少样本视频到视频合成

Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, Bryan Catanzaro
NVIDIA Corporation
{tingchunw,mingyul,atao,guilinl,jkautz,bcatanzaro}@nvidia.com

摘 要

视频到视频合成（vid2vid）旨在将输入的语义视频（例如人体姿势或分割蒙版的视频）转换成输出为具有真实感的视频。尽管 vid2vid 的最新技术已经取得了显著进步，但现有方法共有两个主要局限性。首先，它们需要大量数据。训练需要目标人体或场景的大量图像。第二，学习的模型具有有限的泛化能力。姿势到人的 vid2vid 模型只能合成训练集中单人的姿势，它不能推广到不在训练集中的其他人。为了解决这些局限性，我们提出了一个简单的 vid2vid 框架，该框架通过在测试时利用目标的少量示例图像来学习合成以前没见过的主题或场景的视频。我们的模型通过利用注意力机制的新型网络权重生成模块实现了这种快速的泛化能力。我们进行了广泛的实验验证，并使用几个大型视频数据集（包括人类跳舞视频，头部说话视频和街头现场视频）与强基准进行了比较。实验结果验证了所提出框架在解决现有 vid2vid 方法的两个局限性方面的有效性。代码可在我们的网站获得。

1 介绍

视频到视频合成（vid2vid）是指将输入语义视频转换为输出具有真实感视频的任务。它具有广泛的应用，包括使用人体姿势序列[7、12、57、67]生成人类舞蹈视频，或使用分段蒙版序列[57]生成驾驶视频。通常，要获得这样的模型，首先要收集目标任务的训练数据集。它可以是目标人执行各种动作的一组视频，也可以是通过使用安装在城市中行驶的汽车上的摄像头捕获的一组街头现场视频。然后，该数据集用于训练一个模型，该模型在测试时将全新的输入语义视频转换为相应的真实感视频。换句话说，我们希望针对人类的 vid2vid 模型可以生成同一个人执行某种行为的视频，而这些视频不在训练集中，而街头场景 vid2vid 模型可以依据新型视频的视频场景提供与训练场景相同的视频。随着生成对抗网络（GAN）框架[13]及其图像条件扩展[22，58]的发展，现有的 vid2vid 方法已显示出令人鼓舞的结果。

我们认为，推广到新颖的输入语义视频是不够的。还应该针对一种可以推广到未见过的领域的模型，例如生成训练数据集中未包含的人类主题的视频。更为理想的是，vid2vid 模型应该能够通过仅利用测试时给出的一些示例图像来合成未见过域的视频。如果 vid2vid 模型无法推广到未见过的人或场景样式，则我们必须为每种新的主题或场景样式训练一个模型。此外，如果 vid2vid 模型仅使用少数几个示例图像就无法实现这一领域的泛化能力，则必须为每种新主题或场景样式收集许多图像。这将使模型难以扩展。不幸的是，由于现有的 vid2vid 方法不考虑这种拓展，因此存在这些缺点。

第 33 届神经信息处理系统会议（NeurIPS 2019），加拿大温哥华。

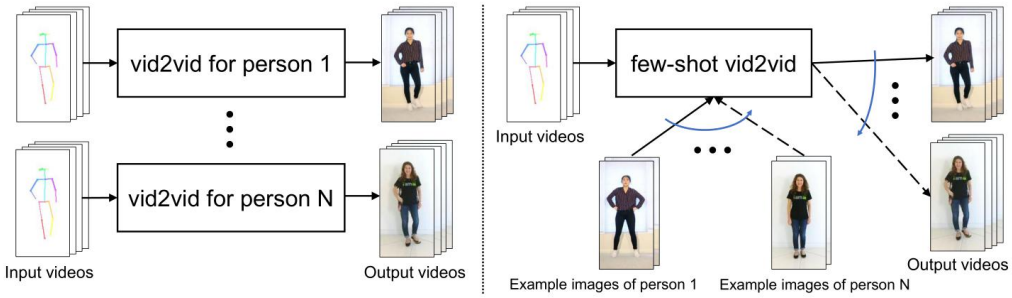


图 1：vid2vid（左）与提出的少样本 vid2vid（右）之间的比较。 现有的 vid2vid 方法[7，12，57]不考虑泛化到未见过的域。经过训练的模型只能用于合成与训练集中的视频相似的视频。例如，vid2vid 模型只能用于生成训练集中人的视频。 为了合成一个新人，需要收集新人的数据集，并使用它来训练新的 vid2vid 模型。 另一方面，我们的少样本 vid2vid 模型没有局限性。 我们的模型可以利用测试时提供的少量示例图像来合成新人的视频。

为了解决这些局限性，我们提出了少样本的 vid2vid 框架。少样本的 vid2vid 框架需要两个输入来生成视频，如图 1 所示。在测试阶段，除了 vid2vid 中的输入语义视频外，它还需要第二个输入，其中包括一些可用的目标域示例图像。注意，在现有的 vid2vid 方法中[7、12、57、67]不存在这种情况。我们的模型使用这几个示例图像通过新颖的网络权重生成机制动态配置视频合成机制。具体来说，我们训练一个模块来依据示例图像生成网络权重。我们精心设计了学习的目标函数，以方便学习网络权重生成模块。

我们使用几个大型视频数据集（包括舞蹈视频，头部说话视频和街头现场视频），与各种基线方法进行比较，进行了广泛的实验验证。实验结果表明，该方法有效地解决了现有 vid2vid 框架的局限性。此外，我们证明了我们模型的性能与训练数据集中视频的多样性以及在测试时可用的示例图像的数量呈正相关。当模型在训练时间内看到更多不同的域时，可以更好地推广处理未见到的域（图 7（a））。当在测试时为模型提供更多示例图像时，合成视频的质量会提高（图 7（b））。

2 相关工作

GANs。提出的少样本 vid2vid 模型基于 GAN [13]。具体来说，我们使用条件 GAN 框架。我们不是通过转换来自某些噪声分布的样本来生成输出[13、42、32、14、25]，而是根据用户输入数据生成输出，从而可以更灵活地控制输出。用户输入数据可以采用各种形式，包括图像[22、68、30、41]，类别标签[39、35、65、4]，文本描述[43、66、62]和视频[7、12， 57，67]。我们的模型属于最后一个。但是，与以视频为唯一数据输入的现有视频条件 GAN 不同，我们的模型还采用了一组示例图像。这些示例图像在测试时提供，我们使用它们通过新颖的网络权重生成模块动态确定视频合成模型的网络权重。这有助于网络生成未见过的域的视频。

图像到图像合成，将输入图像从一个域转移到另一个域中的对应图像[22、50、3、46、68、30、21、69、58、8、41、31、2]，是 vid2vid 的基础。对于视频而言，新的挑战在于生成基于帧的序列图，这些帧序列不仅要具有真实感，而且总体上要与时间保持吻合。最近，提出的 FUNIT [31]通过自适应实例规一化技术[19]生成未见过的域的图像。我们的工作有所不同，我们的目标是视频合成，并通过网络权重生成方案将其推广到未见过的域。我们在实验部分比较这些技术。

视频生成模型可分为三大类，包括 1）无条件视频合成模型[54、45、51]，该模型将随机噪声样本转换为视频片段； 2）未来视频预测模型[48、24、11、34， 33、63、55、56、10、53、29、27、18、28、16、40]，它们基于观察到的视频帧生成未来的视频帧，以及 3）vid2vid 模型[57、7、12、67] ，它将语义输入视频转换为具有真实感的视频。我们的工作属于最后一类，但是

与先前的作品相比，我们的目标是一个 vid2vid 模型，该模型可以通过利用测试时提供的少量示例图像来合成未见过域的视频。

自适应网络是指根据输入数据动态计算部分权重的网络。此类网络与常规网络具有不同的归纳偏置，并已用于多种任务中，包括序列建模[15]，图像过滤[23、59、49]，帧插值[38、37]和神经结构搜索[64]。在这里，我们将其应用于 vid2vid 任务。

人体姿势转移通过利用处于不同姿势的人的图像来合成处于未见过的姿势的人。为了获得高质量的生成结果，现有的人体姿势转移方法在很大程度上利用了人体先验技术，例如人体部位建模[1]或基于人体表面的坐标映射[36]。我们的工作与这些工作的不同之处在于我们的方法更为通用。除了输入的语义视频外，我们不使用特定的人体先验。结果，相同的模型可以直接用于其他 vid2vid 任务，例如街景视频合成，如图 5 所示。此外，我们的模型是为视频合成而设计的，而现有的人体姿势转移方法主要是为静止图像设计的，合成时不考虑时间方面的问题。结果，我们的方法呈现了时间上更一致的结果（图 4）。

3 少样本的视频到视频合成

视频到视频合成旨在学习可以转换语义图像序列的映射功能（见底部 1）：例如输入图像序列 $\mathbf{s}_1^T \equiv \mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_T$，到输出图像序列$\tilde{\mathbf{x}}_1^T \equiv \tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, ..., \tilde{\mathbf{x}}_T$，在给定$\mathbf{s}_1^T$时$\tilde{\mathbf{x}}_1^T$的条件分布类似于给定$\mathbf{s}_1^T$时完全真实图像序列$\mathbf{x}_1^T \equiv \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T$的条件分布。换句话说，它旨在实现$\mathcal{D}(p(\tilde{\mathbf{x}}_1^T|\mathbf{s}_1^T), p(\mathbf{x}_1^T|\mathbf{s}_1^T)) \to 0$，其中 D 是分布散度度量，例如 Jensen-Shannon 散度或 Wasserstein 距离。为了对条件分布进行建模，现有工作做了一个简化的马尔可夫假设，从而得出了序列生成模型：

$$\tilde{\mathbf{x}}_t = F(\tilde{\mathbf{x}}_{t-\tau}^{t-1}, \mathbf{s}_{t-\tau}^t) \tag{1}$$

换句话说，它基于观察到的τ+1 个输入语义图像$\mathbf{s}_{t-\tau}^t$和过去τ生成的图像$\tilde{\mathbf{x}}_{t-\tau}^{t-1}$来生成输出图像$\tilde{\mathbf{x}}_t$。 可以用几种不同的方式对序列生成器 F 进行建模[7、12、57、67]。 一个流行的选择是使用

$$F(\tilde{\mathbf{x}}_{t-\tau}^{t-1}, \mathbf{s}_{t-\tau}^t) = (\mathbf{1} - \tilde{\mathbf{m}}_t) \odot \tilde{\mathbf{w}}_{t-1}(\tilde{\mathbf{x}}_{t-1}) + \tilde{\mathbf{m}}_t \odot \tilde{\mathbf{h}}_t \tag{2}$$

其中符号 **1** 是所有图像，⊙是逐元素乘积运算符，$\tilde{\mathbf{m}}_t$是柔和遮挡图，$\tilde{\mathbf{w}}_{t-1}$是从 t - 1 到 t 的光流，$\tilde{\mathbf{h}}_t$是合成的中间图像。

图 2（a）展示了 vid2vid 架构和消光功能，该图显示了输出图像$\tilde{\mathbf{x}}_t$是通过光流扭曲方式组合最后生成的图像，$\tilde{\mathbf{w}}_{t-1}(\tilde{\mathbf{x}}_{t-1})$，以及合成的中间图像$\tilde{\mathbf{h}}_t$所生成的。柔和遮挡图$\tilde{\mathbf{m}}_t$指示这两个图像在每个像素位置如何组合。直观地，如果在先前生成的帧中观察到像素，则将有利于从变形图像中复制像素值。 实际上，这些变量是通过神经网络的参数化函数 M，W 和 H 生成的：

$$\tilde{\mathbf{m}}_t = M_{\boldsymbol{\theta}_M}(\tilde{\mathbf{x}}_{t-\tau}^{t-1}, \mathbf{s}_{t-\tau}^t), \tag{3}$$

$$\tilde{\mathbf{w}}_{t-1} = W_{\boldsymbol{\theta}_W}(\tilde{\mathbf{x}}_{t-\tau}^{t-1}, \mathbf{s}_{t-\tau}^t), \tag{4}$$

$$\tilde{\mathbf{h}}_t = H_{\boldsymbol{\theta}_H}(\tilde{\mathbf{x}}_{t-\tau}^{t-1}, \mathbf{s}_{t-\tau}^t) \tag{5}$$

其中 M，W 和 H 是可学习的参数。 训练完成后，它们将保持固定。

少样本的 vid2vid。 虽然训练了（1）中的序列生成器用于转换新颖的输入语义视频，但并未训练其用于合成未见过域的视频。例如，针对特定人员训练的模型只能用于生成同一人员的视频。为了使 F 适用于未见过的域，我们让 F 依赖于额外的输入。具体来说，我们让 F 接受另外两个输入参数：一个是一组 K 个来自目标域的示例图像$\{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_K\}$；另一个是一组它们对应的关联语义图像$\{\mathbf{s}_{\mathbf{e}_1}, \mathbf{s}_{\mathbf{e}_2}, ..., \mathbf{s}_{\mathbf{e}_K}\}$。也就是

$$\tilde{\mathbf{x}}_t = F(\tilde{\mathbf{x}}_{t-\tau}^{t-1}, \mathbf{s}_{t-\tau}^t, \{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_K\}, \{\mathbf{s}_{\mathbf{e}_1}, \mathbf{s}_{\mathbf{e}_2}, ..., \mathbf{s}_{\mathbf{e}_K}\}). \tag{6}$$

[1] 例如，分割蒙版或表示人的姿势的图像。

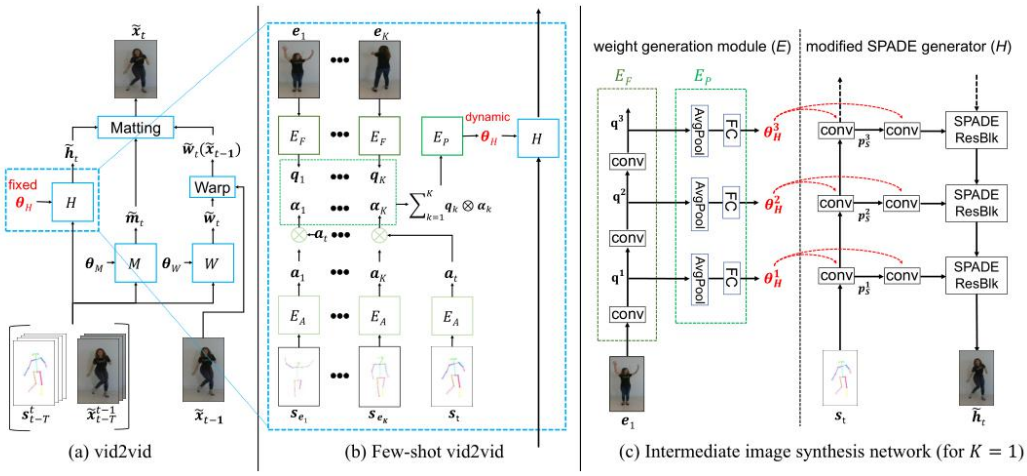


图 2：（a）vid2vid 框架的体系结构[57]。（b）提议的少样本 vid2vid 框架的体系结构。它由网络权重生成模块 E 组成，该模块将示例图像映射到网络权重的一部分以进行视频合成。模块 E 由三个子网组成：E_F，E_P和E_A（当 K> 1 时使用）。子网E_F从示例图像中提取特征 q。当存在多个示例图像（K> 1）时，E_A通过估计柔和注意力图α并平均加权不同的特征来组合提取出的特征。然后将最终表示输入到网络E_P中，以生成图像合成网络 H 的权重$\boldsymbol{\theta}_H$。

这种建模允许 F 利用测试时提供的示例图像来提取一些有用的模式，以合成未见过的域的视频。我们提出了一种网络权重生成模块 E，用于提取模式。具体来说，E 设计为从提供的示例图像中提取模式，并使用它们来计算中间图像合成网络 H 的网络权重$\boldsymbol{\theta}_H$：

$$\boldsymbol{\theta}_H = E(\tilde{\mathbf{x}}_{t-\tau}^{t-1}, \mathbf{s}_{t-\tau}^{t}, \{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_K\}, \{\mathbf{s}_{\mathbf{e}_1}, \mathbf{s}_{\mathbf{e}_2}, ..., \mathbf{s}_{\mathbf{e}_K}\}). \tag{7}$$

注意，网络 E 不生成权重$\boldsymbol{\theta}_M$或$\boldsymbol{\theta}_W$，因为光流预测网络 W 和柔和遮挡图预测网络 W 被设计用于使最后生成的图像变形，并且变形是在域之间自然共享的机制。

我们基于 Wang 等人[57]构建了少样本 vid2vid 框架，这是 vid2vid 任务的最新技术。具体而言，我们重用了他们提出的光流预测网络 W 和柔和遮挡图预测网络 M。中间图像合成网络 H 是条件图像生成器。我们没有采用 Wang 等人提出的架构[57]，而是采用了 SPADE 生成器[41]，它是当前最先进的语义图像合成模型。

SPADE 生成器包含几个空间调制分支和一个主图像合成分支。我们的网络权重生成模块 E 仅生成空间调制分支的权重。这有两个主要优点。首先，它大大减少了 E 必须生成的参数数量，从而避免了过拟合问题。其次，它避免了从示例图像到输出图像的快捷通道，因为所生成的权重仅在空间调制模块中使用，空间调制模块会为主图像合成分支生成调制值。接下来，我们讨论网络 E 的设计细节和学习目标。

网络权重生成模块。如上所述，网络权重生成模块 E 的目标是学习通过控制权重来提取可注入视频合成分支的外观模式。我们首先考虑只有一个示例图像可用的情况（K = 1）。然后，我们扩展讨论以处理多个示例图像的情况。

我们将 E 分解为两个子网：一个示例特征提取器E_F和一个多层感知器E_P。网络E_F由几个卷积层组成，并应用于示例图像$\mathbf{e}_1$上以提取外观表示 q。然后将表示 q 馈入E_P，以在中间图像合成网络 H 中生成权重$\boldsymbol{\theta}_H$。

设置图像合成网络 H 具有 L 层，每层记为H^l，其中$l \in [1, L]$。我们将权重生成网络 E 设计为也具有 L 层，每个E^l生成对应的H^l的权重。具体地，为了产生层H^l的权重$\boldsymbol{\theta}_H^l$，我们首先获取E_F中第 l 层的输出$\mathbf{q}^l$。

然后，我们将$\mathbf{q}^l$平均池化（因为$\mathbf{q}^l$可能仍然是具有空间维度的特征图），然后应用多层感知器E_P^l生成权重$\boldsymbol{\theta}_H^l$。从数学上讲，如果我们定义$\mathbf{q}^0 \equiv \mathbf{e}_1$，则$\mathbf{q}^l = E_F^l(\mathbf{q}^{l-1})$，并且$\boldsymbol{\theta}_H^l = E_P^l(\mathbf{q}^l)$。然后，将这些生成的权重用于卷积当前输入的语义图$\mathbf{s}_t$，以生成 SPADE 中使用的规范化参数（图 2（c））。

对于主 SPADE 生成器中的每一层，我们使用$\boldsymbol{\theta}_H^l$去规范化参数$\boldsymbol{\gamma}^l$和$\boldsymbol{\beta}^l$对输入特征计算去规范化。我们注意到，在原始 SPADE 模块中，比例图$\boldsymbol{\gamma}^l$和偏差图$\boldsymbol{\beta}^l$是由对输入语义图$\mathbf{s}_t$进行操作的固定权重生成的。在我们的设置中，这些映射是由动态权重$\boldsymbol{\theta}_H^l$生成的。此外，$\boldsymbol{\theta}_H^l$包含三组权重：$\boldsymbol{\theta}_S^l$、$\boldsymbol{\theta}_\gamma^l$和$\boldsymbol{\theta}_\beta^l$。$\boldsymbol{\theta}_S^l$充当共享层以提取公共特征，而$\boldsymbol{\theta}_\gamma^l$和$\boldsymbol{\theta}_\beta^l$则采用$\boldsymbol{\theta}_S^l$的输出分别生成$\boldsymbol{\gamma}^l$和$\boldsymbol{\beta}^l$图。对于$G^l$中的每个 BatchNorm 层，我们根据归一化特征$\hat{\mathbf{p}}_H^l$计算非归一化特征$\mathbf{p}_H^l$：

$$\mathbf{p}_S^l = \begin{cases} \mathbf{s}_t, & \text{if } l = 0 \\ \sigma\left(\mathbf{p}_S^{l-1} \circledast \boldsymbol{\theta}_S^l\right), & \text{otherwise} \end{cases} \tag{8}$$

$$\boldsymbol{\gamma}^l = \mathbf{p}_S^l \circledast \boldsymbol{\theta}_{\boldsymbol{\gamma}}^l, \quad \boldsymbol{\beta}^l = \mathbf{p}_S^l \circledast \boldsymbol{\theta}_{\boldsymbol{\beta}}^l \tag{9}$$

$$\mathbf{p}_H^l = \boldsymbol{\gamma}^l \odot \hat{\mathbf{p}}_H^l + \boldsymbol{\beta}^l \tag{10}$$

其中$\circledast$代表卷积，σ是非线性函数。

基于注意力的聚合（K > 1）。另外，我们希望 E 能够从任意数量的示例图像中提取图案。由于不同的示例图像可能带有不同的外观模式，并且它们与不同的输入图像具有不同的相关性，因此我们设计了一种注意力机制[61，52]来聚合提取的外观模式$\mathbf{q}_1,\ldots,\mathbf{q}_K$。

为此，我们构建了一个新的注意力网络E_A，它由几个完全卷积层组成。E_A应用于示例图像$\mathbf{s}_{\mathbf{e}_k}$的每个语义图像。这产生关键向量$\boldsymbol{a}_k \in \mathbb{R}^{C\times N}$，其中 C 是通道数，$N = H \times W$是特征图的空间尺寸。我们还将$E_A$应用于当前输入的语义图像$\mathbf{s}_t$，以提取其关键向量$\boldsymbol{a}_t \in \mathbb{R}^{C\times N}$。然后，我们通过采用矩阵乘积$\boldsymbol{\alpha}_k = (\boldsymbol{a}_k)^T \otimes \boldsymbol{a}_t$来计算注意力权重$\boldsymbol{\alpha}_k \in \mathbb{R}^{N\times N}$。然后，将注意力权重用于计算外观表示$\mathbf{q} = \sum_{k=1}^K \mathbf{q}_k \otimes \boldsymbol{\alpha}_k$，然后将其馈送到多层感知器$E_P$中以生成网络权重（图 2（b））。当不同的示例图像包含主题的不同部分时，此聚合机制很有用。例如，当示例图像同时包含目标人物的正面和背面时，注意力图可以帮助捕获合成过程中相应的身体部位（图 7（c））。

变换示例图像。为了减轻图像合成网络的负担，我们还可以（可选）变换给定的示例图像，并将其与中间的合成输出$\tilde{\mathbf{h}}_t$合并。具体来说，我们使模型估计额外的流$\tilde{\mathbf{w}}_{e_t}$和遮罩$\tilde{\mathbf{m}}_{e_t}$，它们用于将示例图像$\mathbf{e}_1$变换为当前输入语义，类似于我们变换并与先前的帧结合的方式。然后，新的中间图像变为

$$\tilde{\mathbf{h}}_t' = (\mathbf{1} - \tilde{\mathbf{m}}_{e_t}) \odot \tilde{\mathbf{w}}_{e_t}(\mathbf{e}_1) + \tilde{\mathbf{m}}_{e_t} \odot \tilde{\mathbf{h}}_t \tag{11}$$

在多个示例图像的情况下，通过查看注意力权重，我们选择$\mathbf{e}_1$作为与当前帧具有最大相似度得分的图像。在实践中，当示例图像和目标图像在大多数区域类似（例如合成背景保持静态的姿势）时，我们发现这很有用。

训练。我们使用与 vid2vid 框架[57]中相同的学习目标。但是，我们不是使用来自一个域的数据来训练 vid2vid 模型，而是使用来自多个域的数据。在图 7（a）中，我们显示了少样本 vid2vid 模型的性能与训练数据集中包含的域数量成正相关。这表明我们的模型可以从增加的视觉内容中受益。我们的框架在有配对$\mathbf{s}_1^T$和$\mathbf{x}_1^T$可用的监督环境下进行了训练。我们使用从 x 随机采样的 K 个示例图像训练模型，将$\mathbf{s}_1^T$转换为$\mathbf{x}_1^T$。我们采用渐进式训练技术，这会逐渐增加训练序列的长度。最初，我们将 T 设置为 1，这意味着网络仅生成单个帧。之后，我们每隔几个迭代周期将序列长度（T）加倍。

推理。在测试时，我们的模型可以采用任意数量的示例图像。在图 7（b）中，我们表明我们的性能与示例图像的数量呈正相关。此外，我们还可以（可选）使用给定的示例图像微调网络以提高性能。注意，我们仅微调权重生成模块 E 和中间图像合成网络 H，并且使与光流估计（$\boldsymbol{\theta}_M$，$\boldsymbol{\theta}_H$）有关的所有参数保持不变。 我们发现这可以更好地保留示例图像中的人员身份。

4 实验

实验细节。 我们的训练过程遵循 vid2vid 的工作[57]。我们使用 ADAM 优化器[26]，其中 lr = 0.0004 并且$(\beta_1, \beta_2) = (0.5, 0.999)$。 训练是使用带有 8 个 32GB V100 GPU 的 NVIDIA DGX-1 机器进行的。

数据集。 我们采用三个视频数据集来验证我们的方法。

- **YouTube 舞蹈视频。** 它由 1,500 个来自 YouTube 的跳舞视频组成。 我们将它们分为没有重叠主题的训练集和测试集。 每个视频进一步分为连续运动的短片。 这产生约 15,000 个剪辑片段用于训练。 在每次迭代中，我们随机选择一个剪辑片段，并在同一片段中选择一个或多个帧作为示例图像。 在测试时，在训练过程中都看不到示例图像和输入的人体姿势。
- **街头现场视频。** 我们使用来自三个不同地理区域的街道现场视频：1）德国，来自 Cityscapes 数据集[9]，2）波士顿，使用行车记录仪收集，3）纽约市，通过不同的行车记录仪收集。 我们应用预训练的分割网络[60]来获得分割图。同样，在训练期间，我们随机选择与示例图像相同区域的一帧。在测试时，除了来自这三个区域的测试集图像外，我们还对 ApolloScape [20]和 CamVid [5]数据集进行了测试，这些数据集均未包含在训练集中。
- **脸部说话视频。** 我们使用 FaceForensics 数据集[44]中的真实视频，其中包含 854 个来自不同记者的新闻发布会视频。 我们将数据集分为 704 个视频进行训练和 150 个视频进行验证。 我们从类似于 vid2vid 的输入视频中提取草图，然后选择同一视频的一帧作为示例图像，以将草图转换为面部视频。

基线。由于现有的 vid2vid 方法无法使用很少的示例图像来适应未见过的领域，因此我们构建了 3 个强大的基线，这些基线考虑了实现目标泛化能力的不同方法。对于以下所有比较和图片展示，所有方法均使用 1 个示例图像。

- **编码器。** 在这种基线方法中，我们将示例图像编码为样式矢量，然后使用 H 中的图像合成分支对特征进行解码以生成$\tilde{\mathbf{h}}_t$。
- **ConcatStyle。** 在这种基线方法中，我们还将示例图像编码为样式矢量。 但是，不是使用 H 中的图像合成分支直接解码样式矢量，而是将矢量与每个输入语义图像连接起来以生成增强的语义输入图像。 然后将该图像用作我们 H 中空间调制分支的输入，以生成中间图像$\tilde{\mathbf{h}}_t$。
- **AdaIN。**在此基线中，我们在 H 的图像合成分支中的每个空间调制层之后插入一个 AdaIN 归一化层。我们通过将示例图像馈送到编码器来生成 AdaIN 归一化参数，类似于 FUNIT[31]。

除了这些基线，对于人工合成任务，我们还使用作者提供的预训练模型将我们的方法与以下方法进行了比较。

- **PoseWarp** [1]使用示例图像以未见过的姿势合成人。这个想法是假设每个肢体都经历相似性变换。通过将所有变换后的肢体组合在一起，可以获得最终的输出图像。
- **MonkeyNet** [47]被提议用于将运动从序列传输到静止图像。它首先检测图像中的关键点，然后预测它们的变形以使静止图像变形。

评估指标。 我们使用以下指标进行定量评估。

- **Fréchet 初始距离（FID）**[17]测量实际数据的分布与生成的数据之间的距离。它通常用于量化合成图像的保真度。
- **姿势错误。**我们使用 OpenPose [6]估计合成对象的姿势。 这为每个视频帧渲染了一组联合位置。 然后我们计算像素的绝对误差在

Method	YouTube Dancing videos			Street Scene videos			
	Pose Error	FID	Human Pref.	Pixel Acc	mIoU	FID	Human Pref.
Encoder	13.30	234.71	0.96	0.400	0.222	187.10	0.97
ConcatStyle	13.32	140.87	0.95	0.479	0.240	154.33	0.97
AdaIN	12.66	207.18	0.93	0.756	0.360	205.54	0.87
PoseWarp [1]	16.84	180.31	0.83	N/A	N/A	N/A	N/A
MonkeyNet [47]	13.73	260.77	0.93	N/A	N/A	N/A	N/A
Ours	**6.01**	**80.44**	—	**0.831**	**0.408**	**144.24**	—

表 1：我们的方法优于现有的姿势转移方法，以及跳舞和街头场景视频合成任务的基准。对于姿势错误和 FID，值越低越好。对于像素精度和 mIoU，值越高越好。人类的偏爱分数表示受测者偏爱通过我们的方法合成的结果的比例。

图 3：人类视频合成结果的可视化。在给定相同姿势视频但示例图像不同的情况下，我们的方法将合成对象的逼真视频，这些视频在训练期间不会出现。单击图像以在浏览器中播放视频剪辑。

估计的姿势和输入到模型中的原始姿势之间。该度量标准背后的想法是，如果图像合成良好，则训练有素的人体姿势估计网络应该能够恢复用于合成图像的原始姿势。我们注意到在一些先前的工作中[22、58、57]也使用了类似的思想来评估图像合成性能。

• **细分精度。**为了评估街道场景视频的性能，我们对所有竞争方法生成的结果视频运行了最新的街道场景分割网络。然后，我们报告像素精度和平均交并比（IoU）比率。 如上所述，将分割精度用作性能度量的想法遵循了关于使用姿势误差的讨论。

• **人类主观评分。**最后，我们使用 Amazon Mechanical Turk（AMT）评估生成的视频的质量。我们执行 AB 测试，其中我们提供的用户视频来自

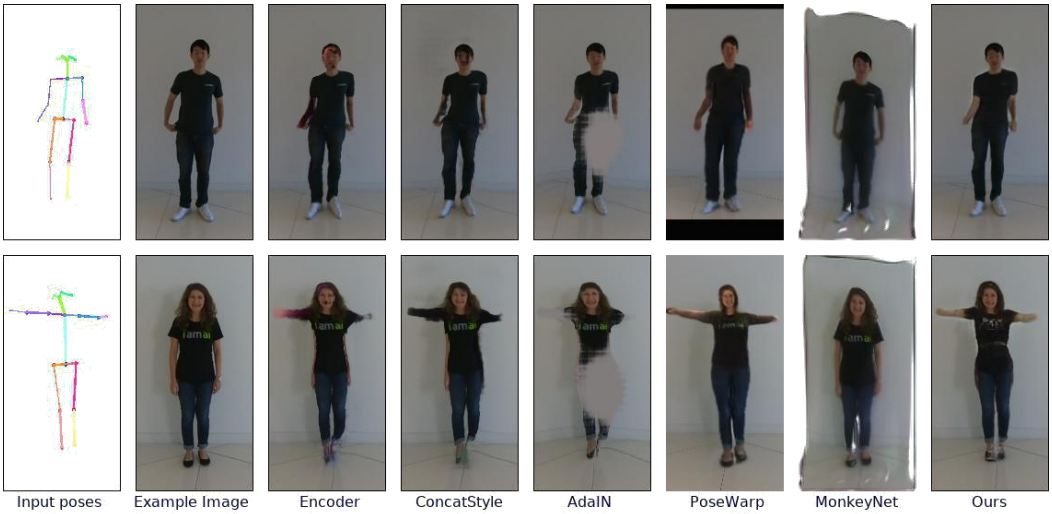


图 4：针对人类运动合成的不同基准进行比较。请注意，竞争方法要么具有许多可见的伪像，要么完全无法传递运动。单击图像以在浏览器中播放视频剪辑。

图 5：街道场景视频合成结果的可视化。即使训练集中未包含的样式，我们的方法也能够合成能够真实反映示例图像样式的视频。 单击图像以在浏览器中播放视频剪辑。

两种不同的方法，请测试员们选择质量更好的一种。对于每对比较，我们都会生成 100 个剪辑片段，每个片段都由 60 名工作人员查看。顺序是随机的。

主要结果。在图 3 中，我们显示了在合成人类时使用不同示例图像的结果。可以看出，我们的方法可以成功地将运动传递到所有示例图像。图 4 显示了我们的方法与其他方法的比较。可以看出，其他方法要么生成明显的伪像，要么无法如实地传递运动。

图 5 显示了使用不同示例图像合成街道场景视频的结果。可以看出，即使使用相同的输入分割图，我们的方法也可以使用不同的示例图像获得不同的视觉效果。

表 1 显示了这两种任务与其他方法的定量比较。可以看出，在所有性能指标上，我们的方法始终比其他方法取得更好的结果。

在图 6 中，我们显示了在合成人脸时使用不同示例图像的结果。我们的方法可以在捕获输入视频中的动作时忠实地保留人的身份。

最后，为了验证我们的假设，即更大的训练数据集有助于提高合成视频的质量，我们进行了一项实验，其中在训练过程中保留了部分数据集。我们改变训练集中的视频数量，并在图 7（a）中绘制结果。我们发现结果支持我们的假设。我们还将评估是否可以访问更多示例图像

图 6：人脸视频合成结果的可视化。 在给定相同的输入视频但示例图像不同的情况下，我们的方法将合成对象的逼真的视频，这些视频在训练期间不会出现。 单击图像以在浏览器中播放视频剪辑。

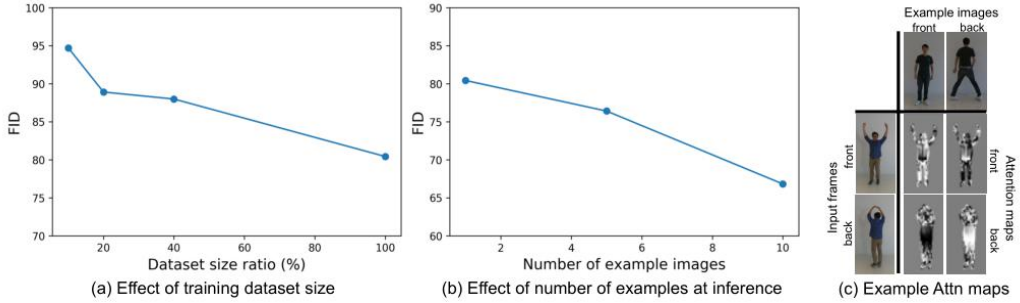


图 7：（a）该图显示了使用较大的数据集训练合成视频时，其质量会提高。种类更多有助于学习更通用的网络权重生成模块，从而提高适应能力。（b）该图显示了我们合成视频的质量与测试时提供的示例图像的数量相关。提出的注意力机制可以利用更大的示例集更好地生成网络权重。（c）当给出多个示例图像时，注意图的可视化。 请注意，在合成目标的前部时，注意图指示网络使用了更多的前部示例图像，反之亦然。

在测试时有助于视频合成性能。 如图 7（b）所示，结果证实了我们的假设。

局限性。尽管我们的网络原则上可以泛化到未见过的域，但是当测试域与训练域相差太大时，它将无法正常运行。例如，当测试看上去与现实世界中的人大不相同的 CG 角色时，网络将陷入困境。另外，由于我们的网络基于语义估计作为输入，例如姿势图或分割图，因此当这些估计失败时，我们的网络也很可能会失败。

5 结论

我们提供了一种视频到视频合成框架，可以在测试时合成未见过的主题或街道场景样式的视频。这是通过我们新颖的自适应网络权重生成方案实现的，该方案根据示例图像动态确定权重。实验结果表明，我们的方法优于竞争方法。

参考文献

[1] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag. Synthesizing images of humans in unseen poses. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[2] S. Benaim and L. Wolf. One-shot unsupervised cross domain translation. In Advances in Neural Information Processing Systems (NIPS), 2018.

[3] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

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

[5] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In European Conference on Computer Vision (ECCV), 2008.

[6] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In arXiv preprint arXiv:1812.08008, 2018.

[7] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros. Everybody dance now. arXiv preprint arXiv:1808.07371, 2018.

[8] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[9] 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 Recognition (CVPR), 2016.

[10] E. L. Denton and V. Birodkar. Unsupervised learning of disentangled representations from video. In Advances in Neural Information Processing Systems (NIPS), 2017.

[11] C. Finn, I. Goodfellow, and S. Levine. Unsupervised learning for physical interaction through video prediction. In Advances in Neural Information Processing Systems (NIPS), 2016.

[12] O. Gafni, L. Wolf, and Y. Taigman. Vid2game: Controllable characters extracted from real-world videos. arXiv preprint arXiv:1904.08379, 2019.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. In Advances in Neural Information Processing Systems (NIPS), 2014.

[14] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein GANs. In Advances in Neural Information Processing Systems (NIPS), 2017.

[15] D. Ha, A. Dai, and Q. V. Le. Hypernetworks. In International Conference on Learning Representations (ICLR), 2016.

[16] Z. Hao, X. Huang, and S. Belongie. Controllable video generation with sparse trajectories. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[17] 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 (NIPS), 2017.

[18] Q. Hu, A. Waelchli, T. Portenier, M. Zwicker, and P. Favaro. Video synthesis from a single image and motion stroke. arXiv preprint arXiv:1812.01874, 2018.

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

[20] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang. The ApolloScape dataset for autonomous driving. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

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

[22] 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 Recognition (CVPR), 2017.

[23] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool. Dynamic filter networks. In Advances in Neural Information Processing Systems (NIPS), 2016.

[24] N. Kalchbrenner, A. v. d. Oord, K. Simonyan, I. Danihelka, O. Vinyals, A. Graves, and K. Kavukcuoglu. Video pixel networks. arXiv preprint arXiv:1610.00527, 2016.

[25] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[26] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations (ICLR), 2015.

[27] A. X. Lee, R. Zhang, F. Ebert, P. Abbeel, C. Finn, and S. Levine. Stochastic adversarial video prediction. arXiv preprint arXiv:1804.01523, 2018.

[28] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang. Flow-grounded spatial-temporal video prediction from still images. In Proceedings of the European Conference on Computer Vision (ECCV), pages 600–615, 2018.

[29] X. Liang, L. Lee, W. Dai, and E. P. Xing. Dual motion GAN for future-flow embedded video prediction. In Advances in Neural Information Processing Systems (NIPS), 2017.

[30] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In Advances in Neural Information Processing Systems (NIPS), 2017.

[31] M.-Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, and J. Kautz. Few-shot unsupervised image-to-image translation. arXiv preprint arXiv:1905.01723, 2019.
[32] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In Advances in Neural Information Processing Systems (NIPS), 2016.
[33] W. Lotter, G. Kreiman, and D. Cox. Deep predictive coding networks for video prediction and unsupervised learning. In International Conference on Learning Representations (ICLR), 2017.
[34] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. In International Conference on Learning Representations (ICLR), 2016.
[35] T. Miyato and M. Koyama. cGANs with projection discriminator. In International Conference on Learning Representations (ICLR), 2018.
[36] N. Neverova, R. Alp Guler, and I. Kokkinos. Dense pose transfer. In European Conference on Computer Vision (ECCV), 2018.
[37] S. Niklaus, L. Mai, and F. Liu. Video frame interpolation via adaptive convolution. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[38] S. Niklaus, L. Mai, and F. Liu. Video frame interpolation via adaptive separable convolution. In IEEE International Conference on Computer Vision (ICCV), 2017.
[39] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier GANs. In International Conference on Machine Learning (ICML), 2017.
[40] J. Pan, C. Wang, X. Jia, J. Shao, L. Sheng, J. Yan, and X. Wang. Video generation from single semantic label map. arXiv preprint arXiv:1903.04480, 2019.
[41] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. Semantic image synthesis with spatially-adaptive normal-ization. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[42] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In International Conference on Learning Representations (ICLR), 2015.
[43] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In International Conference on Machine Learning (ICML), 2016.
[44] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. arXiv preprint arXiv:1803.09179, 2018.
[45] M. Saito, E. Matsumoto, and S. Saito. Temporal generative adversarial nets with singular value clipping. In IEEE International Conference on Computer Vision (ICCV), 2017.
[46] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[47] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe. Animating arbitrary objects via deep motion transfer. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[48] N. Srivastava, E. Mansimov, and R. Salakhudinov. Unsupervised learning of video representations using lstms. In International Conference on Machine Learning (ICML), 2015.
[49] H. Su, V. Jampani, D. Sun, O. Gallo, E. Learned-Miller, and J. Kautz. Pixel-adaptive convolutional neural networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[50] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. In International Conference on Learning Representations (ICLR), 2017.
[51] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz. MoCoGAN: Decomposing motion and content for video generation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems (NIPS), 2017.
[53] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee. Decomposing motion and content for natural video sequence prediction. In International Conference on Learning Representations (ICLR), 2017.
[54] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In Advances in Neural Information Processing Systems (NIPS), 2016.
[55] J. Walker, C. Doersch, A. Gupta, and M. Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In European Conference on Computer Vision (ECCV), 2016.
[56] J. Walker, K. Marino, A. Gupta, and M. Hebert. The pose knows: Video forecasting by generating pose futures. In IEEE International Conference on Computer Vision (ICCV), 2017.
[57] 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 (NIPS), 2018.
[58] 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.
[59] J. Wu, D. Li, Y. Yang, C. Bajaj, and X. Ji. Dynamic sampling convolutional neural networks. In European Conference on Computer Vision (ECCV), 2018.
[60] T. Wu, S. Tang, R. Zhang, and Y. Zhang. Cgnet: A light-weight context guided network for semantic segmentation. arXiv preprint arXiv:1811.08201, 2018.
[61] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. arXiv preprint arXiv:1502.03044, 2015.
[62] 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 Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[63] T. Xue, J. Wu, K. Bouman, and B. Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In Advances in Neural Information Processing Systems (NIPS), 2016.
[64] C. Zhang, M. Ren, and R. Urtasun. Graph hypernetworks for neural architecture search. In International Conference on Learning Representations (ICLR), 2019.
[65] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. In International Conference on Machine Learning (ICML), 2019.
[66] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In IEEE International Conference on Computer Vision (ICCV), 2017.
[67] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg. Dance dance generation: Motion transfer for internet videos. arXiv preprint arXiv:1904.00129, 2019.
[68] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In IEEE International Conference on Computer Vision (ICCV), 2017.
[69] 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 Processing Systems (NIPS), 2017.

A. 与街景基线比较

图 8 显示了与我们的方法与场景序列基线方法的比较。同样，我们的方法是唯一可以在示例图像中真实再现样式的方法，而其他方法会生成伪影以致无法捕获样式。

B. 我们的基准方法的细节

在“实验”部分中，我们介绍了少样本 vid-to-vid 合成任务的三个基准。在这里，我们提供了其体系结构的其他详细信息。请注意，这些基线仅设计用于处理一个示例图像，并且我们在一个示例图像设置上将提出的方法与这些基线的性能进行了比较。

编码器基线。如主要论文中所述，编码器基线由一个图像编码器组成，该图像编码器将示例图像编码为样式潜码，然后将其直接馈入 SPADE 生成器中的主图像合成分支的头部。我们在图 9（a）中可视化编码器基线的体系结构。

ConcatStyle 基线。如图 9（b）所示，在 ConcatStyle 基线中，我们还使用图像编码器将示例图像编码为样式潜码。现在，我们不是通过将样式潜码输入到主图像合成分支的头部中，而是通过广播操作将样式潜码与输入的语义图像连接在一起。连接采用 SPADE 模块的新输入语义图像。

AdaIN 基准。在此基准中，我们使用 AdaIN [19]进行自适应视频到视频合成。具体来说，我们使用图像编码器将示例图像编码为潜在矢量，并使用多层感知器将潜在矢量转换为 AdaIN 运算的均值和方差矢量。AdaIN 参数被馈送到主图像合成分支的每一层。具体来说，我们在 SPADE 规范化层之后添加 AdaIN 规一化层，如图 9（c）所示。

C. 与 AdaIN 的讨论

在 AdaIN 中，来自示例图像的信息表示为缩放矢量和偏置矢量。可以将此操作视为 1x1 卷积，其组大小等于通道大小。从这个角度来看，AdaIN 是所提出的权重生成方案的约束情况，因为我们的方案可以生成组大小等于 1 且内核大小大于 1x1 的卷积内核。而且，所提出的方案可以容易地与 SPADE 模块结合。具体来说，我们使用提出的生成方案来生成 SPADE 层的权重，而权重又会生成空间自适应解调参数。为了证明生成权重的重要性，我们使用加权平均值和注意力模块将其与 AdaIN 进行比较（图 10（a））。

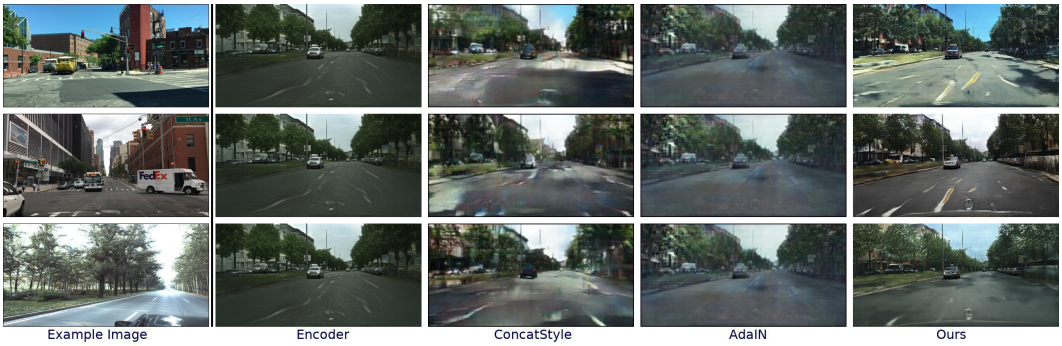


图 8：针对街景综合任务基于不同基线方法的比较。请注意，所提出的方法是唯一可以将样式从示例图像转换为输出视频的方法。 单击图像以在浏览器中播放视频剪辑。

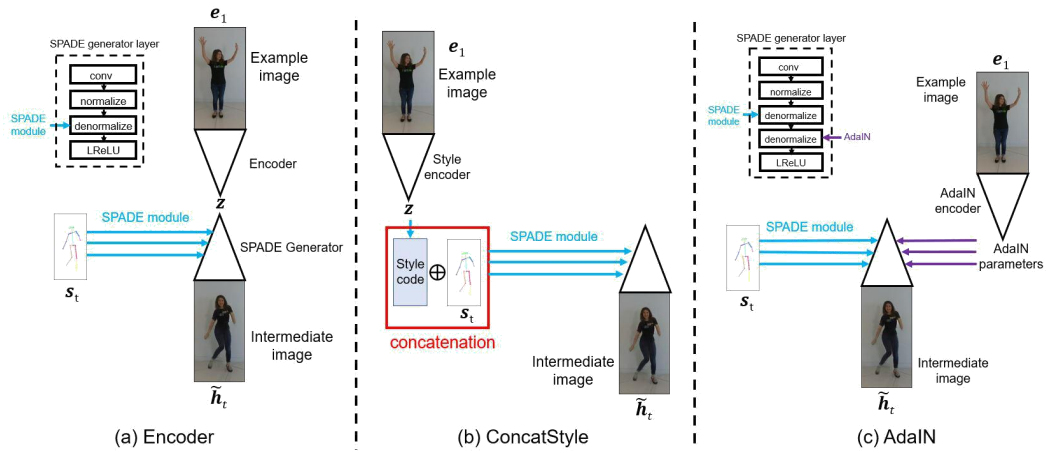


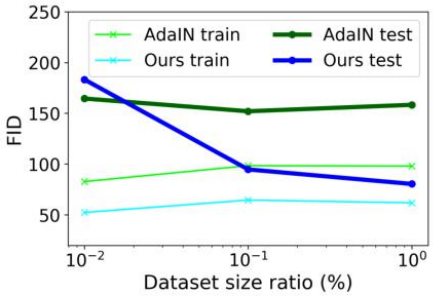
图 9：基线的详细信息。（a）编码器基线尝试通过将样式潜码馈入 SPADE 生成器中的主图像合成分支的头部来尝试输入未见过域的样式信息。（b）ConcatStyle 基线尝试通过将样式潜码与输入的语义图像进行连接来输入未见过域的样式信息，然后将其输入到 SPADE 模块中。（c）AdaIN 基线试图通过使用 AdaIN 调制来输入未见过域的样式信息。

5 example images	AdaIN (avg)	AdaIN (attn)	Ours (attn)
FID	129.90	113.83	***76.53***

(a) Comparison to AdaIN

	Ours (1 example)	Ours (50 examples)	Ours (50 examples + finetune)	vid2vid
FID	56.99	51.10	***43.04***	47.19

(b) Comparison to vid2vid



(c) Performance vs. dataset sizes

图 10：与 AdaIN 和 vid2vid 的比较。（a）我们的注意力机制和权重生成方案都有助于获得更好的图像质量。（b）随着我们在测试时使用更多示例，性能得以提高。 如果允许我们微调模型，我们实际上可以实现与 vid2vid 相当的性能。（c）当数据集较小时，AdaIN 能够获得良好的性能。但是，由于网络的容量有限，因此它将难以处理更大的数据集。

当使用不同的数据集大小时，我们还将与 AdaIN 进行比较。假设当数据集较小时，两种方法都能够捕获数据集中的多样性。但是，随着数据集大小的增大，由于可表达性受到限制，AdaIN 开始失败，如图 10（c）所示。

D. 与 vid2vid 的比较

正如主要论文所讨论的那样，vid2vid 的缺点是针对不同的人或城市需要不同的模型。 例如，他们通常需要几分钟的训练数据和几天的训练时间，而我们的方法只需要一张图像和可忽略的时间即可产生权重。

为了比较我们与 vid2vid 的性能，我们在图 10（b）中显示了用于合成特定人的定量比较。我们发现，即使 K = 1，我们的模型也可以提供可比较的结果。此外，如果根据示例图像进一步微调模型，则可以实现可比甚至更好的性能。