



数字媒体技术基础

Meng Yang

www.smartllv.com

SUN YAT-SEN University

机器智能与先进计算教育部重点实验室





Course Outline

- 14 智能新媒体合成编辑技术
 - 14. 1 图像新媒体合成编辑
 - 14. 2 视频新媒体合成编辑
 - 14. 3 文本生成技术
 - 14. 4 语音生成技术



智能新媒体合成编辑技术

14.1 图像新媒体合成编辑



图像新媒体合成编辑

- 图像编辑
 - 图像编辑是在保持原图像语义的前提下进行对图像的变化。
- 图像合成
 - 图像合成一般是会生成另一幅与输入图像有较大差别的图像。
- 图像新媒体合成编辑通常采用GAN作为模型架构，相较于传统的图像合成编辑技术，会使用到高层的语义信息（卷积神经网络提取到的特征）。

图像新媒体智能抠图 猜猜用到了什么技术？



问题？

电影特效制作



知乎 @王荣

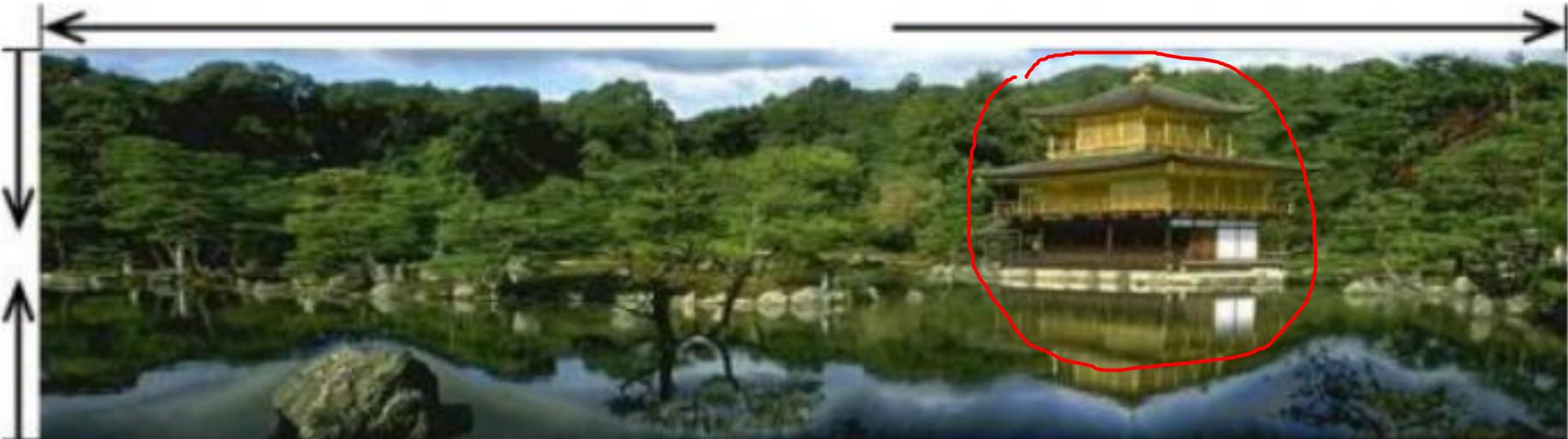
电影特效制作



图像新媒体合成编辑



问题?

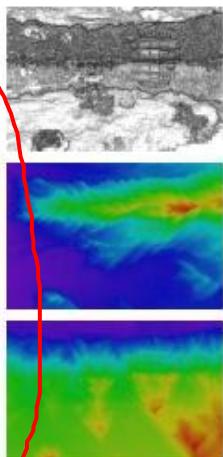


图像新媒体合成编辑

□ 图像编辑

图像变换

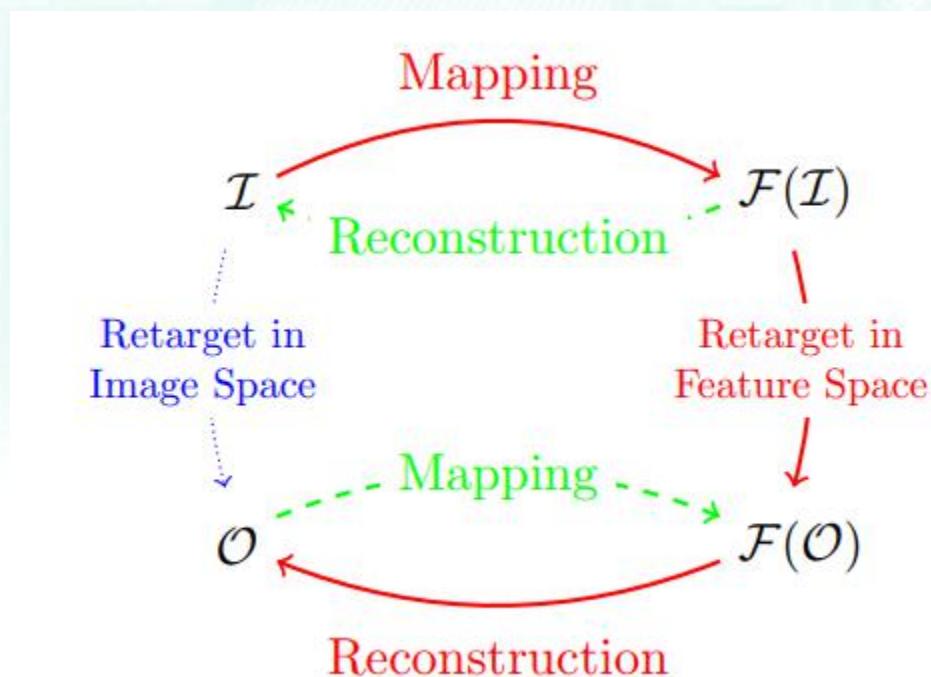
可以让图像在缩小或放大时，让图像变化得更加自然。



图像新媒体合成编辑

□ 基于深度学习的图像变换

深层的神经网络包含图像丰富的重要分割信息，直接调整从预训练分类网络中提取的图像特征图，并使用基于神经网络的优化来重建调整大小的图像。这种方法利用了网络的分层编码，特别是其更深的高级判别能力，可以识别语义对象和区域并允许保持它们的纵横比。



图像新媒体合成编辑

□ 图像着色

目标是在给定灰度输入图像的情况下生成彩色图像。

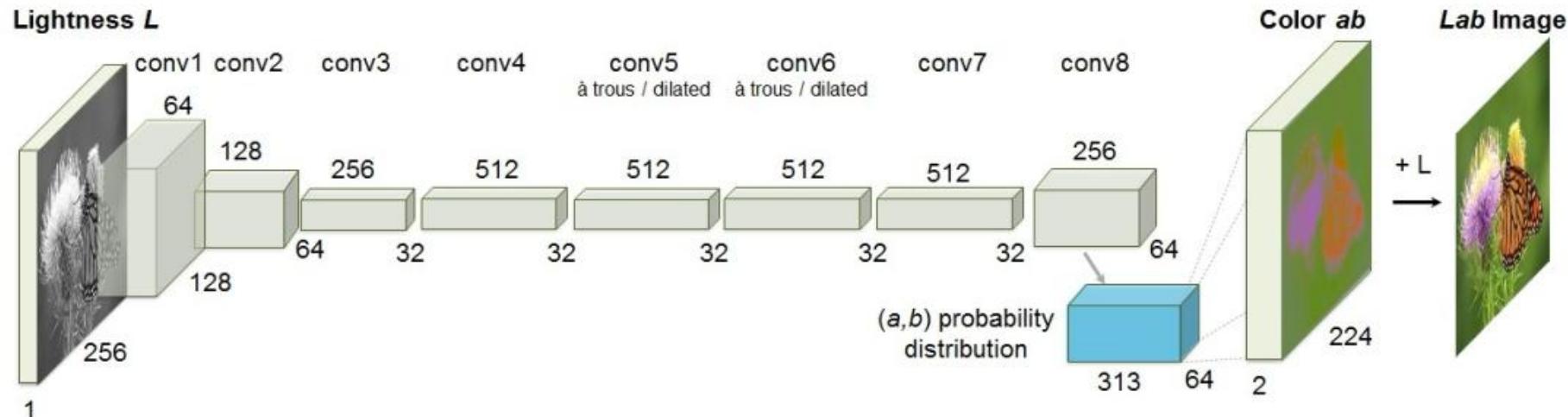


图像新媒体合成编辑



□ 图像着色

训练 CNN 从灰度输入映射到量化颜色值输出的分布。

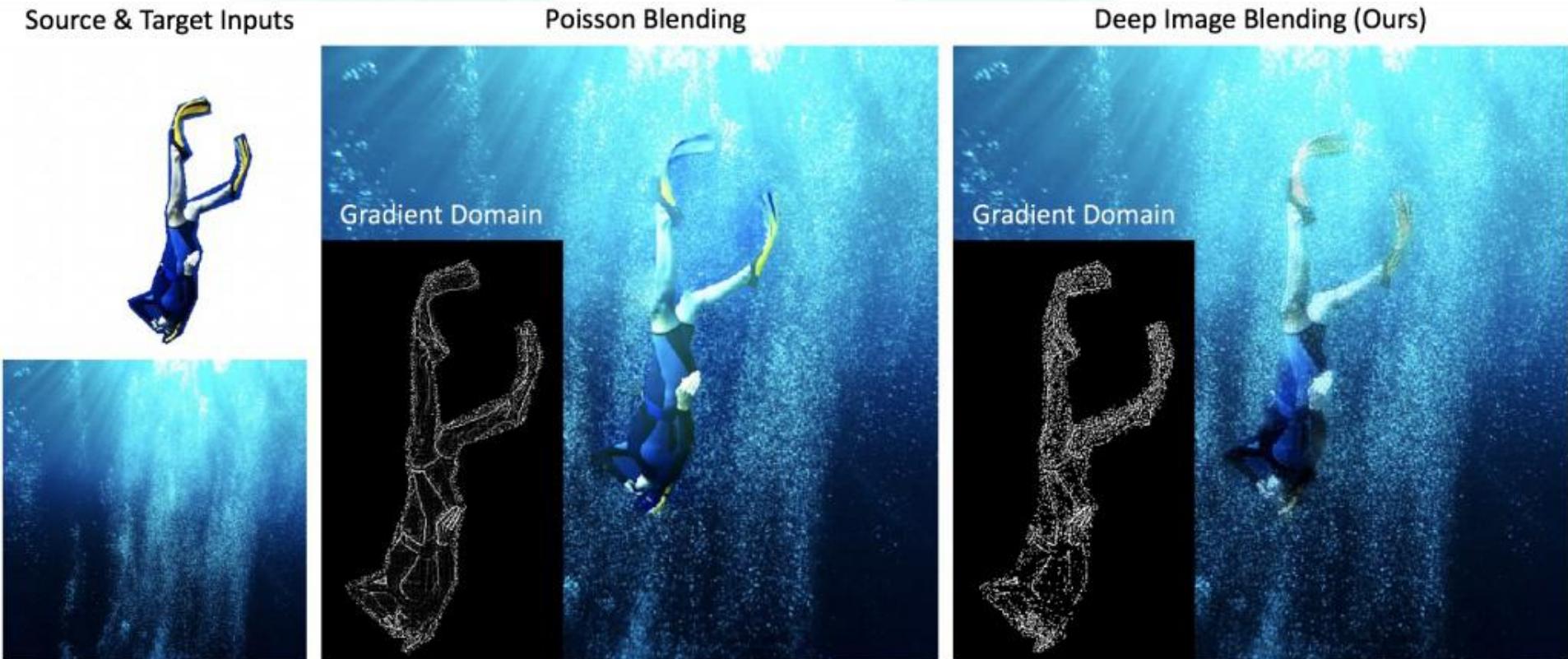


图像新媒体合成编辑



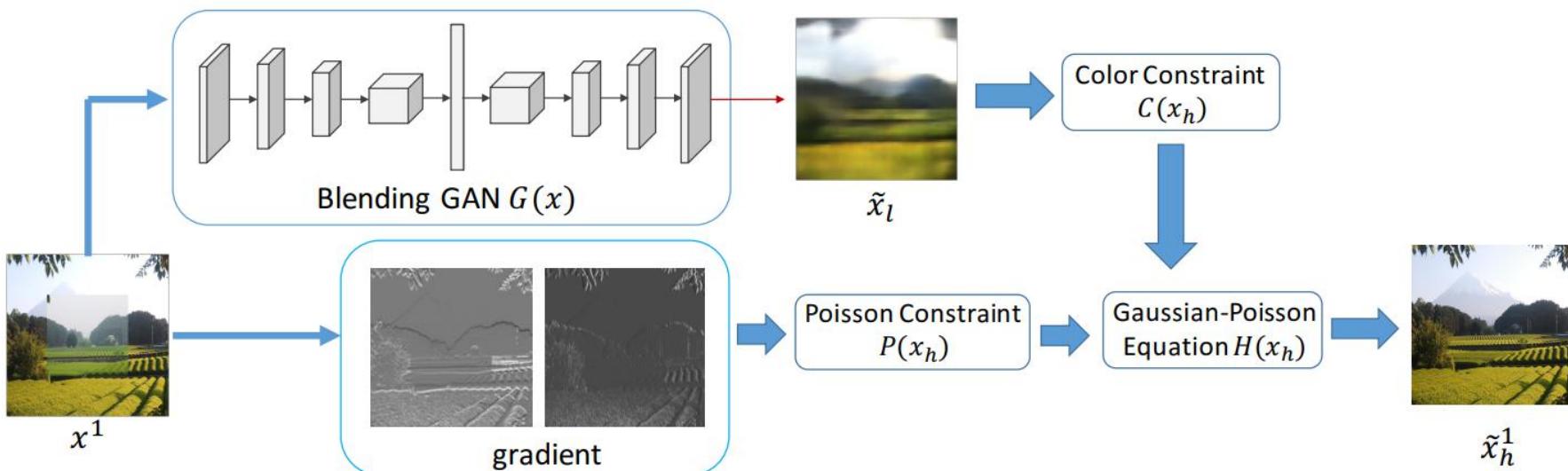
□ 图像拼接

目标是将两幅图像自然地拼接在一起。



□ 图像拼接

Gaussian–Poisson GAN (GP–GAN)，该框架结合了基于经典梯度的方法和 GAN 的优势。特别是，提出高斯泊松方程来制定高分辨率图像混合问题，这是一种受梯度和颜色信息约束的联合优化。



图像新媒体合成编辑

□ 语义分割

语义分割是在像素级别上的分类，属于同一类的像素都要被归为一类，因此语义分割是从像素级别来理解图像的。比如说如下的照片，属于人的像素都要分成一类，属于摩托车的像素也要分成一类，除此之外还有背景像素也被分为一类。



知乎 @stone



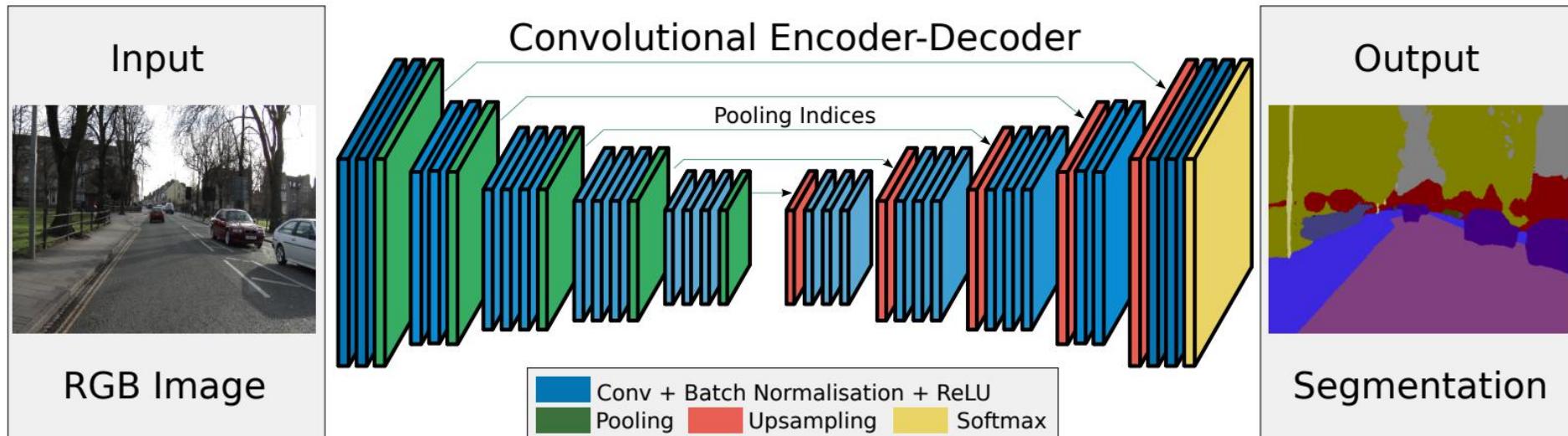
知乎 @stone

图像新媒体合成编辑



□ 语义分割

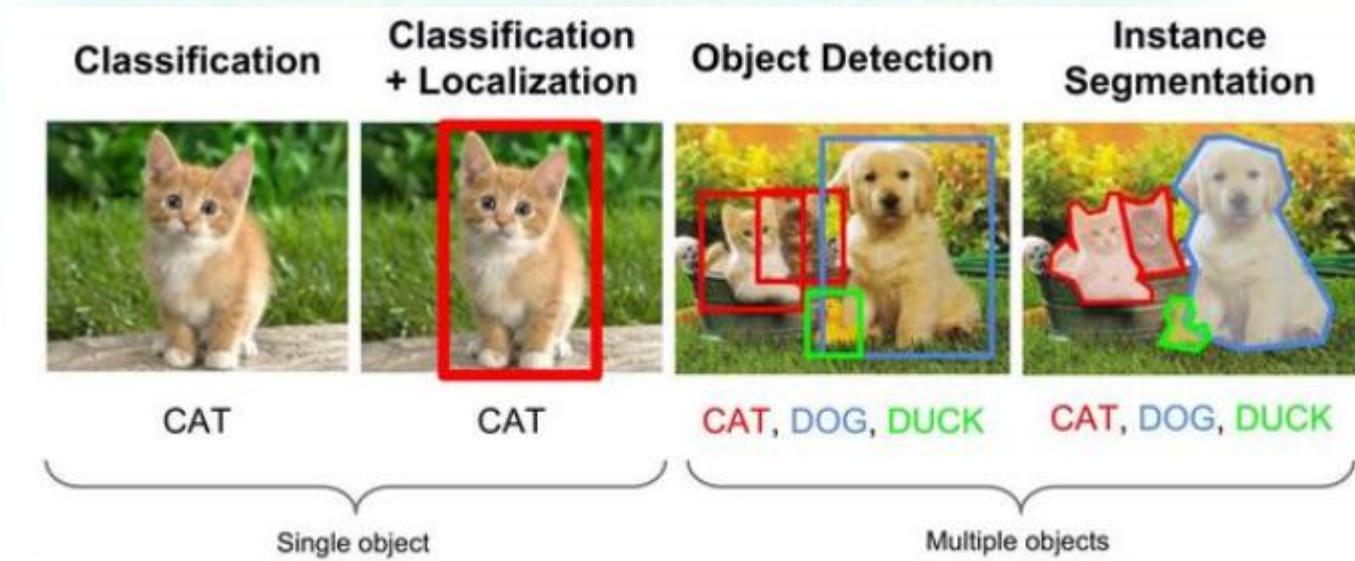
SegNet将一张RGB图片作为输入经过卷积编码器和解码器得到该图片的语义分割图片。



图像新媒体合成编辑

□ 目标检测

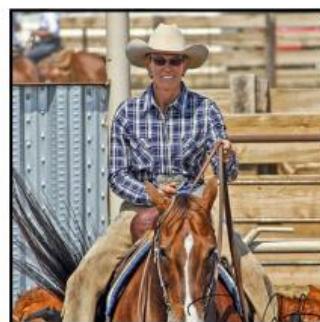
目标检测关注图像中特定的物体目标，要求同时获得这一目标的类别信息和位置信息。



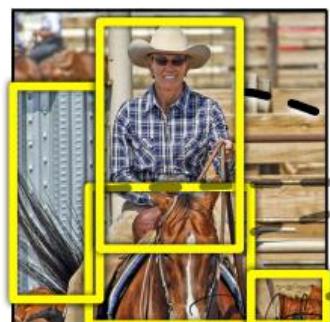
□ 目标检测

R-CNN将检测抽象为两个过程，一是基于图片提出若干可能包含物体的区域（即图片的局部裁剪，被称为Region Proposal；二是在提出的这些区域上运行当时表现最好的分类网络，得到每个区域内物体的类别

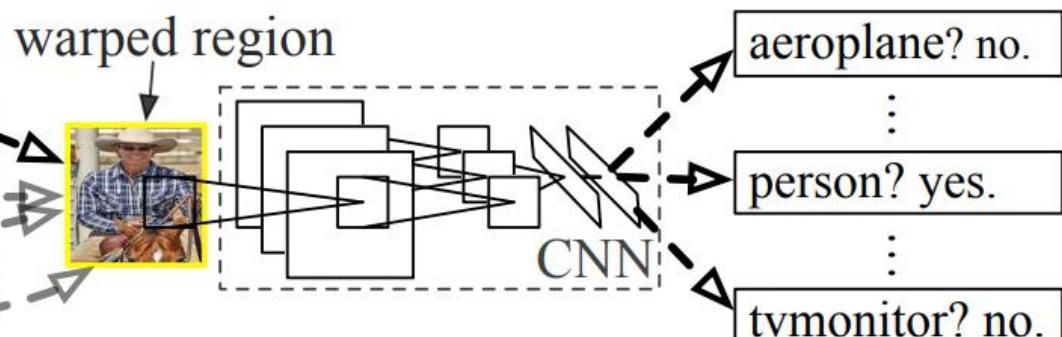
R-CNN: *Regions with CNN features*



1. Input image



2. Extract region proposals (~2k)



3. Compute CNN features

4. Classify regions

图像新媒体合成编辑



问题?

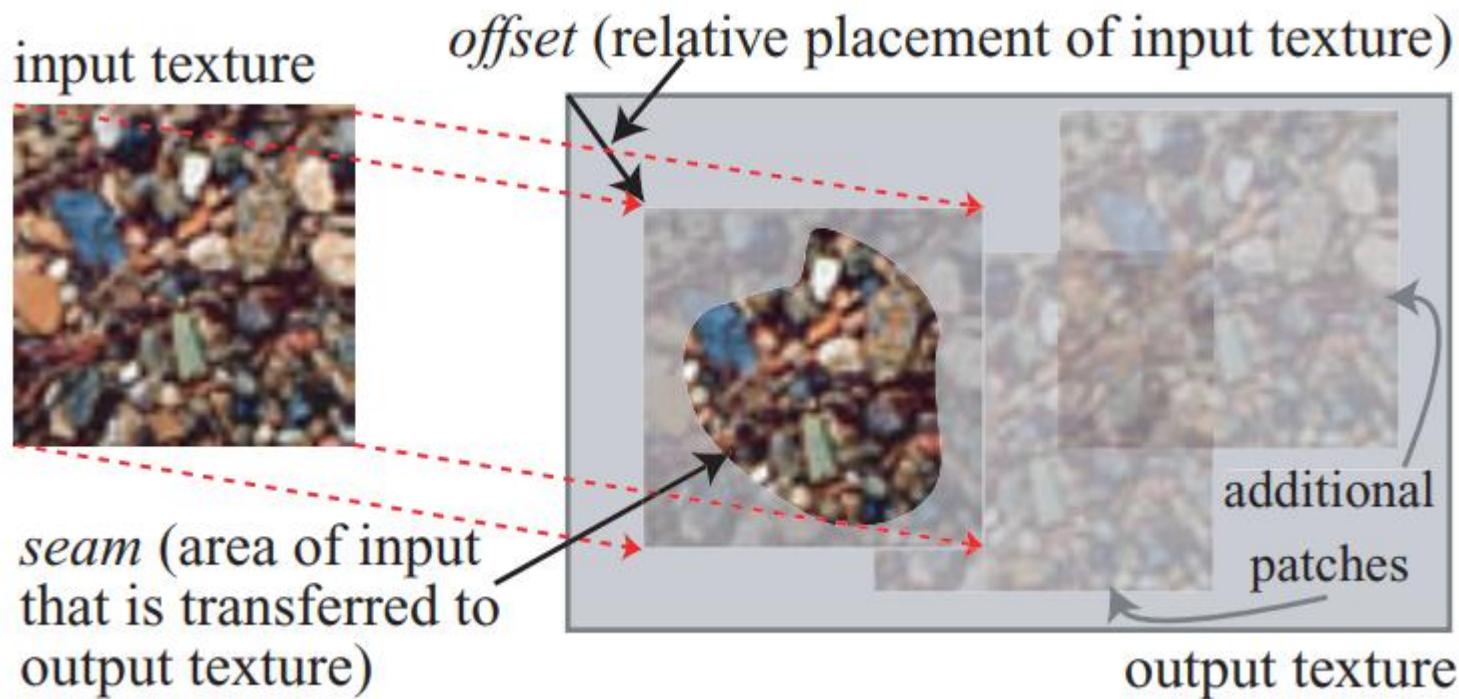
□ 图像合成 纹理合成

让输入图像自我复制和变换，然后沿着最佳接缝拼接在一起以生成新的（通常更大）输出图像。



□ 纹理合成

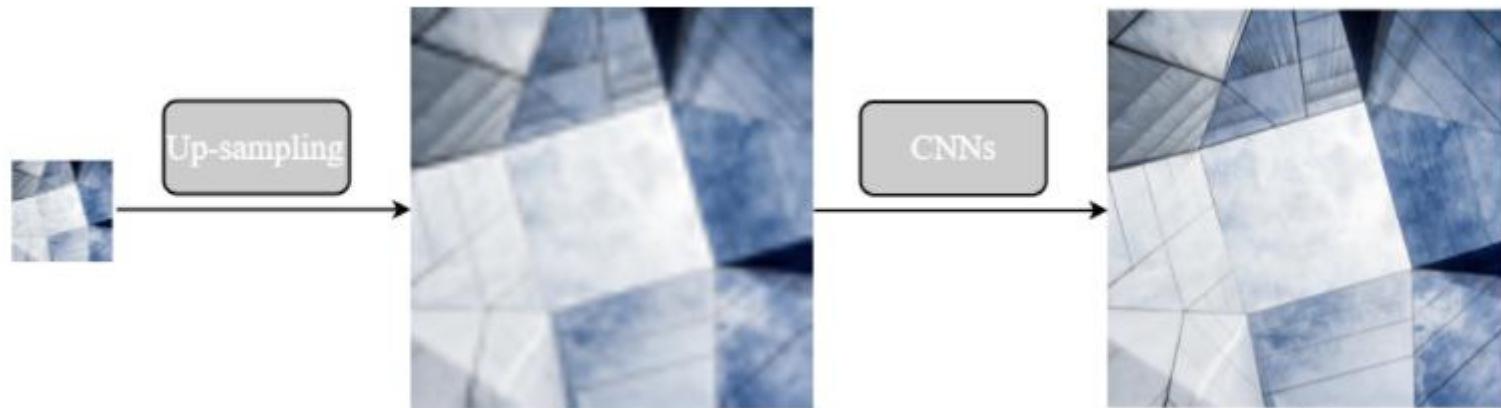
Graphcut Textures通过将不规则形状的区域从样本图像复制到输出图像来合成新纹理。 区域复制过程分两个阶段执行。 首先，通过在候选块和输出图像中已有的像素之间进行比较来选择候选矩形块，其次，计算该矩形的最佳（不规则形状）部分，并且仅将这些像素复制到输出图像。 要复制的区域部分是通过使用图切割算法确定的。



图像新媒体合成编辑

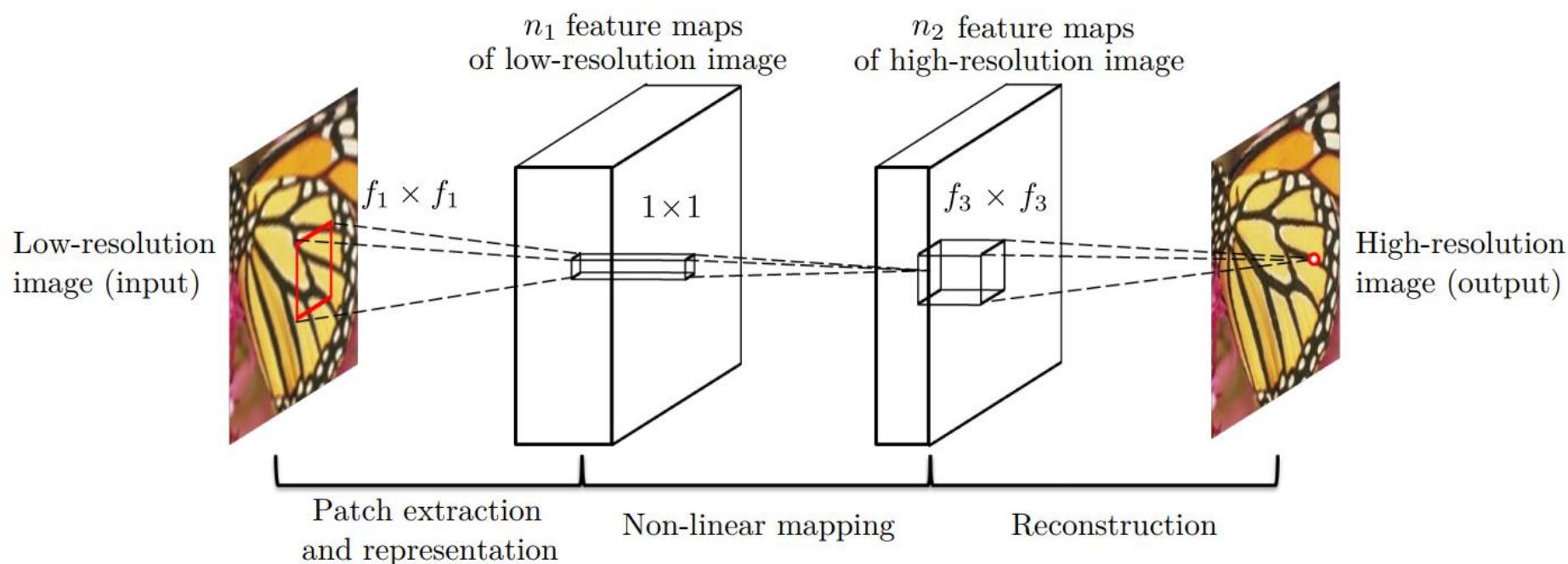
□ 超分辨率

图像超分辨率是指由一幅低分辨率图像或图像序列恢复出高分辨率图像。



□ 超分辨率

SRCNN: 首先使用双三次 (bicubic) 插值将低分辨率图像放大成目标尺寸，接着通过三层卷积网络拟合非线性映射，最后输出高分辨率图像结果。



图像新媒体合成编辑

□ 人脸合成

给定一张人脸图像，输出一张不同状态的人脸图像（表情、角度、妆容等等）。





图像新媒体合成编辑

□ 有约束的图像合成

根据用户的某些指定约束（例如另一个图像、文本描述或边缘图）合成新图像。如文本生成图像和图像生成图像。

文本生成图像

问题？

- 文本生成图像主要任务为从一句描述性文本生成一张与文本内容相对应的图片。

‘This bird is completely red with black wings’

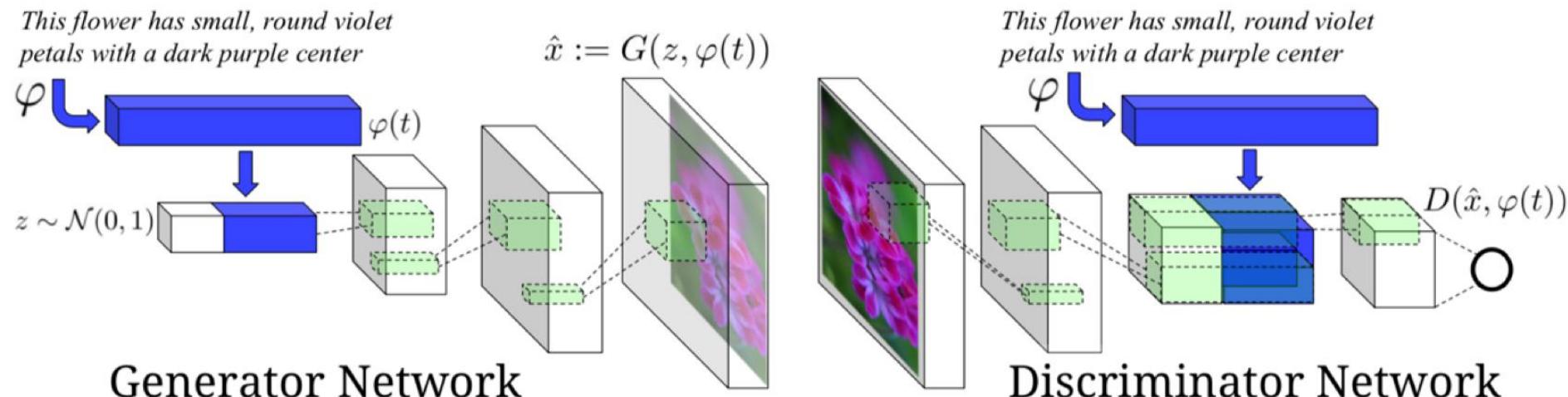


‘this bird is all blue, the top part of the bill is blue, but the bottom half is white’



文本生成图像

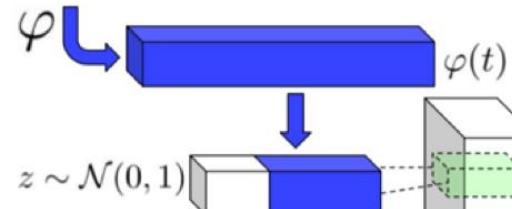
- 近几年，文本生成图像大部分的方法都使用了GAN的思想完成这个任务。
- 如最早使用GAN为模型主体的**GAN-INT-CLS**网络。



https://blog.csdn.net/mohole_zhang

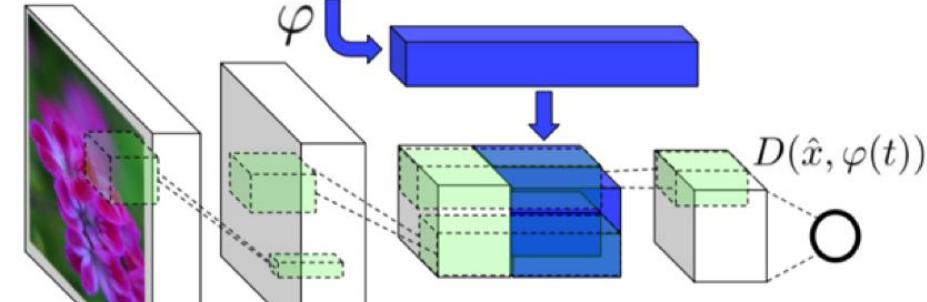
文本生成图像

This flower has small, round violet petals with a dark purple center



$$\hat{x} := G(z, \varphi(t))$$

This flower has small, round violet petals with a dark purple center



Generator Network

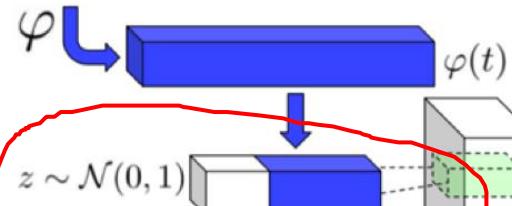
Discriminator Network

https://blog.csdn.net/mohole_zhang

在GAN中生成器Generator根据文本特征生成图片，继而被鉴别器Discriminator鉴定其生成效果，根据鉴别器的鉴定结果生成器再次生成更真实的图片，鉴别器则再次对新图鉴定，以此类推，迭代进行直到网络收敛。

文本生成图像

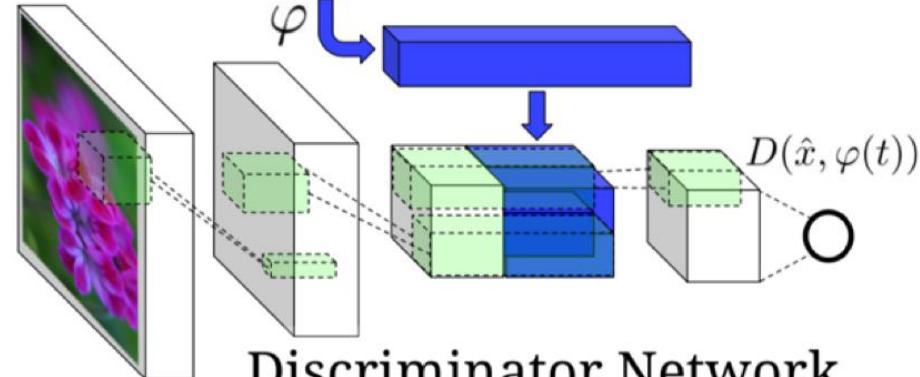
This flower has small, round violet petals with a dark purple center



Generator Network

$$\hat{x} := G(z, \varphi(t))$$

This flower has small, round violet petals with a dark purple center



Discriminator Network

https://blog.csdn.net/mohole_zhang

输入文本会被提取为文本特征；
 在生成器中，文本特征与随机噪声融合后一起输入到生成网络中，生成64x64的图像；
 在鉴别器中，生成图像在下采样之后，跟之前的文本特征连接起来，最后鉴别器根据融合特征进行判定。



文本生成图像

在GAN-CLS中加入了
Matching-aware
discriminator，即在鉴
别器中对错误情况进行
分类，一种是生成的fake图像匹配了正
确的文本，另一种是真
实图像但匹配了错
误文本，利用这种机
制使得鉴别器网络不
仅能够识别图像是否
是生成器生成的。

Algorithm 1 GAN-CLS training algorithm with step size α , using minibatch SGD for simplicity.

- 1: **Input:** minibatch images x , matching text t , mis-matching \hat{t} , number of training batch steps S
 - 2: **for** $n = 1$ **to** S **do**
 - 3: $h \leftarrow \varphi(t)$ {Encode matching text description}
 - 4: $\hat{h} \leftarrow \varphi(\hat{t})$ {Encode mis-matching text description}
 - 5: $z \sim \mathcal{N}(0, 1)^Z$ {Draw sample of random noise}
 - 6: $\hat{x} \leftarrow G(z, h)$ {Forward through generator}
 - 7: $s_r \leftarrow D(x, h)$ {real image, right text}
 - 8: $s_w \leftarrow D(x, \hat{h})$ {real image, wrong text}
 - 9: $s_f \leftarrow D(\hat{x}, h)$ {fake image, right text}
 - 10: $\mathcal{L}_D \leftarrow \log(s_r) + (\log(1 - s_w) + \log(1 - s_f))/2$
 - 11: $D \leftarrow D - \alpha \partial \mathcal{L}_D / \partial D$ {Update discriminator}
 - 12: $\mathcal{L}_G \leftarrow \log(s_f)$
 - 13: $G \leftarrow G - \alpha \partial \mathcal{L}_G / \partial G$ {Update generator}
 - 14: **end for**
-



文本生成图像

在Learning with manifold interpolation (GAN-INT) 中，对于根据描述去生成图片的问题，文本描述数量相对较少是限制合成效果（多样性）的一个重要因素。所以，论文提出通过简单的插值方法来生成大量的新的文本描述。这些插值得到的文本特征是无法直接对应到人工文本标注上的，所以这一部分数据是不需要标注的。想要利用这些数据，只需要在生成器的目标函数上增加这样一项：

$$\mathbb{E}_{t_1, t_2 \sim p_{data}} [\log(1 - D(G(z, \beta t_1 + (1 - \beta)t_2)))]$$

这个公式相当于综合考虑了两个文本特征 t_1 和 t_2 的插值点。通常在实际应用中使用 $\beta=0.5$ 效果就不错了。

文本生成图像

生成效果

GT
this flower is white and pink in color, with petals that have veins.



these flowers have petals that start off white in color and end in a dark purple towards the tips.



bright droopy yellow petals with burgundy streaks, and a yellow stigma.



GAN



GAN - CLS



GAN - INT



GAN - INT - CLS





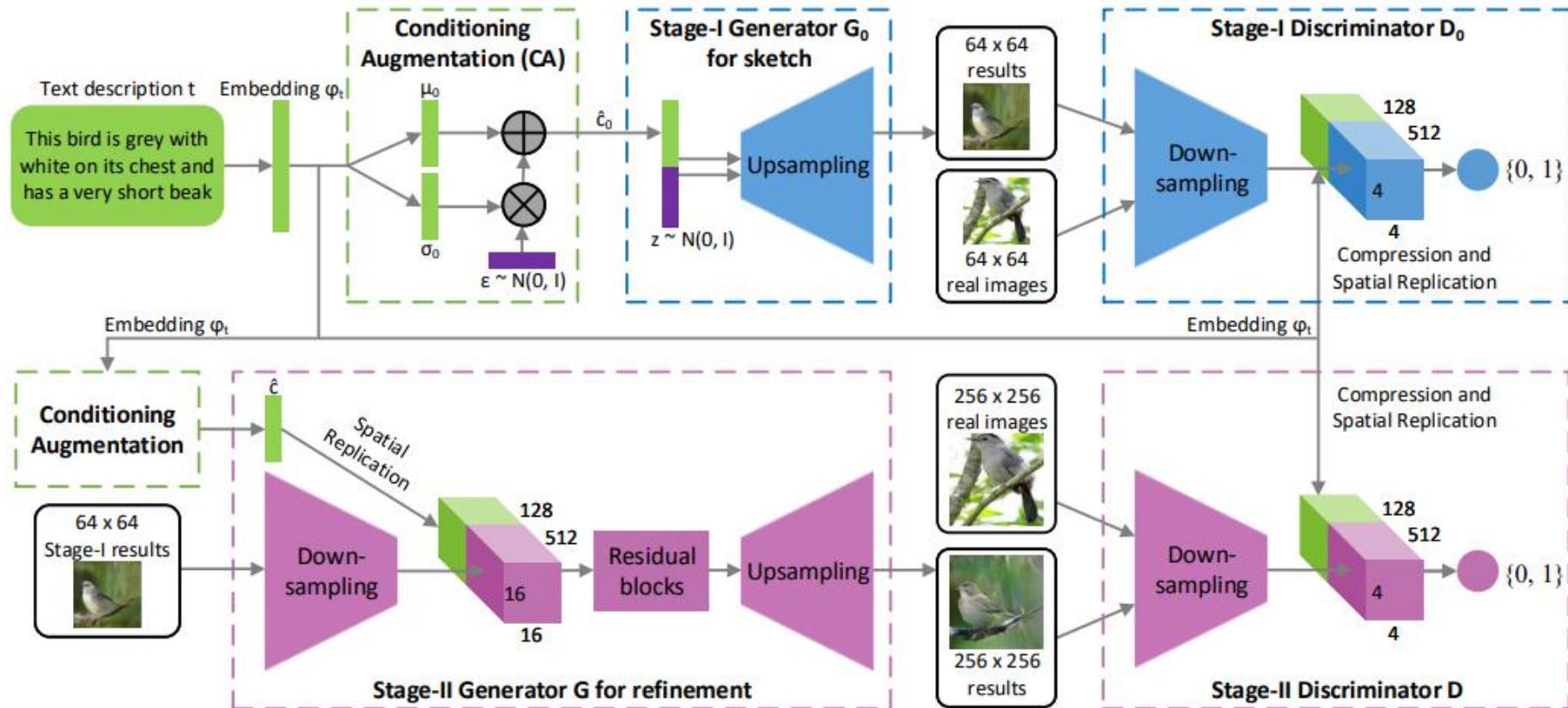
文本生成图像

□ StackGAN

- StackGAN 借鉴别人提出两个GAN叠加在一起的结构，改进了输入条件的部分，以前只能生成 $64*64$ 的图片，而现在可以生成 $256*256$ 的图片。
- 其中两个阶段GAN的作用是：第一阶段的对抗生成网络利用文本描述粗略勾画物体主要的形状和颜色，生成低分辨率的图片。第二阶段的对抗生成网络将第一阶段的结果和文本描述作为输入，生成细节丰富的高分辨率图片。

文本生成图像

□ StackGAN

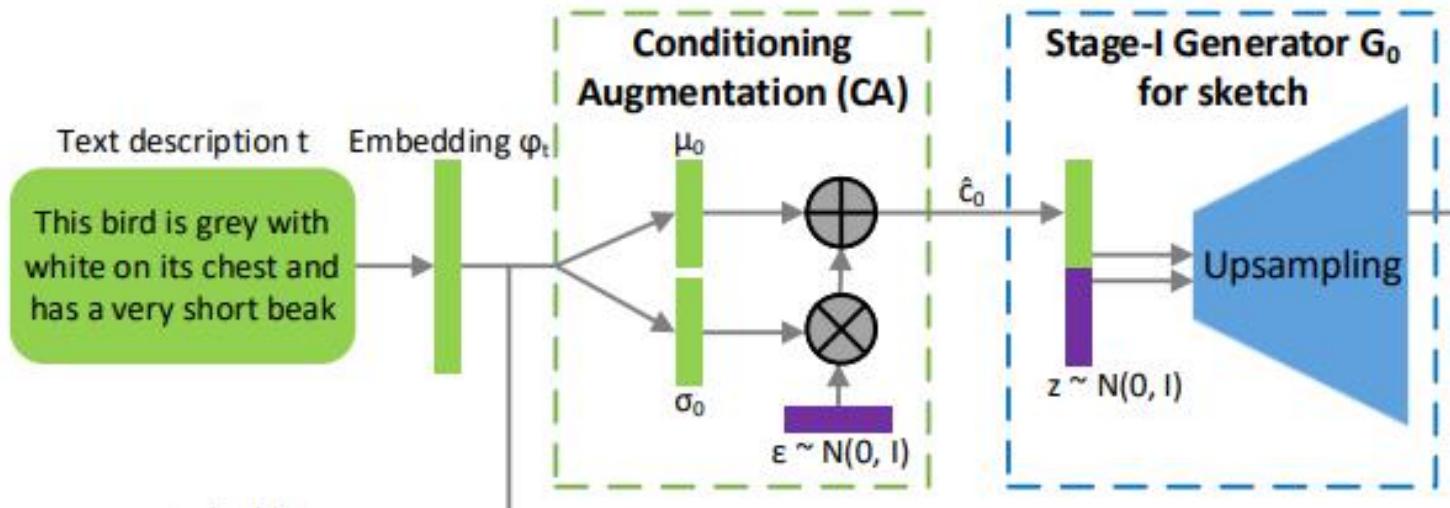


根据噪声 z 和文本描述 φ_t 生成相对粗糙的低分辨率的图像，但是这个图像可能只描绘出物体的形状和颜色以及大致画出背景。

文本生成图像

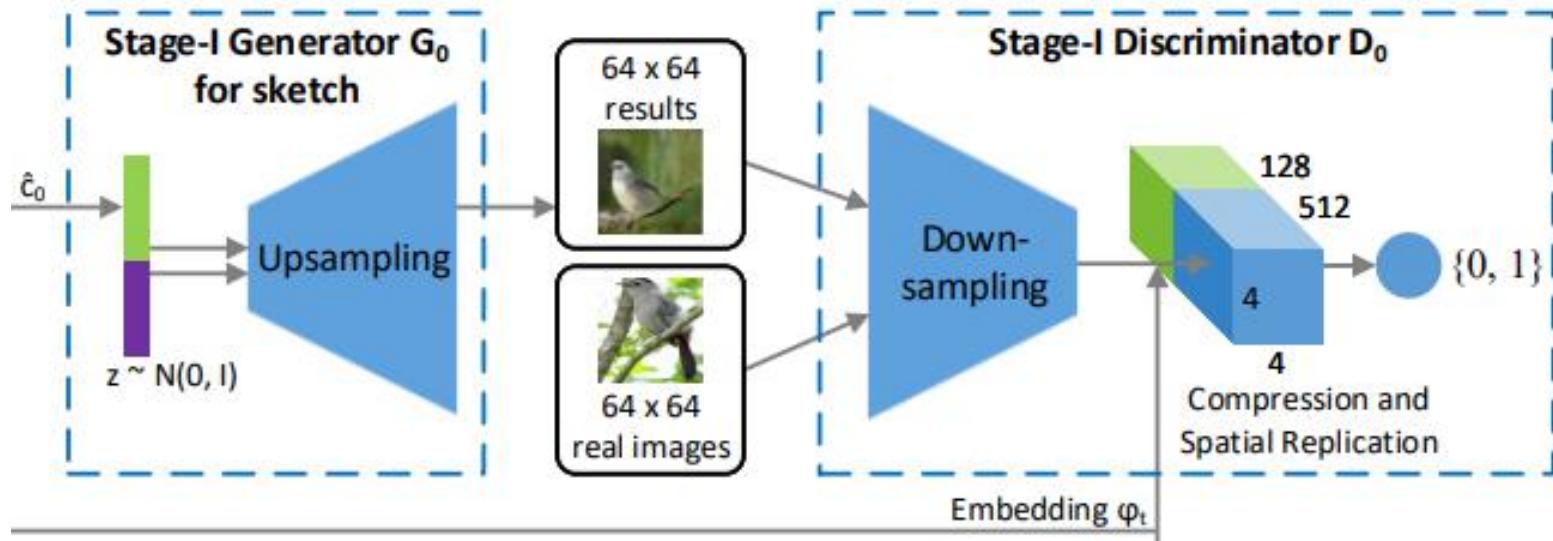


阶段一



- 文本特征的处理：StackGAN 没有直接将文本特征作为 **condition**，而是用文本特征连接了一个 FC 层，学习得到一个独立的高斯分布并随机采样得到的隐含变量作为**condition**。之所以这样做的原因是，文本特征通常比较高维，而相对这个维度来说，文本的数量其实很少，如果将文本特征直接作为 **condition**，那么这个隐变量在隐空间里就比较稀疏，这对训练不利。

文本生成图像

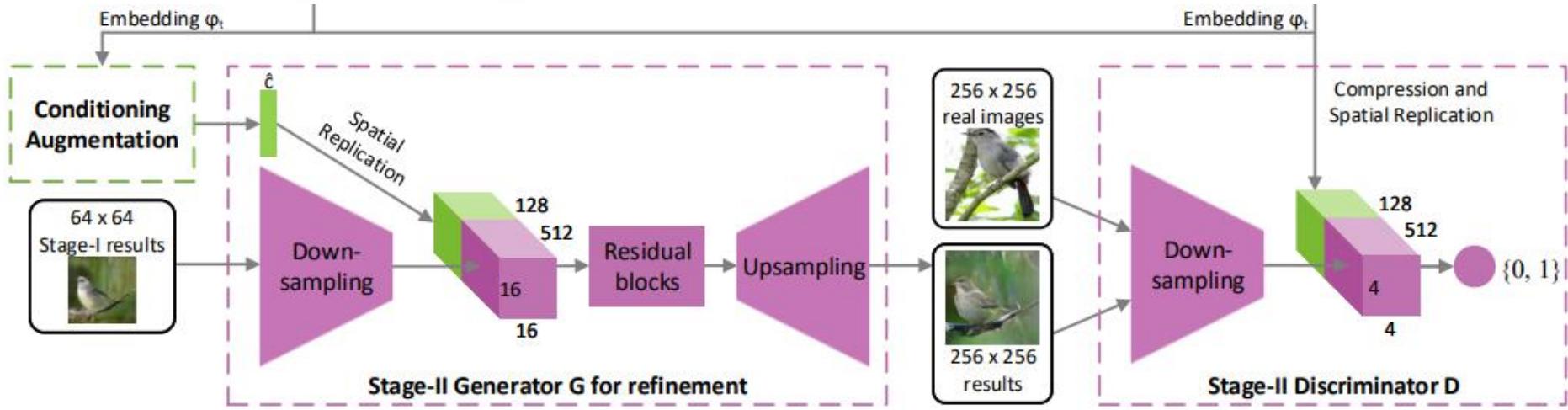


- 判别器：首先文本特征经过一个全连接层被压缩到128维，然后经过复制将其扩成一个 $4*4*128$ 的张量。同时，图像会经过一系列的下采样到 $4*4$ 。然后，将它们的通道连接起来，最后会通过一个二分类器去预测图像真假的概率。

文本生成图像



阶段二



- 前一GAN生成的图形可能会存在物体形状的失真扭曲或者忽略了文本描述中的细节部分，所以再利用一个GAN去根据文本信息修正之前得到的图像，生成更高分辨率含有更多细节信息的图像。

文本生成图像



□ 生成效果

Text description	This flower has a lot of small purple petals in a dome-like configuration	This flower is pink, white, and yellow in color, and has petals that are striped	This flower has petals that are dark pink with white edges and pink stamen	This flower is white and yellow in color, with petals that are wavy and smooth
64x64 GAN-INT-CLS				
256x256 StackGAN				

文本生成图像

□ 阶段一和阶段二生成结果对比

Text description	This bird is blue with white and has a very short beak	This bird has wings that are brown and has a yellow belly	A white bird with a black crown and yellow beak	This bird is white, black, and brown in color, with a brown beak	The bird has small beak, with reddish brown crown and gray belly	This is a small, black bird with a white breast and white on the wingbars.	This bird is white black and yellow in color, with a short black beak
Stage-I images							
Stage-II images							

文本生成图像

□ 两个不同描述语句插值效果

The bird is completely red → The bird is completely yellow



This bird is completely red with black wings and pointy beak →
this small blue bird has a short pointy beak and brown on its wings



图像生成图像

- 图像生成图像其实就是基于一张输入图像得到想要的输出图像的过程，可以看做是图像和图像之间的一种映射。

问题？

Labels to Street Scene



input



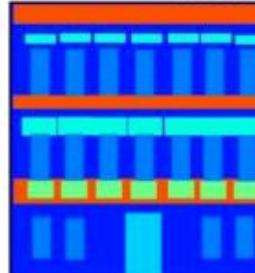
output

Aerial to Map



input

Labels to Facade



input



output

BW to Color



input



output



input



output

Day to Night



input



output

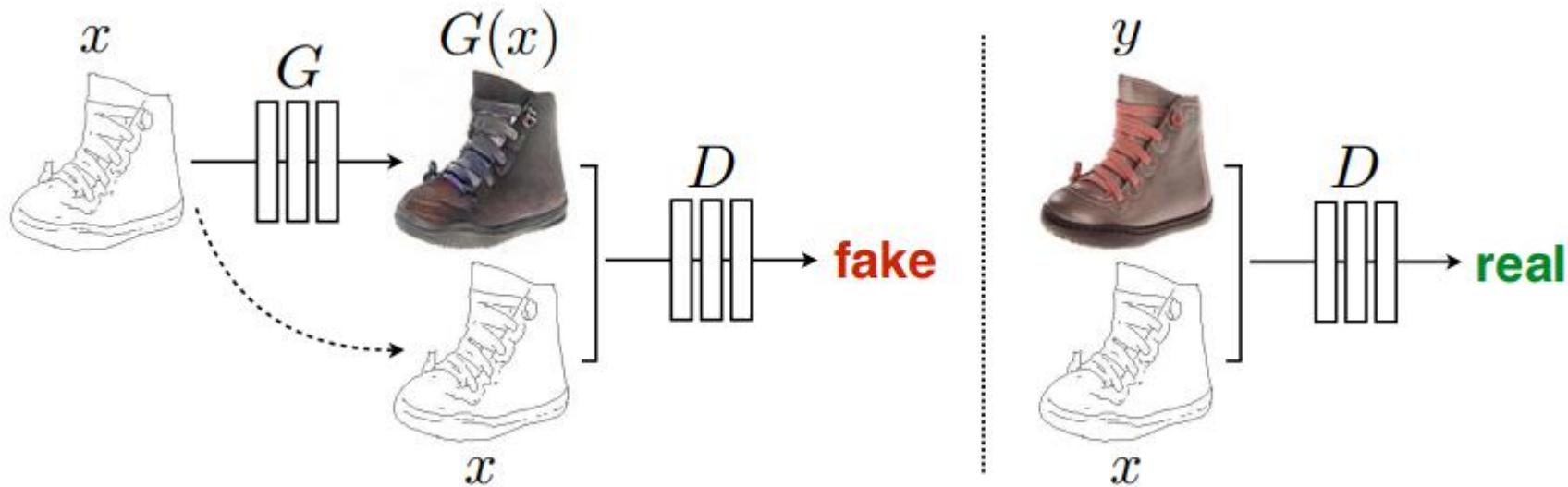
Edges to Photo

图像生成图像



- 图像生成图像经典的方法为Pix2Pix。
- Pix2Pix基于GAN实现图像生成图像，更准确地讲是基于条件GAN，因为条件GAN可以通过添加条件信息来指导图像生成，因此可以将输入图像作为条件，学习从输入图像到输出图像之间的映射，从而得到指定的输出图像。

图像生成图像



- 普通的GAN接收的G部分的输入是随机向量，输出是图像；D部分接收的输入是图像(生成的或是真实的)，输出是对或者错。这样G和D联手就能输出真实的图像。
- Pix2Pix对传统的GAN做了个小改动，它不再输入随机噪声，而是输入用户给的图片。它的G输入显然应该是一张图，输出也是一张图。D的输入是一对图，因为除了要生成真实图像之外，还要保证生成的图像和输入图像是匹配的。



图像生成图像

□ 损失函数

条件GAN损失: $\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z)))]$,

L1损失:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x, z)\|_1].$$

目标函数:

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

x为输入图像，z为随机噪声，y为真实图片。

图像生成图像

□ 应用：标签图像生成图像



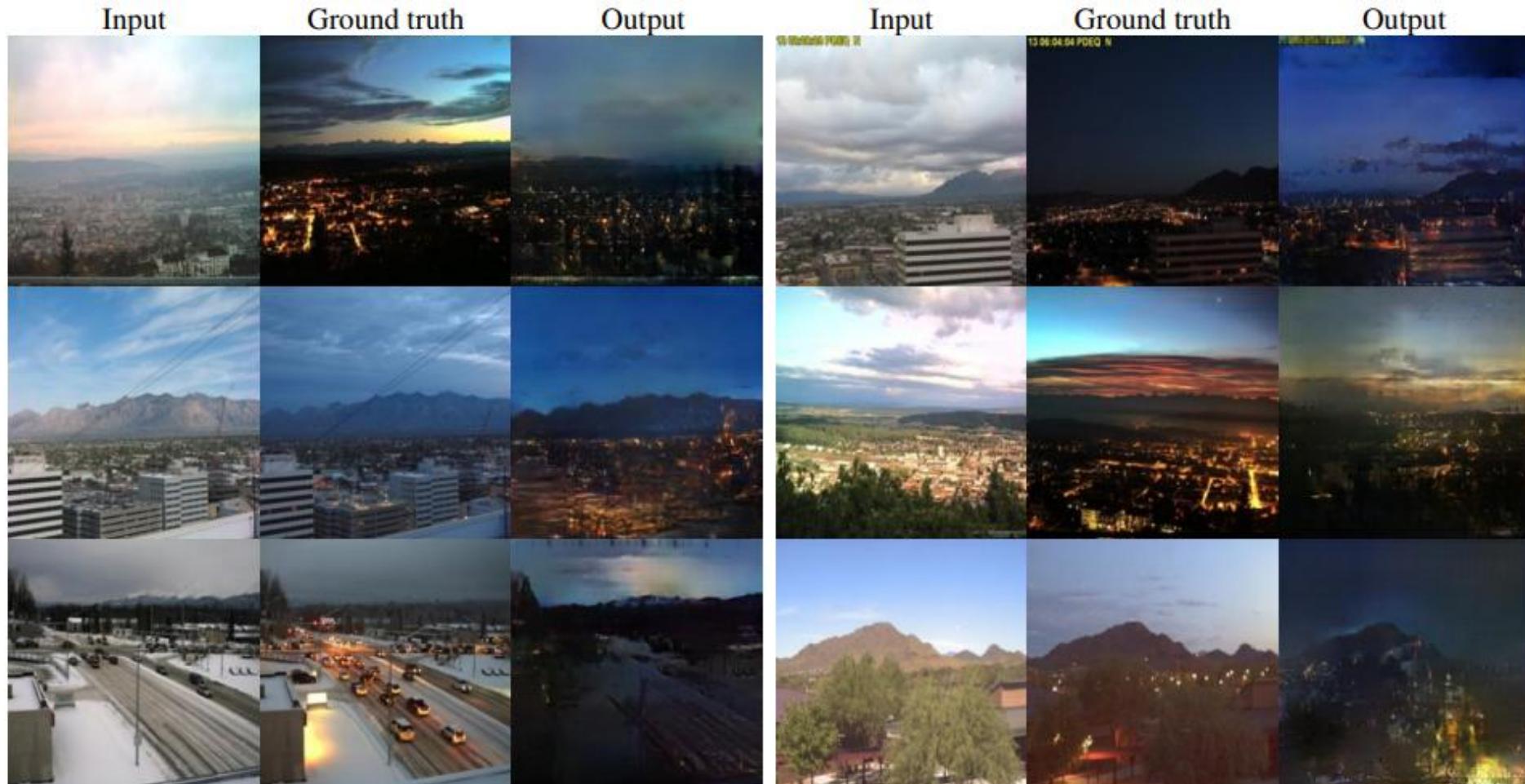
图像生成图像

□ 应用：边缘图像生成图像

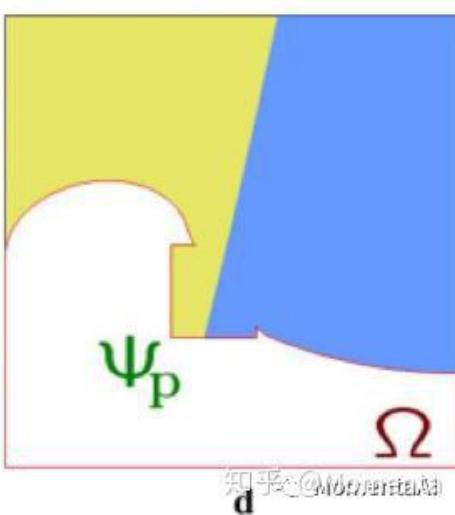
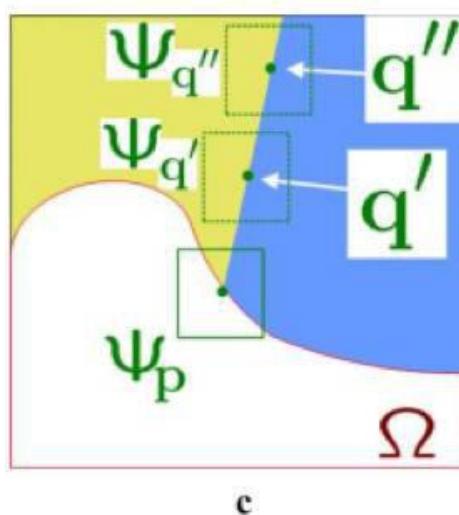
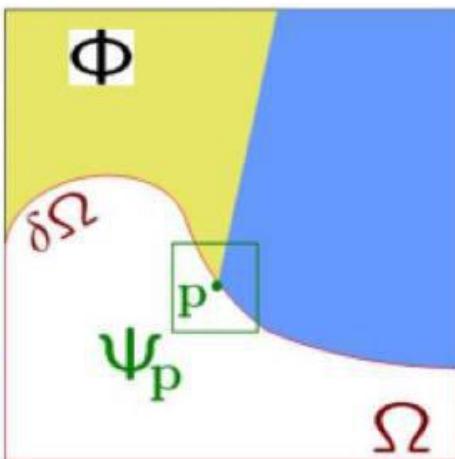
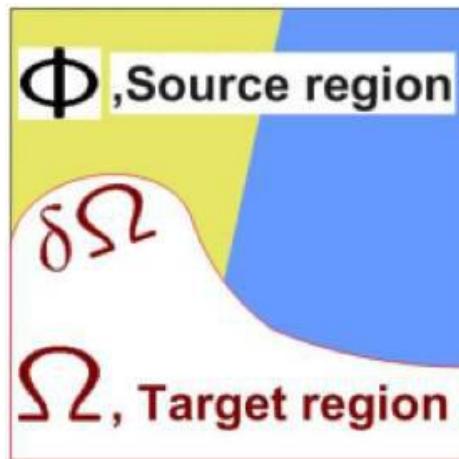


图像生成图像

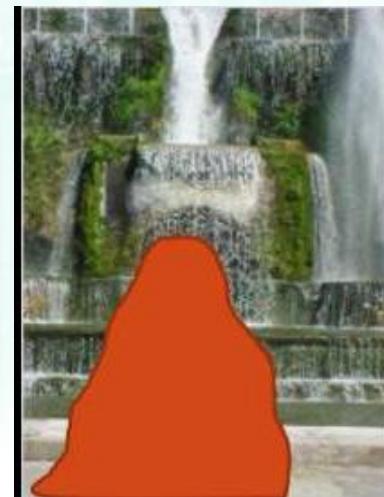
□ 应用：白天图像生成黑夜图像



传统图像修复



知乎 @Momenata

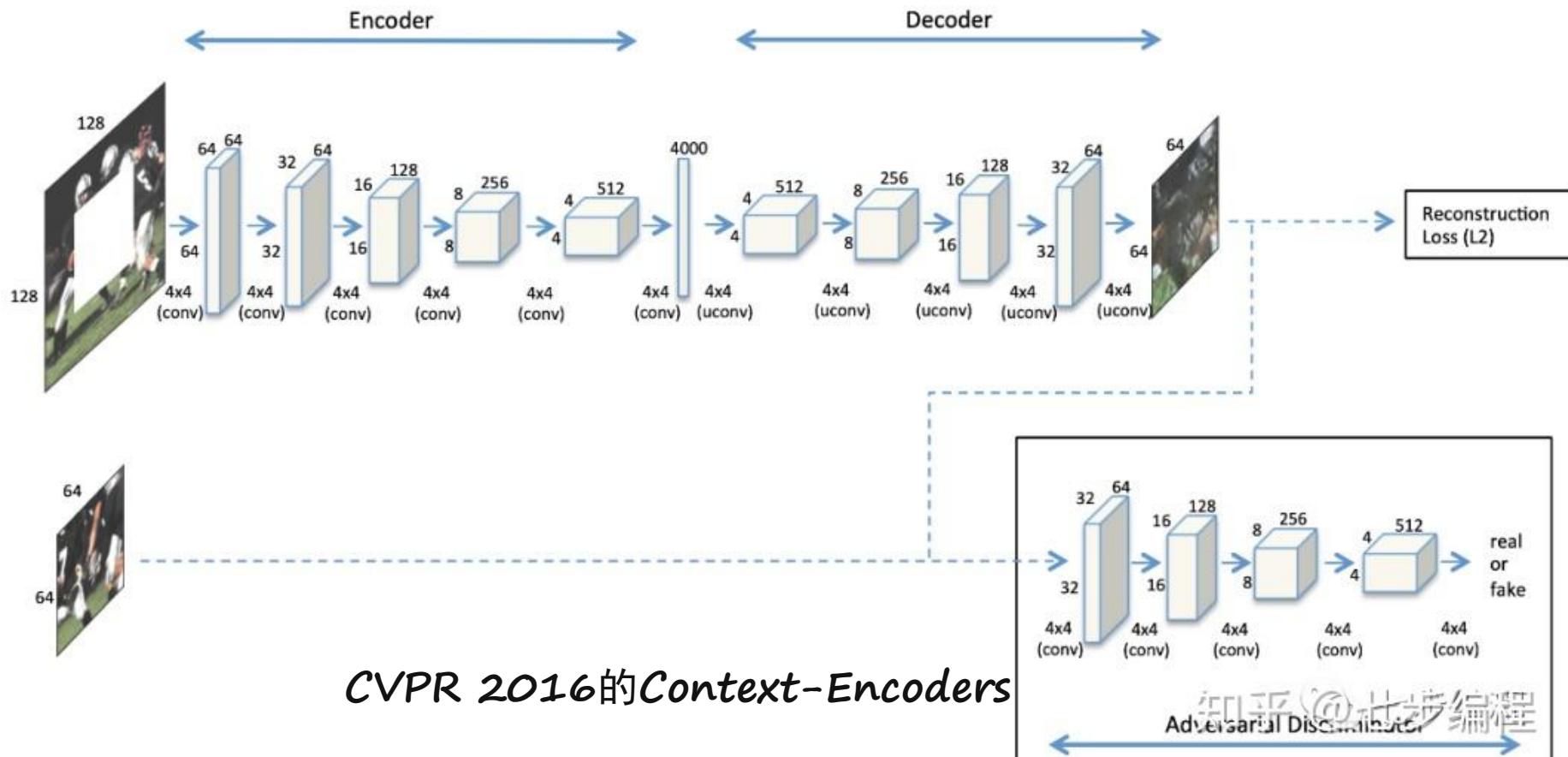


知乎 @Momenata

Region Filling and Object Removal by Exemplar-Based Image Inpainting

这篇文章是2004年的工作，核心思想就是利用图像本身的冗余性(redundancy)，用图像已知部分的信息来补全未知部分。

图像新媒体修复

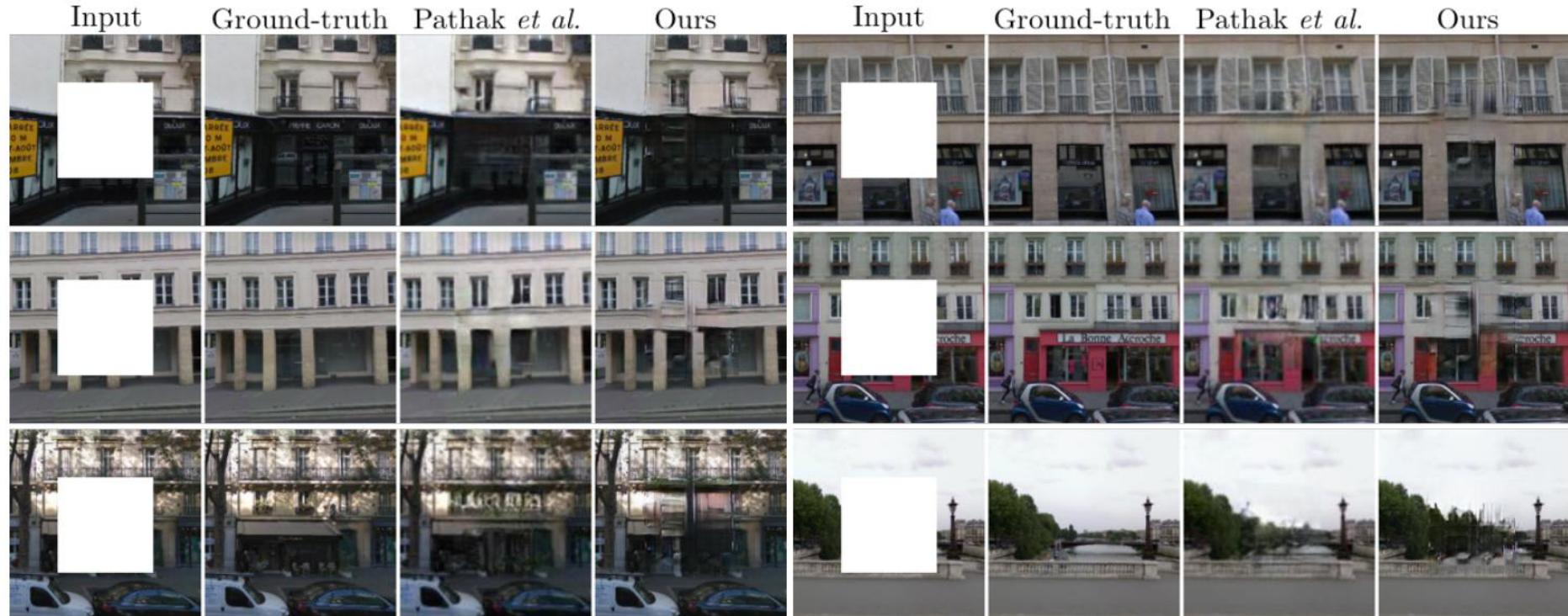


CVPR 2016的Context-Encoders



图像新媒体合成编辑

□ 应用： 图像修复





智能新媒体合成编辑技术

14.2 视频新媒体合成编辑

新媒体中的关键流量

- 高质量的内容，内容是吸引粉丝的关键
- 自媒体流量入口可以通过文字、视频、语音等形式进行分享输出，展示的平台就在各大自媒体



如何让视频新媒体更加亮眼



PR, 你已经是个成熟的软件了
该学会自己拿剪刀了

你写脚本，AI自动剪视频：13分钟完成剪辑师7小时创作

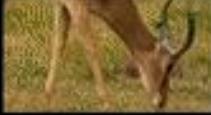


长颈鹿是世界上最高的动物，以其长腿和长脖子闻名于世。它的脖子上有棕色的鬃毛，头上长了两只毛茸茸的角

Text

The giraffe is the world's tallest animal, and well known for its long legs and neck. It has a brown mane on the neck, and its head has two hairy horns. And it is quite interesting that they fight using th...

Theme-related video repository



问题？

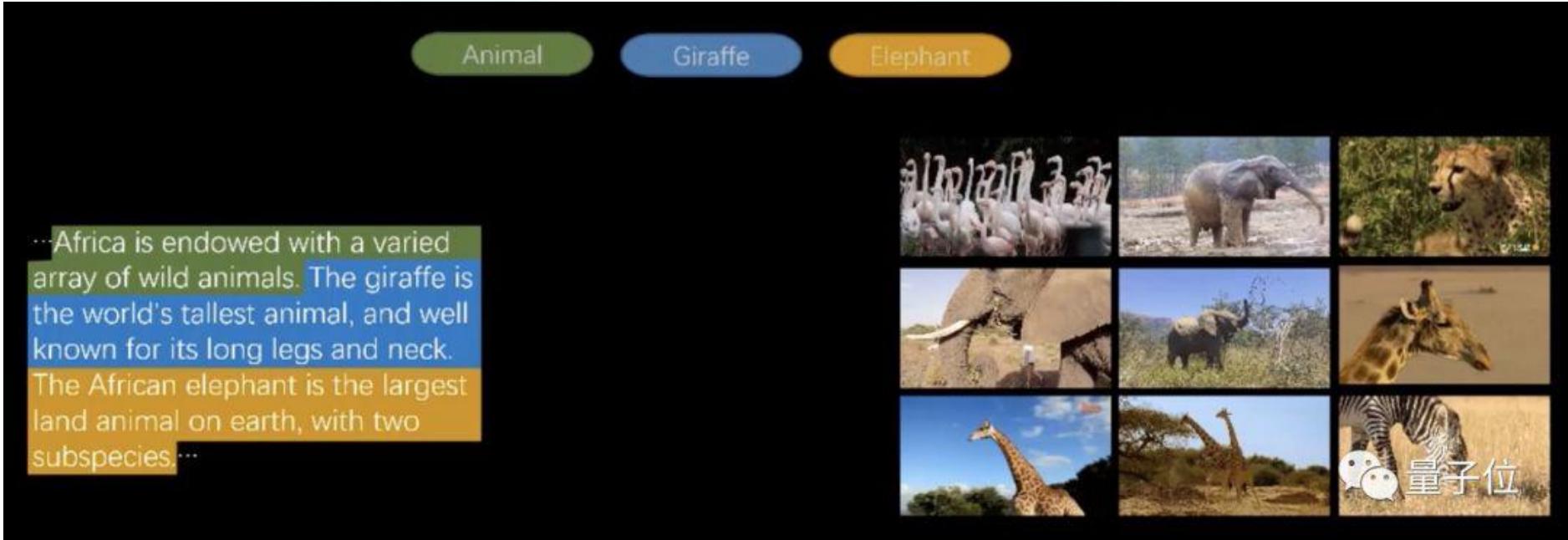


Video montage

先展示一下长颈鹿群的全貌。接着镜头切给一只奔跑中的长颈鹿，展示一下它的长腿长脖子。按照指示，再给鬃毛和犄角来个特写，

文本生成视频

- 第一步，用户以文本的形式提供输入。Write-A-Video会挑选出句子中的关键词



The screenshot shows a user interface for generating a video from text. On the left, there is a text input field containing the following text:

Africa is endowed with a varied array of wild animals. The giraffe is the world's tallest animal, and well known for its long legs and neck. The African elephant is the largest land animal on earth, with two subspecies...

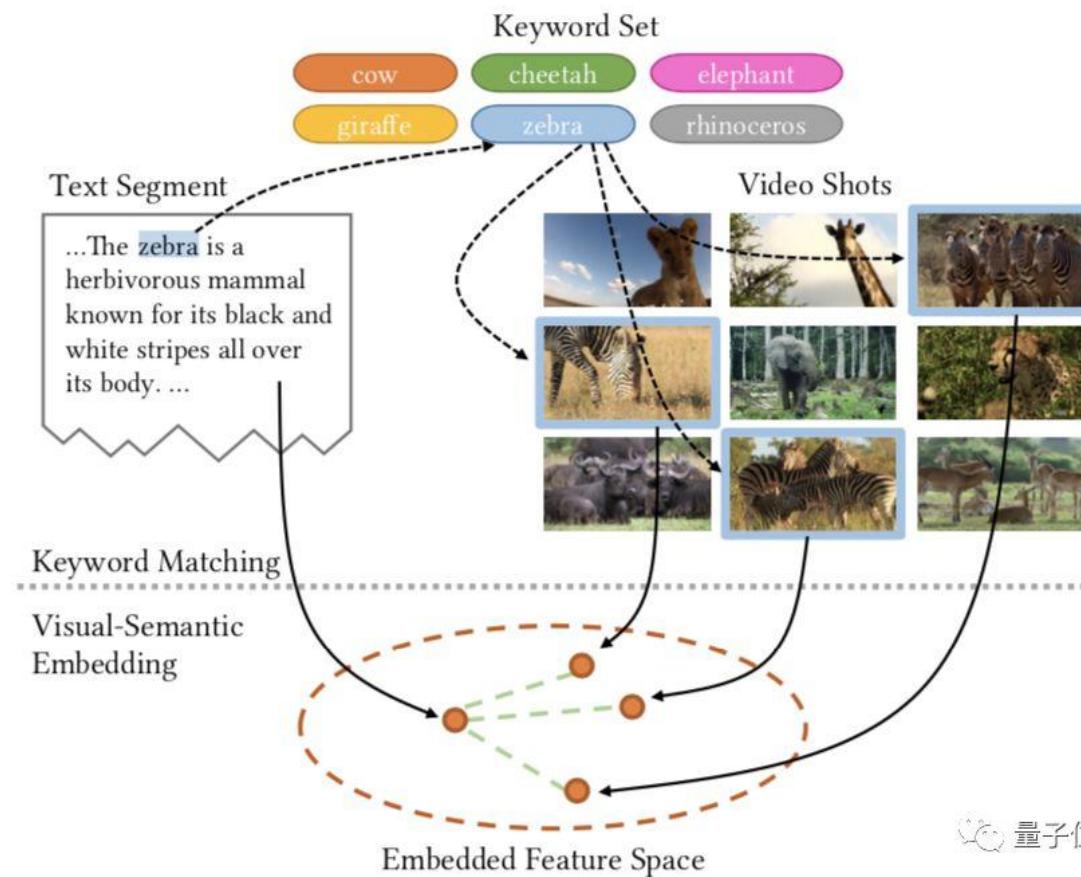
At the top, there are three buttons: "Animal" (green), "Giraffe" (blue, currently selected), and "Elephant" (orange). To the right of the text input, there is a 3x3 grid of nine images. The images are as follows:

- Row 1: A herd of zebras, an African elephant standing, and a cheetah.
- Row 2: An African elephant with tusks, another elephant in a field, and a giraffe's head and neck.
- Row 3: A giraffe standing tall, another giraffe, and a zebra.

In the bottom right corner of the grid, there is a small white icon of a face with large eyes and the text "量子位".

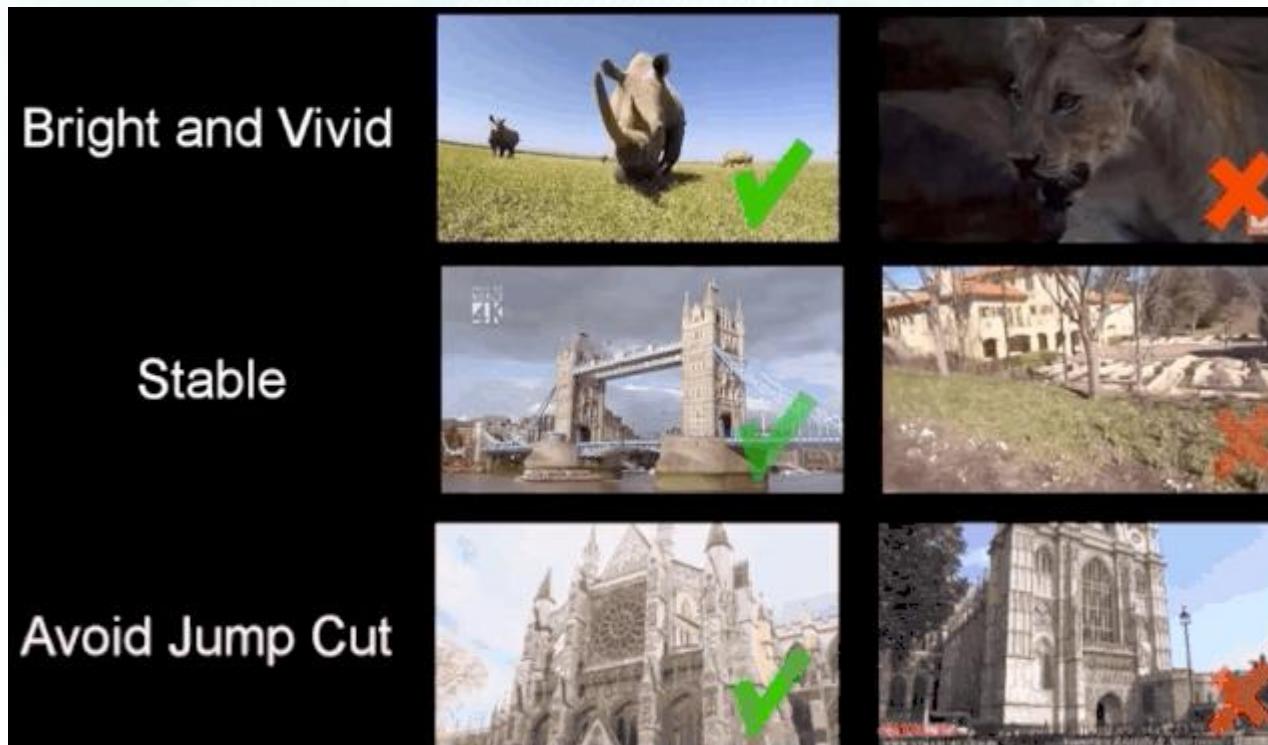
文本生成视频

- 第二步，Write-A-Video会利用关键词，把素材库里与之相匹配的候选片段挑出来。
- 文本和镜头之间的视觉语义匹配主要包括两个步骤：关键字匹配和视觉语义嵌入。



文本生成视频

- 第三步，就是将这些镜头组合在一起，完成视频的剪辑。
- 首先，画面应该是明亮而生动的
- 其次，镜头不能晃动得太厉害。
- 最后，要避免不连贯的跳接（jump cut）和相反的相机运动。





视频新媒体合成编辑

- 视频实际是由图像序列所组成的，图像的合成编辑技术作用在图像序列上，再考虑序列的时域信息，即可达到对视频的合成编辑。
- 同样的，视频新媒体合成编辑为了获得高层语义信息也采用了GAN为模型架构。



音频生成视频

Synthesizing Obama: Learning Lip Sync from Audio

Supasorn Suwajanakorn
Steven M. Seitz
Ira Kemelmacher-Shlizerman

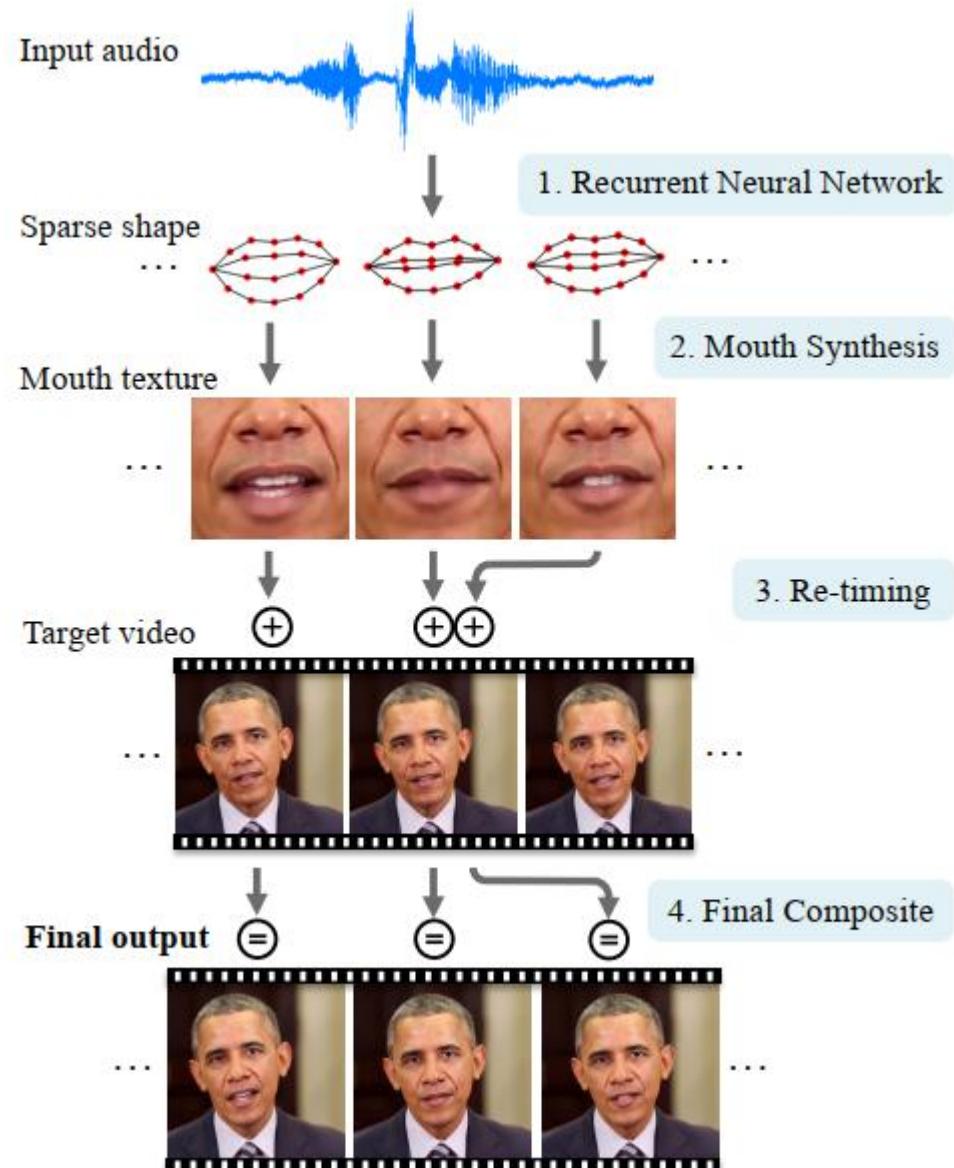
University of Washington

SIGGRAPH 2017
<http://grail.cs.washington.edu/projects/AudioToObama/>

视频新媒体合成编辑

主要是学习一个从音频到视频的序列映射，为了简化问题，论文只关注合成嘴周围区域的内容，其他眼睛、头部、上半身、背景等都完全保留。

给定一段音频序列，作者首先提取特征作为RNN的输入，RNN输出一个稀疏的嘴型对应于每一帧输出的视频，对于每一个稀疏的嘴型，合成嘴和人脸下部的纹理，再将它们嵌入到目标视频中作为输出。





视频新媒体合成编辑

视频生成视频

Video-to-Video Synthesis

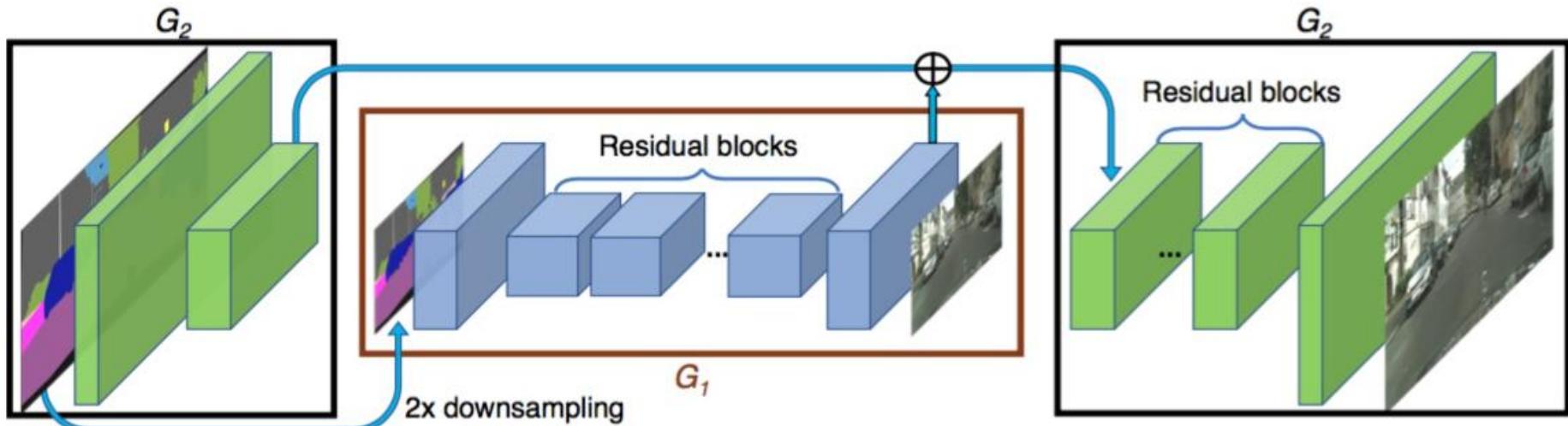
Ting-Chun Wang¹, Ming-Yu Liu¹, Jun-Yan Zhu², Guilin Liu¹,
Andrew Tao¹, Jan Kautz¹, Bryan Catanzaro¹

¹NVIDIA Corporation ²MIT

□ Pix2PixHD

生成器由两部分组成，**G1**和**G2**，其中**G2**又被割裂成两个部分。**G1**和pix2pix的生成器没有差别。**G2**的左半部分提取特征，并和**G1**的输出层的前一层特征进行相加融合信息，把融合后的信息送入**G2**的后半部分输出高分辨率图像。

判别器使用多尺度判别器，在三个不同的尺度上进行判别并对结果取平均。判别的三个尺度为：原图，原图的 $1/2$ 降采样，原图的 $1/4$ 降采样。





视频新媒体合成编辑

□ Vid2Vid

光流约束：视频中一般存在的大量的信息冗余，对于相邻的两帧像，在空间上大部分区域像素都是相同的，而只有少部分存在运动区域的像素有较大的变化，光流是可以用来表示这些区域的变化大小和方向。

在生成第 t 帧视频时，相比与仅使用条件输入第 t 帧语义图 s_t 或者使用前 L 帧视频的完整信息，更好的方式是估算从第 $t - 1$ 帧到第 t 帧的光流 \tilde{w}_{t-1} ，然后作用于 $t - 1$ 帧，从而直接得到当前第 t 帧的预测值，即加入光流约束。

$$\mathcal{L}_W = \frac{1}{T-1} \sum_{t=1}^{T-1} \left(\|\tilde{\mathbf{w}}_t - \mathbf{w}_t\|_1 + \|\tilde{\mathbf{w}}_t(\mathbf{x}_t) - \mathbf{x}_{t+1}\|_1 \right).$$

1. 生成器估计的光流和Flownet2生成的光流之间的pixel-wise误差。

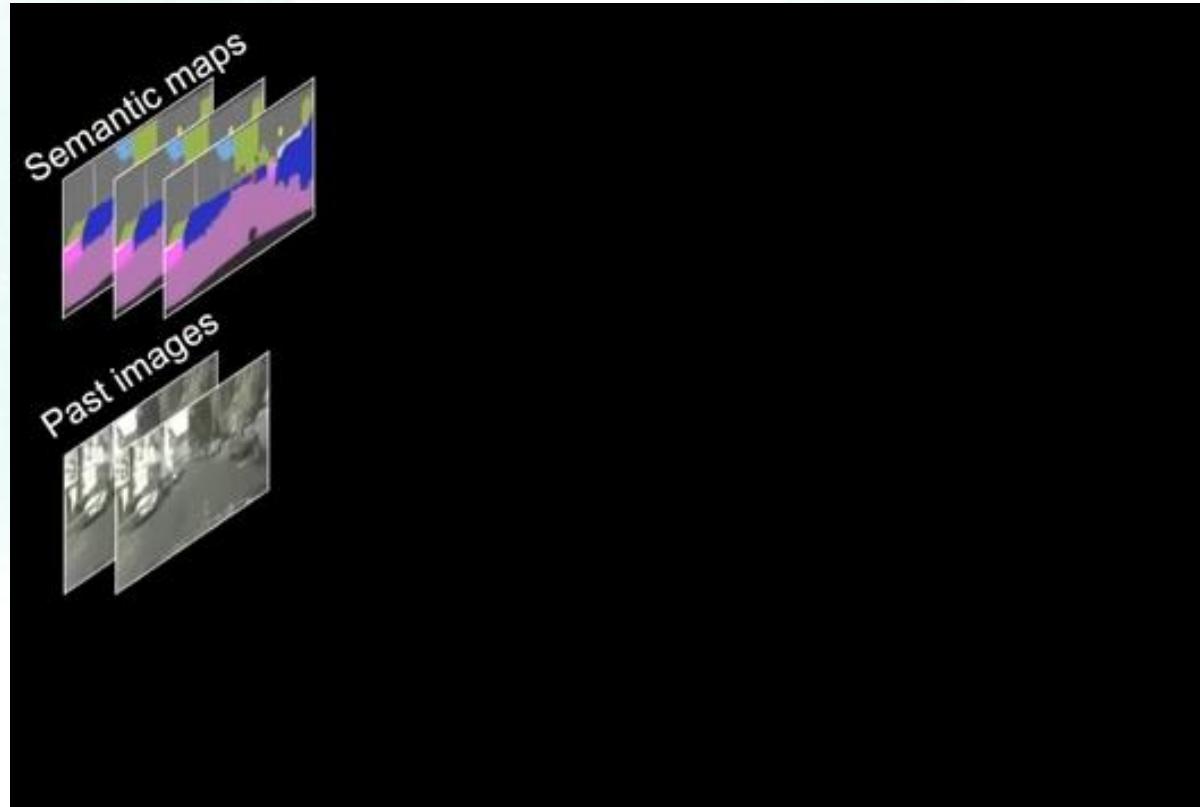
2. 对当前帧的真实图像(real)使用估计的光流得到的下一帧图像与真实的下一帧图像之间的pixel-wise误差。

问题？

视频新媒体合成编辑

□ Vid2Vid生成器

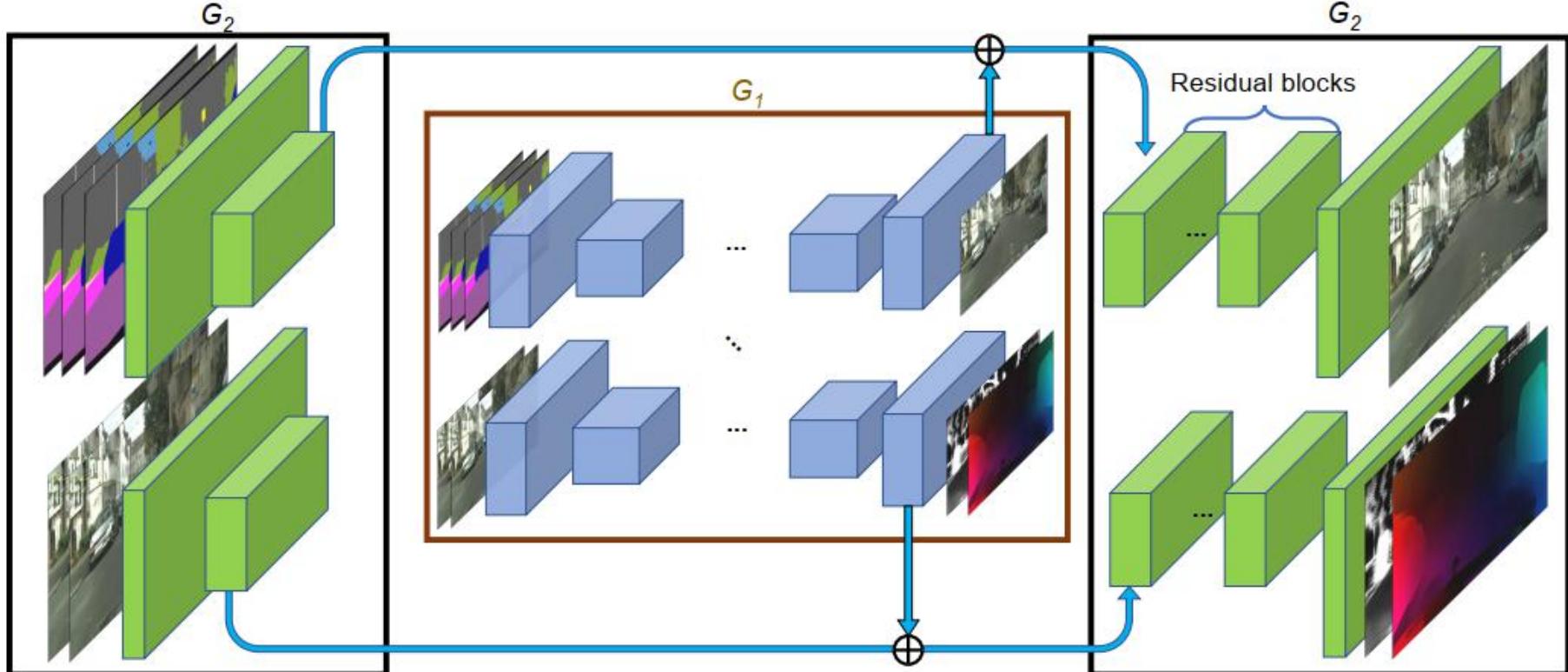
[1] 输入语义图和历史图片经过G1生成器得到低清图片和光流估算。[2] 使用历史图片和光流估计得到Warped扭曲图。[3] 利用扭曲图和低清晰图得到最终的高清图片，并进行下一帧图的构建。



视频新媒体合成编辑

□ Vid2Vid

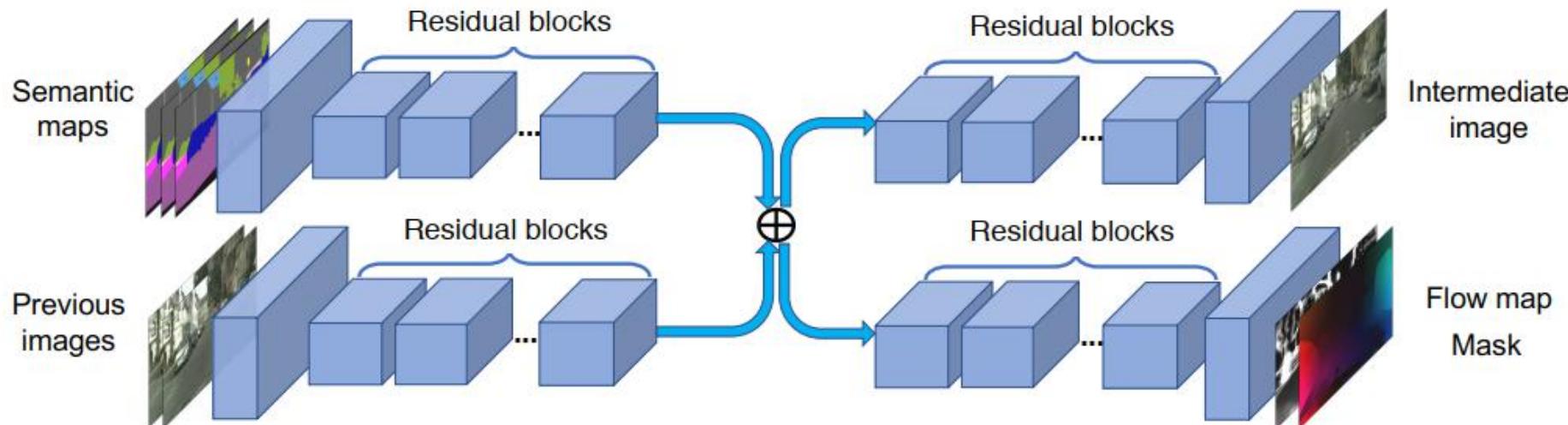
Vid2Vid作为Pix2PixHD的改进版本，重点解决了视频到视频转换过程中的前后帧不一致性问题。它加入前后帧的光流信息作为约束，建立在Pix2PixHD基础之上，加入时序约束，因此可以实现高分辨率视频生成。



视频新媒体合成编辑

□ Vid2Vid

G1的输入是前L帧+当前帧的语义图序列以及之前L帧生成图像的序列，在经过下采样和提特征之后，在网络的中间将两路输出提取入的feature map相加，接着在网络后部又分出两个分支，来生成未加光流约束的原始图像以及光流和权重mask；





智能新媒体合成编辑技术

14.3 文本生成技术



文本生成技术

□ 生成一段对话回复的模型可以简单分为三类：规则模板、检索模型、生成模型。

a. 规则模板。这种回复实际上需要人为设定规则模板，对用户输入进行回复。

优点：1、实现简单，无需大量标注数据；2、回复效果可控、稳定。

不足：1、如果需要回复大量问题，则需要人工设定大量模板，人力工作量大；2、使用规则模板生成的回复较为单一，多样性低。



文本生成技术

□ AIML语言

- AIML，全名为Artificial Intelligence Markup Language（人工智能标记语言），是一种创建自然语言软件代理的XML语言。

```
<category>  
    <pattern>WHAT IS YOUR NAME</pattern>  
    <template>My name is Leo.</template>  
</category>
```

问机器人“WHAT IS YOUR NAME”，机器人便会回答“My name is Leo.”。



文本生成技术

b. 检索模型。利用文本检索与排序技术从问答库中挑选合适的回复。

优点：由于数据来源于已经生成好的回复，或是从已抓取的数据得到的回复，所以语句通顺性高，万能回复少；

不足：1. 不能生成新的回复文本，只能从问答库中得到文本进行回复；2. 当检索或排序时，可能只停留在表面的语义相关性，难以捕捉真实含义。



文本生成技术

- 检索模型一般首先构建一个由大量query-response pair构成的知识库（比如从豆瓣、贴吧等地方抽取），然后将对话中最后一次的回复作为query，通过经典的信息检索方式（TextRank、TFIDF等）来检索若干相关的候选回应。

- 问题1： 答案1
- 问题2： 答案2
- 问题3： 答案3



文本生成技术

c. 生成模型。主要用encoder-decoder结构生成回复。典型技术是Seq2Seq、transformer。

优点：无需规则，能自动从已有对话文本中学习如何生成文本。

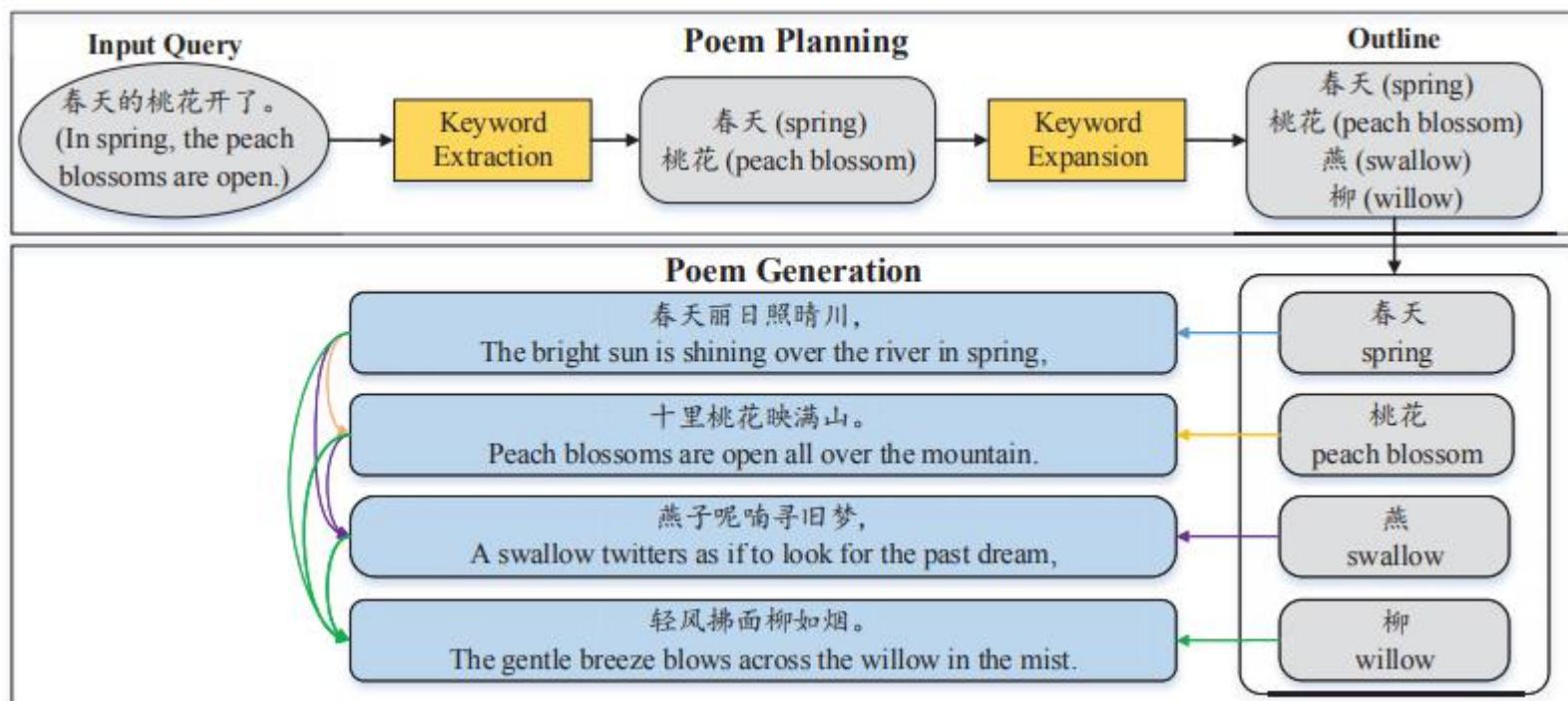
不足：1、生成效果不可控，训练好的模型更像是一个“黑盒”，也无法干预模型的生成效果；2、倾向生成万能回复，如“好的”、“哈哈”等，所以多样性与相关性低。

文本生成技术

□ 诗歌生成方法

第一步：生成诗词主题

根据输入的关键词或句子，提取诗词关键词，根据待生成句子的数量，生成对应数量的关键词。使用TextRank算法结合word2Vec词向量对关键词重要性排列，取最重要的关键词。如果能提取的关键词过少，则使用RNN进行预测新的关键词。

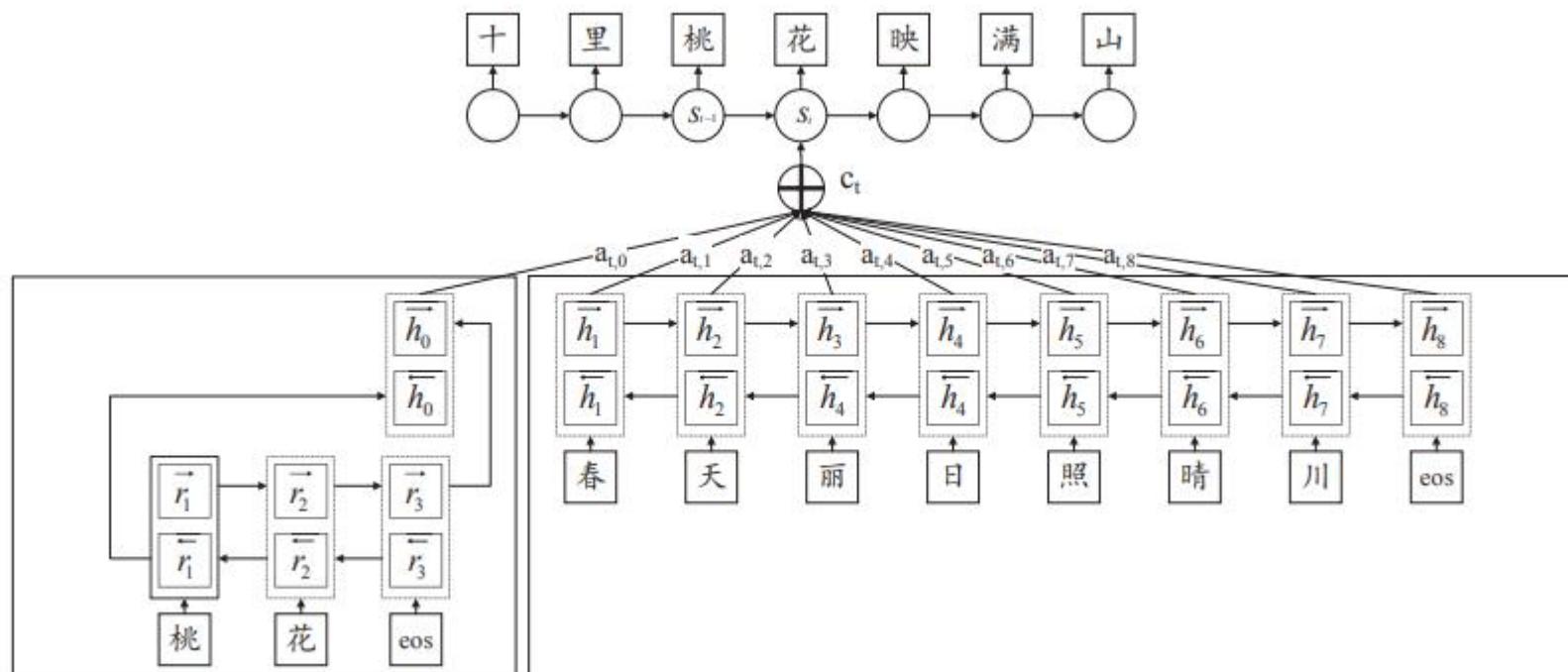


文本生成技术

□ 诗歌生成方法

第二步：生成诗词

第一步生成的关键词结合上一个时间步生成的诗词作为双向RNN的输入，利用Attention机制，生成诗词。

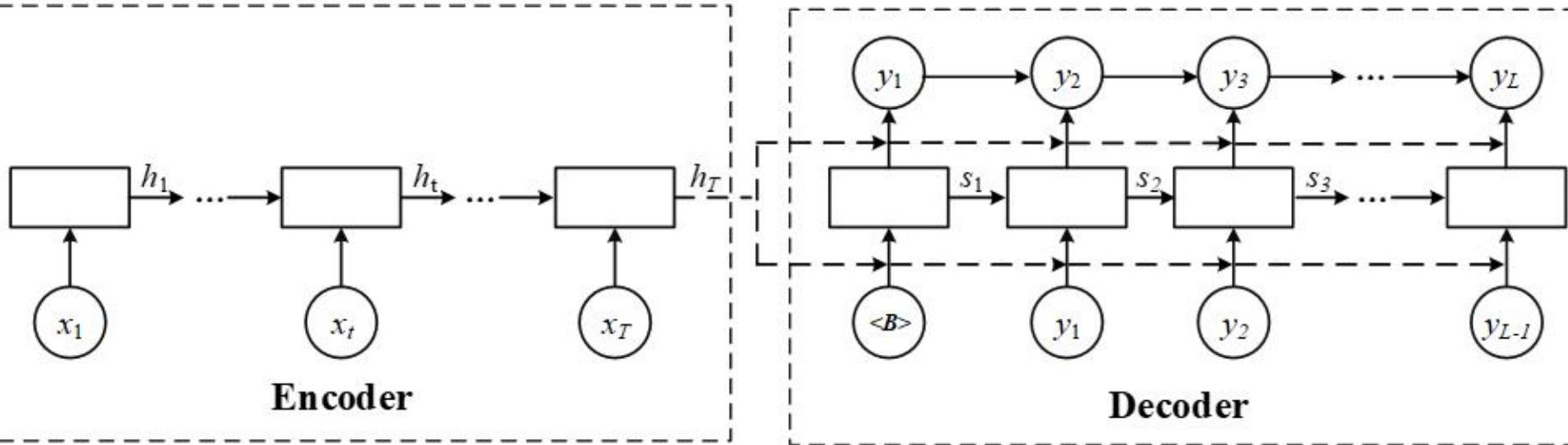


文本生成技术

□ 诗歌生成方法

解码器需要根据生成的第t个字、第t个生成的特征向量和上一次解码器输出的概率，输出第t个字的概率。

$$y_t = \arg \max_y P(y|s_t, c_t, y_{t-1}).$$



$$s_t = f(s_{t-1}, y_{t-1}, h_T)$$

$$y_t = g(s_t, y_{t-1}, h_T)$$

文本生成技术



秋夕湖上

By a Lake at Autumn Sunset

一夜秋凉雨湿衣，

A cold autumn rain wetted my clothes last night,
西窗独坐对夕晖。

And I sit alone by the window and enjoy the sunset.
湖波荡漾千山色，

With mountain scenery mirrored on the rippling lake,
山鸟徘徊万籁微。

A silence prevails over all except the hovering birds.

秋夕湖上

By a Lake at Autumn Sunset

荻花风里桂花浮，

The wind blows reeds with osmanthus flying,
恨竹生云翠欲流。

And the bamboos under clouds are so green as if to flow down.
谁拂半湖新镜面，

The misty rain ripples the smooth surface of lake,
飞来烟雨暮天愁。

And I feel blue at sunset .

啤酒

Beer

今宵啤酒两三缸，

I drink glasses of beer tonight,
杯底香醇琥珀光。

With the bottom of the glass full of aroma and amber light.
清爽金风凉透骨，

Feeling cold as the autumn wind blows,
醉看明月挂西窗。

I get drunk and enjoy the moon in sight by the west window.

冰心

Xin Bing

一片冰心向月明，

I open up my pure heart to the moon,
千山春水共潮生。

With the spring river flowing past mountains.
繁星闪烁天涯路，

Although my future is illuminated by stars,
往事萦怀梦里行。

The past still lingers in my dream.



智能新媒体合成编辑技术

14.4 语音合成技术



语音生成技术

- 让语音助手说话的技术叫 TTS (text-to-speech)，也就是语音合成。TTS 技术本质上解决的是从文本转化为语音的问题，通过这种方式让机器开口说话。

一个从文本转化到语音的问题



你好啊..





语音生成技术

- 比较经典的方法是百度公司的Deep Voice

一个完全由深度神经网络构建的生产质量的文本到语音系统，为真正的端到端神经语音合成奠定了基础。

该系统包括五个主要的模块：定位音素边界的分割模型、字素-音素转换模型、音素时长预测模型、基频预测模型和音频合成模型。

语音生成技术

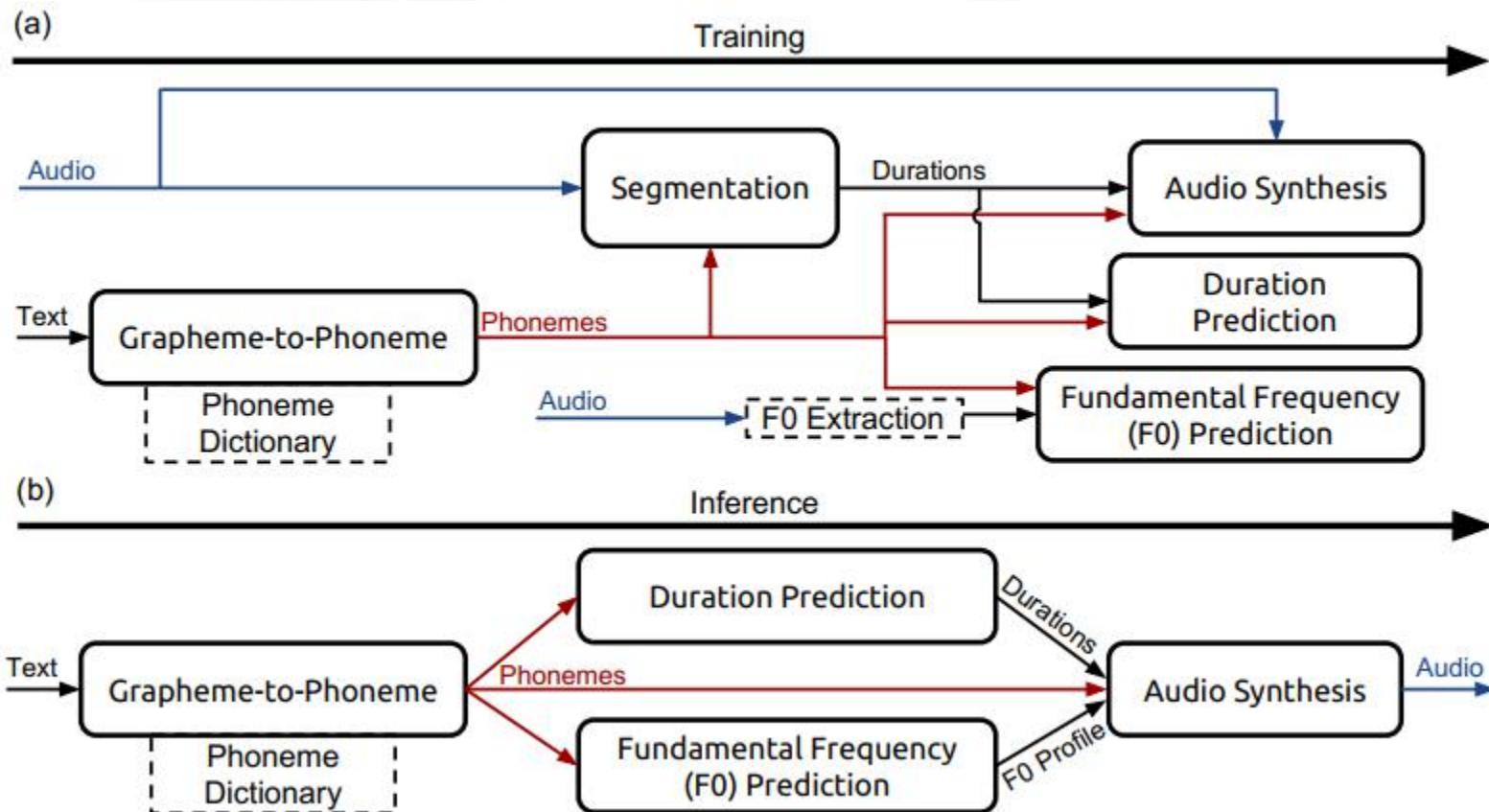


- 系统由五个主要的构建模块组成：
 - a. 字素到音素模型：将书面文本（英语字符）转换为音素（使用像ARPAbet这样的音素字母表编码）。
 - b. 分割模型：对语音数据集中的音素边界进行定位。给定一个音频文件和音频的一个音素逐音素转录，分割模型确定每个音素在音频中的起始和结束位置。
 - c. 音素持续时间模型：预测音素序列（一句话）中每个音素的时间持续时间。
 - d. 基频模型：预测一个音素是否被发声；如果是，该模型预测整个音素持续时间的基频（F0）。
 - e. 音频合成模型：将字素到音素、音素持续时间和基频预测模型的输出组合，并以与所需文本相对应的高采样率合成音频。

语音生成技术



□ Deep Voice的训练过程和推理过程





□ 字素到音素模型

采用了Seq2Seq的编码器-解码器架构。使用了双向 GRU 编码器和同样深度的单向 GRU 解码器。每个解码器层的初始状态初始化为对应编码器转发层的最终隐藏状态。

1. though (和 go 里面的 o 类似)

2. through (和 too 里面的 oo 类似)

3. cough (和 offer 里面的 off 类似)

4. rough (和 suffer 里面的 uff 类似)

相同的拼写，但发音却完全不同

- Input - "It was earky spring"
- Output - [IH1, T, ., W, AA1, Z, ., ER1, L, IY0, ., S, P, R, IH1, NG,.]

音素解决了发音问题



□ 分割模型

分割模型经过训练，输出对齐的给定的话语和目标音素序列。

这个任务类似于语音识别中将语音与文本输出对齐的问题。在该领域中，连接主义者时间分类（CTC）损失函数专注于字符对齐，以学习声音和文本之间的映射。



□ 音素持续时间和基频模型

采用一个结构来共同预测音素持续时间和随时间变化的基频。

该模型的输入是一个带有重音的音素序列，每个音素和重音都被编码为一个one-hot向量。

该架构包括两个全连接层，每个层有 256 个单元，然后是两个单向循环层，每个层有 128 个GRU单元，最后是一个全连接的输出层。

最后一层对每个输入音素产生三种估计：音素持续时间、音素清音的概率(即具有基频)和 20 个时间相关的 F0 值，这些值在预测持续时间内均匀采样。

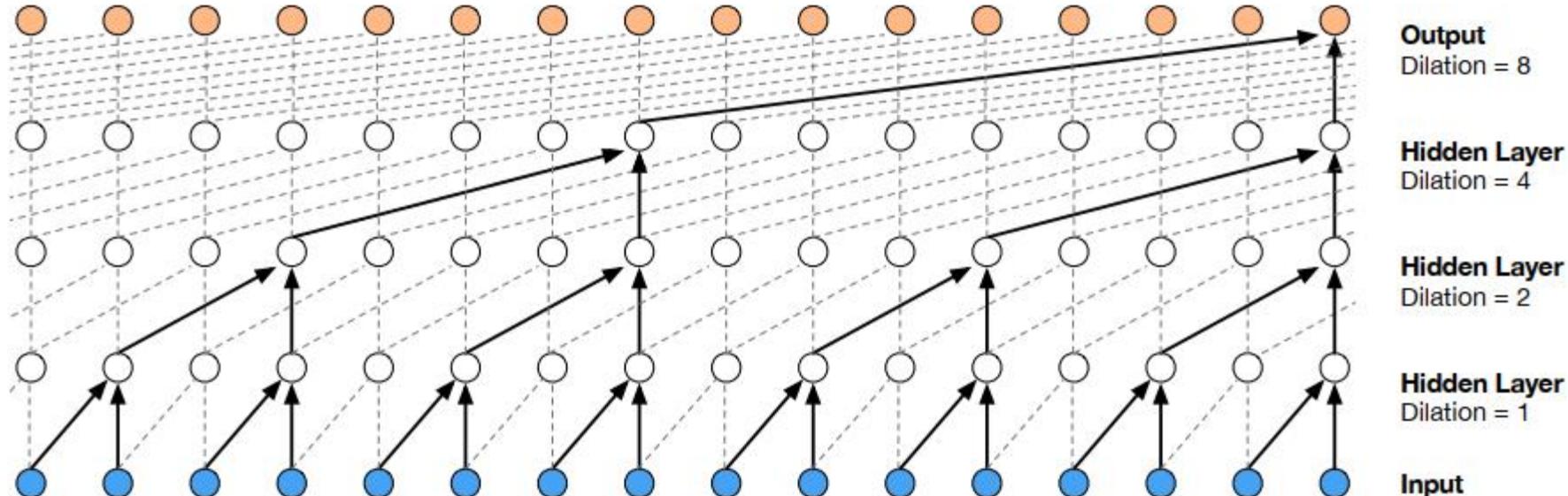
语音生成技术



音频合成模型

参考了WaveNet模型，WaveNet可以根据粗糙的音频生成高质量的音频。
思想是当前的音频可以基于之前的音频序列生成。

$$p(\mathbf{x}) = \prod_{t=1}^T p(x_t | x_1, \dots, x_{t-1})$$



语音生成技术



□ 效果展示

