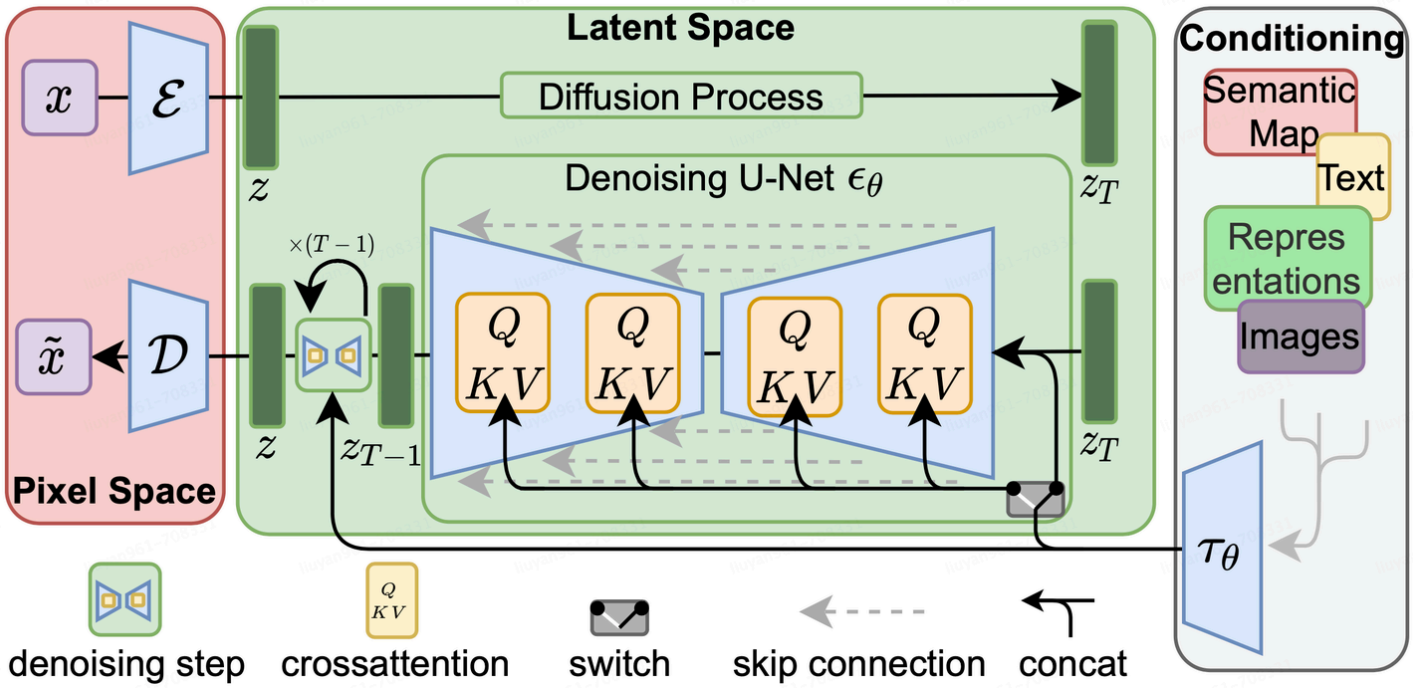
LDM（Latent Diffusion Model）详解



前言

我们这里介绍一篇重要的扩散模型（Diffusion Model）算法：潜空间扩散模型：LDM[1]。LDM是前端时间爆火的图像生成算法Stable Diffusion以及最新备受关注的视频生成模型Sora最核心的理论基础之一。在DDPM[2]的这篇文章中，我们介绍到DDPM是一个基于马尔可夫链的算法，它通过对一个随机噪声进行逐步去噪来实现了图像生成任务。DDPM等算法是直接在图像像素空间中进行操作，并且因为DDPM的链式特性，这造成了它的训练和推理都是非常消耗资源的。为了提升扩散模型的生成效率，LDM提出了将扩散空间从

图像空间转移到潜空间（Latent Space），而这个潜空间的概率分布可以通过训练好的VAE得到。预测概率分布除了能够提升生成效率，还能够避免扩散模型在图像像素上的过度训练，而图像的这些细节我们交给更擅长做这个的VAE去完成，从而显著提升了生成图像的质量。最后，LDM通过交叉注意力模块，支持将不同模态的条件加入到扩散过程中，从而实现了生成模型的生成内容的可控性。

1. 背景知识

1.1 扩散模型

DDPM是最早提出扩散模型的算法之一，它包括前向加噪过程和反向去噪过程，如图1。在前向过程中，我们通过逐渐向一个图像中添加高斯噪声得到一个纯噪声图像。而在反向过程中，我们通过预测每一步添加的噪声来实现对这个图像的还原。而对于一个随机生成的高斯噪声，我们通过对其逐步去噪得到一个无噪声的图像，从而实现了图像的生成。DDPM是一个效率非常低的算法，因此诞生了许多提升DDPM生成效率的算法，例如DDIM[3]提出使用更高效的采样策略来避免计算所有去噪步骤的计算。

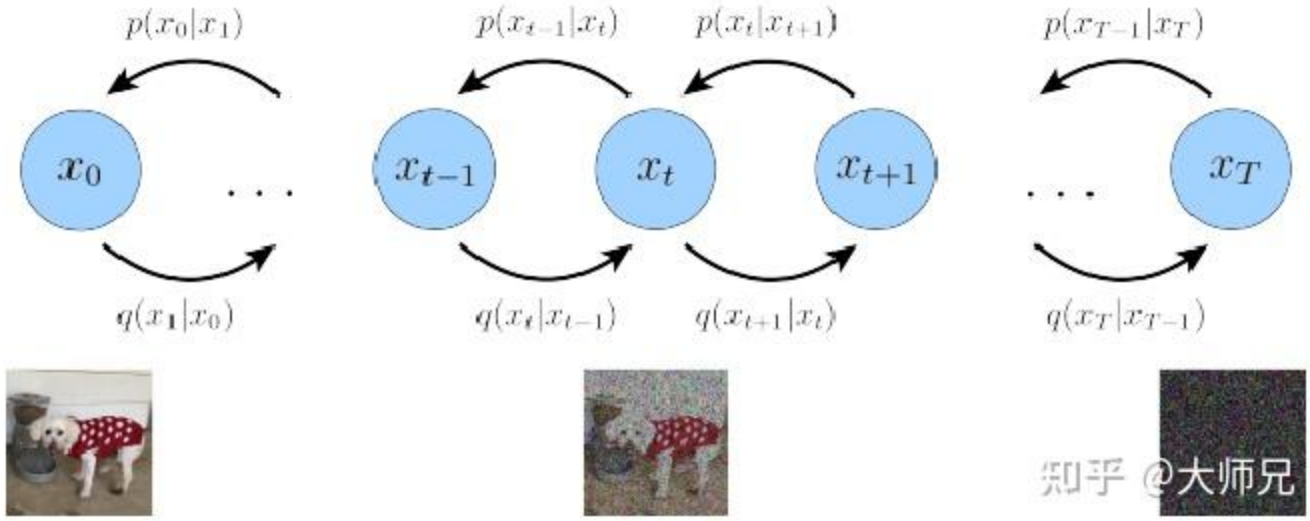


图1： DDPM的前向加噪和后向去噪过程

尽管如此，基于DDPM的扩散模型仍旧是一个效率非常低的算法。其中最重要的原因是DDPM等均是在和输入图像相同尺寸的特征空间上进行操作，而一个高分辨率的图像往往拥有数百万个像素，预测如此密集的像素是DDPM效率低的一个重要原因。LDM提出我们可以通过在特征数更少的潜空间（Latent Space）上计算来提升生成效率。

1.2 VQ-VAE

AE（AutoEncoder）是一个经典的生成模型，它由编码器和解码器组成。VQ-VAE（Vector Quantised - Variational AutoEncoder）[4]与AE最大的不同是AE编码的特征向量是连续的，而VQ-VAE编码的特征向量是离散的，如图2。这种将特征映射为离散值的思想在BEiT v2中也使用过，这个离散值在计算机视觉中被叫做视觉码本（Visual Codebook）或者视觉字典（Visual Dictionary），关于视

觉码本的详细介绍参考我在BEiT v2的第1.1节的介绍。

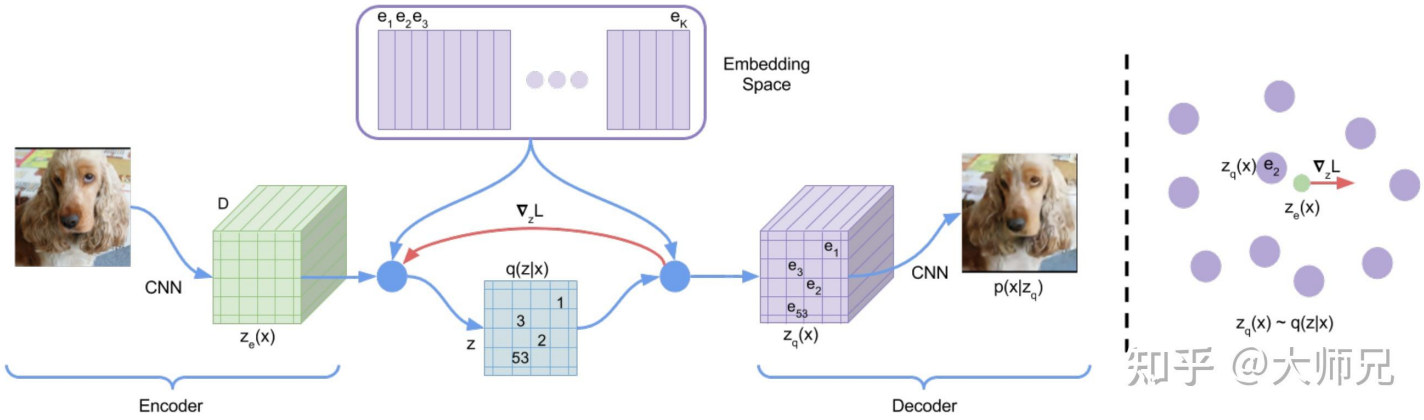


图2：VQ-VAE将特征编码为离散特征

在DDPM中，我们从一个随机高斯噪声还原图像，那么能不能从VQ-VAE得到的离散特征中进行还原呢，LDM就是这么做的。

1.3 VQ-GAN

VQ-GAN[5]是一个改良版的VQ-VAE，相对于VQ-VAE，它做了3点改进，如图3：

- 因为CNN的局部特性无法捕捉较远像素之间的依赖关系，因此VQ-GAN使用了Transformer代替pixelCNN；
- 额外增加了一个PatchGAN作为判别器，并在训练时加入了判别损失。
- 使用了感知损失[6]代替传统的L2损失。

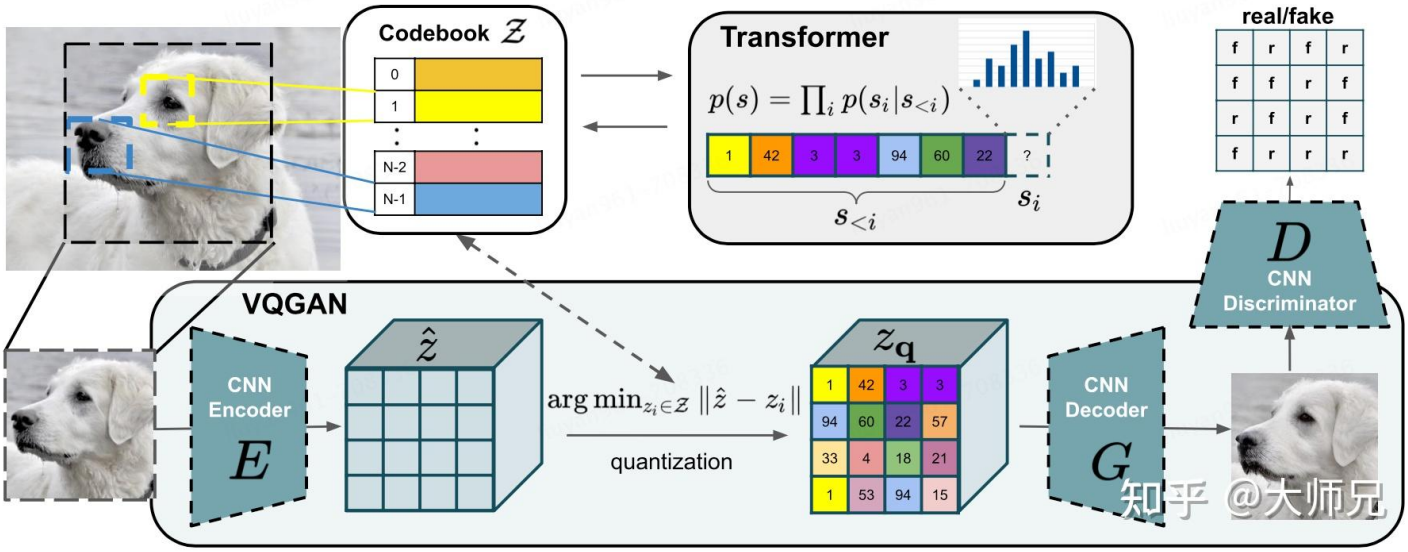


图3：VQ-GAN的计算流程

2. 算法详解

LDM是一个二阶段的模型，包括训练一个VQ-VAE和扩散模型本身，LDM的计算流程如图4所示。

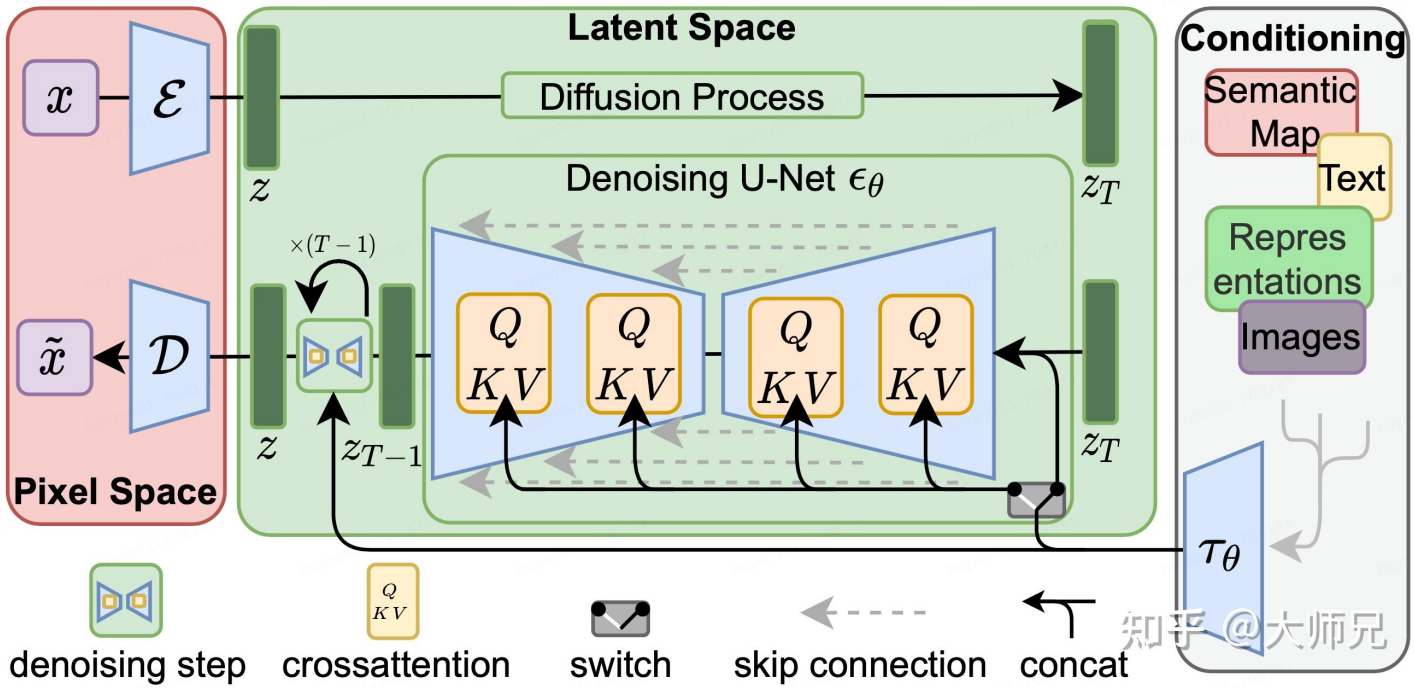


图4：LDM的计算流程

LDM有三个主要模块：

- 感知图像压缩（Perceptual Image Compression）：图3中最左侧红框部分是一个VQ-VAE，用于将输入图像x编码为一个离散特征z。
- LDM：图3的中间绿色部分是在潜变量空间的扩散模型，其中上半部分是加噪过程，用于将特征 z加噪为 z_T。下半部分是去噪过程，去噪的核心结构是一个由交叉注意力（Cross Attention）组成的U-Net，用于将 z_T还原为 z。
- 条件机制（Conditioning Mechanisms）：图3的右侧是一个条件编码器，用于将图像，文本等前置条件编码成一个特征向量 τ_θ，并将其送入到扩散模型的去噪过程中。

3.1 感知图像压缩

具体来讲，在感知图像压缩中我们使用了一个训练好的VQ-GAN。它包括一个编码器 $\mathcal{E}$和一个解码器 $\mathcal{D}$。编码器 $\mathcal{E}$用于将一个RGB彩色图像 $x \in \mathbb{R}^{H\times W\times 3}$压缩到一个特征向量 $z \in \mathbb{R}^{h\times w\times c}$。解码器 $\mathcal{D}$则用于将这个潜变量 z还原为输入图像 $\tilde{x}$。

为了避免潜空间的方差过大，LDM尝试了两种不同的正则方法，分别是KL-reg和VQ-reg。其中KL-reg相当于在潜空间上施加了KL惩罚，使得得到的特征的分布接近正态分布。VQ-reg则是使用了一个向量量化层（vector quantization layer），向量量化层将特征归一化为码本中最近的那个特征，如图5。具体来讲，向量量化层的计算有6步：

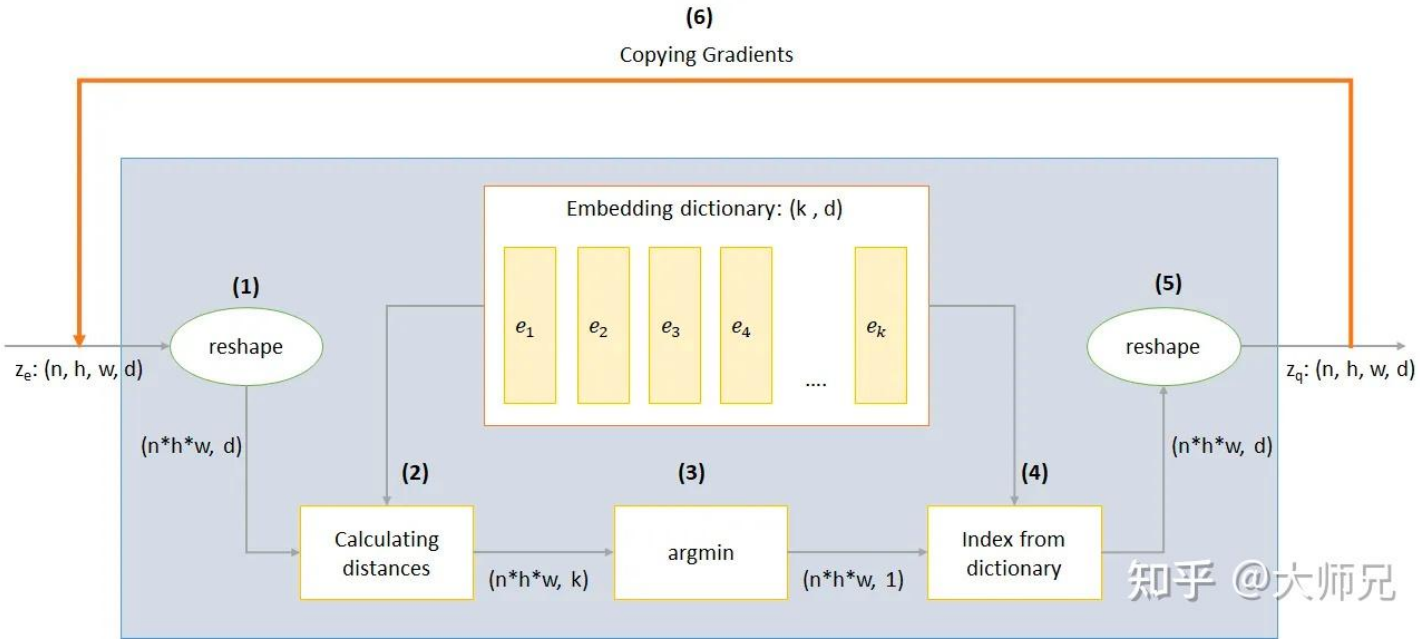


图5：向量量化层的计算流程

1. 特征Reshape：对于输入数据 $z_e \in \mathbb{R}^{n \times h \times w \times d}$，除了最后一个维度外合并所有维度，得到 $n \times h \times w$个长度为 d的特征向量；
2. 计算距离：对于每个特征向量，计算其与视觉字典中 k个向量之间的距离，字典中每个向量的特征数是 d；
3. Argmin：根据得到的 k个距离，根据距离的最小值选择距离最近的特征的索引；

4. 最近特征：根据索引从字典中取最近的特征作为当前特征的具体值，得到 $n \times h \times w$个长度为 d的特征；
5. 还原特征：对特征进行Reshape，得到 $z_q \in \mathbb{R}^{n,h,w,d}$；
6. 复制梯度：因为在第3步中我们使用了argmin，因此这个计算过程是不可导的。向量量化层的方案是直接将 z_q的导数复制到 z_e。

3.2 LDM

有了压缩后的图像潜变量，接下来我们便可以对其进行扩散过程的加噪和去噪了。因为LDM是作用在潜空间，特征的大小要比图像空间小很多，因此LDM的推理速度是要快很多的。并且LDM作用的潜空间是一个具有实际意义的视觉码本，这使得LDM的扩散模型部分更侧重于生成图像的语义信息而非图像的纹理细节。

扩散模型可以理解为一个**时序去噪自编码器**，它的目标是根据输入图像 x在 t时刻的加噪图像 x_t预测一个在其上添加的噪声，则DM的目标函数可以表示为式(1)。其中 t是在序列 $\{1, 2, \cdots, T\}$上的均匀采样。 $\mathcal{L}_{\mathrm{DM}}=\mathbb{E}_{x, \epsilon \sim \mathcal{N}(0,1), t}\left[\|\epsilon-\epsilon_\theta(x_t, t)\|_2^2\right]$而在LDM中，我们是在潜空间中学习，也就是预测在 z_t上添加的噪声，对应的损失函数表示为式(2)。 $\mathcal{L}_{LDM}:=\mathbb{E}_{\mathcal{E}(x), \epsilon \sim \mathcal{N}(0,1), t}\left[\|\epsilon-\epsilon_\theta\left(z_t, t\right)\|_2^2\right]$

3.3 条件机制

类似于其它条件模型，我们可以在LDM中添加不同的条件来让模型生成我们定制的内容。具体来说，我们通过在U-Net引入交叉注意力（cross attention）来引入条件。简单来讲，当注意力的机制的Q，K，V都是使用同一个特征计算来时，这个注意力叫做自注意力；而当Q，K，V来自于不同的数据源时，这个注意力叫做交叉注意力。

在LDM中，我们使用一个领域编码器 $\tau_\theta(y) \in \mathbb{R}^{M \times d_\tau}$将不同模态的条件转化成一个特征向量，例如可以用预训练的BERT转换文本，使用CLIP转换图像等。接着我们使用交叉注意力将条件融入到U-Net的中间层，Q，K，V的实现方式如式(3)。
$Q = W_Q^{(i)} \cdot \varphi_i\left(z_t\right), K = W_K^{(i)} \cdot \tau_\theta(y), V = W_V^{(i)} \cdot \tau_\theta(y)$其中 $\varphi_i(z_t) \in \mathbb{R}^{N \times d_c^i}$是U-Net的一个中间特征，最终加入了条件的LDM的损失函数表示为式(4)。

$$\mathcal{L}_{LDM} := \mathbb{E}_{\mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1), t}\left[\left\|\epsilon - \epsilon_\theta\left(z_t, t, \tau_\theta(y)\right)\right\|_2^2\right] \quad (4)$$

结合了交叉注意力的U-Net的详细结构如图6[7]所示，它相当于在U-Net的降采样部门和上采样部分的残差块之后都添加了一个交叉注意力。

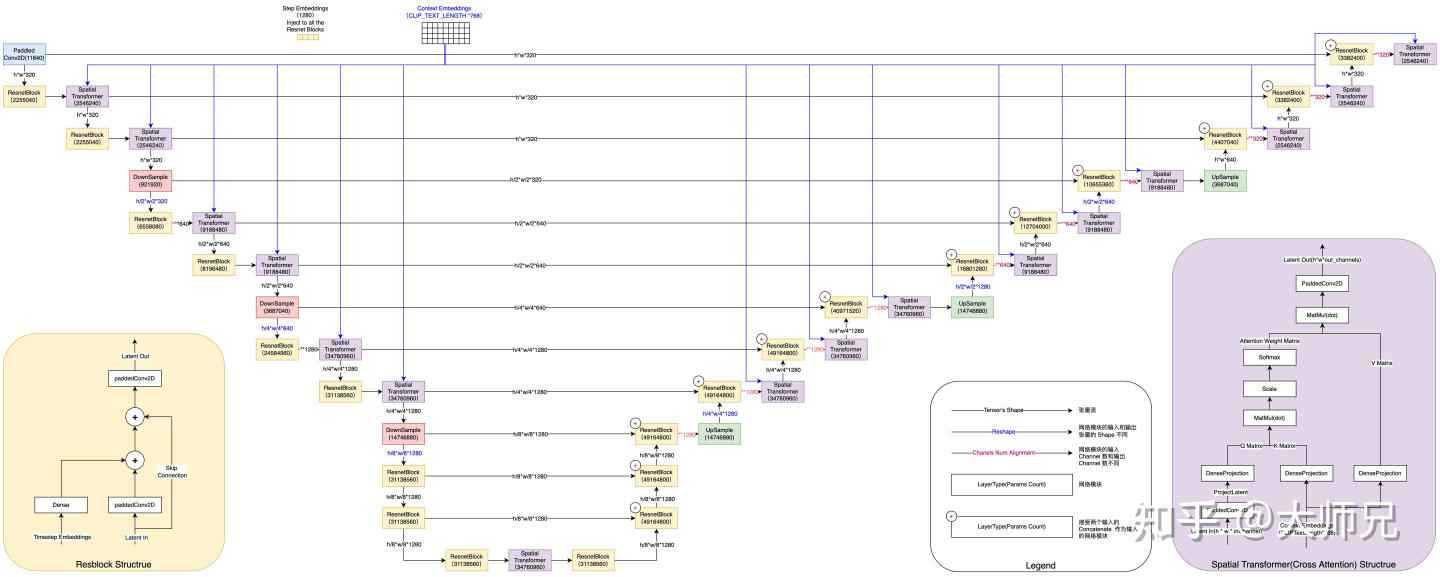


图6：加入了交叉注意力的U-Net

4. 总结

最近备受关注的视频生成模型Sora从技术角度讲可以看做一个视频版本的Stable Diffusion，它们最核心的理论基础之一便是这里介绍的LDM。LDM最重要的改进是将扩散过程从图像空间转移到了潜空间，使得LDM的计算更加高效，从而使得其可以生成更大分辨率的图像。

参考

1. ^Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

2. ^Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in Neural Information Processing Systems* 33 (2020): 6840-6851.
3. ^Song, Jiaming, Chenlin Meng, and Stefano Ermon. "Denoising diffusion implicit models." *arXiv preprint arXiv:2010.02502* (2020).
4. ^Van Den Oord A, Vinyals O. Neural discrete representation learning[J]. Advances in neural information processing systems, 2017, 30.
5. ^Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 12873-12883.
6. ^Johnson, Justin, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution." *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer International Publishing, 2016.
7. ^https://zhuanlan.zhihu.com/p/582266032