arXiv:2301.08846v1 [cs.LG] 21 Jan 2023

再生成学习：一种用于数据生成的学习范式

XuTan[1]*, **Tao Qin**[1], **Jiang Bian**[1], **Tie-Yan Liu**[1], **Yoshua Bengio**[2]1微软研究院
[2]Mila& 蒙特利尔大学[1]{xuta,taoqin,jiabia,tyliu}@microsoft.com[2]
yoshua.bengio@mila.quebec

摘要

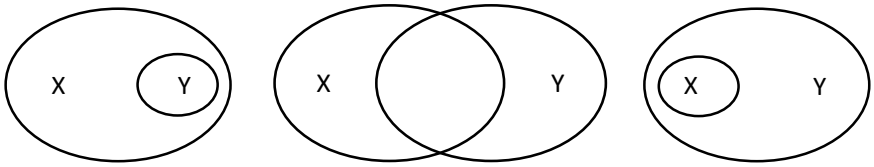
用于条件数据生成的机器学习方法通常会构建从源条件数据 X 到目标数据 Y的映射。目标数据 Y（例如文本、语音、音乐、图像、视频）通常具有高维且复杂的特性，并包含源数据中不存在的信息，这使得源数据与目标数据之间映射的学习难以做到有效且高效。本文提出一种用于数据生成的学习范式，称为再生成学习：首先从 X 生成Y'（ Y的抽象/表示），然后再从 Y'生成 Y。训练期间，通过人工设计规则或自监督学习，从 Y 获得 Y'，并用其学习 $X \to Y'$ 和 $Y' \to Y$。再生成学习将表征学习的概念扩展到数据生成任务，可视为传统表征学习的对应范式，因为：1）再生成学习处理用于数据生成的目标数据 Y的抽象（Y'），而传统表征学习处理用于数据理解的源数据 X 的抽象（X'）；2）再生成学习中的 $Y' \to Y$ 和表征学习中的 $X \to X'$ 都可以通过自监督方式（例如预训练）进行学习；3）再生成学习中从 X 到 Y' 的映射，以及表征学习中从 X' 到 Y的映射，都比从 X 到 Y的直接映射更简单。我们展示了再生成学习可以成为一种广泛适用的数据生成范式（例如文本生成、语音识别、语音合成、音乐创作、图像生成和视频生成），并能为数据生成方法的开发提供有价值的见解。

1 引言

1.1 数据理解与生成

典型的机器学习任务，例如自然语言处理 [49, 13, 77, 18, 8],语音 [2, 34, 51, 79, 71], 计算机视觉 [25, 69, 45, 33, 28], 等领域中的任务，通常处理从源数据 X 到目标数据 Y的映射。例如，在图像分类 [17]中，X 是图像，Y 是类别标签；在风格可控文本生成 [50]中，X 是风格标签，Y是句子；在文本转语音合成 [70, 71]中，X 是文本，Y是语音。
根据 X 和 Y所包含信息的相对多少，这些映射可分为数据理解 [45, 18], 、数据生成 [28, 8], 以及数据理解与生成相结合 [1, 31, 29, 10, 71]。图1展示了这三种任务类型，以及 X 与 Y之间的相对信息量：

*通讯作者：Xu Tan，xuta@microsoft.com



(a) 数据理解　　(b) 数据理解/生成　　(c) 数据生成

类型	信息	任务
理解	$X \gg Y$	图像分类、目标检测 句子分类、阅读理解
生成	$X \ll Y$	根据 ID/类别生成文本或合成图像
理解/生成	$X \not\gg Y$ 和 $X \not\ll Y$	文本转语音、自动语音识别， 文本到图像生成、说话头合成

图1：机器学习中的三种任务，以及源 X与目标 Y之间的相关信息。

• 数据理解任务，其中 X 包含的信息远多于 Y（例如，图像分类 [17, 45], 、目标检测 [27, 60], 、句子分类 [90], 、机器阅读理解 [55]）。

• 数据生成任务，其中 Y包含的信息远多于 X（例如文本生成 [8]或根据类别标签生成图像 [28, 42] ）。

• 数据理解/生成任务，其中 X 包含的信息与 Y 相比没有显著多出或少掉（例如图像迁移 [91], 、文本到图像合成 [57, 58, 86, 64, 11], 、神经机器翻译 [1, 31], 、文本转语音合成 [71, 70], 、自动语音识别 [34]）。在这种情况下，我们既需要对源 X 具备数据理解能力，也需要对目标 Y具备数据生成能力。

X 与 Y之间的信息不匹配，导致解决不同任务时需要采用不同策略。对于数据理解任务，与 Y相比， X 通常具有高维、复杂且冗余的特点；关键在于为 X 学习高度抽象或具有判别性的表示（有时需要去除不必要的信息），以便更好地预测 Y。因此，表征学习 [3][2]，尤其是自监督预训练 [18, 8] ，在过去几年里已成为深度学习研究中最热门的主题之一。对于数据生成任务，与 X相比， Y通常具有高维、复杂且冗余的特点；关键在于如何更好地表示 Y 的分布，并根据 X更好地生成 Y。对于数据理解/生成任务，模型需要兼具理解和生成能力，即从 X 中提取良好的表示，并充分生成 Y中的信息。

1.2 数据生成的挑战

对于数据生成任务以及数据理解/生成任务中的生成部分（在本文余下部分中，我们将这两种类型统称为数据生成任务），它们面临着独特的挑战，而传统的表征学习表述无法应对这些挑战。

• 首先，对于 Y所包含的信息远比 X复杂的生成任务，生成模型会面临严重的一对多映射问题（病态问题或条件不良问题） [5],，从而增加学习难度。例如，在类别条件图像生成中，类别标签“狗”可以对应多张包含狗的不同图像。对病态问题建模不当，可能导致模型在训练集上过拟合，并在测试集上泛化能力较差。

[2]深度学习领域最具影响力的会议之一是 ICLR，它是“国际学习表征会议”（The International Conference on Learning Representations）的简称。

• 其次，对于 X 所包含的信息与 Y相比没有显著多或少的生成任务（例如，语音识别 [34], 语音合成 [70], 说话人头部视频合成 [74]），存在两种情况：1） X 与 Y之间的映射不是一对一的（例如，在自动语音识别中，多个发音相同的词可以对应同一段语音；在文本转语音合成中，语速不同的多段语音可以对应同一个文本序列），这会面临上述相同的问题；2） X 与 Y之间存在一些虚假相关性。例如，在说话人头部视频合成中，源语音中的说话音色与目标视频中的头部姿态 [12] 毫无关联；在歌词到旋律生成中，目标旋律中的某些信息与源歌词 [39]毫无关联。在推理时拟合这些虚假相关性会对生成产生不利影响。

1.3 为什么再生成学习

一些生成模型（例如，生成对抗网络 [28],变分自编码器 [42],自回归模型 [51, 8],归一化流 [62, 41], 扩散模型 [36]）在多种数据生成任务中取得了快速进展。理想情况下，只要生成模型足够强大，就能拟合任意复杂的数据分布。然而在实践中，由于数据映射过于复杂、计算成本过高以及数据稀疏性等诸多原因，它们无法很好地建模复杂分布和一对多映射。类比数据理解任务，尽管理想情况下，学习一个强大的模型（例如卷积神经网络或 Transformer），直接将源数据分类为目标标签，就能达到很高的准确率，但实际仍会受到准确率偏低的问题困扰；而一些先进的表征学习方法，例如大规模自监督预训练 [18]，则可以大幅提升准确率。

本文提出一种用于数据生成任务的学习范式，称为再生成学习 。再生成学习并非直接从源数据 X生成目标数据 Y，而是先生成Y'（ Y的一种表示），再从 X 生成 Y Y'。再生成学习将表征学习的概念扩展到数据生成任务中，并学习目标数据 Y 的良好表示（Y'），以简化生成过程：1） $X \to Y'$ 映射的一对多程度低于$X \to Y$，因为 Y' 是 Y的紧凑/代表性版本；2） $Y' \to Y$映射可以通过自监督方式学习（Y' 由 Y获得），并可借助大规模预训练加以增强，这与传统数据理解任务表征学习中的做法类似（例如 BERT [18]）。

本文其余部分安排如下：第 2 节首先介绍再生成学习的基础表述及其与其他学习方法和范式的联系；第 3 节总结再生成学习的应用；最后，第 4 节列出再生成学习的一些研究机遇。

2 再生成学习的表述

本节将介绍再生成学习，它利用目标数据的中间表示 Y，以弥合 X 与 Y 之间的信息不匹配。再生成学习包含三个步骤：

• 步骤 1：将目标数据 Y 转换为抽象／表征版本 Y'。
• 步骤 2：学习一个模型，根据源数据 X 生成 Y'。
• 步骤 3：学习另一个模型，根据 Y' 生成 Y。

这种学习范式称为再生成学习，原因如下：1）从字面上看，它包含两个生成步骤：先生成 Y'，再生成 Y（即再生成）。这类似于表征学习中“表示”与“呈现”之间的关系[3]。2）从比喻意义上说，类比于表征学习中“表示”的含义，即“用一物来指代另一物”，再生成学习中的“再生成”意为生成一物来指代另一物（即生成 Y' 来指代 Y）。再生成学习有几个优点：1）可以通过自监督的方式学习从Y' 到 Y的映射，因此数据效率高得多；2）从 X 到 Y' 的映射比从 X 到 Y的直接映射更简单。

[3]严格来说，表征学习中的“represent”意为“用一物指代另一物”，这与意为“再次呈现”的“re-present”不同。然而，“represent”的含义与“re-represent”密切相关：假设你用 X' 来指代 X（即用 X' 来表征 X，因此 X' 就是 X的表示），那么在这种情况下，你实际上是用 X'再次呈现了 X。

表述	类别	方法	数据转换 ($Y \to Y'$)
基础	显式	傅里叶变换	语音/图像（例如，波形→频谱图）
		字素到音素转换	文本（例如，learning→lɜːrnɪŋ）
		音乐分析	音乐（MIDI→和弦/节奏）
		三维图像分析	图像（人脸到三维系数）
		下采样	语音/图像（例如，256*256→64*64）
	隐式	分析合成	图像/语音/文本 ($Y \to Z$)
		VAE 向量量化变分自编码器/VQ-GAN	
		扩散自编码器	
扩展	因子分解	AR	图像/语音/文本 ($Y \to Y_{1:t}$)
	扩散	DDPM	图像/语音/文本 ($Y_0 \to Y_t$)
	潜在扩散	变分自编码器 + 去噪扩散概率模型	图像/语音/文本 ($Y \to Z_0, Z_0 \to Z_t$)

表 1：用于 $Y \to Y'$ 转换的不同方法。

我们首先在第 2.1 节介绍再生成学习的三个步骤，并在第 2.2 节讨论再生成学习与现有方法之间的联系，在第 2.3 节讨论再生成学习与表征学习之间的关系。

2.1 基础表述

生成 Y'，目标为 Y。将 Y转换为 Y'时，需遵循三项原则：1） Y' 应比 Y更具抽象性和代表性；2）从 Y转换为 Y' 时移除的信息与源数据 X没有或只有很少的相关性，即 Y' 是 Y 的精简版本，但仍与 X 保持相关；3 ）从 Y 转换为 Y' 应当简单易行，例如可通过简单的变换或提取工具处理；如果需要通过模型学习实现转换，至少不应依赖标注数据。根据上述原则，可采用不同方式将 Y转换为 Y'，如表 1“基础表述”所示:

• 显式变换。我们可以通过一些显式变换方法将 Y转换为 Y'： 1）数学变换，例如傅里叶变换或小波变换。例如，我们可以使用短时傅里叶变换（STFT）将语音波形转换为线性刻度或梅尔刻度频谱图序列 [79, 65]。2）模态变换，例如字素到音素转换。例如，我们可以使用字素到音素转换模型 [68]，将文本/字符/字素序列转换为音素序列。3）数据分析。例如，我们可以使用一些音乐分析工具，从旋律序列中提取音乐模板（和弦、节奏等），以获得音乐的抽象表示 [39]；也可以使用三维人脸模型 [6]从人脸图像 [47] 中提取三维人脸参数。4）下采样。例如，我们可以将图像从 $256 * 256$ 分辨率简单地下采样至 $64 * 64$，或将语音序列的采样率从 48kHz 降至 24kHz。这些变换方法通常基于成熟的方法或工具，且由 Y转换得到的 Y' 通常采用显式的数据格式。

• 隐式变换。不同于显式变换通过构建完善的规则、算法或模型，将 Y转换为 Y'，并采用显式数据格式，端到端学习通过分析—合成流程或重建来学习中间的隐式表示，从而实现这种转换。常用模型包括自编码器（AE）、去噪自编码器（DAE）、变分自编码器（VAE） [42], 向量量化自编码器（VQ-VAE） [76, 59], 扩散自编码器（DiffsuionAE） [52], 等。除了学习将 Y转换为 Y'的编码器之外，这些模型还可以进一步学习一个将 Y' 转换回 Y 的解码器，该解码器可用于再生成学习的第三步。

从 Y' 源生成 X。 </style}从 X 到 Y' 的生成可以采用任何机器学习方法来实现，类似于用于建模 $X \to Y$的方法（例如，自回归模型、生成对抗网络、变分自编码器、流模型、扩散）。由于 Y' 是从满足上述三项原则的 Y中提取的，因此 $X \to Y'$ 任务比 $X \to Y$任务更简单。例如，在文本转语音合成中，从文本生成梅尔频谱图远比生成波形简单；生成三维人脸参数

比在说话人头部视频合成中根据语音生成的人脸图像简单得多；根据源条件生成表示序列（潜在编码序列），也比在条件图像或视频生成中生成像素级图像简单得多。

生成目标 Y 从 Y'。 由于我们可以借助工具或模型轻松地从 Y 中提取 Y'，而无需依赖任何配对训练数据（即 X 和 Y），因此我们可以通过自监督方式训练模型，根据 Y' 预测 Y；配对训练数据（Y'，Y）可以低成本地大规模收集。我们还可以仅使用大量非配对数据 Y 进行预训练，以增强 $Y' \to Y$ 的能力。请注意，使用自编码方法将 Y 转换为 Y' 时，我们已经可以得到一个将 Y' 转换为 Y 的解码器，无需再训练另一个模型。

通过 Y' 从 X 推断 Y。我们讨论如何将两个生成步骤 $X \to Y'$ 和 $Y' \to Y$ 结合起来，从 X 生成 Y。请注意，由于 $X \to Y$ 映射是一对多的，因此根据 X 预测 Y 需要以分布方式进行，即 $Y \sim P(Y|X)$，而不是逐点预测。同样，由于 Y' 是 Y 的紧凑版本，会造成信息损失，因此 $Y' \to Y$ 也是一种一对多映射，也应以分布方式建模，即 $Y \sim P(Y|Y')$。此外，尽管与 $X \to Y$ 中的问题相比，$X \to Y'$ 中的不适定映射问题已大幅缓解，但总体而言，$X \to Y'$ 映射仍是一对多的，因此也应以分布方式建模，即 $Y' \sim P(Y'|X)$。另一方面，$Y \to Y'$ 通常是逐点的，因为它由确定性函数转换而来。不过，它也可以是分布式的，例如使用 VAE 编码器，根据 Y 求出 Y' 的均值和方差。然而，无论 $Y \to Y'$ 是逐点还是分布式，$X \to Y'$ 和 $Y' \to Y$ 都应以分布方式建模。我们通常利用深度生成模型（例如自回归模型、生成对抗网络、变分自编码器、归一化流和去噪扩散概率模型）来学习条件分布 $P(Y'|X)$ 和 $P(Y|Y')$。之后，给定一个数据样本 X，我们可以先从条件分布 $P(Y'|X)$ 中采样得到 Y'，然后给定采样得到的数据 Y'，再从条件分布 $P(Y|Y')$ 中采样得到 Y，即 $Y' \sim P(Y'|X)$，$Y \sim P(Y|Y')$。

2.2 与其他方法的联系

本小节将讨论一些与再生成学习相关的方法，包括再生成学习的基础版本和扩展版本，以及不属于再生成学习但与之相关的方法。

再生成学习的基础版本 根据第 2.1 节中的表述，有些方法可视为再生成学习的基础版本，例如：

• 基于模板的方法从目标数据中提取具体模板，然后根据由源数据生成的模板再生成目标数据 [39]。

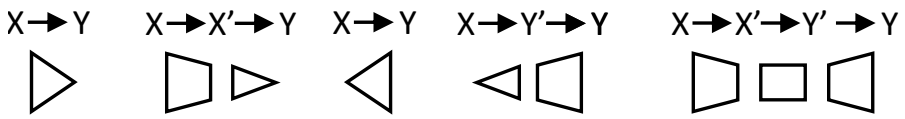
• 声码器方法使用傅里叶变换将语音波形转换为频谱图，并根据由源文本或语音生成的频谱图再生成波形 [79,61, 43]。

• 字素/音素转换将文本/字符序列转换为音素序列，并根据由源语音生成的音素序列再生成文本/字符序列 [87]。

• 将目标数据转换为表示，并从由源数据生成的表示中再生成目标数据的方法 [76, 59, 57, 71]。

再生成学习的扩展版本 有些方法并不严格遵循再生成学习的三步定义，但可视为再生成学习的扩展版本。我们在表 1 的“扩展表述”中展示了这些 $Y \to Y'$ 转换方法，并作如下介绍。

• 在自回归生成中，我们使用结构因子分解将目标数据 Y 分解到每个步骤 $Y_{1:\,t}$，以自回归的方式进行。我们可以将部分序列 $Y_{<t}$ 视为 $Y_{<t+1}$ 的简单/紧凑版本，并将所有中间部分序列 $Y_{<1}$，$Y_{<2}$，$...Y_{<T}$ 视为 $Y_{1:\,T}$ 的简单/紧凑版本，其中 T 是目标序列 Y 的长度。在这种情况下，再生成包含多个生成步骤，每一步都由 $Y_{<t}$ 生成 $Y_{<t+1}$。



(a) (b) (c) (d) (e)

图（a）：呈现（$X \to$）。图（b）：表示（$X \to X' \to$）。图（c）：生成（$\to Y$）。图（d）：再生成（$\to Y' \to Y$）。图（e）：表示 + 再生成（$X \to X' \to Y' \to Y$）。

范式	原始	紧凑	自监督学习	易于映射
（b）表征学习	X	X'	$X \to X'$	$X' \to Y$
(d) 再生成学习	Y	Y'	$Y' \to Y$	$X \to Y'$
(e) 组合	X, Y	X', Y'	$X \to X', Y' \to Y$	$X' \to Y'$

图2：再生成学习与传统表征学习在$X \to Y$上的比较。图(a)-(e)中的三角形、梯形和四边形表示信息的变化。图(a)：数据理解任务，其中 X 包含的信息比 Y更多。图(b)：采用表征学习的数据理解任务，其中 X' 是 X的更紧凑版本。图(c)：数据生成任务，其中 Y 包含的信息比 X更多。图(d)：采用再生成学习的数据生成任务，其中 Y' 是 Y的更紧凑版本。图(e)：结合表征学习与再生成学习的数据理解/生成任务（例如语音到说话人头部视频合成等序列到序列学习任务），其中 X' 和 Y' 包含的信息量相当，且分别是 X 和Y的紧凑版本（例如， X' 可以是通过自监督模型学习到的语音表征，而 Y'可以是在说话人头部视频合成中由三维人脸模型提取的三维系数）。

• 在去噪扩散概率模型中， [36], 我们可以逐步向原始数据 Y（Y_0）添加噪声，得到数据的扩散版本 Y_t [36]。与自回归生成类似，扩散模型也是再生成学习的一种扩展形式，其中 Y_{t+1} 可以视为 Y_t的紧凑版本。基于迭代的非自回归序列生成 [26] 也可以本着相同的理念视为再生成学习。

• 有些方法将自编码器与扩散模型结合起来 [75, 63], ，先使用变分自编码器或向量量化变分自编码器将目标数据 Y 转换为潜向量 Z （记作 Z_0），然后逐步向 Z_0 添加噪声，得到 Z_t。这些方法首先通过扩散模型生成 Z_0，然后使用变分自编码器或向量量化变分自编码器的解码器再生成目标 Y。

与再生成学习相关的方法 我们介绍一些与再生成学习相关、但不属于再生成学习的方法。

• 生成 Y后，可以使用后精炼方法进一步提升生成的Y的质量。但我们不将其视为再生成学习，因为这些方法不采用 Y的抽象版本，而是直接将 Y作为生成目标。

• 在扫视机制 [53, 37], 中，会扫视 Y中的一些词元，并将其作为输入来预测 Y。我们也不将这种机制视为再生成学习，因为学习过程中没有使用 Y'。

• 在知识蒸馏 [9, 35, 40], 中，学生模型将教师模型的输出（Y'）作为学习目标，这似乎属于再生成学习。然而，知识蒸馏由于多种原因并不属于再生成学习。首先，知识蒸馏中的学生模型不会从 Y'再生成 Y。其次，知识蒸馏可以是分布级 [35] 或点级（例如，序列级蒸馏 [40]）；在分布级蒸馏中， Y' 比 Y更丰富，而不像再生成学习那样， Y' 是 Y的抽象版本。

我们将在未来工作中研究再生成学习与 GFlowNet [4] 等其他基于迭代的学习方法之间的关系。

Task	X	Y	Y'	$Y \rightarrow Y'$ & $Y' \rightarrow Y$
语音合成	Text	波形	频谱图 / 编码	短时傅里叶变换 & 声码器 / 编解码器
语音识别	语音	字符	音素	字素转音素 & 音素转字素
文本生成	文本/知识	Text	模板	文本转模板与模板转文本
歌词/视频转旋律	歌词/视频	旋律	音乐模板	音乐分析与生成
说话头合成	语音	视频	三维人脸参数	3D人脸分析与渲染
图像/视频/声音生成	类别/文本	图像/视频/声音	潜在编码	编解码器提取与生成

表 2：利用再生成学习的典型数据生成任务（文本、语音、音乐、声音、图像和视频）。

2.3 再生成学习与表示学习

再生成学习将表征学习的概念扩展到数据生成，因此可以将其视为一种用于数据生成的特殊表征学习。此外，我们可以将再生成学习视为传统表征学习的对应形式，因为：1）再生成学习处理目标数据 Y 的抽象（Y'），用于数据生成，而传统表征学习处理源数据 X 的抽象（X'），用于数据理解；2）再生成学习中的 $Y' \rightarrow Y$ 过程和传统表征学习中的 $X \rightarrow X'$ 过程都可以通过自监督方式（例如预训练）进行学习；3）再生成学习中从 X 到 Y' 的映射，以及传统表征学习中从 X' 到 Y 的映射，都比从 X 到 Y 的直接映射更简单。图2展示了再生成学习与传统表征学习之间的比较。

3 再生成学习的应用

条件数据生成中的各种任务（例如文本生成、语音识别、语音合成、音乐创作、图像生成和视频生成）都可以从这种再生成学习范式中受益。我们在表 2 中列出了一些典型的生成任务，并介绍如下运用再生成学习的各项任务。

• 文本转语音合成。在文本转语音合成中（例如 Tacotron 2 [65], FastSpeech [61]），目标波形通过短时傅里叶变换（STFT）转换为梅尔频谱序列，并使用声学模型根据源文本预测梅尔频谱序列，再使用声码器根据预测的梅尔频谱序列生成波形。还有一些研究 [32, 48]使用向量量化变分自编码器，将语音波形转换为连续向量或离散词元，以衔接文本与语音之间的映射。

• 自动语音识别。一些自动语音识别领域的研究 [78, 87, 89] 首先使用字素到音素转换（G2P）将目标字符序列转换为音素序列，并使用声学模型从源语音序列预测音素序列，然后使用另一个模型根据预测的音素序列生成字符序列。

• 文本生成。一些文本生成领域的研究 [80, 84] 首先使用模板提取或句法分析将目标句子转换为模板或句法序列，然后根据源文本或从头开始预测模板或句法序列，并根据预测的模板或句法序列生成目标文本。

• 旋律生成。一些旋律生成领域的研究 [39, 81, 15, 92, 19] 首先从目标旋律序列中提取音乐模板，然后根据源条件（例如歌词、背景视频、舞蹈视频）预测模板，并根据预测的模板生成旋律，例如 TeleMelody [39]。

• 说话人头部视频合成。在说话人头部视频合成中，一些用于高质量人脸合成的流程 [74, 38, 85, 46, 67,47, 72] 通常会通过三维人脸模型从目标人脸图像中提取三维人脸参数 [6, 73, 24], ，再根据源语音或文本生成三维人脸参数，最后根据生成的三维人脸参数生成图像。

• 图像/视频/声音生成。在图像/视频/声音生成中，由于图像或音频过于复杂，会带来过多的计算开销，因此许多研究 [57, 20, 82, 56, 7, 83, 44] 首先使用 VQ-VAE/VQ-GAN [76, 23, 88]将图像/波形转换为离散视觉/音频词元，或使用 CLIP [54, 58], 将其转换为连续隐表示，然后生成这些视觉/音频

词元或隐藏表示，生成时以文本为条件（例如 DALL-E 1/2 [57, 58], StableDiffusion [63], AudioGen [44] ），或从前缀/头开始生成（例如 AudioLM [7], VideoGPT [82], LVT [56]），最后使用先前 VQ-VAE/VQ-GAN 中的解码器，或 GAN、变分自编码器、归一化流或扩散模型等独立解码器，将生成的词元重新生成图像/视频/音频。

一般来说，只要符合以下某些情况，条件数据生成任务就可以利用再生成学习：

• 目标数据的维度过高、过于复杂而难以生成，或生成时计算成本过高，例如文本转语音合成中的波形生成 [65, 61, 70],以及图像/视频/声音生成 [57, 20, 82, 56, 63, 44]。在这种情况下，将目标数据 Y 转换为更紧凑的表示 Y'，能够大幅降低计算成本，并使模型得以专注于如何生成目标数据的高层抽象信息或语义信息，而非细枝末节。
• 源数据 X 与目标数据 Y 包含过多彼此无关的信息（即 $X \cap Y \ll X \cup Y$），例如条件旋律生成中的歌词/视频与旋律 [39, 81, 15, 92, 19], ，以及说话人头部视频合成中的语音与人脸图像 [74, 38, 85, 46, 67, 47]。直接学习 X 与 Y 之间的映射会导致过拟合。因此，将 Y 转换为更紧凑的表示 Y'，能够使 X 与 Y' 之间的映射不那么不适定，从而减轻模型的学习负担。
• 缺少配对的 X 和 Y，因此可以利用再生成学习，仅基于目标数据 Y，通过大规模自监督学习来训练 $Y' \to Y$。

4 再生成学习的研究机遇

我们探讨了一些研究机遇，以增强再生成学习解决各种数据生成任务的能力，并在各个小节和表 3 中列出了一些相应的研究问题（RQ），主要从三个视角展开：1）如何获得 Y'；2）如何学习 $X \to Y'$ 和 $Y' \to Y$ 之间的映射；3）如何减少再生成学习流程中的训练-推理不匹配。

4.1 如何获得 Y'

如何找到合适的 Y'，对于 $X \to Y'$ 和 $Y' \to Y$ 之间的映射至关重要。例如，在文本转语音合成中，梅尔频谱图和梅尔频率倒谱系数（MFCCs）都可以作为目标波形 Y 的 Y'。然而，研究表明，梅尔频谱图远优于 MFCCs，因为 MFCCs 过于抽象，丢失了大量细粒度信息，因此难以从 Y' 重建 Y。下文列出了若干可能的研究方向，以获得更好的 Y'。

更优的隐式学习。 除了使用一些人工设计规则或成熟工具来寻找Y'之外，我们还可以从 Y 中自动学习 Y'。典型方法基于分析合成流程，例如变分自编码器（VAE） [42] 和向量量化变分自编码器（VQ- VAE） [76]。例如，在文本转语音合成 [14, 71, 48], 中，NaturalSpeech [71] 利用 VAE，而 DelightfulTTS 2 [48] 利用 VQ-VAE，从语音波形 Y 中学习中间表示 Y'，并从 Y' 重建 Y。其动机在于，常用的中间表示（例如梅尔频谱图）是通过 SFTF算法提取的，这些算法可能并非最优；而由 VAE 学得的表示可能更能表征波形，并使生成过程更容易。在图像和视频领域，VQ-VAE [76] 被广泛用于学习表示图像和视频的离散视觉词元。然而，另一些工作尝试通过引入层次化学习（例如 VQ-VAE-2 [59]）、添加对抗损失（例如 VQ-GAN [23]）和感知损失（例如 PECO [22]）、使用残差量化器 [30, 88, 16], ，或利用扩散模型学习隐藏表示 [52]，进一步改进 VQ-VAE 的离散词元提取。

• **研究问题1**：如何设计更好的分析合成方法（超越变分自编码器、向量量化变分自编码器、扩散自编码器，等）来学习 Y'?
• **研究问题2**：除了分析合成之外，如何设计更好的学习范式来学习 Y'?

视角	ID	研究问题
$Y \to Y'$	1	如何设计更好的分析合成方法（超越 VAE、向量量化变分自编码器、扩散自编码器等）来学习 Y'?
	2	如何设计分析合成之外的更好学习范式来学习 Y'?
	3	如何利用非配对数据 Y和/或配对数据（X，Y）来学习 Y'?
	4	学习 Y'时，如何更好地权衡 $X \to Y'$和 $Y' \to Y$映射的难度?
	5	如何将语义含义与感知细节解耦，从而学习更注重语义而非细节的 Y'?
	6	对于每个数据生成任务，如何确定 Y' 的离散或连续格式?
$X \to Y'$ $Y' \to Y$	7	如何设计更好的生成模型，以学习 $X \to Y'$ 和 $Y' \to Y$ 之间的映射?
	8	如何利用语义转换和细节渲染的假设来设计更好的方法?
	9	如何利用大规模自监督学习实现 $Y' \to Y$ 映射?
$X \to Y' \to Y$	10	如何减少再生成学习中的训练-推理不匹配?

表 3：关于再生成学习的研究问题。

学习 Y' 时考虑 X。 </style}一种直觉是，抽象式表示 Y' 不应仅取决于 Y，还应与 X 相关联，以便根据 X更好地预测 Y'。为此，我们将学习策略表述如下。我们将目标数据 Y输入目标编码器，得到 $Y' = f_Y(Y)$，并使用目标解码器重构 $Y \sim P(Y|Y')$。我们进一步将源数据 X 输入源编码器，得到 $X' = f_X(X)$，并通过使用对比损失或 L1/L2 损失等损失函数，鼓励 X'和 Y' 彼此包含信息：$||X' - f_{Y'}(Y')||_2^2$和 $||Y' - f_{X'}(X')||_2^2$。此外，我们使用损失 $-\log P(Y|f_{X'}(X'))$，促使目标解码器适应根据 X'预测得到的 Y'，这也能解决第 4.3 节中提到的问题。总损失如下：

$$L = -\log P(Y|Y') + ||X' - f_{Y'}(Y')||_2^2 + ||Y' - f_{X'}(X')||_2^2 - \log P(Y|f_{X'}(X')), \quad (1)$$

其中 $Y' = f_Y(Y)$ 和 $X' = f_X(X)$。我们可以利用非配对数据 Y和配对数据（X，Y），根据公式 1 学习 Y'。使用非配对数据 Y进行学习时，我们只使用第一个损失项。我们将公式 1 的理论和实证研究留作未来工作，同时也欢迎有兴趣的研究人员探讨这一课题。

除了利用配对数据（X，Y）学习优质的 Y'，我们还可以在选择如何将 Y转换为 Y'时，利用关于 X 的一些先验知识。例如，在自动语音识别中，我们知道源 X 是语音，因此会将文本序列 Y转换为音素序列 Y'，以便于将 X 转换为 Y'。在文本到图像合成（例如 DALL-E 2 [58]）中，图像会被转换为接近文本表示的 CLIP [54] 特征，以便于进行文本到图像的转换。

然而，$X \to Y'$ 与 $Y' \to Y$的学习难度之间存在权衡，因为在学习 Y' 时考虑更多的 X 会让 $X \to Y'$ 更容易学习，却让 $Y' \to Y$更难学习。我们应当做好权衡，以获得更好的整体表现。

• **研究问题3**：如何利用非配对数据 Y和/或配对数据（X，Y）来学习 Y'?
• **研究问题4**：在学习Y'时，如何更好地权衡 $X \to Y'$ 与 $Y' \to Y$映射之间的难度?

将语义与感知细节解耦。一般来说，$X \to Y'$ 更关注从源到目标的语义映射/转换，而 $Y' \to Y$更关注渲染感知细节，以获得目标 Y。例如，在文本到图像合成中，$X \to Y'$将源文本转换为描述目标图像语义含义的离散视觉词元，而 $Y' \to Y$则根据视觉词元渲染图像细节。如何设计一种学习机制，更好地将语义含义与感知细节解耦，并让 Y' 专注于语义含义，将是一个很好的研究机会。

• **研究问题5**：如何解耦语义含义与感知细节，从而学习更具语义性而非细节性的 Y'?

离散与连续 Y'。离散和连续的 Y' 都可用于建立 X 与 Y之间的映射。例如，DelightfulTTS 使用类似向量量化变分自编码器的模型对语音波形进行量化，并将连续向量用作 Y'；而 DiscreTalk [32] 则使用离散

词元作为 Y'。哪种表示形式（离散还是连续）更适合 Y'，也是一个值得探讨的问题。

• **研究问题6**：如何为每个数据生成任务确定 Y' 的离散或连续格式?

4.2 如何学习 $X \to Y'$ 和 $Y' \to Y$

再生成学习将条件数据生成任务 $X \to Y$ 分解为数据转换和数据渲染两个过程。数据转换过程将源数据 X 转换到目标域，同时保留其具体语义，但不一定包含细粒度细节。数据渲染过程则进一步渲染目标数据的细粒度细节，从而实现高质量的数据生成。粗略地说，$X \to Y'$ 更多承担数据转换的作用，而 $Y' \to Y$ 更多承担数据渲染的作用。

如何设计更好的训练方法来学习 $X \to Y'$ 和 $Y' \to Y$，对于再生成学习中 $X \to Y$ 映射的最终性能至关重要。自回归模型 [51, 8], 变分自编码器 [42], 生成对抗网络 [28], 归一化流 [21, 62], 以及扩散模型 [66, 36] 等先进生成模型可以发挥重要作用。此外，考虑到 $Y' \to Y$ 映射通常可以通过大规模自监督方法学习，我们在学习 $Y' \to Y$ 时应更多利用非配对数据。

• **研究问题7**：如何设计更好的生成模型来学习 $X \to Y'$ 和 $Y' \to Y$ 之间的映射?

• **研究问题8**：如何利用 $X \to Y'$ 中的语义转换假设和 $Y' \to Y$ 中的细节渲染来设计更好的方法?

• **研究问题9**：如何利用大规模自监督学习来学习 $Y' \to Y$ 映射?

4.3 如何减少 $X \to Y' \to Y$ 中的训练-推理不匹配

$Y' \to Y$ 模型以自监督方式训练，训练时从 Y 中提取 Y'。然而，推理时则根据 X 预测 Y'，这会导致训练-推理不匹配。如何减少这种不匹配，对于确保该级联系统（$X \to Y'$ 和 $Y' \to Y$）的性能至关重要。一种直接的方法是在 $X \to Y'$ 和 $Y' \to Y$ 之间设计端到端优化方法，同时仍将 Y' 保留为中间表示。例如，NaturalSpeech [71]利用变分自编码器和归一化流，通过双向先验/后验优化实现端到端学习。我们能否设计其他方法来减少再生成学习中的训练-推理不匹配?

• **研究问题10**：如何减少再生成学习中的训练-推理不匹配?

5 结论

本文提出了一种用于数据生成任务的学习范式，称为再生成学习。顾名思义，它意味着将数据生成两次：首先从源数据 X 生成中间表示 Y'，然后从 Y' 生成目标数据 Y。从比喻意义上说，它意味着生成 Y' 作为中间表示，以表征 Y，这与表征学习类似。再生成学习可视为传统表征学习的对应范式：再生成学习处理数据生成中目标数据 Y 的抽象（Y'），而表征学习处理数据理解中源数据 X 的抽象（X'）；再生成学习中的 $Y' \to Y$ 过程和表征学习中的 $X \to X'$ 过程，均可通过自监督方式学习（从字面意义上说，它们也相互对应：呈现 $\to$ 表示与生成 $\to$ 再生成）。我们探讨了再生成学习与其他方法之间的联系，展示了多种能够从再生成学习中受益的数据生成任务，并进一步指出了再生成学习的一些研究机遇。再生成学习可成为一种广泛应用于高质量数据生成的范式，并为开发数据生成方法提供有价值的见解。

致谢

我们感谢微软亚洲研究院机器学习组的同事们参与讨论。感谢孟奇和刘畅为本文提供了有益的见解和建议。

参考文献

[1] Dzmitry Bahdanau、Kyunghyun Cho 和 Yoshua Bengio。通过联合学习对齐与翻译实现神经机器翻译。*arXiv*预印本 *arXiv:1409.0473*，2014年。[2] Jacob Benesty、M Mohan Sondhi、Yiteng Huang 等。《*Springer*语音处理手册》，第1卷。Springer，2008年。[3] Yoshua Bengio、Aaron Courville 和 Pascal Vincent。表征学习：综述与新视角。*IEEE*模式分析与机器智能汇刊，35(8)：1798–1828，2013。[4] Yoshua Bengio、Tristan Deleu、Edward J Hu、Salem Lahlou、Mo Tiwari 和 Emmanuel Bengio。GFlowNet基础。*arXiv*预印本 *arXiv:2111.09266*，2021年。[5] Mario Bertero、Tomaso A Poggio 和 Vincent Torre。早期视觉中的不适定问题。*IEEE*会刊，76(8)：869–889，1988年。[6] Volker Blanz 和 Thomas Vetter。用于合成三维人脸的可变形模型。载于第*26*届计算机图形学与交互技术年度会议论文集，第187–194页，1999年。[7] Zalán Borsos、Raphaël Marinier、Damien Vincent、Eugene Kharitonov、Olivier Pietquin、Matt Sharifi、Olivier Teboul、David Grangier、Marco Tagliasacchi 和 Neil Zeghidour。AudioLM：一种用于音频生成的语言建模方法。*arXiv*预印本 *arXiv:2209.03143*，2022年。[8] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。语言模型是少样本学习者。神经信息处理系统进展，33：1877–1901，2020年。[9] Cristian Bucila、Rich Caruana 和 Alexandru Niculescu-Mizil。模型压缩。2006年。[10] William Chan、Navdeep Jaitly、Quoc Le 和 Oriol Vinyals。聆听、关注与拼写：一种用于大词汇量会话语音识别的神经网络。载于*2016*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），第4960–4964页。IEEE，2016年。[11] Huiwen Chang、Han Zhang、Jarred Barber、AJ Maschinot、Jose Lezama、Lu Jiang、Ming-Hsuan Yang、Kevin Murphy、William T Freeman、Michael Rubinstein 等。Muse：通过掩码生成式变换器实现文本到图像生成。*arXiv*预印本 *arXiv:2301.00704*，2023。[12] Lele Chen、Ross K Maddox、Zhiyao Duan 和 Chenliang Xu。采用动态逐像素损失的分层跨模态说话人脸生成。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第7832–7841页，2019年。[13] Ronan Collobert、Jason Weston、Léon Bottou、Michael Karlen、Koray Kavukcuoglu 和 Pavel Kuksa。自然语言处理（几乎）从零开始。机器学习研究期刊，12（文章）：2493–2537，2011年。[14] Jian Cong、Shan Yang、Lei Xie 和 Dan Su。Glow-WaveGAN：从基于生成对抗网络的变分自编码器中学习语音表征，以实现高保真、基于流的语音合成。*arXiv*预印本 *arXiv:2106.10831*，2021年。[15] Shuqi Dai、Zeyu Jin、Celso Gomes 和 Roger B Dannenberg。通过分层音乐结构表示实现可控深度旋律生成。*arXiv*预印本 *arXiv:2109.00663*，2021年。

[16] Alexandre Défossez、Jade Copet、Gabriel Synnaeve 和 Yossi Adi。高保真神经音频压缩。*arXiv*预印本 *arXiv:2210.13438*，2022年。[17] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。ImageNet：大型分层图像数据库。载于*2009*年*IEEE*计算机视觉与模式识别会议，第248–255页。IEEE，2009年。[18] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向Transformer预训练。*arXiv*预印本 *arXiv:1810.04805*，2018年。[19] Shangzhe Di、Zeren Jiang、Si Liu、Zhaokai Wang、Leyan Zhu、Zexin He、Hongming Liu 和 Shuicheng Yan。利用可控音乐Transformer生成视频背景音乐。载于第*29*届*ACM*国际多媒体会议论文集，第2037–2045页，2021年。[20] Ming Ding、Zhuoyi Yang、Wenyi Hong、Wendi Zheng、Chang Zhou、Da Yin、Junyang Lin、Xu Zou、Zhou Shao、Hongxia Yang 等。CogView：借助Transformer精通文本到图像生成。*arXiv*预印本 *arXiv:2105.13290*，2021年。[21] Laurent Dinh、David Krueger 和 Yoshua Bengio。NICE：非线性独立成分估计。*arXiv* 预印本 *arXiv:1410.8516*，2014 年。[22] Xiaoyi Dong、Jianmin Bao、Ting Zhang、Dongdong Chen、Weiming Zhang、Lu Yuan、Dong Chen、Fang Wen 和 Nenghai Yu。PECO：用于视觉Transformer的BERT预训练的感知码本，2021年。[23] Patrick Esser、Robin Rombach 和 Bjorn Ommer。驾驭Transformer实现高分辨率图像合成。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第12873–12883页，2021年。[24] Yao Feng、Haiwen Feng、Michael J Black 和 Timo Bolkart。从野外图像中学习可动画化的精细三维人脸模型。*ACM*图形学汇刊（*TOG*），40(4)：1–13，2021年。[25] David Forsyth 和 Jean Ponce。《计算机视觉：一种现代方法。》Prentice Hall，2011年。[26] Marjan Ghazvininejad、Omer Levy、Yinhan Liu 和 Luke Zettlemoyer。Mask-Predict：条件掩码语言模型的并行解码。载于*2019*年自然语言处理经验方法会议与第*9*届自然语言处理国际联合会议论文集（*EMNLP-IJCNLP*），第6112–6121页，2019年。[27] Ross Girshick。Fast R-CNN。载于*IEEE*国际计算机视觉会议论文集，第1440–1448页，2015年。[28] Ian Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio。生成对抗网络。神经信息处理系统进展，27，2014年。[29]Alex Graves。利用循环神经网络进行序列转换。*arXiv*预印本 *arXiv:1211.3711*，2012年。[30] Robert Gray。矢量量化。*IEEE ASSP*杂志，1(2)：4–29，1984年。[31] Hany Hassan、Anthony Aue、Chang Chen、Vishal Chowdhary、Jonathan Clark、Christian Feder- mann、Xuedong Huang、Marcin Junczys-Dowmunt、William Lewis、Mu Li 等。实现自动中英新闻翻译的人类水平。*arXiv*预印本 *arXiv:1803.05567*，2018年。[32] Tomoki Hayashi 和 Shinji Watanabe。DiscreTalk：将文本转语音视为机器翻译问题。*arXiv*预印本 *arXiv:2005.05525*，2020年。[33] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集，第770–778页，2016年。

[34] Geoffrey Hinton、Li Deng、Dong Yu、George E Dahl、Abdel-rahman Mohamed、Navdeep Jaitly、Andrew Senior、Vincent Vanhoucke、Patrick Nguyen、Tara N Sainath等。用于语音识别声学建模的深度神经网络：四个研究团队的共同观点。*IEEE*信号处理杂志，29(6):82–97，2012年。[35] Geoffrey Hinton、Oriol Vinyals、Jeff Dean等。蒸馏神经网络中的知识。*arXiv*预印本 *arXiv:1503.02531*，2(7)，2015年。[36] Jonathan Ho、Ajay Jain和Pieter Abbeel。去噪扩散概率模型。*arXiv*预印本 *arXiv:2006.11239*，2020年。[37] Chenyang Huang、Hao Zhou、Osmar R Zaïane、Lili Mou和Lei Li。通过逐层预测和深度监督实现非自回归翻译。*arXiv*预印本 *arXiv:2110.07515*，2021年。[38] Xinya Ji、Hang Zhou、Kaisiyuan Wang、Wayne Wu、Chen Change Loy、Xun Cao和Feng Xu。音频驱动的情感视频肖像。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第14080–14089页，2021年。[39] Zeqian Ju、Peiling Lu、Xu Tan、Rui Wang、Chen Zhang、Songruoyao Wu、Kejun Zhang、Xiangyang Li、Tao Qin和Tie-Yan Liu。TeleMelody：采用基于模板的两阶段方法实现歌词到旋律生成。*arXiv*预印本 *arXiv:2109.09617*，2021年。[40] Yoon Kim和Alexander M Rush。序列级知识蒸馏。载于*2016*年自然语言处理实证方法会议论文集，第1317–1327页，2016年。[41] Diederik P Kingma和Prafulla Dhariwal。Glow：具有可逆 1×1 卷积的生成流。载于第*32*届神经信息处理系统国际会议论文集，第10236–10245页，2018年。[42] Diederik P Kingma和Max Welling。自编码变分贝叶斯。*arXiv*预印本 *arXiv:1312.6114*，2013年。[43] Jungil Kong、Jaehyeon Kim和Jaekyoung Bae。HiFi-GAN：用于高效、高保真语音合成的生成对抗网络。神经信息处理系统进展，33:17022–17033，2020年。[44] Felix Kreuk、Gabriel Synnaeve、Adam Polyak、Uriel Singer、Alexandre Défossez、Jade Copet、Devi Parikh、Yaniv Taigman和Yossi Adi。AudioGen：文本引导的音频生成。*arXiv*预印本 *arXiv:2209.15352*，2022年。[45] Alex Krizhevsky、Ilya Sutskever和Geoffrey E Hinton。使用深度卷积神经网络进行ImageNet分类。神经信息处理系统进展，25:1097–1105，2012年。[46] Avisek Lahiri、Vivek Kwatra、Christian Frueh、John Lewis和Chris Bregler。LipSync3D：利用姿态和光照归一化，从视频中数据高效地学习个性化3D说话人脸。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第2755–2764页，2021年。[47] Jun Ling、Xu Tan、Liyang Chen、Runnan Li、Yuchao Zhang、Sheng Zhao和Li Song。StableFace：分析并提升说话人脸生成的运动稳定性。*arXiv*预印本 *arXiv:2208.13717*，2022年。[48] Yanqing Liu、Ruiqing Xue、Lei He、Xu Tan和Sheng Zhao。DelightfulTTS 2：采用对抗式向量量化自编码器的端到端语音合成。*arXiv*预印本 *arXiv:2207.04646*，2022年。[49] Christopher Manning和Hinrich Schutze。统计自然语言处理基础。麻省理工学院出版社，1999年。[50] Kathleen McKeown。文本生成。剑桥大学出版社，1992年。

[51] Aaron van den Oord、Sander Dieleman、Heiga Zen、Karen Simonyan、Oriol Vinyals、Alex Graves、Nal Kalchbrenner、Andrew Senior和Koray Kavukcuoglu。WaveNet：一种原始音频生成模型。*arXiv*预印本 *arXiv:1609.03499*，2016年。[52] Konpat Preechakul、Nattanat Chatthee、Suttisak Wizadwongsa和Supasorn Suwajanakorn。扩散自编码器：迈向有意义且可解码的表示。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第10619–10629页，2022年。[53] Lihua Qian、Hao Zhou、Yu Bao、Mingxuan Wang、Lin Qiu、Weinan Zhang、Yong Yu和Lei Li。用于非自回归神经机器翻译的Glancing Transformer。*arXiv*预印本 *arXiv:2008.07905*，2020年。[54] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark等。通过自然语言监督学习可迁移的视觉模型。载于国际机器学习会议，第8748–8763页。PMLR，2021年。[55] Pranav Rajpurkar、Jian Zhang、Konstantin Lopyrev和Percy Liang。SQuAD：100,000+ 个用于文本机器阅读理解的问题。*arXiv*预印本 *arXiv:1606.05250*，2016年。[56] Ruslan Rakhimov、Denis Volkhonskiy、Alexey Artemov、Denis Zorin和Evgeny Burnaev。潜在视频变换器。*arXiv*预印本 *arXiv:2006.10704*，2020年。[57] Aditya Ramesh、Mikhail Pavlov、Gabriel Goh、Scott Gray、Chelsea Voss、Alec Radford、Mark Chen和Ilya Sutskever。零样本文本到图像生成。*arXiv*预印本 *arXiv:2102.12092*，2021年。[58] Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu和Mark Chen。利用CLIP潜变量实现分层文本条件图像生成。*arXiv*预印本 *arXiv:2204.06125*，2022年。[59] Ali Razavi、Aaron van den Oord和Oriol Vinyals。使用VQ-VAE-2生成多样化的高保真图像。载于神经信息处理系统进展，第14866–14876页，2019年。[60] Joseph Redmon、Santosh Divvala、Ross Girshick和Ali Farhadi。You Only Look Once：统一的实时目标检测。载于*IEEE*计算机视觉与模式识别会议论文集，第779–788页，2016年。[61] Yi Ren、Yangjun Ruan、Xu Tan、Tao Qin、Sheng Zhao、Zhou Zhao和Tie-Yan Liu。FastSpeech：快速、稳健且可控的文本转语音。*arXiv*预印本 *arXiv:1905.09263*，2019年。[62] Danilo Rezende和Shakir Mohamed。使用归一化流进行变分推断。载于国际机器学习会议，第1530–1538页。PMLR，2015年。[63] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser和Björn Ommer。使用潜在扩散模型进行高分辨率图像合成。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第10684–10695页，2022年。[64] Chitwan Saharia、William Chan、Saurabh Saxena、Lala Li、Jay Whang、Emily Denton、Seyed Kamyar Seyed Ghasemipour、Burcu Karagol Ayan、S Sara Mahdavi、Rapha Gontijo Lopes等。具备深度语言理解能力的照片级真实感文本到图像扩散模型。*arXiv*预印本 *arXiv:2205.11487*，2022年。[65] Jonathan Shen、Ruoming Pang、Ron J Weiss、Mike Schuster、Navdeep Jaitly、Zongheng Yang、Zhifeng Chen、Yu Zhang、Yuxuan Wang、Rj Skerrv-Ryan等。通过以梅尔频谱图预测为条件的WaveNet实现自然的文本转语音合成。载于*2018*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），第4779–4783页。IEEE，2018年。[66] Jascha Sohl-Dickstein、Eric Weiss、Niru Maheswaranathan和Surya Ganguli。使用非平衡热力学进行深度无监督学习。载于国际机器学习会议，第2256–2265页。PMLR，2015年。

[67] Luchuan Song、Bin Liu、Guojun Yin、Xiaoyi Dong、Yufei Zhang 和 Jia-Xuan Bai。Tacr-net：深度视频和语音肖像编辑。载于*ACM*第*29*届国际多媒体会议论文集，第478–486页，2021年。[68] Hao Sun、Xu Tan、Jun-Wei Gan、Hongzhi Liu、Sheng Zhao、Tao Qin 和 Tie-Yan Liu。用于字素到音素转换的词元级集成蒸馏。*arXiv*预印本 *arXiv:1904.03446*，2019年。[69] Richard Szeliski。计算机视觉：算法与应用。Springer Science & Business Media，2010年。[70] Xu Tan、Tao Qin、Frank Soong 和 Tie-Yan Liu。神经语音合成综述。*arXiv*预印本 *arXiv:2106.15561*，2021年。[71] Xu Tan、Jiawei Chen、Haohe Liu、Jian Cong、Chen Zhang、Yanqing Liu、Xi Wang、Yichong Leng、Yuanhao Yi、Lei He 等。NaturalSpeech：实现人类水平质量的端到端文本转语音合成。*arXiv*预印本 *arXiv:2205.04421*，2022年。[72] Anni Tang、Tianyu He、Xu Tan、Jun Ling、Runnan Li、Sheng Zhao、Li Song 和 Jiang Bian。记忆是缓解说话人脸生成中一对多映射问题的手段。*arXiv*预印本 *arXiv:2212.05005*，2022年。[73] Justus Thies、Michael Zollhofer、Marc Stamminger、Christian Theobalt 和 Matthias Nießner。Face2face：RGB视频的实时人脸捕捉与再现。载于*IEEE*计算机视觉与模式识别会议论文集，第2387–2395页，2016年。[74] Justus Thies、Mohamed Elgharib、Ayush Tewari、Christian Theobalt 和 Matthias Nießner。神经语音木偶：音频驱动的人脸再现。载于欧洲计算机视觉会议，第716–731页。Springer，2020年。[75] Arash Vahdat、Karsten Kreis 和 Jan Kautz。潜在空间中的基于得分的生成建模。神经信息处理系统进展，34：11287–11302，2021年。[76] Aaron van den Oord、Oriol Vinyals 和 Koray Kavukcuoglu。神经离散表示学习。载于第*31*届神经信息处理系统国际会议论文集，第6309–6318页，2017年。[77] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力就是一切。载于神经信息处理系统进展，第5998–6008页，2017年。[78] Xiong Wang、Zhuoyuan Yao、Xian Shi 和 Lei Xie。级联 RNN-Transducer：借助音节到汉字转换器实现基于音节的流式端侧普通话语音识别。载于*2021*年*IEEE*口语技术研讨会（*SLT*），第15–21页。IEEE，2021年。[79] Yuxuan Wang、RJ Skerry-Ryan、Daisy Stanton、Yonghui Wu、Ron J Weiss、Navdeep Jaitly、Zongheng Yang、Ying Xiao、Zhifeng Chen、Samy Bengio 等。Tacotron：迈向端到端语音合成。*arXiv*预印本 *arXiv:1703.10135*，2017年。[80] Sam Wiseman、Stuart M Shieber 和 Alexander M Rush。学习用于文本生成的神经模板。载于*2018*年自然语言处理经验方法会议论文集，第3174–3187页，2018年。[81] Jian Wu、Xiaoguang Liu、Xiaolin Hu 和 Jun Zhu。PopMNet：利用神经网络生成结构化流行音乐旋律。人工智能，286：103303，2020年。[82] Wilson Yan、Yunzhi Zhang、Pieter Abbeel 和 Aravind Srinivas。VideoGPT：使用向量量化变分自编码器和Transformer生成视频。*arXiv*预印本 *arXiv:2104.10157*，2021年。[83] Dongchao Yang、Jianwei Yu、Helin Wang、Wen Wang、Chao Weng、Yuexian Zou 和 Dong Yu。DiffSound：用于文本到声音生成的离散扩散模型。*arXiv*预印本 *arXiv:2207.09983*，2022年。

[84] Jian Yang、Shuming Ma、Dongdong Zhang、Zhoujun Li 和 Ming Zhou。通过软模板预测改进神经机器翻译。载于计算语言学协会第*58*届年会论文集，第5979–5989页，2020年。[85] Ran Yi、Zipeng Ye、Juyong Zhang、Hujun Bao 和 Yong-Jin Liu。通过学习实现个性化头部姿态的音频驱动说话人脸视频生成。*arXiv*预印本 *arXiv:2002.10137*，2020年。[86] Jiahui Yu、Yuanzhong Xu、Jing Yu Koh、Thang Luong、Gunjan Baid、Zirui Wang、Vijay Vasudevan、Alexander Ku、Yinfei Yang、Burcu Karagol Ayan 等。扩展自回归模型，实现内容丰富的文本到图像生成。*arXiv*预印本 *arXiv:2206.10789*，2022年。[87] Jiahong Yuan、Xingyu Cai、Dongji Gao、Renjie Zheng、Liang Huang 和 Kenneth Church。解耦普通话自动语音识别中的识别与转录。载于*2021*年*IEEE*自动语音识别与理解研讨会（*ASRU*），第1019–1025页。IEEE，2021年。[88] Neil Zeghidour、Alejandro Luebs、Ahmed Omran、Jan Skoglund 和 Marco Tagliasacchi。Soundstream：端到端神经音频编解码器。*IEEE/ACM*音频、语音与语言处理汇刊，30：495–507，2021年。[89] Shuai Zhang、Jiangyan Yi、Zhengkun Tian、Ye Bai、Jianhua Tao 等。解耦端到端代码转换自动语音识别中的发音与语言。载于*ICASSP 2021——2021*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），第6249–6253页。IEEE，2021年。[90] Ye Zhang 和 Byron Wallace。句子分类卷积神经网络的敏感性分析（及实践者指南）。*arXiv*预印本 *arXiv:1510.03820*，2015年。[91] Jun-Yan Zhu、Taesung Park、Phillip Isola 和 Alexei A Efros。使用循环一致性对抗网络进行无配对图像到图像转换。载于*IEEE*国际计算机视觉会议论文集，第2223–2232页，2017年。[92] Yi Zou、Pei Zou、Yi Zhao、Kaixiang Zhang、Ran Zhang 和 Xiaorui Wang。Melons：利用Transformer和结构图生成具有长期结构的旋律。*arXiv*预印本 *arXiv:2110.05020*，2021年。