计算机科学
COMPUTER SCIENCE

基于潜在注意力的高性能视频超分辨率技术

王宇骥, 董昊呈, 龚雪鸾, 陈艳姣

引用本文

王宇骥, 董昊呈, 龚雪鸾, 陈艳姣. 基于潜在注意力的高性能视频超分辨率技术[J]. 计算机科学, 2023, 50(11A): 221100156-10.

WANG Yuji, DONG Haocheng, GONG Xueluan, CHEN Yanjiao. Efficient Video Super-Resolution with Latent Attention [J]. Computer Science, 2023, 50(11A): 221100156-10.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221100156

基于潜在注意力的高性能视频超分辨率技术

王宇骥[1]　董昊呈[1]　龚雪鸾[2]　陈艳姣[3]
1 武汉大学国家网络安全学院　武汉 430070
2 武汉大学计算机学院　武汉 430070
3 浙江大学电气工程学院　杭州 310058
(2020302181008@whu.edu.cn)

摘　要　为了解决视频超分辨率的问题,可以对视频中的时空相关性信息加以利用,这是将低分辨率视频重建为高分辨率视频的一种行之有效的方法。之前的相关工作主要集中在利用运动补偿来捕捉视频生成中的时间依赖性,这种阶段性重建策略是低效的。相比运动补偿,注意力模型更能在寻找时空相关性中发挥作用。为了使注意力模型可以被应用于视频超分辨率问题,利用基于摊销变分推理的注意力估计构建潜在注意力模型,并设计了长程注意力模块和短程注意力模块两个有效的注意力功能模块。在此基础上构建出一个新型深度学习网络模型,它可以有效地捕捉视频超分辨率的时空相关性,并允许端到端学习。通过在公共视频数据集的广泛实验,可以证明该方法相比当前最先进的几种方法如SPMC,DUF-16L等具有更优越的性能。

关键词:超分辨率;深度学习;潜在注意力;变分推理;高性能
中图法分类号　TP391.41

Efficient Video Super-Resolution with Latent Attention

WANG Yuji[1],DONG Haocheng[1],GONG Xueluan[2] and CHEN Yanjiao[3]
1 School of Cyber Science and Engineering,Wuhan University,Wuhan 430070,China
2 School of Computer Science,Wuhan University,Wuhan 430070,China
3 College of Electrical Engineering,Zhejiang University,Hangzhou 310058,China

Abstract　To solve the problem of video super-resolution,the spatio-temporal correlation information in videos can be utilized,which is an effective method for reconstructing low resolution videos into high-resolution videos. Prior works mainly focus on utilizing motion compensation to capture temporal dependency in video generation,leading to inefficient stage-wise modeling strategies. Compared to motion compensation,attention model is more efficient in the search for spatio-temporal correlation. In this paper,we formulate a latent attention model for attention estimation with amortized variational inference and instantiate two effective attention modules for video super-resolution. Based on it,a novel deep network model,which can capture spatio-temporal correlations efficiently for video super-resolution and admit end-to-end learning,is presented. Extensive experiments on public video datasets demonstrate the superior performance of our approach over several state-of-the-art methods like SPMC,DUF-16L.

Keywords　Super-resolution,Deep learning,Latent attention,Variational inference,Efficient

1　引言

近年来,深度学习模型因表现卓越而被广泛应用于解决图像或视频的上采样问题[1-4]。关于视频超分辨率技术,先前的大多数工作[5-9]分为两阶段:先计算运动估计和运动补偿,然后执行上采样操作。这些方法依赖于准确的运动估计,通常被表述为一个优化问题——我们认为相邻帧运动的变化较小,即相邻帧之间有较强相关性,所以在预测某一帧时,将其相邻帧的信息作为输入。

尽管上述基于运动补偿的方法取得了一定的效果,但它们仍然存在几个缺陷。首先,预测的子像素位移可能是不准确的。在以下情况下,该问题尤为明显:1)物体的位移较大;2)在不同目标函数下分别进行预测,最终的预测值取均值而被平滑化。类似地,由于对像素值取平均,因此,在一组经过运动补偿的帧上进行操作也会导致帧图像模糊的问题。其次,当需要对较长序列进行重建时,这类模型的运行效率明显较低。它们接收相邻的多个帧而仅预测一个帧,这种朴素的滑动窗口思想的计算成本非常高。基于此,一些工作[10-11]采用递归的框架,虽然优于滑动窗口的方法,但是这类因果系统仅限于逐帧处理给定的视频。

为了规避以上问题,我们设计了一种注意力估计策略,提出一种运用潜在注意力的超分辨率方法,该方法可以在图像和视频任务中捕捉到时空相关性信息。我们的工作为以下问题提供了较好的解决方案:1)如何准确估计注意力? 2)如何在设计中高性能地估计注意力?

针对第一个问题,我们建立了一个概率模型,将视频注意力作为一个潜变量,用变分推理来解决这一问题。这个方法有两个优势:1)注意力的估计更加准确、更加多样化,因为概率模型将注意力设定在一个区域而不是一个单一的点,这避免了运动估计被平滑化,我们提出的概率模型将按照特定分

通信作者:龚雪鸾(xueluangong@whu.edu.cn)

布产生潜变量的平均值和方差。2)作为模型潜变量,注意力有助于以概率性原则的方式推理出具有可解释性的依赖关系[12]。针对第二个问题,我们提出了两种注意力功能块的设计,专门用于捕捉视频中的时空相关性,其灵感来自于文献[13-14]。这两个功能块被结合起来形成一个强大的注意力估计器,该模块能够兼顾邻近范围或大范围区域内的注意力。

本文提出了一个潜在注意力模型,它能够有效地捕捉视频超分辨率中的时空相关性,同时具有较高的性能。此处的潜在注意力指先从低分辨率视频学习到注意力知识,再一起作为模型的输入以完成重建。这个模型具有以下 3 个优点:1)很好地实现了实时性,因为我们用潜在注意力模型取代了运动估计,并且帧序列可以被并行处理,没有任何滑动窗口或自回归行为;2)通过变分推理以及引入噪声分布和蒙特卡洛梯度估计器,所提出的概率模型是可微分的,因此可以进行端到端的联合训练,而无需像 VESPCN[5] 和 SPMC[6] 那样预先训练特定的功能块;3)所提出的注意力功能块可以通过残差连接实现即插即用,使其适用于视频超分辨率以外的所有视觉任务。

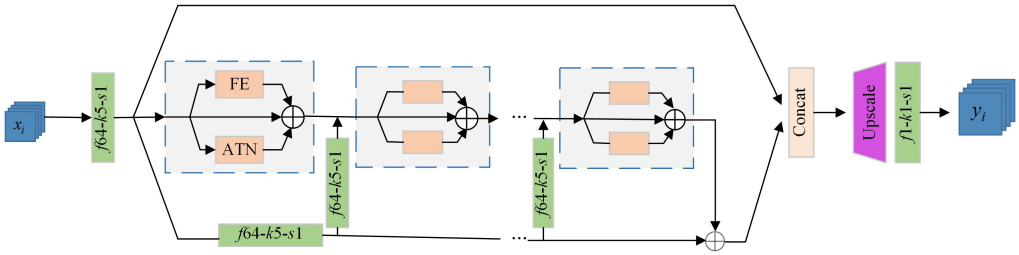
本文提出的视频超分辨率网络在公共视频数据集上进行了广泛的评估,取得了非常好的效果。

本文的主要工作及贡献有如下 3 个方面:

1)提出了一个用于视频超分辨率的深度网络模型,该模型是可以端到端训练的,并能有效且高性能地捕获时空相关性。实验表明,相比于其他工作,所提模型在运用注意力模块的情况下生成的序列视觉上较为清晰,同时 PSNR(Peak Signal to Noise Ratio,峰值信噪比)指标能达到最先进的效果。最重要的优点是,所提模型的效率比其他模型有了明显的提升,运行时间大大缩短。这使得我们的模型具有更好的实用价值。

2)建立了一个潜在注意力模型,利用摊销变分推理,实现了准确和多样化的注意力估计。摊销变分推理使得模型参数是可微分的,也使得模型可以进行端到端的训练。

3)提出了一个有效的视频注意力模块和两个功能块,即长程注意力和短程注意力。长程注意力能够发现全时空范围内帧序列的相关性,短程注意力能够发现相邻帧之间的相关性,两者形成互补的关系。



注:FE,ATN 和 Upscale 分别代表特征提取器、注意力估计器和空间升频器,这是 3 个主要模块。Concat 表示连接操作,绿色块表示卷积层,标签 f-k-s 表示特征通道、核尺寸和步长。⊕表示残差连接的加法运算。

图 1 本文提出的视频超分辨率模型的概览

Fig. 1 Overview of the proposed model for video super resolution

2 相关工作

在视觉领域的早期研究中,超分辨率就引起了人们的关注,有很多关于图像和视频超分辨率的文献。由于内容的限制,这里我们主要关注基于深度学习的方法。

2.1 单图像超分辨率

在图像高分辨率研究领域,之前的研究大多集中在利用各种图像先验,如空间相关性、相似图像的模式或学习到的图像正则化,来提升静态图像的空间分辨率。

受深度网络在语义视觉中的成功的启发,SRCNN[1] 首次将卷积神经网络用于图像超分辨率。后续的工作采用残差网络[15-17] 和递归网络[18-19],这使他们能够采用更深的网络结构和更多的参数共享来提高生成图像的质量。为了降低插值效应,一些方法在低分辨率图像空间引入了增强特征表示学习[20-22]。最新的一些方法[23-24] 利用记忆网络来达到更好的图像特征表示。本文以残差网络结构为基础来捕捉特征,特别是空间相关性,来完成图像的重建。

2.2 视频超分辨率

不同于图像领域,视频超分辨率有一个额外的时间维度,这为我们提供了可以利用的数据冗余。为了合成时间上一致的高分辨率帧,最新的工作找到了提取除上述空间相关性之外的时间相关性的方法。

一种常见的提取时间相关性的策略是预测光流并利用它来计算帧融合的运动补偿。其中,VSRnet[7] 首次引入了基于 CNN 的视频超分辨率模型,但它在插值时产生了昂贵的计算成本。VESPCN[5] 将空间变换网络用于运动补偿,并将亚像素卷积层[20] 用于实现高性能的超分辨率。RobustVSR[8] 也用于学习并补偿帧之间的运动,它在不同的时间半径下学习,综合产生输出。SPMC[6] 对 VESPCN[5] 提出的运动补偿变换进行了调整,增加了亚像素卷积的同步上采样这一步骤。补偿后的帧被送入一个中间有 Conv-LSTM[25] 的编码器-解码器结构中。

以上所有模型都遵循先运动补偿然后上采样的工作流程,在这过程中,一个时间半径的帧序列只能完成一个中心帧的重建,即滑动窗口方法。时间半径越大,运动估计越准确。所以这些模型不可避免地要在效率和性能之间进行权衡,也就是说,为了获得更好的精度,我们需要设置更大的时间半径,这会导致昂贵的计算成本。然而,本文方法没有直接建立运动补偿模型,避免了上述问题。

FRVSR[26] 采用递归网络来缓解这一系列的运动补偿方法存在的问题,该网络在最新的预测上估计相邻两帧之间的运动,以此来产生当前的预测。此外,文献[25]也提出了一个递归网络,通过 3 种类型的卷积来捕捉时空相关性。递归网络具有自回归行为,它将最新的输出作为当前的输入,使帧序列的预测严格按照顺序进行。相比之下,我们的模型是可以并行的,因为我们通过注意力捕捉时间上的相关性,并保持网络的前馈性。

2.3 注意力模型

作为模型的一部分,注意力使得模型能够关注输入数据的特定部分。它常常被用于从序列到序列的问题,如机器翻译[27]和图像描述生成[28]。文献[29]表示,利用自注意力可以获得最先进的神经网络机器翻译(Neural Machine Translation,2NMT)效果——它关注整个序列来综合考虑某一位置的结果。文献[13]的作者将自注意力应用到视觉任务上,并达到了预期的效果。受到这一成果的启发,我们为视频任务制定了注意力模型,并提出了两种合适的功能块。

3 基础知识

在这一节中,我们简要地介绍了 VAE 和注意力机制,这是本文提出的模型的基础。

3.1 VAE

Variational Auto-Encoder(VAE)是一类生成模型,它在自动编码器的结构中执行摊销变分推理。VAE 做出这样的假设:一个潜变量 z 可以表示观察到的数据 y,即从 z 可以解码出 y。为了预估 z 的分布,它进一步假设 z 依赖于 y,也就是说,从 y 可以对 z 进行编码。

我们将观测到的数据记为 $\hat{y}$,$\hat{y}$ 是一个和 y 同分布的样本,它被生成用于近似估计 y。核心问题是推断后验概率 $p(z|y)$。在给定一个数据集 $\hat{y}_{1\cdots n}$ 的情况下,后验概率可用于产生 z,进一步将 z 编码为 $\tilde{y}=\hat{y}$。

VAE 采用摊销变分推理法,通过优化来解决后验推理问题。它首先建立一个变分分布 $q(z;v)$,v 表示待优化的变分参数集。然后,VAE 在变分参数上进行搜索,找到最接近 $p(z|y)$ 的 $q(z;v*)$,其中 $v*$ 是优化后的变分参数集。由于 v 是由 y 产生的,即 $v=a_\phi(y)$,其中 $a_\phi(\cdot)$ 是一个可以优化的参数化函数,所以符号 $q(z;v)$ 实际上等同于 $q_\phi(z|y)$。

推理问题就转化为最小化下面这个式子:

$$KL(q_\varphi(z|y)\parallel p(z|y))=E[\log p_\theta(z|y)-E[\log p_\theta(z|y)\ p(z)]+\log p(y)$$

其中,$KL(\cdot|\cdot)$ 表示 KL 散度,用于衡量两种分布之间的差异。上面的公式可以改写为:

$$ELBO(\theta,\phi)=-E[\log q_\varphi\phi z|y)]+E[\log p_\theta(y|z)p(z)] \tag{1}$$

$$\log p(y)=KL(q_\phi(z|y)\parallel p(z|y))+ELBO(\theta,\phi) \tag{2}$$

$$KL(q_\phi(z|y)\parallel p(z|y))=\log p(y)-ELBO(\theta,\phi) \tag{3}$$

上述公式中的 ELBO 在文献中被称作置信下界,它是置信度的对数形式 $\log p(y)$ 的下界,正如式(1)所示。当 $q_\varphi(z|y)=p(z|y)$ 时,这一约束是严格的。

解决推理问题,即优化 $q(z;v)$ 使得 KL 散度最小化,也就是最大化 ELBO,因为 $p(y)$ 与 $q(z;v)$ 无关。另外,最大化观测数据的对数似然即 $\log p(y)$,可以通过最大化 ELBO 来实现,即最小化 KL 散度,如式(2)和式(3)所示。

由此,为了解出 q_φ 和 p_θ,我们只需要最大化 ELBO。

$$\log p(y=\hat{y})\geqslant E[\log q_\phi(z|\hat{y})]+E[\log p_\theta(y=\hat{y}|z)p(z)]$$

传统的变分推理会用每个观测值的参数对 q_ϕ 作出调整,例如 $z_i\sim N(\mu_i,\sigma_i)$,$v=\{\mu_i,\sigma_i\}_n$,并用坐标下降法优化目标,这与深度学习模型是不兼容的。而摊销变分推理用神经网络对分布进行参数化,例如,$z_i\sim N(\mu_\phi(y_i),\sigma_\phi(y_i))$。基于此,VAE 可以通过 ELBO 进行微分优化。

VAE 在训练中的生成过程是:编码器 q_ϕ 用给定的 $\hat{y}$ 对 z 进行编码,解码器 p_θ 将 z 解码成 $\tilde{y}$,我们希望 $\tilde{y}$ 和 $\hat{y}$ 是相同的。

然而,在许多应用中,我们关心的是将数据从源域变换到目标域,例如将视频从低分辨率转变成高分辨率。本文中,我们通过对 x 的全分布进行调节来扩展 VAE,并将潜变量 z 设置为注意力。更多细节将在 4.1 节中展示。

3.2 注意力机制

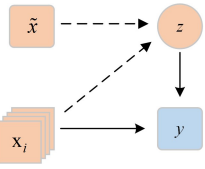
注意力机制[30]是受到生物感知过程的启发而产生的,使我们能够整合来自多个来源的相关信息。它被应用在许多视觉和语言任务上,例如神经网络机器翻译(NMT)[27,31]、视频字幕[32]和图像分类[33]。在自然语言处理领域,在描述注意力机制时,使用键-查询-值(key-query-value)的术语。图 2 是文献[29]中使用的一个典型的方法,它被用于计算源序列和目标序列之间的对应关系。

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}})\boldsymbol{V} \tag{4}$$

其中,$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 分别表示查询、键和值。查询代表目标序列的位置,键和值是源序列中的位置的表示。首先,我们计算出查询和各个键之间的相关性,并标准化以形成源序列整体位置的分布,其中每个位置作为值的权重系数。

上述的加权总和被称为软注意力,因为值的权重是平滑变化的。相比之下,硬注意力是从分布中获取单一样本,它在训练上更具挑战性。

本文提出了一个视频注意力的通用框架,它从上述形式中演变而来。



注:给定来自视频集的键、值和查询,对应高分辨率框架可以通过注意力来产生。虚线和直线分别代表编码和解码过程。

图 2 视频超分辨率的潜在注意力模型

Fig. 2 Latent attention model for video super-resolution

4 模型结构设计

本节首先介绍视频超分辨率的问题,然后针对我们的特定模型的学习目标进行实例化,最后讨论神经网络的结构。

4.1 视频超分辨率的潜在注意力

本文的目标是从低分辨率版本的视频序列 $x=\{x_i\}_L$ 来重建一个长度为 L 的高分辨率视频序列 $\tilde{y}=\{y_i\}_L$,它看起来和真实序列 $\hat{y}=\{y_i\}_L$ 一样可信。我们通过对尺寸为 $sH\times sW\times C$ 的序列执行下采样操作获得尺寸为 $H\times W\times C$ 的低分辨率序列,其中 s 表示的是缩放系数,H 代表高度,W 代表宽度,C 代表图像通道。在本次实验中,在不失视频一般性的情况下,我们用双三次插值法来获取低分辨率序列,并且设定 $s=4$,即原序列的尺寸会被缩小 4×4 倍。本文致力于利用丰富的时空相关性信息来恢复视频破损版本的分辨率,并不关注从时间维度上提高分辨率的问题。

为了提取视频中的时空相关性以重建高分辨率视频,我们利用注意力的思想,以有效模拟实体间的成对关系。我们希望学习到这样一种注意力机制——在重建某一帧区域时,自适应地聚合时空范围内其他帧区域的相关特征。然而,这

种注意力的显式监督学习是难以实现的，因为我们几乎无法确定哪些区域与需要重建的目标区域相关。

先前的大多数工作都是通过优化重建目标来模糊地学习这种注意力模型，现在我们从概率的角度提供了一个替代方案。我们将时空相关性的标注，或者说“注意力”，作为潜变量 z，并且在给定解释性数据 x 的情况下学习推断 z，因为注意力没有相应的标签，不能直接观察，因此一个很自然的想法是将注意力作为潜变量来确定其后验分布。从形式上看，在引入 z 后，原始目标可以写成如下式子，并推导出一个代理损失函数 $Q(x,y)$，它是观测数据置信度的变分边界。

$$\begin{aligned}\log p(y|x)&=\log\sum_z p(y,z|x)\geqslant Q(x,y)\\&=E_{q_\phi(z|x,y)}[\log p_\theta(y|z,x)]-\\&\quad KL[q_\phi(z|x,y)\parallel p_\omega(z|x)]\end{aligned}\tag{5}$$

其中，θ,ϕ,ω 都是神经网络函数逼近器中可学习的参数。这里我们引入后验概率 $q_\phi(z|x,y)$ 来逼近未知的先验概率 $p(z|x,y)$，该先验概率模拟了视频时空范围内区域之间的相关性。$p_\omega(z|x)$ 作为一个信息瓶颈，可使 $q_\phi(z|x,y)$ 规范化。

上述公式与 3.1 节提到的 VAE 相似，不同点在于这里所有的参数都以输入的低分辨率序列 x 为条件。

然而，式(5)将 y 作为 $q(z|x,y)$ 中的一个条件，这在测试中是不可行的。参考文献[34]，我们通过假设 $y=y(x)$ 来消除推理网络中的 y，即 $q(z|x,y)=q(z|x,y(x))\doteq q(z|x)$。此外，我们令 $q_\phi(z|x,y)=p_\omega(z|x)$，就好像我们在推理网络和编码器网络之间共享参数一样。这种简化有两个好处：1)不需要训练一个不同的网络 $p_\omega(z|x)$，该网络在测试过程中不使用；2)变分约束被简化为第一项，而不必担心高维向量上的 KL 项。

因为 VSR 是不适定问题，从高分辨率降为低分辨率的过程中，信息有损失，所以 $y=y(x)$ 不成立。尽管如此，我们根据经验发现，这一近似仍然取得了良好的效果。

式(5)给出了最终损失函数，它可以进一步产生一个简洁的式子，对应于常用的 L1 和 L2 范数重构损失。

$$\begin{aligned}Q(x,y)&\approx E_{q_\phi(z|x)}[\log p_\theta(y|z,x)]\\&=\sum_{(x_i,y_i)\in(x,y)}\log p(y_i,x_i)\\&=\sum_{(x_i,y_i)\in(x,y)}\alpha\parallel y_i-f_\varepsilon(x_i;\theta,\phi)\parallel\end{aligned}\tag{6}$$

其中，α 是某个常量，$f_\varepsilon(\cdot)$ 表示整个模型，参数 $\varepsilon\sim N(0,1)$ 引入了随机性。如果假定为拉普拉斯分布，对数概率可以简化为 L1 范数，如果假定为对角高斯分布，则为 L2 范数。

在变分自动编码器中，式(6)以编码-解码的方式工作：q_ϕ 将低分辨率序列 x 映射到辅助上下文特征 z 中，该特征包含了区域之间的相关性。x 与 z 共同生成了高分辨率序列 y。

式(6)中的 ϕ 出现在期望值上。为了计算式(6)的梯度，我们通过隐含地或明确地引入一个随机性的辅助变量来实施重参数技巧，具体细节见 4.2 节。

作为经典的策略梯度算法，REINFORCE 算法(Reward Increment = Nonnegative Factor times Offset Reinforcement times Characteristic Eligibility)[35]为我们提供了损失函数梯度的通用形式，而不需要任何进一步的假设。在设计损失函数时，我们从该算法中获得了启发。

$$\begin{aligned}\nabla_\phi Q(x,y;\theta,\phi)&=E_{q_\phi}\log[p_\theta(y|z,x)\nabla_\varphi\log q_\phi(z|x)]\\&=-\alpha E_{q_\phi}[\parallel y-f_\varepsilon(x_i;\theta,\phi)\parallel\nabla_\phi\log q_\phi(z|x)]\end{aligned}\tag{7}$$

这表明，优化操作促进了 q_ϕ 将更多的概率质量放在潜在的上下文特征(即“注意力”)所在空间中的区域，这能够减少重构误差。这个逻辑符合我们的直觉，在理论上保证了我们的公式有可能获得准确的注意力。

4.2 注意力机制的设计

在上一节中，我们展示了注意力是如何依赖于低分辨率视频输入以及如何和输入视频序列一起生成高分辨率视频。在这一节中，我们讨论在给定输入的情况下，注意力应该采取何种具体的数学形式。我们提出了一个通用的公式：

$$z=h(t(\tilde{\boldsymbol{x}}',\boldsymbol{x}'),g(\boldsymbol{x}'))\tag{8}$$

其中，$\boldsymbol{x}'$ 和 $\tilde{\boldsymbol{x}}'$ 分别表示源序列和目标序列。采用键-查询-值的术语，这里 $\boldsymbol{x}'$ 是键，$\tilde{\boldsymbol{x}}'$ 是查询，$g(\boldsymbol{x}')$ 是键的转换后的版本。函数 $t(\cdot|\cdot)$ 计算键和查询之间的相关性，$h(\cdot|\cdot)$ 根据键和查询之间的相关性来决定值的效果。

3.2 节中提到的注意力形式可以被看作是上述公式的实例化，令 $\boldsymbol{x}'=rep^{\text{src}}$，$\tilde{\boldsymbol{x}}'=rep^{\text{des}}$，$t(\boldsymbol{x}',\boldsymbol{y}')=\text{softmax}(\boldsymbol{x}'\boldsymbol{y}'^{\mathrm{T}})$，$g(\boldsymbol{x}')=\boldsymbol{x}'$，$h(\boldsymbol{x}',\boldsymbol{y}')=\boldsymbol{x}'\boldsymbol{y}'$，$\boldsymbol{y}'$ 是形式参数，rep^{src} 和 rep^{des} 代表每个源序列和目标序列的隐含观测量，典型的尺寸是 $L\times C$。

而在视频注意力的情境下，注意力的表征往往是四维的，即 $L\times H\times W\times C$。我们想要计算时空相似性，这是更加复杂的。

我们为视频注意力设计了两个功能块，它们的计算复杂性和侧重有所不同。

1)长程注意力。借鉴键-查询-值(K-Q-V)注意力的思想，我们将视频的四维张量变成二维矩阵，即 $LHW\times C$，并以类似的矩阵乘法计算相关性。我们称这种类型的注意力为“长程”，因为它考虑了 $L\times H\times W$ 的完整时空范围内的相关性，相比其他注意力关注更长的范围。

$$t(\boldsymbol{x}',\boldsymbol{y}')=\text{softmax}(\phi(\boldsymbol{x}')^{\mathrm{T}}\theta(\boldsymbol{y}'))\tag{9}$$

$$h(\boldsymbol{x}',\boldsymbol{y}')=\boldsymbol{x}'\boldsymbol{y}'\tag{10}$$

$$g(\boldsymbol{x}')=\omega_\gamma\boldsymbol{x}',\theta(\boldsymbol{y}')=\omega_\alpha\boldsymbol{y}',\phi(\boldsymbol{x}')=\omega_\beta\boldsymbol{x}'\tag{11}$$

为了计算长程注意力，我们首先用不同的权重矩阵对键、查询和值进行线性转换；然后通过矩阵乘法获得键-查询的相关性，接着将乘积作为 softmax 函数的输入；最后用另一个矩阵乘法将相关性应用于值。

$t(\boldsymbol{x}',\boldsymbol{y}')$ 可以视为带有线性嵌入式高斯的标准化欧氏距离。我们参考了文献[13]，最终选择了该函数。

假设 $\boldsymbol{x}'$ 是一个和 $\tilde{\boldsymbol{x}}'$ 来自同一视频但不同帧序列的表示，其尺寸为 $N\times H\times W\times C$。矩阵乘法，例如键和查询之间的乘法，需要 $O(LN(HWC)^2)$ 的复杂度，输出大小为 $LHW\times NHW$，每一个元素可以看作是扁平化的 $\tilde{x}'$ 中第 i 个响应和 $\boldsymbol{x}'$ 中的第 j 个响应之间的相关系数。通过这种矩阵乘法，我们得以进行全范围的相关性计算，尽管其代价是二次复杂度。

对于长程注意力，我们采用 $kernel=3$，$stride=1$ 的卷积层和 $kernel=2$，$stride=2$ 的池化层进行下采样，并且采用 $kernel=1$，$stride=1$ 的卷积层进行线性变换，最后使用 $kernel=4$，$stride=2$ 的去卷积层来恢复空间分辨率。

我们采用了两个技巧来降低计算成本。如图 3 所示：(1)在计算注意力之前，我们在空间分辨率上对输入的张量进行下采样；(2)在计算过程中，我们只对空间分辨率上的键和值进行下采样。这相当于考虑一个更稀疏的注意力版本，其中的相关性是在空间块而不是像素上计算的。

为了和提出的潜在注意力模型联系起来，这里引入 $z=$

$z_{i:n}$，它表示在其他位置的特征 $\boldsymbol{v}'$ 已知的情况下，存储每个时空像素的增强特征的集合，$z_i=\sum_{v'\in V} v'\cdot f(v',v_i)$，其中 $V=f(x)$ 是确切计算的，这似乎与 z 是一个随机变量的假设不一致。然而，$z=f(x;\varepsilon)$，$\varepsilon\in p(\varepsilon)$ 可以看作是通过从训练样本中引入 $x\in\chi(\varepsilon)$ 来应用重参数技巧，因为文献[35]表明，与高斯等其他分布相比，χ 分布的随机性实现了类似的测试可能性。

值得一提的是，从结构设计的角度来看，非局部[13]功能块和长注意力功能块是相似的。它们的区别在于，针对视频处理，我们在注意力功能块中加入了下采样组件以减少开销。从训练过程的角度来看，我们的注意力功能块有一个随机抽样的步骤，这使得模型能够得到有效的训练。

2）短程注意力。为了进一步缓解密集计算，我们提出了一种自上而下和自下而上相结合的掩码注意力，它侧重于计算时空邻域之间的相关性。这个思想是受到文献[14]的启发。

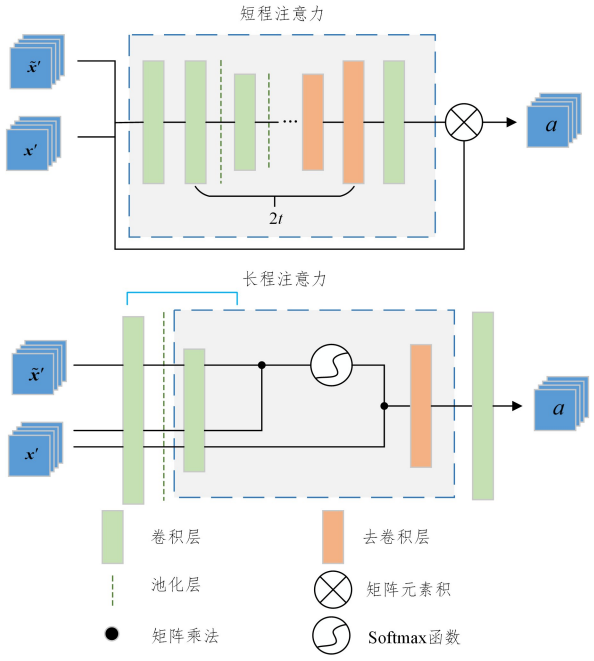
$$t(\boldsymbol{x}',\boldsymbol{y}')=M([\omega_o\boldsymbol{x}',\omega_\beta\boldsymbol{y}'];\theta) \tag{12}$$

$$g(\boldsymbol{x}')=\omega_\gamma\boldsymbol{x}' \tag{13}$$

$$h(\boldsymbol{x}',\boldsymbol{y}')=\boldsymbol{x}'\odot\boldsymbol{y}' \tag{14}$$

其中，$[\cdot,\cdot]$ 表示跨通道连接，$\odot$ 表示矩阵元素积。

我们把这种类型的注意力称为掩码注意力，因为我们首先通过 $M(\cdot\mid\cdot)$ 获取一个基于键 $\boldsymbol{x}'$ 和查询 $\tilde{\boldsymbol{x}}'$ 的张量掩码，并将它按元素应用于线性转换后的值 $\omega_\gamma\boldsymbol{x}'$。



注：x' 表示键和值，$\tilde{x}'$ 表示查询，a 表示注意力功能块的输出。超参数 t 表示短程注意力中卷积或去卷积层的数量。

图 3　两种注意力功能块的结构

Fig. 3　Structure of two attention modules

对于短程注意力，所有的线性变换都是通过 $kernel=1$，$stride=1$ 的卷积实现的。$M(\cdot\mid\cdot)$ 将线性嵌入的键和查询的连接作为输入，将它们送入几个堆叠的 $kernel=3$，$stride=1$ 的卷积和 $kernel=2$，$stride=2$ 的池化层，降低空间分辨率，然后送入堆叠的 $kernel=4$，$stride=2$ 的去卷积层，恢复空间分辨率。我们称其为短程注意力，因为单个功能块的感受野是小于全时空范围的。在我们的实验中，我们为自上而下-自下而上的功能块各取两个堆栈，因此功能块本身就可以在 $5\times5\times5$ 的时空分辨率下检查相关性。但这并不一定意味着短程注意力不能捕捉长距离的关系。当插入到神经网络的深层时，功能块的感受野应加在其上层的感受野之上，这使得实际的感受野可能非常大。

4.3　模型结构

这一小节，我们讨论能够有效表示函数形式的神经网络结构。图 1 给出了本文提出的模型架构概览。

本文模型的输入形式为 $x\in[0,255]^{L\times H\times W\times C}$，输出为 $y\in[0,255]^{L\times sH\times sW\times C}$ 以及潜在注意力 z。本文模型与其他 VSR 相关工作的区别在于，本文模型不涉及任何自回归行为，即所有帧和时空相关性是可以并行计算的。

本文模型的结构分为 3 个主要模块，它们可以并行处理任意大小和长度的输入。

1）特征提取器。特征提取器的目的是从输入中提取出有用的特征用于视频重构，它由一系列三维卷积层构成，每一层可以表示为如下形式：

$$F_l(\boldsymbol{x}^l;\theta_l)=\sigma(W_l * F_{l-1}(\boldsymbol{x}^{l-1}\theta_{l-1})),x^0\doteq x \tag{15}$$

其中，$*$ 表示卷积操作，$\sigma(\cdot\mid\cdot)$ 表示一个非线性函数，W_l 表示第 l 层的卷积核，F_l 表示第 l 层的特征提取器，$\boldsymbol{x}^l$ 表示第 l 层的特征图。在本文的实验中，我们使用 $kernel=3$，$stride=1$ 的卷积层，后面跟一个 leakyRelu 激活函数。从图 4 中可以看到更多的细节。

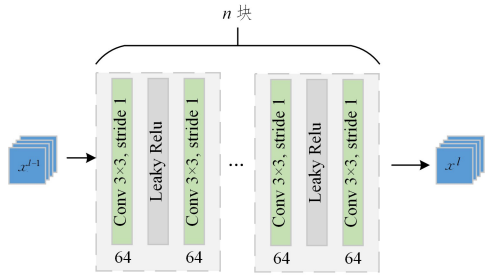


图 4　特征提取器的结构

Fig. 4　Structure of feature extractor

2）注意力估计器。这一模块的目的是提取时空相关性以补充前一个模块获得的框架性特征。长程注意力和短程注意力功能块都可以提取时空相关性作为增强的特征，且我们可以利用以下特点进行选择：长程注意力考虑完整的时空相关性，短程注意力着重于计算邻域的时空相关性。通过所提出的注意力机制，时空相关性被有效地从输入帧序列中提取出来。为了将注意力与特征提取器相结合，我们利用了残差连接。这样一个残差网络可以表示为如下形式：

$$A_l(\boldsymbol{x}^l;\phi_l)=h(t(\tilde{\boldsymbol{x}}^l,\boldsymbol{x}^l),g(\boldsymbol{x}^l)) \tag{16}$$

$$O_l(\boldsymbol{x}^l,\theta^l,\varphi_l)=A_l(\boldsymbol{x}^l;\phi_l)+F_l(\boldsymbol{x}^l;\theta_l) \tag{17}$$

其中，下标 l 表示第 l 层，A 表示注意力的特征图，O 表示一个层的输出。需要注意的是，我们将参数分离为 θ 和 ϕ，只是为了区分特征提取器和注意力估计器，但它们可以共同优化。

残差连接保证了功能块的灵活性，使注意力功能块达到即插即用的效果。也就是说，没有此类模块的模型可以被视为将该功能块的参数设置为 0。当需要启动注意力时，我们要添加连接，使得注意力功能块与预训练的模型共同训练。

$A_l(\boldsymbol{x}^l;\phi_l)$ 能够接受包括提出的实例在内的多种形式。在本文中，我们验证了不同形式下特征图的有效性，并且以特定顺序安排如图 3 所示的两种注意力功能块以测试效果。通过实验我们发现，这样的安排使得两种注意力功能块产生了

互补的作用。

尽管 $\boldsymbol{x}$ 应该是一条视频的完整序列，但鉴于计算资源是有限的，这在具体实验上很难实现。在本文实验中，网络的输入是一个固定长度的帧序列，它是一段视频的连续帧子集。当使用短程注意力时，我们只接受需要处理的帧序列，并且设置查询 $\tilde{\boldsymbol{x}}=\boldsymbol{x}$，因为序列中的相邻帧的相关性能够更好地被该功能块识别。对于长程注意力，我们希望它找到短程注意力所不能达到的较长范围的依赖关系，所以待实现超分辨率的查询 $\tilde{\boldsymbol{x}}$ 需要是输入 $\boldsymbol{x}$ 的一部分。注意，这里的“程”一词包括时间性和空间性。具体来说，对于一个尺寸为 $L\times H\times W\times C$ 的输入帧序列，“程”对于长程注意力来说意味着 $L\times H\times W$，对于短程注意力来说意味着 $5\times5\times5$。换言之，输入的关键帧序列 $\boldsymbol{x}$ 可能是 $\tilde{\boldsymbol{x}}$ 和其他帧序列的组合，只要它们来自同一视频序列即可。这是基于如下的直觉：在恢复高分辨率视频时，我们需要的是细节信息，也可以理解为某种特定的模式，在同一视频中，这样的特定模式往往无处不在，而不仅仅存在于相邻帧中。从这一角度来看，即使参考帧与当前帧有一定的距离，它仍然可以提供一些有用的信息。

为了适用于本文提出的潜在注意力模型，我们引入了多输出结构，用于生成 VAE 中需使用的潜变量 $\boldsymbol{z}$。需要注意，我们利用神经网络生成高斯分布的均值和方差参数，这可用于潜在注意力的采样。

虽然我们可以直接为均值和方差分别建立两个独立的网络，但这种方法会产生大量参数。为了减少参数，我们在两个网络之间进行参数共享，并让这两个网络只在最后几层出现差别。在实验中，我们在上述功能块的基础上增加了两个额外的平行层，一个用于平均数，另一个用于方差，如图 5 所示。采样是通过取一个额外的噪声分布 ε 来完成的，令 $z=\mu(x)+\varepsilon\sigma(x)$ 作为 VAE 中的潜变量。

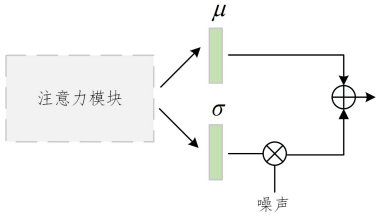


图 5 注意力估计器的多输出结构

Fig. 5 Two-head structure of attention estimator

在后面的章节，我们将用于生成均值参数的层称为均值层，生成方差参数的层称为方差层。

综上所述，注意力估计器有以下特点：

(1)利用注意力功能块可以并行地处理帧序列。

(2)通过两种类型的注意力获得时空相关性，这些注意力在多个分辨率下进行。

(3)注意力功能块是可以端到端训练的。

3)空间升频器。之前的所有操作都是在低空间分辨率下进行的。该模块执行以下操作，将空间分辨率提高一个系数 r。

$$\boldsymbol{x}^1=w*\boldsymbol{x} \tag{18}$$

$$\boldsymbol{x}^r=F_{\text{up}}(\boldsymbol{x}^1) \tag{19}$$

函数 F_{up} 可以被表示为如下形式：

$$F_{\text{up}}(x^1)_{i,j,k}=\boldsymbol{x}^1_{s_i+k \bmod s,s_j+(\frac{\cdot}{\cdot}) \bmod s,\frac{\cdot}{\cdot}} \tag{20}$$

$$F_{\text{up}}:R^{L\times H\times W\times r^2C}\rightarrow R^{L\times rH\times rW\times C} \tag{21}$$

5 模型训练过程

在这一节中，我们会总结有效学习所需的损失函数，并给出一个算法来描述学习过程。

5.1 序列重建损失

由于我们旨在重建一个与真实值尽可能相似的高分辨率序列，因此首要的目标是使生成的序列与真实值之间的欧氏距离最小。

$$L_{\text{re}}=E_x[\,\|\hat{y}-\tilde{y}\|_{L_2}\,] \tag{22}$$

其中，$\hat{y}$ 是 x 对应的的真实值，$\tilde{y}$ 是生成的高分辨率序列。

5.2 VAE 损失

正如 4.1 节所提到的，本文中变分损失被简化为：

$$\begin{aligned}L_{\text{vae}}&=E_{z\sim q_\phi(z|x)}[\log p_\theta(y|z,x)]\\&=\sum_{(x_i,y_i)\in(x,y)}\alpha\,\|y_i-f_\theta(x_i,\mu_\varphi(x_i)+\varepsilon\sigma_\theta(x_i)\|\end{aligned} \tag{23}$$

其中，$\varepsilon\sim N(0,1)$，f_θ 根据指定的分布直接给出概率密度，本文指定为高斯分布。

上述损失函数可以直接对 θ 和 φ 进行微分，由此得到的梯度估计是一个无偏估计。为了近似期望值，我们对每个数据点 $(x_i,\hat{y}_i)$ 使用一个蒙特卡洛样本。

5.3 模型训练

学习过程从计算变分损失开始。为了提高稳定性，我们将学习过程分成两个阶段：1)只打开注意力的均值端，根据 L_{re} 对生成器网络进行预训练；2)引入一个噪声分布并打开注意力的方差端，引入超参数 λ，使用软加权损失 $L_{\text{re}}+\lambda L_{\text{vae}}$ 进行训练。

需要注意的是，注意力是从潜在变量的分布中采样的，如式(23)所示，采样是在计算 L_{re} 的同时进行的。在训练的第二阶段，L_{re} 和 L_{vae} 同时被计算。由于整个训练过程是端到端的，注意力功能块后面的部分只需要传播一次。

具体步骤如算法 1 所示。

算法 1 潜在注意力模型的训练方案

设置学习率为 η，批量大小为 B，VAE 损失的权重为 λ

输入：x(帧序列)，$\hat{y}$(真实值)

输出：潜在注意力模型

1. 关闭注意力的方差端

2. 利用均值端进行预训练

2.1. 获取一个批次的输入：$\{x,\hat{y}\}_B$

2.2. 利用公式 $\nabla_{\theta,\phi}L_{\text{re}}$ 更新参数 θ,φ

2.3. 对以上操作进行 K 次迭代

3. 打开注意力的方差端

4. 引入噪声分布进行训练

4.1. 获取一个批次的输入：$\{x,\hat{y}\}_B$

4.2. 每个注意力功能块在 $\varepsilon\sim N(0,1)$ 进行采样

4.3. 利用公式 $\nabla_{\theta,\phi}L_{re}+\lambda L_{vae}$ 更新参数 θ,ϕ

4.4. 在损失收敛前重复以上操作

6 消融实验

本章进行了几个实验来探讨长程注意力功能块和短程注意力功能块的效果。

6.1 单一注意力功能块的性能

为了测试使用单一注意力功能块的效果，我们使用 Res16 作为基线模型。Res16 是一个 16 层的残差网络，正如 4.3 节中所述，它只包含由一定数量的 3-D 卷积层所堆叠而成的特征提取器。Satn10 是一个 10 层的残差网络，其中插入了 5 个短程注意力功能块；Latn10 也是一个 10 层的残差网络，其中有 5 个长程注意力功能块。为了进行公平的对比，这

些模型已经用相同数量的参数(170 万±2 万)进行了构建，同时 Satn10 与 Latn10 拥有相同的序列长度 4。对于长程注意力，需要取一个长度为 N 的外部帧序列。目前，在长程注意力和短程注意力中都将$\tilde{x}$设置为 x。

如表 1 所列，Satn10 和 Latn10 在 PSNR 指标方面优于基线。二者在“正常”数据集上比基线模型的 PSNR 值高出 0.20 dB，在“远程”数据集上高出 0.70 dB。从最早的 SRCNN 到现在，在测试集 Vid4 上，视频超分辨率的 PSNR 指标改善不超过 3 dB，而最近提出的每一种新方法的 PSNR 指标改善都小于 1 dB。因此，本文提出的方法对于 PSNR 指标的改善是相当显著的。

表 1　插入注意力功能块的性能

Table 1　Performance of plugging in attention modules

数据集	模型	PSNR/dB
正常	Res16	29.14
正常	Satn10	29.43
正常	Latn10	29.34
远程	Res16	27.81
远程	Satn10	28.22
远程	Latn10	28.51

6.2　外部帧序列的性能

在实验中，一段视频将被切割成几个固定长度的帧序列。外部帧序列是指来自同一视频但不同序列的关键帧，在关于“远程”数据集的实验中会用到。由于外部关键帧中包含空间和时间的相关性，它应该会为注意力模块带来更好的效果。

如表 1 所列，本文提出的方法在“远程”数据集上的 PSNR 指标的提升程度要高于“正常”数据集，因为后者只包含具有长距离依赖性的视频，在这些视频上常常很难使用超分辨率技术。

我们认为，Latn10 在“正常”数据集上表现略差于 Satn10 的原因是序列长度太短，长程注意力无法充分发挥作用。

为了证明这一点，我们用不同长度的序列训练 Latn10 和 Satn10，序列长度的范围是[3，11]。这里的记号通过添加后缀来表示输入序列的长度为 n。从图 6 中我们可以看到，随着序列长度的增加，长程注意力拥有更好的性能增长效果。在下面的实验中，我们使用长度为 7 的序列，即图 6 中两条折线的交点。对于一个长度大于 7 的视频序列，我们将整个视频划分成长度为 7 的较短视频序列，最后一段序列的长度若不足 7，则用 0 填充至长度为 7。对应于被填充过的输入帧的输出帧将被丢弃。

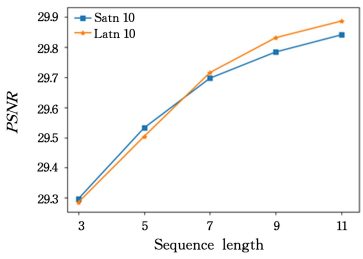


图 6　不同序列长度下的效果

Fig. 6　Performance with different sequence length

我们认为长程注意力能够检测到较远范围内的时空相关性，并通过用外部关键帧对模型性能进行测试来检验这一想法。对于每个视频片段，关键帧是由 ffmpeg 从同一范围内提取的。我们在这里使用的验证数据集是根据以下标准收集得到的：1) 有强烈的运动；2) 有较大的场景变化。结果如表 1 所列，Latn10 的表现明显更好。我们没有重新训练模型，因为 Latn10 学到的规则已经从一个以短距离关联为主的数据集转移到了一个需要更多长程注意力的数据集上。

6.3　两种注意力功能块结合的性能

在本小节中，我们考察将两种注意力功能块结合是否能带来更好的性能，并设计了以下 5 种方案：1)将两种注意力功能块交替放置，表示为 $LSatn_x$；2)将短程注意力功能块放在长程注意力功能块之前，表示为 $LSatn_{sl}$；3)将长程注意力功能块放在短程注意力功能块之前，表示为 $LSatn_{ls}$；4)将短程注意力功能块放在中间，其余为长程注意力功能块，表示为 $LSatn_{lsl}$；5)将长程注意力功能块放在中间，其余为短程注意力功能块，表示为 $LSatn_{sls}$。实验结果如表 2 所列，$LSatn_{sls}$ 效果最好，我们对此提出以下猜想：1)我们的验证集中描述短程相关性的特征比描述长程相关性的特征要多，如果现实情况下只有少数视频确切地需要长程注意力以获得更好的超分辨率，那么这个猜想就是合理的；在大多数情况下，深度神经网络中只设置短程注意力就足够了；2)这个架构类似于一个典型的编码-解码器，在这个架构中间的长程注意力功能块集体处理来自中等大小的感受区的识别特征。

表 2　两种注意力功能块结合后的效果

Table 2　Performance of combining two attention modules

模型	$LSatn_x$	$LSatn_{ls}$	$LSatn_{lsl}$	$LSatn_{sl}$	$LSatn_{sls}$
PSNR/dB	29.59	29.7	29.74	29.95	30.03

7　对比实验

我们使用几个公开的视频数据集来评估基于注意力的超分辨率网络。在本节中，我们通过实验来回答以下问题：1)所提出的基于注意力的实例在提取视频中的时空相关性方面是否有效？2)基于注意力的概率模型能给我们带来什么好处？3)与 VSR 领域的其他工作相比，所提模型能带来哪些好处？下面首先描述实验设置和实施细节。

我们从 Harmonic[38]，CDVL[39] 和 SJTU4K[40] 这 3 个数据集中收集了 230 个高清晰度的视频片段，其中 200 个用于训练，30 个用于验证。每个视频片段包含一个长度为 105 的连续帧序列。我们将这个数据集命名为“正常”数据集。

考虑到一个长的相似帧序列可能会对视频超分辨率产生更高的效益，我们特别构建了另一个数据集，命名为“远程”，它只包含数据集“正常”中具有长距离依赖性的视频，即某一帧可以利用来自远距离帧的信息实现超分辨率。

通过这两个数据集，我们可以检验同时利用短距离和长距离时空关联的注意力功能块的性能。

在我们的实验中，为了获得一个大小为 $L\times H\times W\times C$ 的输入样本，我们从一个视频片段中抽取一个长度为 L 的连续帧序列，用尺寸为 $sH\times sW$ 的相同边框裁剪每一帧，裁剪中坐标是均匀采样的，并通过双三次插值将其下采样为 $H\times W$ 大小。这一预处理过程用 Scipy 实现。

我们用 2 块 1080-Ti GPU 在 Tensorflow 框架中实现所提模型，并通过随机梯度下降法(SGD)来训练网络，在训练中，使用 Adam 算法更新网络参数。使用 Adam 算法的优化器各参数初始化为：学习率 1×10^{-4}，$\beta_1=0.9$，$\beta_2=0.999$。在训练过程中，采用最大全局范数为 3 的梯度裁剪来稳定训练

程序。样本的批量大小被设置为 GPU 容量和模型复杂度约束下的最大值。

除了进行线性变换使用大小 1×1 且不激活的卷积层，所有其他卷积层的卷积核大小为 3，采用 leakyRelu 激活函数。

7.1 使用概率模型的优势

我们使用 L_{vae} 损失，基于预训练的 $Latn10_7$ 和 $Satn10_7$ 进行训练，并将它们表示为 $VLatn10_7$ 和 $VSatn10_7$，相关结果如表 3 所列。和预期一样，采用变分法的确带来了性能的增长，可能的原因是变分损失提供了更准确的注意力。

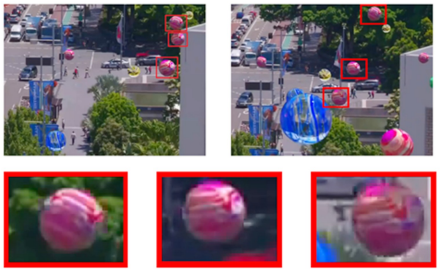
表 3 使用概率模型的效果

Table 3 Performance of employing probabilistic model

模型	$Latn10_7$	$Satn10_7$	$VLatn10_7$	$VSatn10_7$
PSNR/dB	29.72	29.70	29.87	29.95

尽管长程和短程的变分注意力都带来了性能的提升，但短程注意力表现出了 0.1 的额外增长。为了解释这一现象，我们在输入的低分辨率序列上将注意力可视化为热图。对于短程注意力，我们考查一个 15×15 的感受区，如图 3 所示，它是由第二个注意力功能块进行检查的。对于长程注意力，相关性被计算为一个 $LHW\times LHW$ 矩阵，其中第 ijk 行表示整个 $L\times H\times W$ 区域分布上的第 ijk 个区域，i,j,k 分别指代 L，H，W 3 个维度的索引。

图 7 是长程注意力的可视化。左上方的图片是当前帧，图片中红色矩形中的气球代表着查询。右上方的图片是帧序列中的另一帧，它具有可以利用的时空相关性，而该帧中的 3 个气球是 3 个关联性最强的键值对。最下面的 3 个气球是右上方图片中键和值的细节。



注：左上图红色矩形中的气球代表查询，右上图的气球是与键和值对应的关联性最强的 3 个气球。下方的细节是前 3 个最相关的键值对。

图 7 长程注意力可视化(电子版为彩图)

Fig. 7 Visualization of long attention

图 8 显示了短程注意力的多样性，它是在 $VSatn10_7$ 模型中通过采样得到的。在同一实例中，确定性的注意力表现地更加嘈杂，这可能是由于它对可能存在的注意力进行了平均。而概率性的注意力则表现得更加清晰和多样。

(a)原始输入 (b)确定性注意力

(c)概率性注意力

图 8 短程注意力的可视化

Fig. 8 Visualization of short attention

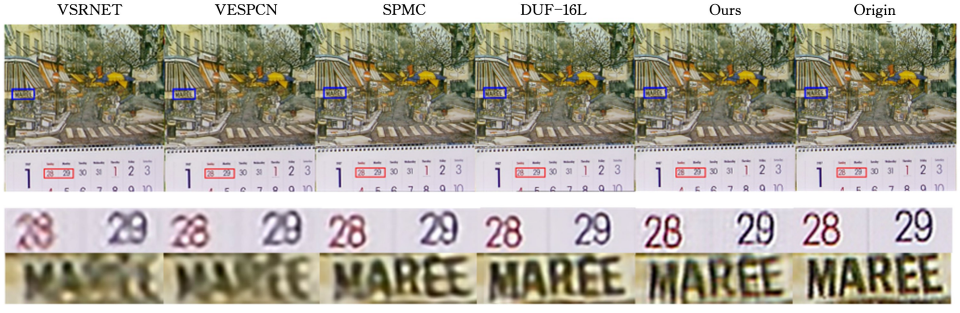
7.2 与其他现有模型相比的优势

本小节从两个方面回答性能和效率的问题。我们使用的测试集在文献[37]中被称为 Vid4，它包含 4 个有代表性的视频序列：日历、城市、树叶和行走。Vid4 数据集已被先前的工作广泛用作通用基准，因此我们选择该数据集来实现公平比较。在数据集上运行本文方法时，为了适应设定的序列长度，在需要时我们会对数据集进行填充。为了与 VSRNET[7] 进行公平的比较，我们删除了第一帧和最后两帧，并在边界处裁剪了 8 个像素。性能评估的结果见表 4 和图 9。结果表明，本文模型在 PSNR 指标方面达到了最先进的视频超分辨率结果。

表 4 与其他超分辨率方法的效果的比较

Table 4 Performance comparison with video SR methods

模型	PSNR/dB
SRCNN[41]	24.68
VSRNET[7]	24.73
VESPCN[5]	25.34
Robust VSR[8]	25.24
SPMC[6]	26.02
FRVSR[26]	26.68
DUF-16L[42]	26.79
Ours	**26.44**



注：本文模型生成的视频质量只比 DUF16L 稍差，但在运行时间上优于 DUF16L，并且在细节上明显比其他带有运动补偿的模型更加清晰。

图 9 本文方法与基线方法在 Vid4 上的定性比较

Fig. 9 Qualitative comparison of our methods with baseline on Vid4

表 5 列出了包括效果最好的工作在内的先前工作的参数数量和每帧的运行时间，其中参数是以数量计算而不是以字

节计算的,运行时间是通过对同一测试集的运行时间进行平均而得到的。VESPCN[5]和 FRVSR[26]没有提供任何代码或预训练的模型,所以无法与之比较。通过分析这两个与最新工作比较的表格,可以得出结论:本文模型在运行速度上获得了巨大的提升,而在效果上只有很小的牺牲,这是从我们的潜在注意力设计中取得的效益。

表 5　与其他超分辨率方法的效率比较

Table 5　Efficiency comparison with video SR methods

模型	参数数量	运行时间/s
SPMC	1722161	0.104
DUF-16L	1900640	0.115
Ours	4108739	**0.063**

除了使用 Res16 作为基线模型外,我们还对更优的基线模型 Res101 进行了一些实验,该模型的网络层数更深。但我们发现,相比原先的模型,注意力功能块在更深的模型中带来的 PSNR 指标的增益更少。而且由于参数数量的增加,运算的效率降低,因此其不一定适用于实时视频处理的场景,同时也会对本文方法的其中一个主要特点造成损害。造成这个现象可能的原因是,当使用更好的基线模型时,由于归纳的偏差,我们的方法带来的效益可能相对有限。

结束语　本文提出了一个用于视频超分辨率的潜在注意力模型,它是一个可微分的模型,并且可以以端到端的方式进行训练。我们的框架是可伸缩的,它可以接受任意大小和长度的视频;并且该模型是可并行的,不涉及自回归行为。正如实验结果所示,所提注意力功能块可以有效且高效地捕捉时空相关性,而且变分的注意力是准确的、敏锐的和多样化的。本文模型以更低的模型复杂度和更高的效率实现了最先进的性能。同时,本文注意力模块可以做到即插即用,它同样适用于其他视觉任务,可以用于提高超分辨率以外的视觉工作的效果。但是,本文方法在较复杂的模型中带来的效果较为有限,我们希望在未来的工作中改进所提方法并探索将其应用于更复杂的模型的可能性。

参 考 文 献

[1] DONG C,LOY C C,HE K,et al. Image super-resolution using deep convolutional networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2015,38(2):295-307.

[2] LIANG M,WANG X. Semantic segmentation model for remote sensing images combining super resolution and domain adaptation[J]. Chinese Journal of Computers,2022,45(12):2619-2636.

[3] HE P H,YU Y,XU C Y. Image super-resolution reconstruction network based on dynamic pyramid and subspace attention [J]. Computer Science,2022,49(S2):423-430.

[4] WU J,YE X J,HUANG F,et al. A review of single image super-resolution reconstruction based on deep learning[J]. Chinese Journal of Electronics,2022,50(9):2265-2294.

[5] CABALLERO J,LEDIG C,AITKEN A,et al. Real-time video super-resolution with spatio-temporal networks and motion compensation[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2017:4778-4787.

[6] TAO X,GAO H,LIAO R,et al. Detail-revealing deep video super-resolution[J]. arXiv:1704.02738. 2017.

[7] KAPPELER A,YOO S,DAI Q,et al. Video super-resolution with convolutional neural networks[J]. IEEE Transactions on Computational Imaging,2016,2(2):109-122.

[8] LIU D,WANG Z,FAN Y,et al. Robust video super-resolution with learned temporal dynamics[C]// IEEE International Conference on Computer Vision. 2017:2507-2515.

[9] FU L H,SUN X W,ZHAO Y,et al. Fast video super-resolution reconstruction method based on motion feature fusion[J]. Pattern Recognition and Artificial Intelligence,2019,32(11):1022-1031.

[10] SHI X J,CHEN Z,WANG H,et al. Convolutional lstm network:A machine learning approach for precipitation nowcasting [C]// Annual Conference on Neural Information Processing Systems. 2015:802-810.

[11] FUOLI D,GU S,TIMOFTE R. Efficient video super-resolution through recurrent latent space propagation[C]// ICCVW. 2019.

[12] DENG Y,KIM Y,CHIU J,et al. Latent alignment and variational attention[C]// Advances in Neural Information Processing Systems. 2018:9712-9724.

[13] WANG X,GIRSHICK R,GUPTA A,et al. Non-local neural networks[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2018:7794-7803.

[14] WANG F,JIANG M,QIAN C,et al. Residual attention network for image classi? cation[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2017:3156-3164.

[15] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2016:770-778.

[16] KIM J,LEE J K,LEE K M. Accurate image super-resolution using very deep convolutional networks[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2016:1646-1654.

[17] LEDIG C,THEIS L,HUSZ'AR F,et al. Photo-realistic single image super-resolution using a generative adversarial network [J]. arXiv:1609.04802,2016.

[18] KIM J,LEE J K,LEE K M. Deeply-recursive convolutional network for image super-resolution[J]. IEEE Conference on Computer Vision and Pattern Recognition,2016:1637-1645.

[19] TAI Y,YANG J,LIU X. Image super-resolution via deep recursive residual network[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[20] SHI W,CABALLERO J,HUSZ′AR F,et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2016:1874-1883.

[21] LIM B,SON S,KIM H,et al. Enhanced deep residual networks for single image super-resolution[C]// IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017.

[22] PARK S J,SON H,CHO S,et al. Srfeat:Single image super-resolution with feature discrimination[C]// European Conference on Computer Vision. 2018:439-455.

[23] TAI Y,YANG J,LIU X,et al. Memnet:A persistent memory network for image restoration[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2017:4539-4547.

[24] ZHANG Y,TIAN Y,KONG Y,et al. Residual dense network for image super-resolution[J]. arXiv:1802.08797,2018.

[25] SAJJADI M S,VEMULAPALLI R,BROWN M. Frame-recurrent video super-resolution[J]. arXiv:1801.04590,2018.

[26] HUANG Y, WANG W, WANG L. Bidirectional recurrent convolutional networks for multi-frame super-resolution[C]// Annual Conference on Neural Information Processing Systems. 2015: 235-243.

[27] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv: 1409.0473, 2014.

[28] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: A neural image caption generator[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3156-3164.

[29] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017: 5998-6008.

[30] MNIH V, HEESS N, GRAVES A, et al. Recurrent models of visual attention[C]// Annual Conference on Neural Information Processing Systems. 2014: 2204-2212.

[31] LUONG M T, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation[J]. arXiv: 1508.04025, 2015.

[32] YAO L, TORABI A, CHO K, et al. Describing videos by exploiting temporal structure[C]// IEEE International Conference on Computer Vision. 2015: 4507-4515.

[33] WANG F, JIANG M, QIAN C, et al. Residual attention network for image classification[J]. arXiv: 1704.06904, 2017.

[34] ZHOU C, NEUBIG G. Morphological in? ection generation with multi-space variational encoder-decoders [C] // CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection. 2017: 58-65.

[35] WILLIAM S, RONALD J. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning[J]. Machine Learning, 1992, 8(3/4): 229-256.

[36] MAKHZANI A, SHLENS J, JAITLY N, et al. Adversarial autoencoders[J]. arXiv: 1511.05644, 2015.

[37] LIU C, SUN D. A bayesian approach to adaptive video super resolution[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2011: 209-216.

[38] HARMONIC I. Free 4K Demo Footage Center[OL]. https://www.harmonicinc.com/free-4k-demo-footage/.

[39] PINSON M H. The consumer digital video library [J]. IEEE Signal Processing Magazine, 2013, 30(4): 172-174.

[40] SONG L, TANG X, ZHANG W, et al. The sjtu 4k video sequence dataset[C]// Quality of Multimedia Experience. IEEE, 2013: 34-35.

[41] DONG C, LOY C C, HE K, et al. Learning a deep convolutional network for image super-resolution[C]// European Conference on Computer Vision. Springer, 2014: 184-199.

[42] JO Y, WUG OH S, KANG J, et al. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3224-3232.

WANG Yuji, born in 2002, undergraduate. His main research interests include deep learning and computer vision.

GONG Xueluan, born in 1996, Ph.D candidate. Her main research interest is network security.