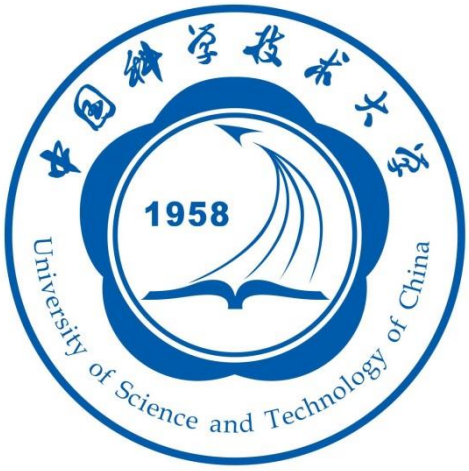
中国科学技术大学

《计算机图形学前沿》结课作业



学生姓名：黄瑞轩

学生学号：PB20111686

学　　院：计算机科学与技术学院

一级学科：计算机科学与技术

图像跨域转换综述

黄瑞轩 PB20111686

（中国科学技术大学　计算机科学与技术学院，计算机科学与技术系）

摘　要　图像跨域转换指从一副图像到另一副图像的转换，比如语义分割图转换为真实街景图、灰色图转换为彩色图等。本文将以图像跨域转换技术的发展为主题做一综述。

关键词　图像跨域转换，对抗学习，Image Translation

1 引言[1]

图像跨域转换，又称图像翻译（Image Translation），是一种旨在将源域的图像转换为目标域的图像的技术，具体来说是使生成图像在保持源图像的结构（轮廓、姿态等）的同时具有目标图像的风格（纹理、颜色等）。图像跨域转换技术在视觉领域有着广泛的应用，如照片编辑和视频特效制作。近年来，该技术在深度学习尤其是生成式对抗网络（GAN）的基础上得到了飞速发展，图 1 是当前图像跨域转换的一些实例[2]。

图像跨域转换的目标主要是学习能够将源域的图像映射到目标域对应图像的函数。在图像跨域转换的研究场景下，源域和目标域往往具有相同的内容结构和语义相关性，比如人脸图像中的年轻域和年老域、室外场景的白天域和夜晚域。跨域转换后生成的图像保持输入的源域图像的内容结构，同时其图像风格应具有目标域特有的风格属性。图像跨域转换方向得到了深度学习和计算机视觉领域研究人员的广泛关注，因为它已广泛应用于图像风格转换、图像编辑、图像超分辨率和图像彩色化的工作中了。

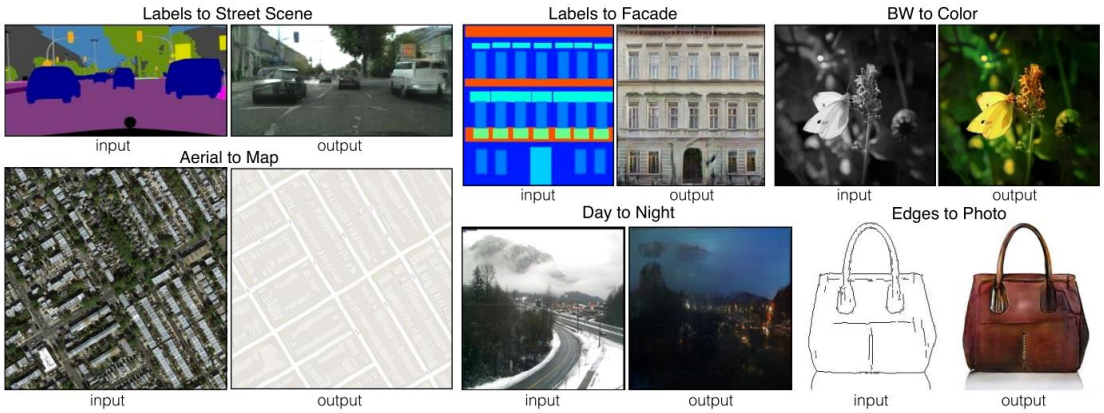


图 1　图像跨域转换的一些实例

2 基于条件对抗网络的图像跨域转换

传统上，图像跨域转换任务中的每一项都是用单独的专用机器来处理的，并且设置总是相同的：从像素预测像素。Zhu[2]等为所有这些问题开发了一个通用框架。

2.1 提出思路

卷积神经网络（CNN）已成为各种图像预测问题背后的常用工具。CNN 学习的目标是使损失函数最小化。尽管学习过程是自动的，但在设计有效损失方面仍然需要大量的手动工作。即我们仍然需要告诉 CNN 我们希望它最小化什么。但如果我们采取一种天真的方法，比如要求 CNN 使预测和真实像素之间的欧几里德距离最小化，它将倾向于产生**模糊**的结果[3, 4]。这是因为欧几里德距离通过平均所有可能的输出来做最小化，这会导致出现模糊。如何提出一个恰当的损失函数迫使 CNN 做我们真正想做的事情——例如输出清晰、逼真的图像——是一个开放的问题，通常需要非常专业的知识。

最近提出的生成性对抗网络（GAN）使得我们可以只指定一个高级别的目标，比如“使输出与现实不可区分”，然后自动学习一个适合于实现该目标的损失函数。GANs 学习一个损失函数，试图分类输出图像是真是假，同时训练生成模型以使该损失最小化。

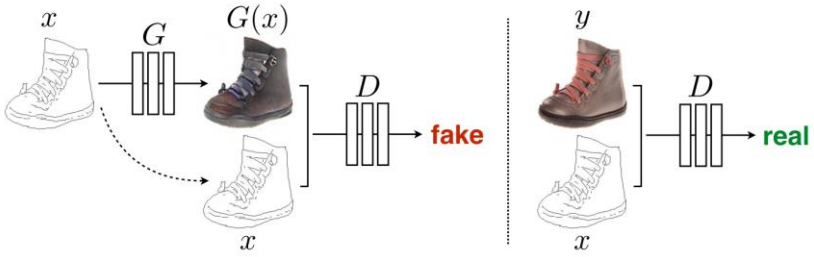


图 2　基于 GAN 的图像跨域转换

从上面的讨论中得出最终的损失函数构成：

$$G^* = \arg\min_G \max_D \mathcal{L}_{C-GAN}(G,D) + \lambda \mathcal{L}_{L1}(G)$$

GAN 损失函数使用 PatchGAN 进一步保证生成图像的清晰度。

2.2 相关工作

图像建模中的结构损失。图像到图像的翻译问题通常根据像素分类或回归[2]来制定。这些公式将输出空间视为“非结构化”的，即在给定输入图像的情况下，每个输出像素被视为条件独立于所有其他像素。相反，条件 GAN（C-GAN）将学习结构性损失，结构性损失会影响输出的联合配置。大量文献考虑了此类损失，方法包括条件随机场、SSIM 度量、特征匹配、非参数损失、卷积伪先验和基于匹配协方差统计的损失[2]。

Conditional GANs。之前的工作[2]已将 GAN 限制在离散标签、文本和图像上。图像条件模型处理了来自正态映射的图像预测、未来帧预测、产品照片生成和来自稀疏注释

的图像生成。其他的学者也用GAN进行图像到图像的映射，但仅是无条件地应用GAN，同时依赖其他技术（如 L2 回归）强制使输出以输入为条件。这些论文在修复、未来状态预测、用户约束引导的图像处理、风格转换和超分辨率方面取得了大量成果。

如上所述的方法都是为特定应用量身定制的。Zhu 等的框架不同之处在于其是通用的，不针对特定于应用场景。与过去的工作不同[2]，对于生成器，Zhu 等使用基于“U-Net”的架构，对于鉴别器，Zhu 等使用卷积“PatchGAN”分类器。Li 等提出了一种类似的 PatchGAN 架构，用于捕获局部风格的统计数据[5]。Zhu 等证明了这种方法在更广泛的问题上是有效的，并且研究了改变面片大小的影响。

2.3 改进工作

以前针对该领域问题的许多解决方案[4, 6, 7, 8, 9]都使用了编码器-解码器网络[10]。在这样一个网络中，输入通过一系列逐步降低采样的层，直到瓶颈层。这样的网络需要让所有信息流通过所有层。对于许多图像翻译问题，输入和输出之间共享了大量低级信息，因此希望将这些信息直接在网络上传递。例如，在图像着色的情况下，输入和输出共享突出边缘的位置。为了给生成器一种绕过此类信息瓶颈的方法，Zhu 等按照“U-Net”的一般形状添加了跳过连接[11]。

L2 和 L1 在图像生成问题上产生模糊结果[3,4]。对于低频情况下的问题，不需要一个全新的框架来在低频率下强制正确执行。为了模拟高频，只需将注意力限制在局部图像块中的结构即可。因此，Zhu 等设计了 PatchGAN，只在补丁规模上惩罚结构。该鉴别器尝试对图像中的每个 N×N 面片的真假进行分类。在图像上卷积运行该鉴别器，对所有结果取平均以提供其最终输出。测试结果如图 3 所示。

图 3　PatchGAN 与其他网络的执行结果对比

2.4 展示与总结

Zhu 等的工作解决了再图像跨域转换中已有的输出模糊的问题。并且，Zhu 等的结果表明，对于许多图像跨域转换任务，特别是涉及高度结构化图形输出的任务，条件对抗网络是一种很有前途的方法。这些网络学习适应于我们需要的任务和数据的损耗函数，这使得它们适用于各种环境。

图 4、5、6 是 Zhu 等工作的成果展示。

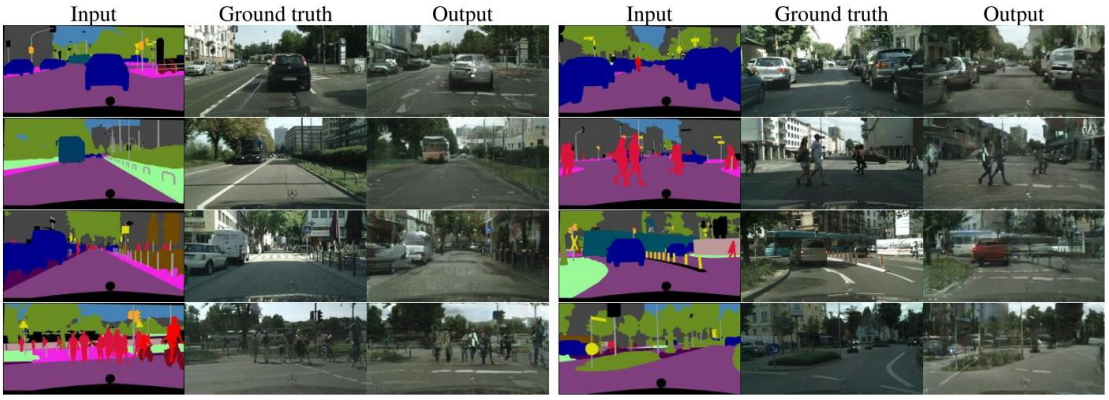


图 4　城市景观标签上的示例

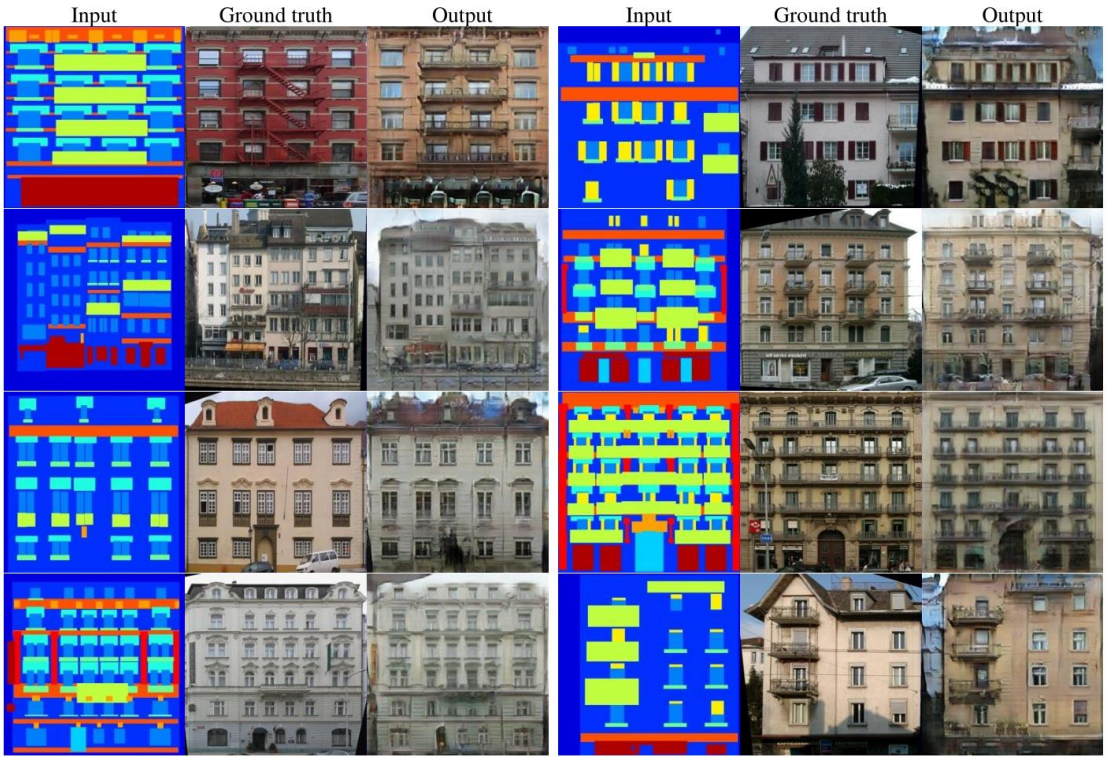


图 5　立面标签上的示例

图 6　自动检测边缘上的示例

3 基于 C-GANs 的高分辨率图像合成和语义处理

在 Zhu 等的工作[2]的基础上，Wang 等使用多尺度的生成器以及判别器等方式从而生成高分辨率图像[12]。他们使用了一种非常巧妙的方式，实现了对于同一个输入，产生不同的输出，并且实现了交互式的语义编辑方式。

Wang 等提出了一种从语义标签地图生成高分辨率图像的新方法。该方法具有广泛的应用。例如，可以使用它创建用于训练视觉识别算法的合成训练数据，因为为所需场景创建语义标签要比生成训练图像容易得多。使用语义分割方法，可以将图像转换为语义标签域，编辑标签域中的对象，然后将其转换回图像域。该方法还为更高级别的图像编辑提供了新的工具，例如，向图像中添加对象或更改现有对象的外观。

3.1 提出思路

为了从语义标签合成图像，可以使用 pix2pix 方法，这是一种图像到图像的转换框架[2]，在条件设置中利用生成性对抗网络。最近，Chen 和 Koltun[13]提出，对抗训练可能不稳定，并且在高分辨率图像生成任务中容易失败。相反，他们采用改进的感知损失来合成图像，这些图像具有高分辨率，但通常缺乏精细细节和真实纹理。

上述方法的两个主要问题：

（1）用 GANs 生成高分辨率图像的困难；

（2）在以前的高分辨率结果中缺乏细节和真实纹理。

Wang 等表明，通过一个新的、强大的对抗性学习目标以及新的多尺度生成器和鉴别器架构，可以合成 2048×1024 分辨率的照片级真实感图像，这比以前方法计算的图

像更具视觉吸引力。首先，仅通过对抗性训练获得结果，而不依赖任何手工制作的损失或预训练网络进行感知损失。然后可以证明，如果预训练网络可用，添加预训练网络的感知损失可以在某些情况下略微改善结果。这两个结果在图像质量方面都大大优于以前的工作。

结果得到的损失函数由三部分组成：一是 GAN 得到的损失函数，这里与[2]中的方法相同（使用 PatchGAN）；二是特征匹配损失函数，将生成的样本和真实照片分别送入判别器提取特征，然后对特征做 Element-wise 损失；三是内容损失函数，将生成的样本和真实照片分别送入 VGG16 提取特征，然后对特征做 Element-wise 损失。最终结果为：

$$G^* = \min_G \left(\left(\max_{D_1,D_2,D_3} \sum_{k=1,2,3} \mathcal{L}_{GAN}(G, D_k) \right) + \lambda \sum_{k=1,2,3} \mathcal{L}_{FM}(G, D_k) \right)$$

3.2 相关工作[12]

生成对抗网络。GAN 旨在通过强制生成的样本与自然图像无法区分来模拟自然图像分布。GAN 可以实现多种应用，例如图像生成、表示学习、图像处理、对象检测和视频应用。已经提出了各种从粗到精的方案，以在无条件设置下合成较大的图像（例如 256×256）。受其成功经验的启发，Wang 等提出了一种新的从粗到精的生成器和多尺度鉴别器架构，适用于更高分辨率的条件图像生成。

深度视觉处理。最近，深度神经网络在各种图像处理任务中取得了很好的结果，如风格转换、修复、着色和恢复。然而，这些作品大多缺乏用户调整当前结果或探索输出空间的界面。为了解决这个问题，Zhu 等人[13]开发了一种基于 GANs 学习的先验知识编辑对象外观的优化方法。最近的一些工作还提供了用户界面，用于从颜色和草图等低级线索创建新的图像。之前的所有工作都报告了低分辨率图像的结果。Wang 等的系统与过去的工作具有相同的原理，但他们专注于对象级语义编辑，允许用户与整个场景交互并操纵图像中的单个对象。因此，用户可以用最小的代价快速创建一个新场景。

3.3 改进工作

为了生成高分辨率图像，Wang 等主要从三个层面做了改进：一是模型结构；二是损失函数设计；三是使用 Instance-map 的图像进行训练。其模型结构如图 7 所示。Wang 等首先在低分辨率图像上训练残差网络 G_1。然后，将另一个残差网络 G_2 附加到 G_1，并在高分辨率图像上联合训练这两个网络。具体来说，G_2 中剩余块的输入是 G_2 的特征图和 G_1 的最后一个特征图的元素总和。

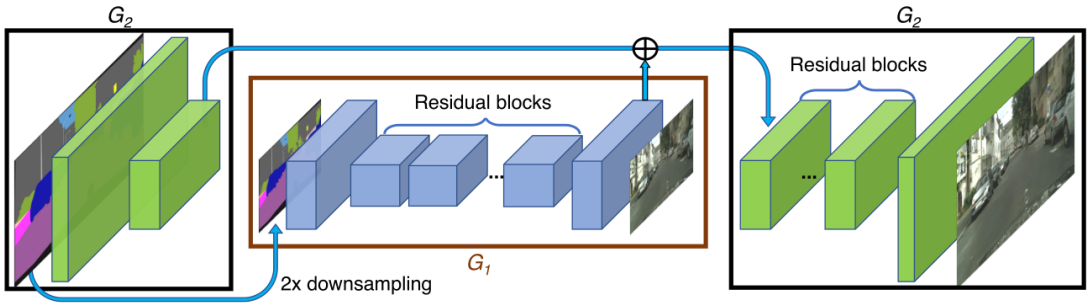


图 7　Wang 等的模型结构

Zhu 等的工作[2]采用语义分割的结果进行训练，可是语义分割结果没有对同类物体进行区分，导致多个同一类物体排列在一起的时候出现模糊，这在街景图中尤为常见。Wang 等使用个体分割（Instance-level segmention）的结果来进行训练，因为个体分割的结果提供了同一类物体的边界信息。这一过程的示意图如图 8 所示。

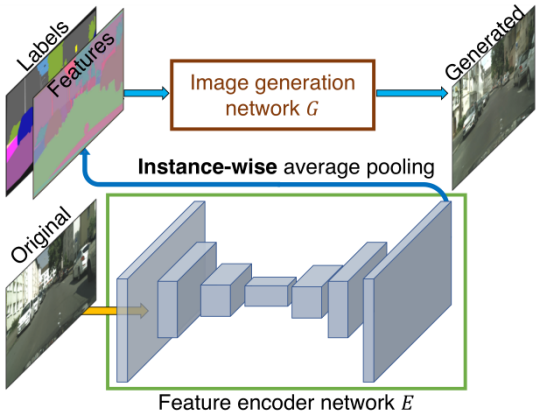


图 8　使用 Instance-map 来生成图像

具体的改进思路是：首先训练一个编码器，然后利用编码器提取原始图片的特征。根据标签信息特征。这个特征的每一类像素的值都代表了这类标签的信息。如果输入图像有足够的多，那么特征的每一类像素的值就代表了这类物体的先验分布。对所有输入的训练图像通过编码器提取特征，然后进行 K-means 聚类，得到 K 个聚类中心，以 K 个聚类中心代表不同的颜色，纹理等信息。实际生成图像时，除了输入语义标签信息，还要从 K 个聚类中心随机选择一个，即选择一个颜色/纹理风格。

3.4 展示与总结

Wang 等提出了生成高分辨率图像的多尺度网络结构，包括生成器，判别器。还提出了通过特征损失提取和 VGG 损失函数提升图像的分辨率的方法——通过学习隐变量达到控制图像颜色，纹理风格信息。图 9、10、11、12 是 Wang 等研究成果的展示示例。

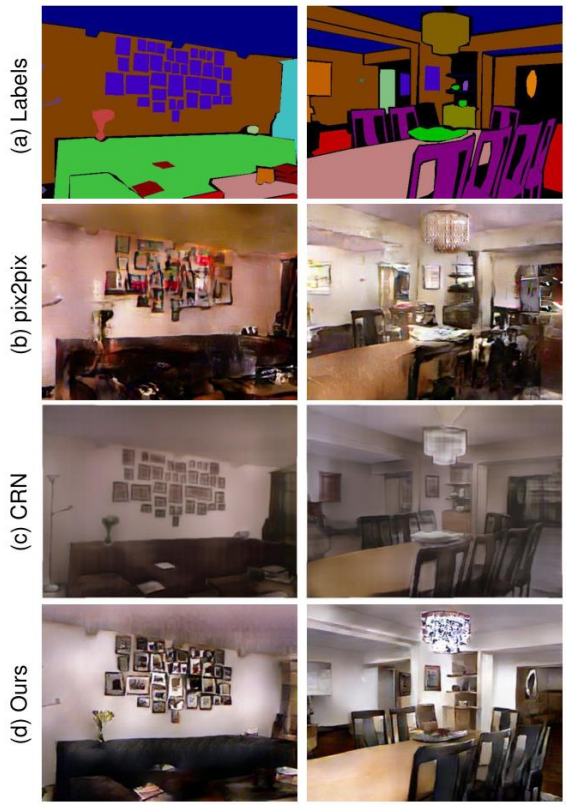


图 9　与纽约大学数据集的比较

图 10　在 Cityscapes 数据集上与 CRN 的比较

图 11　在 ADE20K 数据集上的结果

图 12　在海伦人脸数据集上的不同结果

Wang 等还表明，可以扩展图像到图像的合成管道，以产生不同的输出，并在给定适当的训练输入输出对（例如，我们的示例中的实例映射）的情况下实现交互式图像操作。我们的模型在不知道“纹理”是什么的情况下，学习将不同的对象样式化，这也可以推广到其他数据集（即，使用一个数据集中的纹理在另一个数据集合成图像）。

4 视频到视频的转换[15]

建模和重建视觉世界动态的能力对于构建智能中介至关重要。除了纯粹的科学兴趣外，学习合成连续的视觉体验在计算机视觉、机器人和计算机图形学中有着广泛的应用。例如，在基于模型的强化学习中，视频合成模型可用于近似世界的视觉动力学，以训练具有较少真实经验数据的智能中介。使用学习的视频合成模型，可以生成逼真的视频，而无需明确指定场景几何体、材料、照明和动力学，这将很麻烦，但在使用标准图形渲染引擎时是必要的。

视频合成问题以各种形式存在，包括未来视频预测和无条件视频合成。Wang 等研究了一种新的形式：视频到视频合成[15]。其核心目标仍是学习一种映射函数，该函数可以将输入视频转换为输出视频。Wang 等在几个数据集上进行了大量实验，将一系列分割遮罩转换为照片级真实感视频。定量和定性结果都表明，其合成画面看起来比来自强基线的画面更真实。其进一步证明，该方法可以生成长达 30 秒的 2K 分辨率的照片级真实感视频。其方法还允许用户对视频生成结果进行灵活的高级控制。例如，用户可以轻松地在街景视频中用树木替换所有建筑物。此外，其方法适用于其他输入视频格式，如人脸草图和身体姿势，支持从人脸交换到人体运动传输的许多应用。

4.1 提出思路

Wang 等将视频到视频合成问题转化为分布匹配问题，目标是训练一个模型，使给定输入视频的合成视频的条件分布类似于真实视频的条件分配。为此，其学习了条件生成对抗模型。给定成对的输入和输出视频，使用精心设计的生成器和鉴别器，以及新的学习目标，学习合成高分辨率、照片级真实感、时间相关的视频。

Wang 等通过在生成器加入光流约束、判别器加入光流信息以及对前景、背景分别建模的改进思路实现了高分辨率的视频生成。示例如图 13 所示。

图 13　从城市景观上的输入分割地图视频生成照片级真实感视频

4.2 相关工作

生成对抗网络。Wang 等在 GANs 上建立了模型。在 GAN 训练期间，生成器和鉴别器进行博弈。生成器旨在生成真实的合成数据，以便鉴别器无法区分真实数据和合成数据。除了噪声分布外，还可以使用各种形式的数据作为生成器的输入，包括图像、分类标签和文本描述。这种条件模型称为 C-GAN，允许灵活控制模型的输出。其方法属于 GANs 条件视频生成的范畴。然而，其方法不是在当前观察到的帧上预测未来的视频条件，而是在可操作的语义表示（例如分割掩码、草图和姿势）上合成照片级真实感视频条件。

无条件视频合成。最近的一些工作[16, 17, 18]扩展了用于无条件视频合成的 GAN 框架，该框架学习用于将随机向量转换为视频的生成器。VGAN 使用时空卷积网络。TGAN 将潜代码投影到一组潜图像代码，并使用图像生成器将这些潜图像代码转换为帧。MoCoGAN 将潜在空间分解为运动子空间和内容子空间，并使用递归神经网络生成运动代码序列。由于无条件设置，这些方法通常生成低分辨率和短长度视频。

未来视频预测。根据观察到的帧，训练视频预测模型以预测未来帧。这些模型中的许多都是在图像重建损失的情况下训练的，由于经典的回归均值问题，通常会产生模糊的视频。此外，即使进行对抗性训练，他们也无法生成长时间的视频。视频到视频合成问题有很大不同，因为它不试图预测对象运动或摄像机运动。相反，Wang 等的方法以现有视频为条件，可以在不同领域生成高分辨率和长视频。

4.3 改进方法

对生成器加入光流约束。设输入图像序列为 $s_1{}^T$，目标图像序列为 $x_1{}^T$，生成的图像序列为 $\xi_1{}^T$。则视频到视频的转换问题可以建模为一个条件分布：

$$p(\xi_1^T|s_1^T) = \prod_{t=1}^{T} p(\xi_t|\xi_{t-L}^{t-1}, s_{t-L}^{t})$$

那么训练一个 CNN，将条件分布建模为$\xi_t = F(\xi_{t-L}^{t-1}, s_{t-L}^{t}) \coloneqq (1 - m_t) \odot w_{t-1}(\xi_{t-1}) + m_t \odot h_t$，此公式中的未知量均是通过学习一个 CNN 得到的。其中w_{t-1}表示 $t-1$ 到 t 帧的光流；$w_{t-1}(\xi_{t-1})$表示利用 $t-1$ 帧的光流信息预测得到的第 t 帧的输出ξ_t；h_t表示当前帧的输出结果；m_t表示当前输出结果的模糊程度。

对判别器加入光流约束。Wang 等使用了两个判别器：一是图像粒度的判别器，即使用 C-GAN。二是视频粒度的判别器，即输入为视频序列及其光流信息，同样输入到 C-GAN。

对前景，背景分别建模。对于语义地图转换为街景图这个任务，Wang 等还分别对前景，背景进行建模，以加快收敛速度。具体来说，可以把语义地图中的“行人”，“车

辆”当做前景，“树木”，“道路”当做背景。背景通常都是不动的，因此光流计算会很准，所以得到的图像也会很清晰。因此，我们可以控制前景和背景的透明度。具体公式如下：

$$F(\xi_{t-L}^{t-1}, s_{t-L}^{t}) = (1 - m_t) \odot w_{t-1}(\xi_{t-1}) + m_t \odot \left((1 - m_{B,t}) \odot h_{F,t} + m_{B,t} \odot h_{B,t} \right)$$

其中，$h_{F,t}$和$h_{B,t}$分别代表前景和背景，二者的计算也是通过 CNN；$m_{B,t}$是背景的不透明度。

4.4 展示与总结

Wang 等提出了一种基于条件 GAN 的通用视频合成框架。通过精心设计的生成器和鉴别器以及时空对抗目标，可以合成高分辨率、真实感和时间一致的视频。大量实验表明，Wang 等的结果明显优于最先进方法的结果。其方法也优于竞争的视频预测方法。图 14、15 是其结果的展示。

图 14　示例多模式视频合成结果

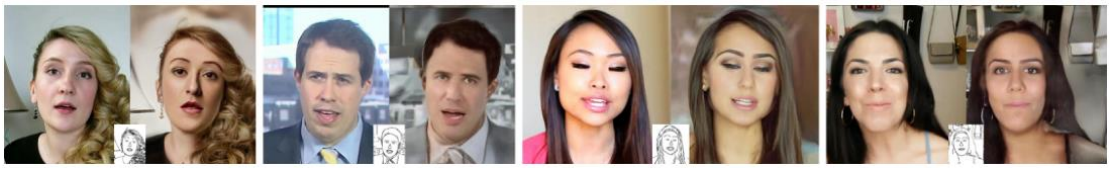


图 15　人脸→轮廓→人脸的示例

虽然其方法优于以前的方法，但在几种情况下仍然会失败。例如，由于标签地图中的信息不足，其模型难以合成转弯车辆。Wang 等认为，这可能通过添加额外的 3D 提示来解决，例如深度图。此外，其模型仍然不能保证对象在整个视频中具有一致的外观。偶尔，汽车可能会逐渐变色。如果使用对象跟踪信息强制同一对象在整个视频中共享相同的外观，则此问题可能会得到缓解。最后，当执行语义操作（例如树转换为建筑物）时，偶尔会出现可见的瑕疵，因为建筑物和树具有不同的标签形状。Wang 等认为，如果用较粗糙的语义标签训练模型可能会解决，因为训练后的模型对标签形状不太敏感。

5 基于自相似性与对比学习的图像跨域转换算法[1]

目前的多样化跨域转换算法还存在以下问题：

（1） 转换结果图像的内容结构与输入的源域图像的内容结构存在显著差异，无法满足对图像编辑前后结构保持要求严格的应用要求。

（2） 转换结果图像和参考图像（来自目标域）之间的风格差异导致颜色模式崩溃（仅学习一些显著的颜色模式），这意味着转换结果图像的颜色内容不够丰富，没有将参考图像的色彩空间全部学习到。

为了解决上述问题，受自注意力机制[19]的启发，赵磊等提出了一种称为 SSAL-GAN 的新算法，其中 SSAL 代表自结构注意力损失。自结构注意力损失函数确保转换图像的内容与源域图像的内容高度一致。此外，他们还设计了一个基于统计的颜色损失函数，以提高转换图像的颜色丰富性，具体包含内容损失，风格损失、对比损失和对抗损失。其算法运行的结果如图 16、17、18 所展示。

图 16　不同算法在 summer2winter 数据集上的实验结果对比

图 17　SSAL-GAN 算法在 summer2winter 数据集和 monet2photo 数据集上的实验结果

图 18　本文算法在 night2day 和 MWI 上的实验结果

6 总结

近年来，图像跨域转换在深度学习尤其是生成式对抗网络的基础上得到了飞速发展。许多学者提出了不同的损失函数和其他优化方法以改善图像跨域转换的效果和性能。未来，更多好的框架和算法的提出将进一步改善这一技术的应用效果，提升其应用水平并扩展其应用领域。

参考文献

[1] 基于自相似性与对比学习的图像跨域转换算法, 赵磊等, *计算机研究与发展*, 2022.

[2] Image-to-Image Translation with Conditional Adversarial Networks, Zhu et al., In *CVPR*, 2016.

[3] Context encoders: Feature learning by inpainting, Pathak et al., In *CVPR*, 2016.

[4] Colorful image colorization, Zhang et al., In *ECCV*, 2016.

[5] Precomputed real-time texture synthesis with markovian generative adversarial networks, Li et al., In *ECCV*, 2016.

[6] Generative image modeling using style and structure adversarial networks, Wang et al., In *ECCV*, 2016.

[7] Perceptual losses for real-time style transfer and super-resolution, Johnson et al., In *ECCV*, 2016.

[8] Learning temporal transformations from time-lapse videos, Zhou et al., In *ECCV*, 2016.

[9] Pixel-level domain transfer. Kim et al., In *ECCV*, 2016.

[10] Reducing the dimensionality of data with neural networks, Hinton et al., *Science*, 313(5786):504–507, 2006.

[11] U-net: Convolutional networks for biomedical image segmentation. Ronneberger et al., In *MIC-CAI*, 2015.

[12] High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs, Wang et al., In *CVPR*, 2017.

[13] Photographic image synthesis with cascaded refinement networks, Chen et al., In *ICCV*, 2017.

[14] Generative visual manipulation on the natural image manifold, Zhu et al., In *ECCV*, 2016.

[15] Video-to-Video Synthesis, Wang et al., In *CVPR*, 2018.

[16] Temporal generative adversarial nets with singular value clipping, Saito et al., In *ICCV*, 2017.

[17] MoCoGAN: Decomposing motion and content for video generation, Tulyakov et al., In *CVPR*, 2018.

[18] Generating videos with scene dynamics, Vondrick et al., In *NIPS*, 2016.

[19] Self-attention generative adversarial networks, Zhang et al., In *ICML*, 2019.