合成结果。 图像是从 Flickr 上的照片的语义布局生成的。

方法在所有数据集中，都大大优于当前最先进的方法。对于 COCO-Stuff，我们的方法实现了 35.2 的 mIoU 评分，这比之前的领先方法好大约 1.5 倍。我们的 FID 也比以前的领先方法好 2.2 倍。我们注意到 SIMS 模型产生较低的 FID 分数，但在 Cityscapes 数据集上的分割性能较差。这是因为 SIMS 通过首先拼接来自训练数据集的图像块来合成图像。使用真实图像补丁时，得到的图像分布可以更好地匹配真实图像的分布。然而，因为不能保证数据集中存在完美查询（例如，特定姿势中的人），所以它倾向于复制具有不匹配段的对象。

定性结果。在图 5 和图 6 中，我们提供了竞争方法的定性比较。我们发现我们的方法产生的结果具有更好的视觉质量和更少的伪像，特别是对于 COCO-Stuff 和 ADE20K 数据集中的不同场景。当训练数据集大小较小时，SIMS 模型还呈现具有良好视觉质量的图像。然而，所描绘的内容经常偏离输入分割掩模（例如，图 6 的第二行中的游泳池的形状）。

在图 7 和图 8 中，我们显示了来自 Flickr Landscape 和 COCO-Stuff 数据集的更多示例结果。提出的方法可以生成具有高保真度图像的各种场景

Dataset	Ours vs. CRN	Ours vs. pix2pixHD	Ours vs. SIMS
COCO-Stuff	79.76	86.64	N/A
ADE20K	76.66	83.74	N/A
ADE20K-outdoor	66.04	79.34	85.70
Cityscapes	63.60	53.64	51.52

表 2：用户偏好研究。 这些数字表示赞成所提议方法的结果而非竞争方法的用户百分比。

更多结果包含在附录中。

人的评价。 我们使用 Amazon Mechanical Turk（AMT）来比较我们的方法与现有方法的视觉保真度。 具体来说，我们为 AMT 工作人员提供输入分段掩码和来自不同方法的两个合成输出，并要求他们选择看起来更像是分段掩码的相应图像的输出图像。 工人有无限的时间进行选择。 对于每次比较，我们为每个数据集随机生成 500 个问题，每个问题由 5 个不同的工作人员回答。 对于质量控制，只有终身任务批准率大于 98%的员工才能参与我们的评估。

表 2 显示了评估结果。 我们发现用户非常青睐我们在所有数据集上的结果，特别是在具有挑战性的 COCO-Stuff 和 ADE20K 数据集上。 对于

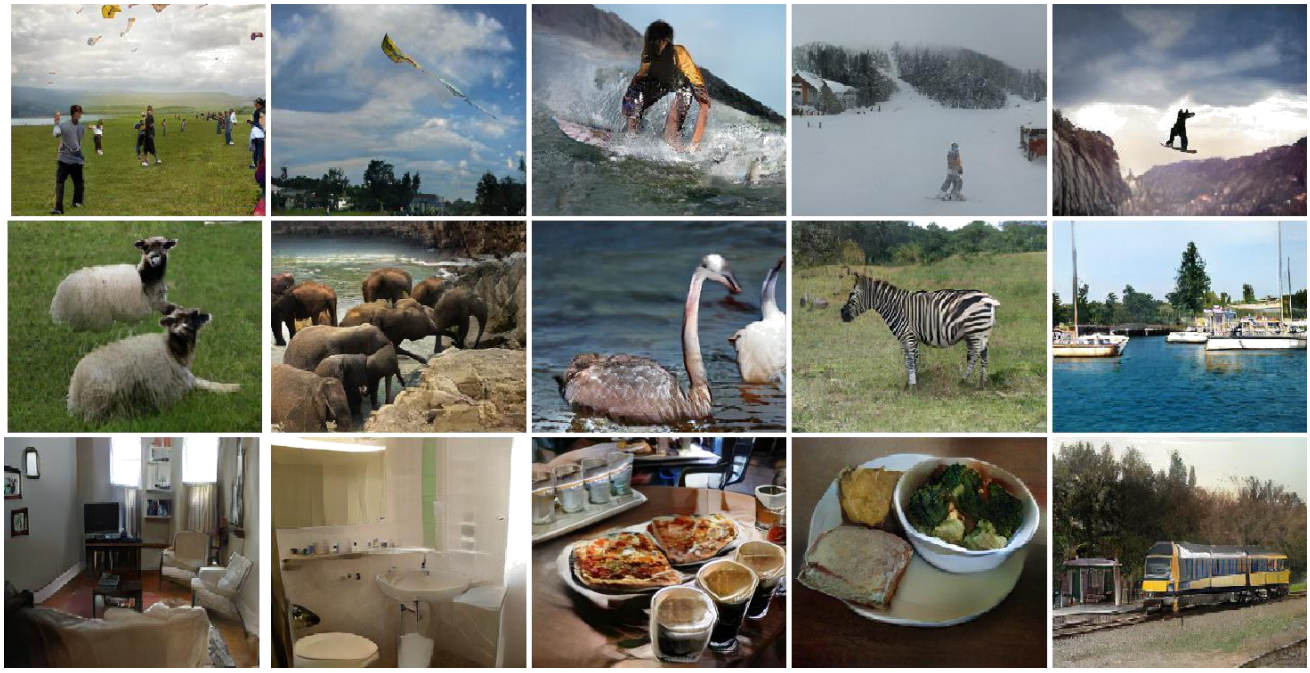

图 8：COCO-Stuff 上的语义图像合成结果。 我们的方法成功地在从动物到体育活动的各种场景中生成逼真的图像。

Method	#param	COCO.	ADE.	City.
decoder w/ SPADE (Ours)	96M	**35.2**	38.5	62.3
compact decoder w/ SPADE	61M	**35.2**	38.0	**62.5**
decoder w/ Concat	79M	31.9	33.6	61.1
pix2pixHD++ w/ SPADE	237M	34.4	**39.0**	62.2
pix2pixHD++ w/ Concat	195M	32.9	38.9	57.1
pix2pixHD++	183M	32.7	38.3	58.8
compact pix2pixHD++	103M	31.6	37.3	57.6
pix2pixHD [40]	183M	14.6	20.3	58.3

表 3：当使用 SPADE 层时，对于 pix2pixHD++的解码器架构（图 4）和编码器 - 解码器架构（我们在 pix2pixHD 上的改进基线 [40]），mIoU 得分得到提升。 另一方面，简单地在每一层连接语义输入是不能这样做的。 此外，我们在所有层上具有较小深度的紧凑型模型优于所有基线。

Method	COCO	ADE20K	Cityscapes
segmap input	35.2	38.5	62.3
random input	35.3	38.3	61.6
kernelsize 5x5	35.0	39.3	61.8
kernelsize 3x3	35.2	38.5	62.3
kernelsize 1x1	32.7	35.9	59.9
#params 141M	35.3	38.3	62.5
#params 96M	35.2	38.5	62.3
#params 61M	35.2	38.0	62.5
Sync Batch Norm	35.0	39.3	61.8
Batch Norm	33.7	37.9	61.8
Instance Norm	33.9	37.4	58.7

表 4：SPADE 生成器使用不同的配置。 我们改变了生成器的输入，作用于分割图的卷积核大小，网络的容量以及无参数归一化方法。 本文中使用的设置以粗体显示。

城市风景，即使所有竞争方法都达到高图像保真度，用户仍然更喜欢我们的结果。

SPADE 的有效性。 为了研究 SPADE 的重要性，我们引入了一个强大的基线 pix2pixHD++，它结合了我们发现的所有技术，除了 SPADE 之外，还有助于提高 pix2pixHD 的性能。 我们还通过在通道方向上串联（pix2pixHD++ w / Concat）来训练在所有中间层接收分段掩码输入的模型。 最后，将强基线与 SPADE 组合的模型表示为 pix2pixHD++ w / SPADE。 另外，我们通过在生成器中使用不同数量的卷积核来比较具有不同容量的模型。

如表 3 所示，在图 4 中描述的解码器式架构和 pix2pixHD 中使用的更传统的编码器 - 解码器架构中，具有所提出的 SPADE 的架构始终优于其对应物。 我们还发现，在所有中间层连接分段掩码，SPADE 提供语义信号的直观替代方法，并没有达到与 SPADE 相同的性能。 此外，解码器式 SPADE 生成器实现了更好的性能

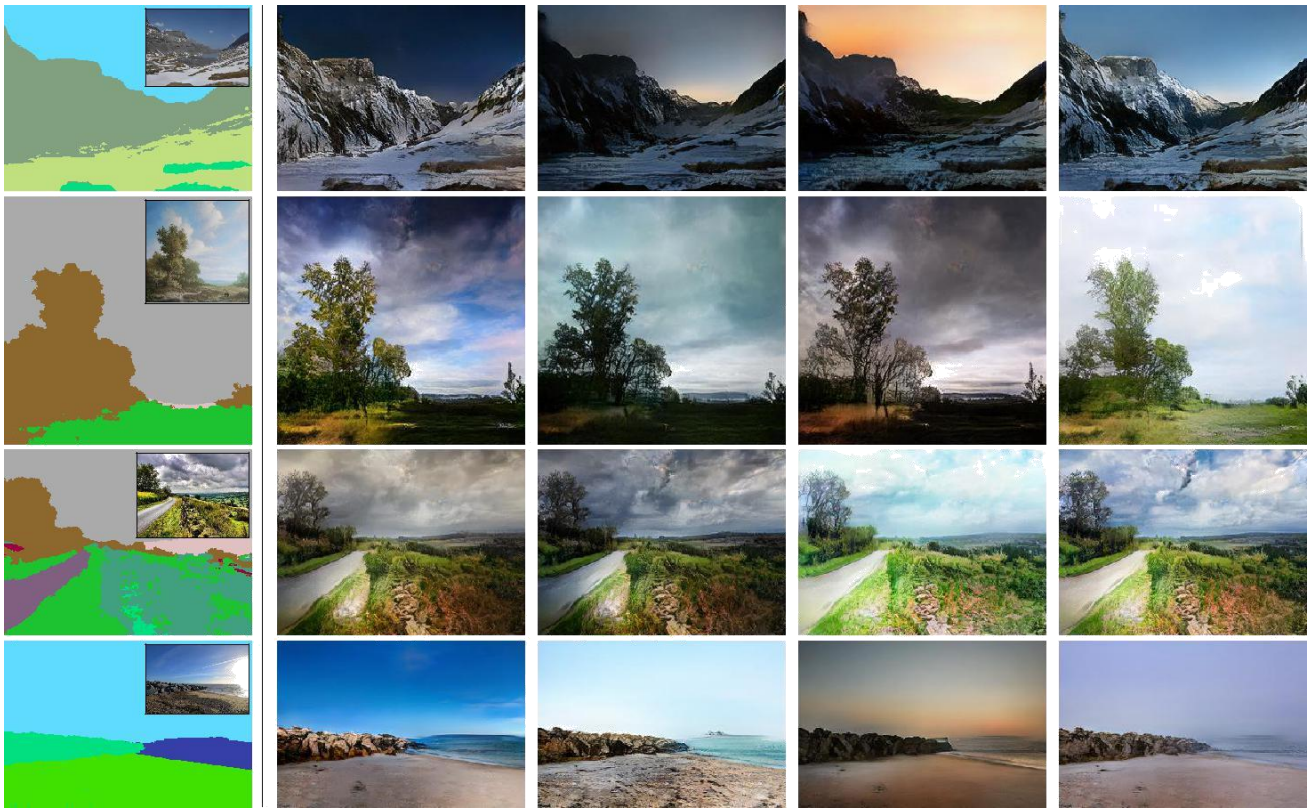

图 9：我们的模型在使用图像编码器进行训练时可获得多模态合成功能。 在部署期间，通过使用不同的随机噪声，我们的模型合成具有不同外观的输出，但都具有输入掩码中描述的相同语义布局。 作为参考，完全真实图像显示在输入分割掩模内。

相比较于需要强大的基线而言，即使使用较少数量的参数。

SPADE 生成器的变化。表 4 报告了我们的生成器变化的性能。首先，我们将两种类型的输入与生成器进行比较：随机噪声或下采样分割图。我们发现两者都具有相似的性能，并得出结论，SPADE 单独的调制提供了关于输入掩模的足够信号。其次，我们在应用调制参数之前改变无参数归一化层的类型。我们观察到 SPADE 在不同的归一化方法中可靠地工作。接下来，我们改变作用于标签映射的卷积内核大小，并发现 1x1 的内核大小会损害性能，可能是因为它禁止使用标签的上下文。最后，我们通过改变卷积核的数量来修改生成器网络的容量。我们在附录中提供了更多变化和消融研究，以便进行更详细的调查。

多模态合成。在图 9 中，我们在 Flickr Landscape 数据集上显示了多模态图像合成结果。对于相同的输入分段掩码，我们采用不同的噪声输入来实现不同的输出。更多结果包含在附录中。

语义操作和引导图像合成。 在图 1 中，我们展示了一个用户绘制不同分割蒙版的应用程序，我们的模型渲染了相应的景观图像。 此外，我们的模型允许用户选择外部样式图像来控制输出图像的全局外观。 我们通过用图像编码器计算的样式图像的嵌入向量替换输入噪声来实现它。

5.结论

我们已经提出了空间自适应归一化，其利用输入语义布局，同时在归一化层中执行仿射变换。 所提出的归一化导致第一语义图像合成模型，其可以为包括室内，室外，风景和街道场景的各种场景产生照片级真实输出。 我们进一步证明了它在多模态合成和引导图像合成中的应用。

致谢 我们感谢 Alexei A. Efros 和 Jan Kautz 提供的富有洞察力的建议。 Taesung Park 在 NVIDIA 实习期间为这项工作做出了贡献。 他的博士学位由三星奖学金支持。

参考文献

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gen-erative adversarial networks. In International Conference on Machine Learning (ICML), 2017. 3

[2] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016. 2

[3] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Gold-man. Patchmatch: A randomized correspondence algorithm for structural image editing. In ACM SIGGRAPH, 2009. 1

[4] A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis. In Inter-national Conference on Learning Representations (ICLR), 2019. 1, 2

[5] H. Caesar, J. Uijlings, and V. Ferrari. Coco-stuff: Thing and stuff classes in context. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 2, 4

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully con-nected crfs. IEEE Transactions on Pattern Analysis and Ma-chine Intelligence (TPAMI), 40(4):834–848, 2018. 4

[7] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In IEEE International Con-ference on Computer Vision (ICCV), 2017. 1, 2, 4, 5, 13, 14, 15, 16, 17, 18

[8] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In IEEE Conference on Computer Vision and Pattern Recog-nition (CVPR), 2016. 2, 4

[9] H. De Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville. Modulating early visual processing by language. In Advances in Neural Information Process-ing Systems (NeurIPS), 2017. 2

[10] V. Dumoulin, J. Shlens, and M. Kudlur. A learned repre-sentation for artistic style. In International Conference on Learning Representations (ICLR), 2016. 2, 3

[11] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intel-ligence and statistics, pages 249–256, 2010. 12, 13

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Gen-erative adversarial nets. In Advances in Neural Information Processing Systems (NeurIPS), 2014. 2

[13] J. Hays and A. A. Efros. Scene completion using millions of photographs. In ACM SIGGRAPH, 2007. 1

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 3

[15] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In Advances in Neural Information Processing Systems (NeurIPS), 2017. 4, 5, 13

[16] S. Hong, D. Yang, J. Choi, and H. Lee. Inferring seman-tic layout for hierarchical text-to-image synthesis. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 2

[17] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In IEEE Inter-national Conference on Computer Vision (ICCV), 2017. 2, 3

[18] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. European Confer-ence on Computer Vision (ECCV), 2018. 2, 3

[19] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International Conference on Machine Learning (ICML), 2015. 2, 3

[20] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In IEEE Conference on Computer Vision and Pattern Recogni-tion (CVPR), 2017. 1, 2, 3, 11, 12

[21] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In International Conference on Learning Rep-resentations (ICLR), 2015. 4

[22] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In International Conference on Learning Representa-tions (ICLR), 2014. 2, 4, 11, 12

[23] A. Kolliopoulos, J. M. Wang, and A. Hertzmann. Segmentation-based 3d artistic rendering. In Rendering Techniques, pages 361–370, 2006. 2

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet clas-sification with deep convolutional neural networks. In Ad-vances in Neural Information Processing Systems (NeurIPS), 2012. 2

[25] J. H. Lim and J. C. Ye. Geometric gan. arXiv preprint arXiv:1705.02894, 2017. 3, 11

[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ra-manan, P. Dollar,´ and C. L. Zitnick. Microsoft coco: Com-mon objects in context. In European Conference on Com-puter Vision (ECCV), 2014. 2, 4

[27] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In Advances in Neural Informa-tion Processing Systems (NeurIPS), 2017. 2

[28] X. Mao, Q. Li, H. Xie, Y. R. Lau, Z. Wang, and S. P. Smol-ley. Least squares generative adversarial networks. In IEEE International Conference on Computer Vision (ICCV), 2017. 3, 11

[29] L. Mescheder, A. Geiger, and S. Nowozin. Which training methods for gans do actually converge? In International Conference on Machine Learning (ICML), 2018. 2, 3, 11

[30] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spec-tral normalization for generative adversarial networks. In In-ternational Conference on Learning Representations (ICLR), 2018. 3, 4, 11

[31] T. Miyato and M. Koyama. cGANs with projection discrim-inator. In International Conference on Learning Representa-tions (ICLR), 2018. 2, 3, 11

[32] K. Nakashima. Deeplab-pytorch. https://github.com/kazuto1011/deeplab-pytorch, 2018. 4

[33] A. Odena, C. Olah, and J. Shlens. Conditional image synthe-sis with auxiliary classifier GANs. In International Confer-ence on Machine Learning (ICML), 2017. 2

[34] E. Perez, H. De Vries, F. Strub, V. Dumoulin, and A. Courville. Learning visual reasoning without strong priors. In International Conference on Machine Learning (ICML), 2017. 2

[35] X. Qi, Q. Chen, J. Jia, and V. Koltun. Semi-parametric image synthesis. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 4, 5, 13, 17, 18

[36] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In In-ternational Conference on Machine Learning (ICML), 2016. 2

[37] T. Salimans and D. P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neu-ral networks. In Advances in Neural Information Processing Systems (NeurIPS), 2016. 2

[38] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. arxiv 2016. arXiv preprint arXiv:1607.08022, 2016. 2, 3

[39] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. In Advances in Neural Information Processing Systems (NeurIPS), 2018. 1, 4

[40] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 1, 2, 3, 4, 5, 7, 11, 12, 13, 14, 15, 16, 17, 18

[41] Y. Wu and K. He. Group normalization. In European Con-ference on Computer Vision (ECCV), 2018. 2

[42] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun. Unified perceptual parsing for scene understanding. In European Conference on Computer Vision (ECCV), 2018. 4

[43] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In IEEE Confer-ence on Computer Vision and Pattern Recognition (CVPR), 2018. 2

[44] F. Yu, V. Koltun, and T. Funkhouser. Dilated residual net-works. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 4

[45] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. arXiv preprint arXiv:1805.08318, 2018. 1, 2, 3, 11

[46] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. Stackgan: Text to photo-realistic image synthe-sis with stacked generative adversarial networks. In IEEE International Conference on Computer Vision (ICCV), 2017. 1, 2

[47] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. IEEE Transac-tions on Pattern Analysis and Machine Intelligence (TPAMI), 2018. 1

[48] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In IEEE Conference on Computer Vision and Pattern Recog-nition (CVPR), 2017. 2, 4

[49] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial net-works. In IEEE International Conference on Computer Vi-sion (ICCV), 2017. 2

[50] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. Toward multimodal image-to-image translation. In Advances in Neural Information Pro-cessing Systems (NeurIPS), 2017. 2, 3

A. 更详细的实现细节

生成器。 生成器的体系结构由一系列提出的 SPADE ResBlks 组成，具有最近邻上采样。 我们同时使用 8 个 GPU 训练我们的网络，并使用批量归一化的同步版本。 我们将谱归一化[30]应用于生成器中的所有卷积层。 提出的 SPADE 和 SPADE ResBlk 的架构分别在图 10 和图 11 中给出。 生成器的架构如图 12 所示。

判别器。 判别器的体系结构遵循 pix2pixHD 方法[40]中使用的结构，该方法使用具有实例归一化（IN）的多尺度设计。 唯一的区别是我们将谱归一化应用于判别器的所有卷积层中。

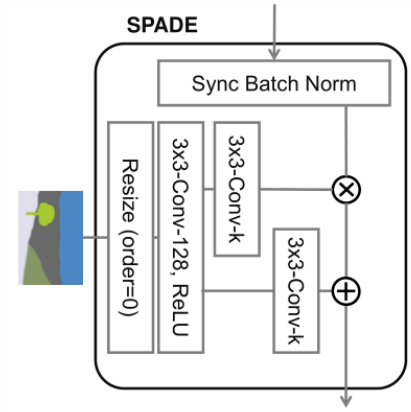

图 10：SPADE 设计。 术语 3×3-Conv-k 表示具有 k 个卷积核的 3×3 卷积层。 调整分割图的大小以匹配使用最近邻下采样的对应特征图的分辨率。

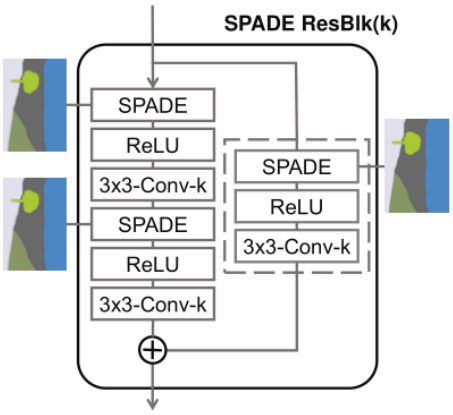

图 11：SPADE ResBlk。 残差块设计很大程度上遵循[29]和[31]中的设计。 我们注意到，对于残差块之前和之后的通道数量不同的情况，也学习了跳过连接（图中的虚线框）。

判别器体系结构的细节如图 13 所示。

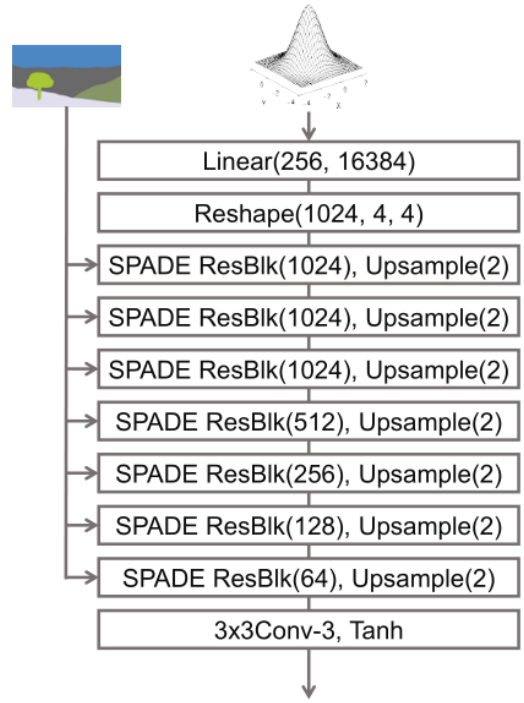

图 12：SPADE 生成器。 与现有的图像生成器[20,40]不同，语义分割掩模通过图 11 中提出的 SPADE ResBlks 传递给生成器。

图像编码器。 图像编码器由 6 个步幅-2 的卷积层和 2 个线性层组成，以产生输出分布的均值和方差，如图 14 所示。

学习目标。 我们在 pix2pixHD 工作中使用学习目标函数[40]，除了我们用铰链损失项[25,30,45]替换它的 LS-GAN 损失项[28]。 我们在目标函数中使用与 pix2pixHD 工作中相同的权重加权。

当使用用于多模态合成和样式引导图像合成的图像编码器训练所提出的框架时，我们包括 KL 散度损失：

$$\mathcal{L}_{\mathrm{KLD}} = \mathcal{D}_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$$

其中先验分布 p（z）是标准高斯分布，变分分布 q 由均值向量和方差向量完全确定[22]。 我们使用重新伪造技巧[22] 将梯度从生成器进行反向传播到图像编码器。KL 散度损失的权重为 0.05。

在图 15 中，我们概述了训练数据流。 图像编码器将实像编码为均值向量和方差向量，它们用于通过重参数化技巧计算生成器的噪声[22]。

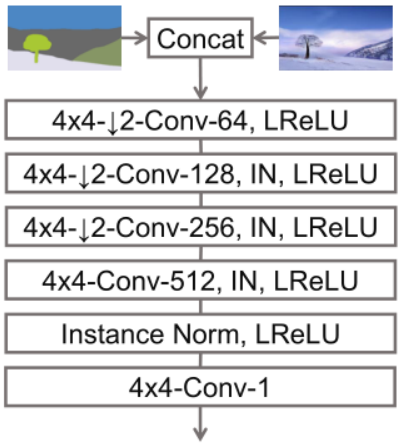

图 13：我们的判别器设计很大程度上遵循 pix2pixHD [40]。 它将分割图和图像作为输入连接起来。 它基于 Patch-GAN [20]。 因此，判别器的最后一层是卷积层。

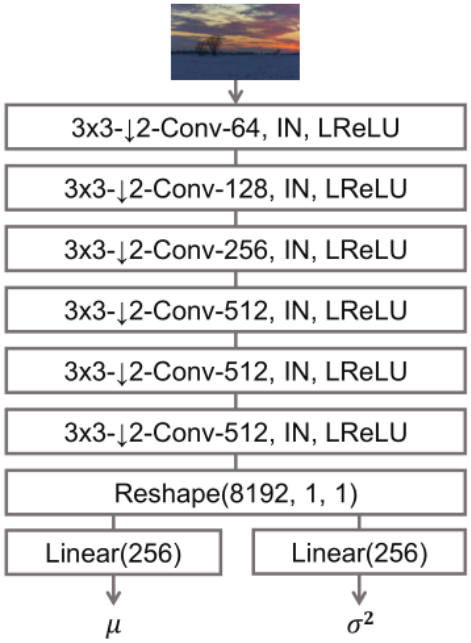

图 14：图像编码器由一系列卷积层组成，其中步长 2 后跟两个线性层，输出平均向量μ和方差向量σ。

生成器还采用了分割掩码关于

使用建议的 SPADE ResBlks 输入图像作为输入。判别器将分割掩码和来自生成器的输出图像串联作为输入，并且旨在将其分类为伪造的。

训练细节。 我们对 Cityscapes 和 ADE20K 数据集进行了 200 个时期的训练，对 COCO-Stuff 数据集进行了 100 个时期的训练，并对 Flickr Landscapes 数据集进行了 50 个时期的训练。 图像尺寸为 256x256，城市景观除为 512x256 外。 对于 Cityscapes 和 ADE20K 数据集，我们将学习率从 epoch 100 到 epoch 200 线性衰减到 0。 批量大小为 32。我们使用 Glorot 初始化方法[11]初始化网络权重。

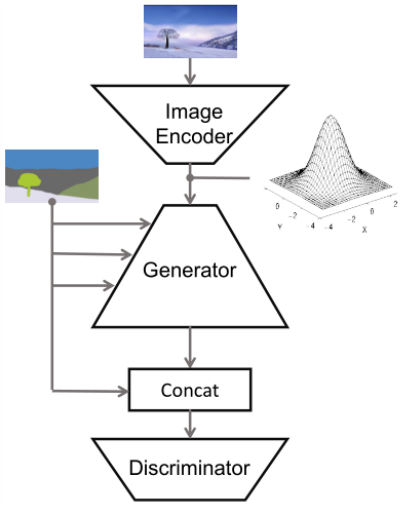

图 15：图像编码器将实像编码为潜在表示，以生成平均向量和方差向量。 它们用于通过重参数化技巧计算生成器的噪声输入[22]。 生成器还通过提议的 SPADE ResBlks 将输入图像的分段掩码作为输入。 判别器将分割掩码和来自生成器的输出图像串联作为输入，并且旨在将其分类为伪造的。

B. 额外的消融研究

Method	COCO.	ADE.	City.
Ours	35.2	38.5	62.3
Ours w/o Perceptual loss	24.7	30.1	57.4
Ours w/o GAN feature matching loss	33.2	38.0	62.2
Ours w/ a deeper discriminator	34.9	38.3	60.9
pix2pixHD++ w/ SPADE	34.4	39.0	62.2
pix2pixHD++	32.7	38.3	58.8
pix2pixHD++ w/o Sync Batch Norm	27.4	31.8	51.1
pix2pixHD++ w/o Sync Batch Norm, and w/o Spectral Norm	26.0	31.9	52.3
pix2pixHD [40]	14.6	20.3	58.3

表 5：关于 mIoU 评分的额外消融研究结果：该表显示感知损失和 GAN 特征匹配损失项是重要的。 使判别器更深不会导致性能提升。 该表还显示，所提出的方法中使用的组件（同步批量归一化，谱归一化，TTUR，铰链损失和 SPADE）也有助于我们强大的基线 pix2pixHD++。

表 5 提供了额外的消融研究结果，分析了提出的方法中各个组分的贡献。 我们首先发现从 pix2pixHD [40]的学习目标函数继承的感知损失和 GAN 特征匹配损失都是重要的。 删除其中任何一个都会导致性能下降。 我们还发现，通过在 pix2pixHD 判别器的顶部插入一个以上的卷积层来增加判别器的深度不会导致性能提升。

在表 5 中，我们还分析了我们的强基线中使用的每个组件的有效性，pix2pixHD++方法，源自 pix2pixHD 方法。我们发现谱范数，同步批次范数，TTUR [15]和铰链损失都有助于提高性能。 但是，通过将 SPADE 添加到强基线，性能会进一步提高。 注意 pix2pixHD++ w/o Sync Batch Norm 和 w/o Spectral Norm 仍然与 pix2pixHD 不同，在于它使用铰链损耗，TTUR，大批量和 Glorot 初始化[11]。

C. 更多结果

在图 16,17 和 18 中，我们展示了 COCO-Stuff 和 ADE20K 数据集上提出的方法的其他合成结果，并与 CRN [7]和 pix2pixHD [40]方法进行了比较。

在图 19 和 20 中，我们在 ADE20K-outdoor 和 Cityscapes 数据集上显示了所提方法的其他综合结果，并与 CRN [7]，SIMS [35]和 pix2pixHD [40]方法进行了比较。

在图 21 中，我们显示了所提出方法的其他多模态合成结果。作为从标准多变量高斯分布中采样不同的 z，我们合成了不同外观的图像。

在随附的视频中，我们展示了我们的语义图像合成界面。我们展示了用户如何通过在画布上绘制语义标签来创建逼真的风景图像。我们还展示了用户如何针对相同的语义分割掩模合成大小不同外观的图像，以及将所提供的样式图像的外观转移到合成的样式图像。

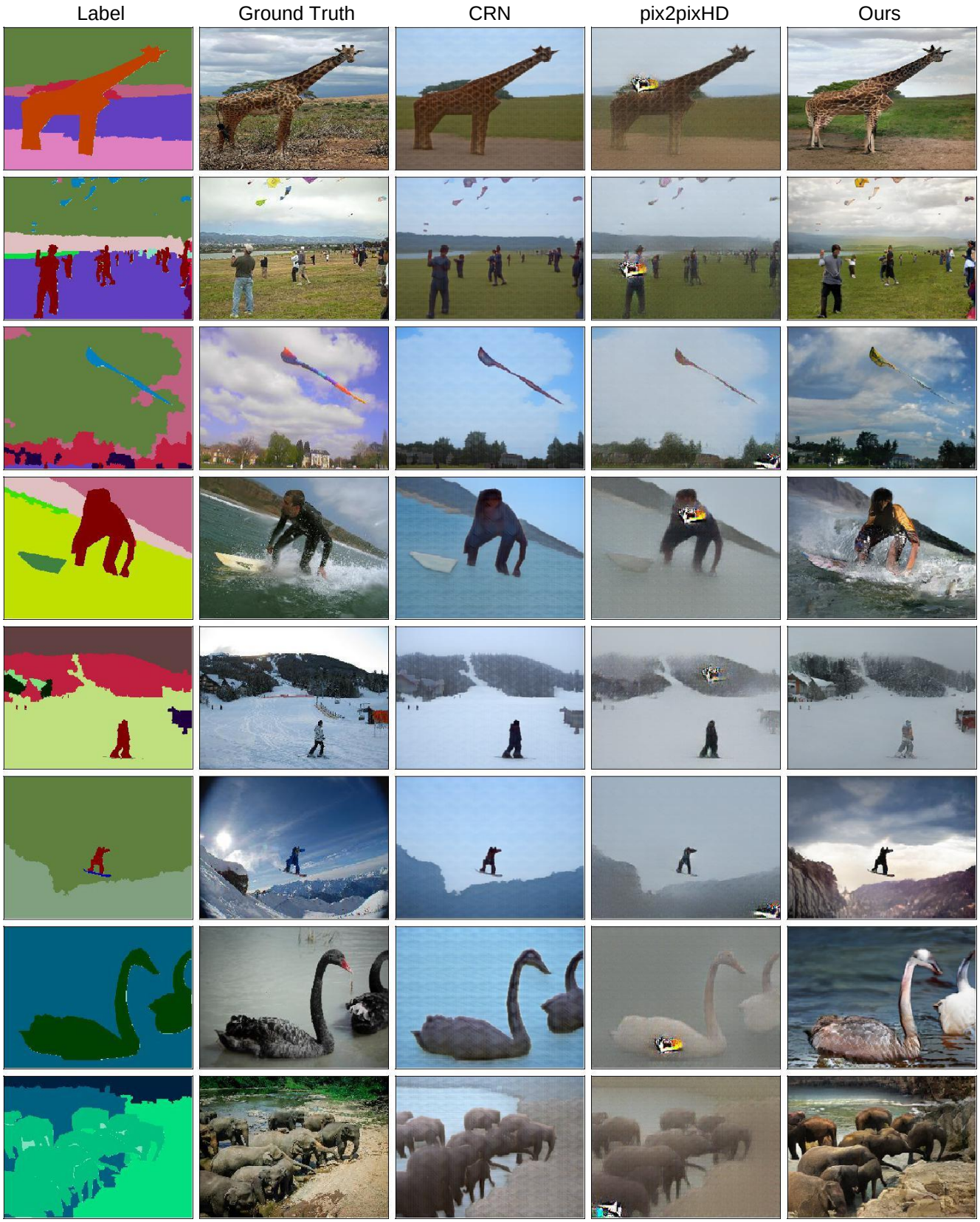

图 16：与 COCO-Stuff 数据集上的 CRN [7]和 pix2pixHD [40]方法的结果进行比较的附加结果。

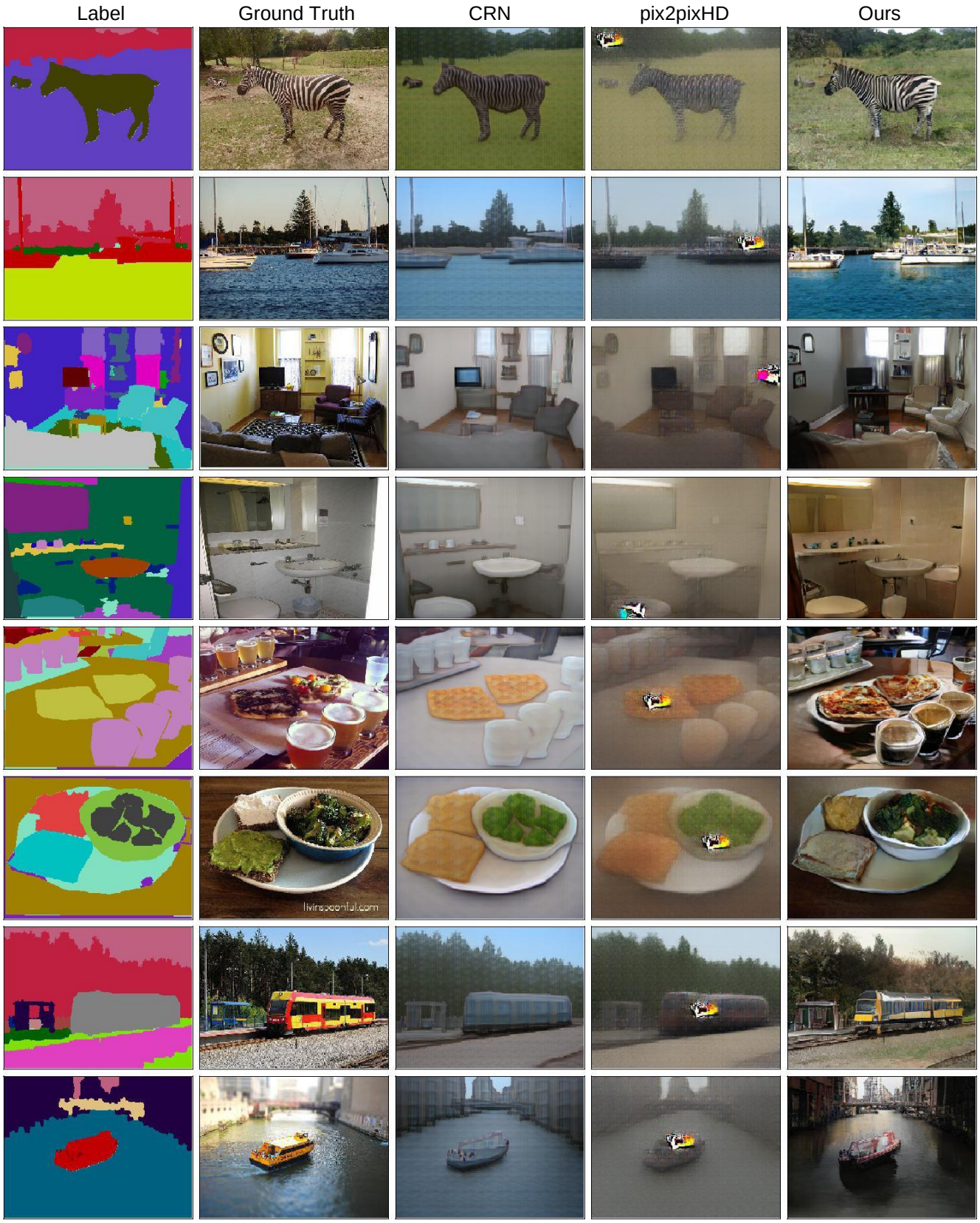

图 17：与 COCO-Stuff 数据集上的 CRN [7]和 pix2pixHD [40]方法的结果进行比较的其他结果。

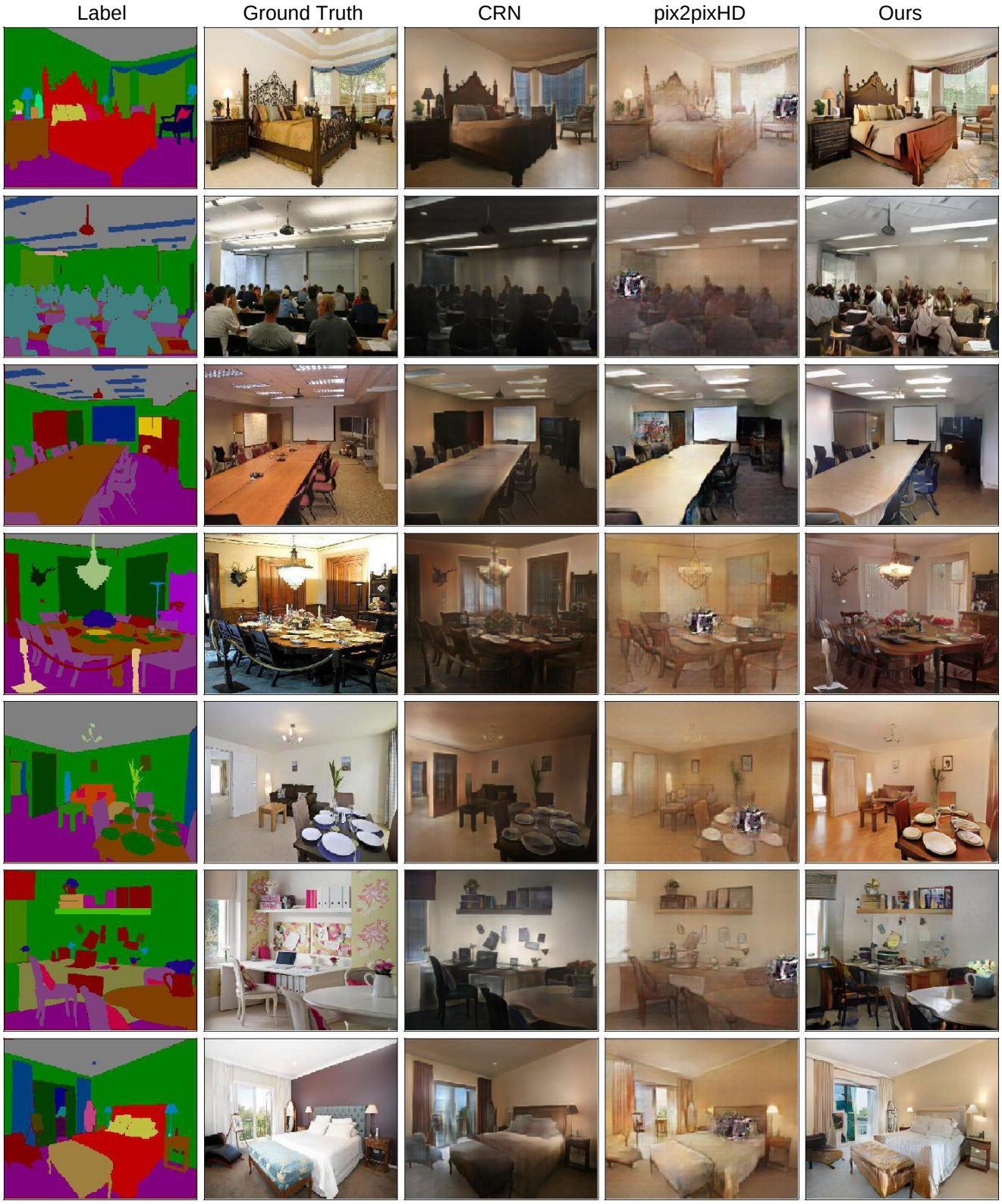

图 18：与 ADE20K 数据集上 CRN [7]和 pix2pixHD [40]方法的结果进行比较的其他结果。

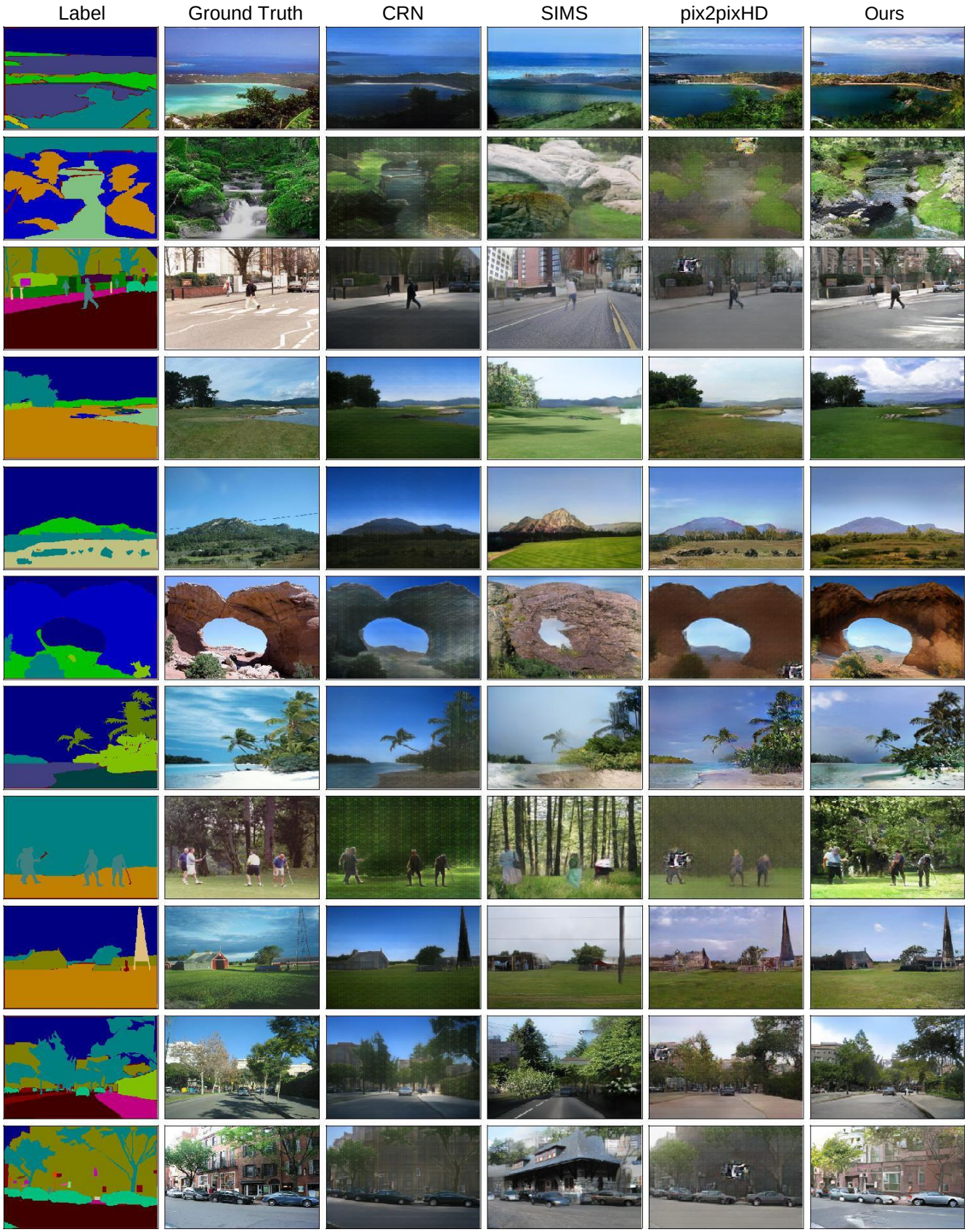

图 19：与 ADE20K 室外数据集上 CRN [7]，SIMS [35]和 pix2pixHD [40]方法的结果进行比较的其他结果。

图 20：与 Cityscapes 数据集上的 CRN [7]，SIMS [35]和 pix2pixHD [40]方法的结果进行比较的其他结果。

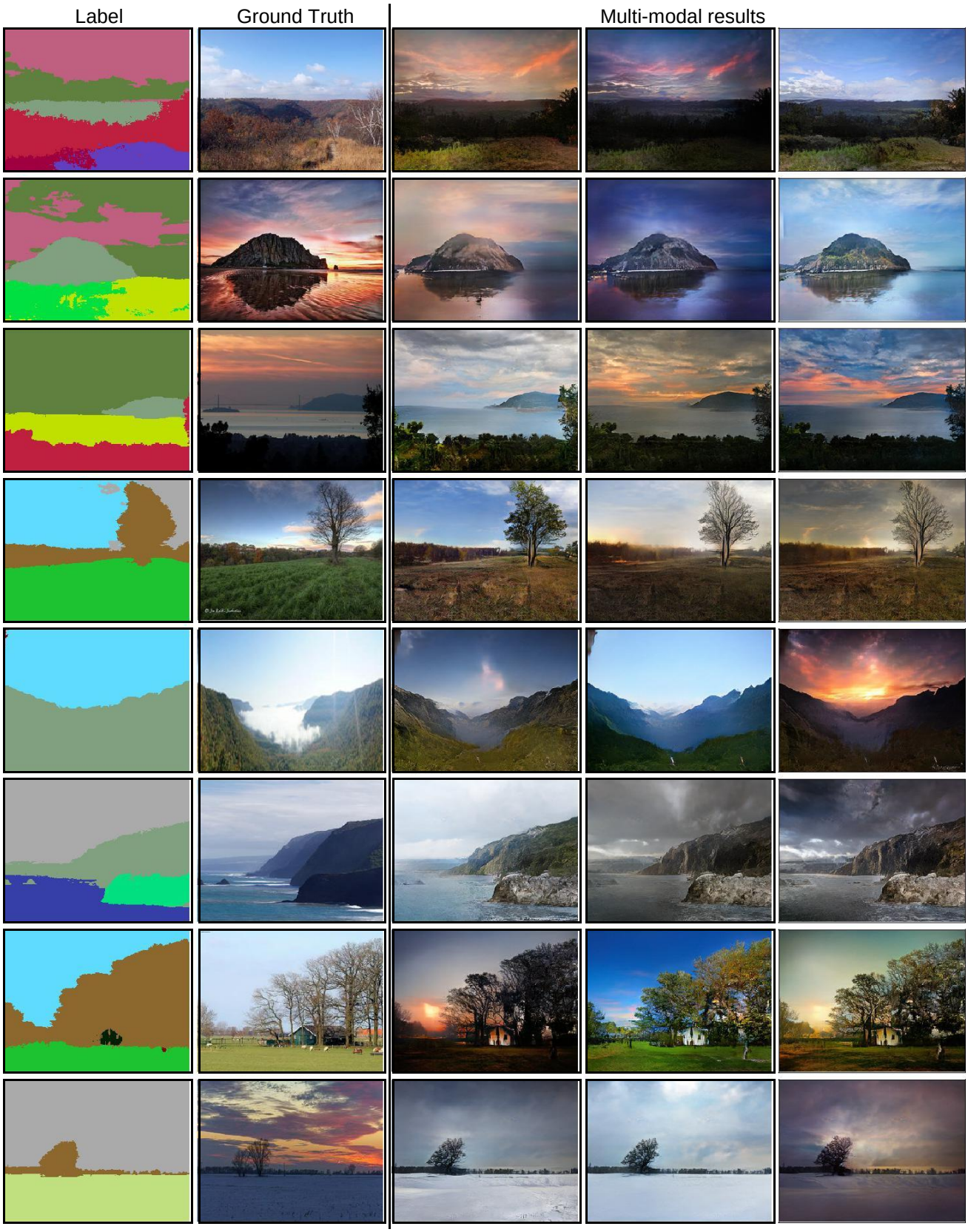

图 21：Flickr Landscapes Dataset 上的附加多模态合成结果。 通过从标准高斯分布中采样潜在向量，我们合成了不同外观的图像。