计算机科学
COMPUTER SCIENCE

基于隐式对齐的视频超分辨率模型

王凤玲, 魏爱敏, 庞雄文, 李智, 谢景明

引用本文

王凤玲, 魏爱敏, 庞雄文, 李智, 谢景明. 基于隐式对齐的视频超分辨率模型[J]. 计算机科学, 2025, 52(8): 232-239.

WANG Fengling, WEI Aimin, PANG Xiongwen, LI Zhi, XIE Jingming. Video Super-resolution Model Based on Implicit Alignment [J]. Computer Science, 2025, 52(8): 232-239.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240500069

基于隐式对齐的视频超分辨率模型

王凤玲[1]　魏爱敏[2]　庞雄文[3]　李　智[1]　谢景明[4]
1 华南师范大学人工智能学院　广东 佛山 528000
2 广州番禺职业技术学院建筑工程学院　广州 511483
3 华南师范大学计算机学院　广州 510555
4 广州商学院信息技术与工程学院　广州 511363
(wfl314159@163.com)

摘　要　视频帧之间不仅具有空间相关性,还存在时间相关性。根据低分辨率视频重建高分辨率视频时,可以利用相邻的多帧信息对齐到目标帧,以指导当前帧的恢复。相邻帧之间的对齐一般采用光流指导的可变形卷积进行显式对齐,这种方法克服了可变形卷积的不稳定性,但会影响帧中高频信息的恢复,降低对齐信息的准确性并放大伪影。为解决上述问题,提出了一种基于隐式对齐的视频超分模型 IAVSR(Implicit Alignment Video Super-Resolution)。IAVSR 通过偏移量和原始值将光流编码到特定像素位置,以此计算光流预对齐的信息而不是利用插值函数插值获得,随后利用光流指导的可变形卷积对计算后的预对齐特征进行重对齐,以帮助高频信息的恢复。在双向传播中利用前两帧传播的信息进行对齐来指导当前帧的恢复,并引入残差网络结构,在提高对齐信息准确性的同时避免引入过多的参数。在 REDS4 公开数据集上的实验结果表明,IAVSR 的峰值信噪比(PSNR)比基准模型提高了 0.6dB,且模型训练时的收敛速度提升了 20%。

关键词:视频超分辨率;可变形卷积;重采样;隐式对齐;光流
中图分类号　TP391

Video Super-resolution Model Based on Implicit Alignment

WANG Fengling[1],WEI Aimin[2],PANG Xiongwen[3],LI Zhi[1] and XIE Jingming[4]
1 School of Artificial Intelligence,South China Normal University,Foshan,Guangdong 528000,China
2 School of Architectural Engineering,Guangzhou Panyu Polytechnic College,Guangzhou 511483,China
3 School of Computer Science,South China Normal University,Guangzhou 510555,China
4 School of Information Technology & Engineering,Guangzhou College of Commerce,Guangzhou 511363,China

Abstract　Video contains both intra-frame spatial correlation and inter-frame temporal correlation. When reconstructing high-resolution video from low-resolution video,adjacent multi-frame information can be aligned to guide the current frame recovery. Deformable convolution guided by optical flow is commonly used for explicit frame-by-frame alignment,this method overcomes the instability of deformable convolution,but will affect the recovery of high-frequency information in the frame,reduce the accuracy of the alignment information and magnify artifacts. To address these issues,this paper proposes IAVSR(Implicit Alignment Video Super-Resolution),a video super-resolution model based on implicit alignment. IAVSR encodes optical flow to specific pixel positions using offset and original values,calculating pre-alignment information instead of interpolating. Deformable convolution is used to realign pre-aligned features and recover high-frequency information. Bidirectional propagation uses information from the first two frames to guide current frame recovery,while a residual network structure improves alignment accuracy and avoids excessive parameter introduction. Experimental results on the REDS4 public dataset show that IAVSR achieves 0.6dB higher PSNR value than the benchmark models and improves model convergence speed by 20% during training.

Keywords　Video super resolution,Deformable convolution,Re-sampling,Implicit alignment,Optical flow

1　引言

将低分辨率的视频重建(恢复)为高分辨率的视频的过程,称为视频超分(辨率)。视频超分和图像超分过程存在显著的不同,图像超分只能针对单张图像进行处理,视频超分可以利用视频中帧之间的相关性提供时间和空间信息,改善

到稿日期:2024-05-20　返修日期:2024-09-06
基金项目:2022 年广州市科技局基础与应用基础研究项目(20220101185)
This work was supported by the 2022 Guangzhou Science and Technology Bureau Basic and Application Basic Research Project(20220101185).
通信作者:谢景明(32959247@qq.com)

超分效果。因此，视频超分[1-3]（VSR）模型和图像超分模型之间的差别主要是对时间依赖性的建模。视频本质上是长时间序列图像，目前对长时间序列建模[4-6]分为滑动窗口和循环恢复两种策略。滑动窗口[5]将多帧信息限制在一个窗口内，可利用的信息受窗口尺寸限制，窗口大小的变化会影响模型性能；循环恢复分为单向传播和双向传播。针对滑动窗口尺寸影响性能的问题，Isobe 等[7]提出了类似时间传播的单向传播，从第一帧开始，逐步向后传播特征，后续帧可以利用前一帧传播的信息（前一个时间步的信息）。但是单向传播会导致不同帧接收到的信息不均衡，最后一帧将获得整个视频的信息，而第一帧只能获得与自身相关的信息，从而影响模型的性能。BasicVSR[6]采用允许特征在时间上独立地向前和向后传播的双向传播方法，每一帧都能收到前后两个方向传播过来的视频特征信息，能更好地理解视频序列的动态特征。BasicVSR++[8]在延续双向传播架构的基础上，引入了二阶网格传播和光流指导的可变形卷积，这些改进使得网络能够更好地处理视频中的运动信息和细节，从而提高了超分辨率重建的效果。

在视频中，相机视角或物体本身可能会发生运动，需将不同帧之间位置发生变化的同一物体对齐到同一位置，使物体的特征都处在相同位置，即帧对齐。帧对齐在视频超分中至关重要，是跨帧信息交流中不可或缺的一环[8-9]。如图 1 所示，帧对齐能提供额外的子像素信息，使对齐后的图像更清晰。帧对齐包括运动估计和运动补偿两个步骤。光流网络常用于估计光流运动，然后向后执行变换（Warp）操作进行补偿运动[6]，即通过显式插值函数对光流值进行重采样。Tian 等[9]提出使用可变形卷积（DCN）进行对齐，以可变形注意力的形式将运动估计和运动补偿合并，这样可以同时学习到相邻帧的多处特征。Kim 等[10]尝试使用 3D 卷积来聚合时空信息，尽管这种方法可以快速聚合不同邻域的信息，但其融合了大量时间冗余的特征，可能会降低模型在重构任务中的性能。Lin 等[11]提出光流指导的可变形卷积，首先使用光流进行预对齐，然后通过预对齐特征辅助可变形卷积进行进一步对齐，这样可以在保证对齐准确性的同时，减少因光流不准确引起的误差，提升了帧对齐的效果。Rota 等[12]通过 VAE 解码器将光流从潜在空间转换到像素域，然后应用运动补偿来实现帧对齐。

图 1 未对齐图像与已对齐图像对比

Fig. 1 Unaligned image compared to aligned image

图像由不同的频率部分组成。低频成分代表图像中缓慢变化的部分，即图像中灰度值变化比较平缓的区域。高频成分代表图像中急剧变化的部分，即图像中灰度值变化较快的区域。目前上述方法只关注图像中低频信息的恢复，忽略了高频信息，这会降低对齐信息的准确性。在循环恢复策略下，前一帧的错误信息会影响当前帧对齐信息的准确性，并逐步累加，导致最终恢复的图像出现伪影，如图 2 所示。

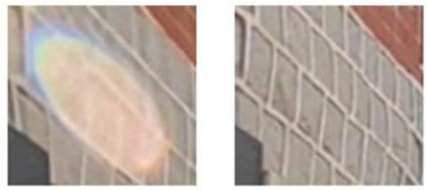

图 2 伪影

Fig. 2 Artifact

本文提出了一种基于隐式对齐的视频超分辨率模型 IAVSR。它采用双向传播机制，能够聚合前后两个方向的信息。由于帧对齐的效果会极大地影响模型的性能，因此使用隐式对齐[13]方法。该方法不通过任何显式插值函数对光流进行插值，而是通过偏移量和原始值计算得出，没有对插值施加平滑度约束，避免了对一些高频信息产生平滑效应。通过结合隐式对齐能恢复高频信息和光流指导的可变形卷积，可学习偏移多样性的特点，设计了新的隐式对齐模块。在双向传播中，利用前两帧传播的信息进行对齐，提高对齐信息的准确性，避免引入过多的参数，并使用了残差网络结构减轻负载，加快训练过程。最后设计了一个重建模块，充分利用多帧信息指导当前帧的恢复。实验表明，IAVSR 恢复的图像更清晰，可减少传播过程中出现的伪影，并在 PSNR 指标上超过基准模型 0.6 dB。

本文的工作如下：

1）利用光流指导的可变形卷积和隐式对齐方法的优势，设计了一个新的隐式对齐模块，能够有效地利用相邻多帧的时空信息。同时，为了提高模型捕捉信息的能力，提出了一个新的图像重建模块，将双向传播的特征与当前帧进行融合，重建出高分辨率的图像。

2）相比于 IconVSR[6]，TCNet[14]等模型，提出的视频超分辨率模型 IAVSR 在保持一定参数量的情况下能够获得更好的性能。

2 相关工作

2.1 图像插值

参考帧与目标帧对齐需要对图像上的子像素值进行重新插值，以估计缺失或未知像素的值。在插值过程中，原始数据会受到低通滤波器的影响，插值方法的选择会影响最终结果的平滑程度。最邻近插值无平滑度要求，只是简单地选择最接近的像素值，而不考虑周围像素的信息。双线性插值强制 L0 平滑，双三次插值强制 L1 平滑[15]。双线性（或双三次）插值考虑了更多周围像素的信息，通过使用复杂的插值函数，可以减少放大图像时出现的锯齿状边缘效应，获得更平滑的结果，进而产生更加连续的图像，从而保证得到的图像表面具有 L0（或更高阶）的平滑度。但它不会保留边缘信息，因此图像

会更加模糊。本文模型中的隐式对齐模块并未采用任何显式插值函数进行插值，而是通过偏移量和原始值计算获得，因此不会对高频信息进行平滑度约束。

2.2 视频超分辨率

图像超分辨率[16-18]旨在从低分辨率图像中恢复高分辨率图像，未考虑帧间的连续性，只能捕捉空间信息，不能聚合相邻帧的时间信息，应用于视频会导致帧之间的时序性较差，因此研究人员将目标转向了 VSR 模型，以便更好地捕捉相邻帧的特征信息和时序性。VSR 面临的一个主要挑战是帧的准确对齐，这对于保持时间连贯性并避免重建视频中的伪影至关重要。早期方法[19]没有考虑帧与帧之间的空间对齐。最初，VSR 方法使用光流法来对齐相邻图像，然而不准确的光流会导致其性能下降。最近，对齐策略已经从对齐图像[6]转变为对齐特征图，或者使用光流指导的可变形卷积[8]和可变形注意力方案[20]。为了提高对不准确光流的鲁棒性，Shi 等[21]提出了补丁对齐，通过平均预定义网格内的运动来对齐块。本文使用的隐式对齐模块能有效捕捉高频信息，并学习偏移多样性。同时，两时间步对齐法利用前两帧传播的信息而非前一帧传播的信息进行对齐，充分利用相邻多帧的时间信息，进一步提升了对齐信息的准确性。

3 本文方法

本文提出的基于隐式对齐的视频超分辨率模型 IAVSR 如图 3 所示。该模型基于 BasicVSR[6]，是一个简单易用的基准模型，主要分为 4 个步骤：传播、对齐、聚合和上采样。具体来说，IAVSR 采用双向传播的方法，在反向传播中，首先使用光流网络估计相邻帧之间的光流，然后利用隐式对齐方法对特征进行对齐，将对齐后的特征送入可变形卷积中做进一步对齐。为了充分利用相邻帧的信息，IAVSR 利用前两帧传播过来的信息来对齐当前帧，并通过残差网络结构减轻网络的负载，在多个残差块中提取特征后传播到下一帧进行对齐。在正向传播中，执行相同的操作，同时将正向传播、反向传播以及当前帧的特征送入重建模块，以最终恢复当前帧的高分辨率图像。

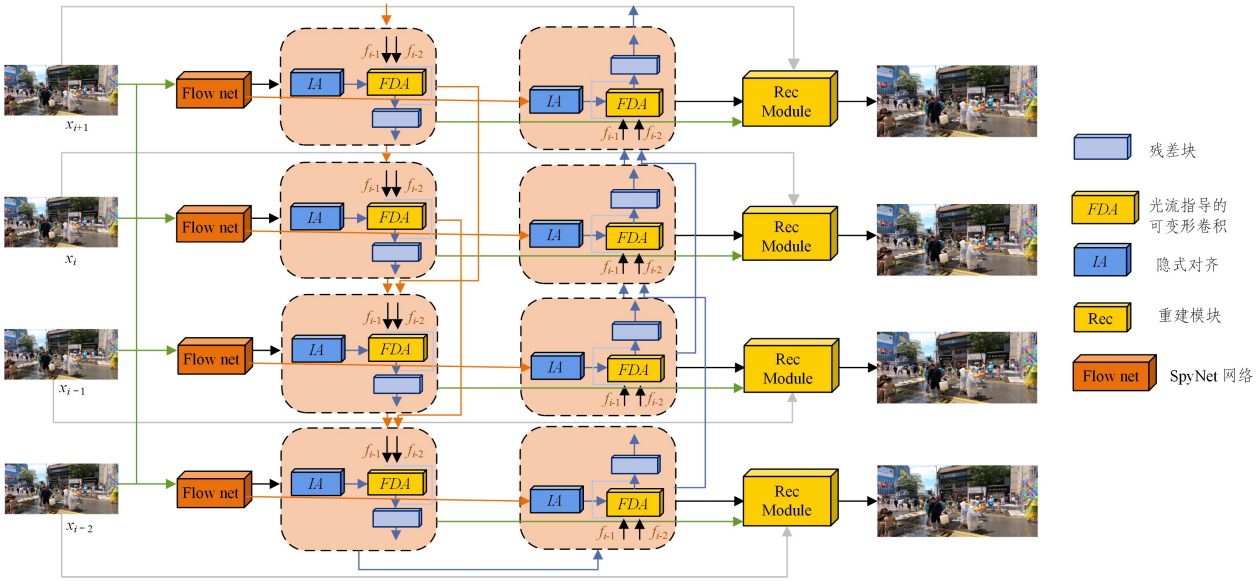

图 3 IAVSR 模型

Fig. 3 IAVSR model

3.1 隐式对齐方法

最常见的对齐方法是光流 warp[6]。给定光流 F，其中 $F(x,y)=(dx,dy)$。warp 过程通常为：将特征值从源帧 I' 传播到目标 I，然后在各个位置进行计算。

$$I^{\wedge}=W_{of}(I',dx,dy)=\sum_{(i,j)}G((i,j),(x+dx,y+dy))I'(i,j) \tag{1}$$

由于 (dx,dy) 是连续的，但 $I(\cdot,\cdot)$ 是在离散空间中定义的，因此需要重新采样。常见的重采样方法有双线性插值、双三次插值等。在插值过程中，$G((i,j),\cdot)$ 被用作插值内核，以聚合像素 $I'(x,y)$ 的权重。为构造插值内核，通常会对平滑性进行假设。最邻近插值无平滑度要求，双线性插值强制 L0 平滑，双三次插值强制 L1 平滑，在插值过程会对原始数据施加低通滤波器。本文使用一种隐式对齐[13]方法，其中重采样的值不通过任何显式插值函数获得，而是通过偏移量和原始值计算得出，如图 4 所示。

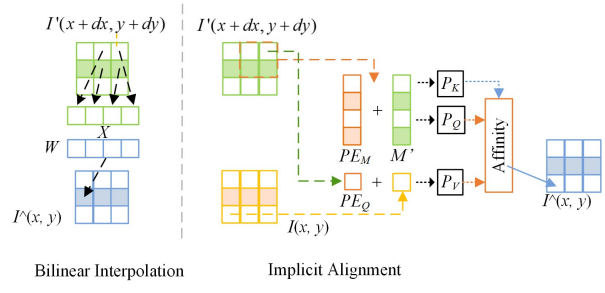

图 4 隐式对齐方法

Fig. 4 Implicit alignment method

对于目标帧 t 中的每个像素 $I(x,y)$ 和源帧 t' 中的 $I'(x,y)$，给定从 t 到 t' 的估计运动 f，其中 $f(x,y)=(dx,dy)$ 表示从帧 t 到 t' 的像素偏移量：

$$I^{\wedge}(x,y)=W_{ia}(I',x+dx,y+dy)=A(\boldsymbol{Q},\boldsymbol{K}')\boldsymbol{V}' \tag{2}$$

A 可以通过式(3)计算得到：

$$A(\boldsymbol{Q},\boldsymbol{K}')=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}'^{\mathrm{T}}}{\sqrt{C_p}}\right) \tag{3}$$

其中，$\boldsymbol{Q}$ 是 $I(x,y)$ 的 query，$\boldsymbol{K}'$ 和 $\boldsymbol{V}'$ 是源帧的 key 和 value，C_p 是 query 和 key 的通道数，$\boldsymbol{Q}$，$\boldsymbol{K}'$ 和 $\boldsymbol{V}'$ 的定义如下。

Query：$\boldsymbol{Q}$ 由编码器 P_Q 根据原始值 $I(x,y)$ 和运动估计 (dx,dy) 的位置编码的子像素偏移量来计算，同时使用正弦函数将偏移编码为位置编码 PE_Q 和 PE_M：

$$\boldsymbol{Q}=P_Q(I(x,y)+PE_Q) \tag{4}$$

Key 和 Value：此处仅对查询位置 $(x+dx,y+dy)$ 周围的窗口中的像素进行编码。在此过程中，h 和 w 是特征的高度和宽度，所选像素 M' 通过对 i 和 j 迭代获得：

$$M'(i,j)=I'([d+dx]-\left[\frac{h}{2}\right]+i,[y+dy]-\left[\frac{w}{2}\right]+j) \tag{5}$$

K' 和 V' 是通过编码及其位置编码到具有 P_K 和 P_V 的 key 空间来计算的：

$$\boldsymbol{K}'=P_K(M+PE_M) \tag{6}$$

$$\boldsymbol{V}'=P_V(M+PE_M) \tag{7}$$

其中，P_K 和 P_V 分别是 key 和 value。对 P_q，P_k 和 P_v 使用单个全连接层。

$$PE_Q=f(\Delta x,\Delta y) \tag{8}$$

$$\Delta x=dx-[dx],\Delta y=dy-[dy] \tag{9}$$

位置编码将窗口内的网格索引转换为 C 维正弦表示：

$$PE_M(i+wj)=f\left(i-\left[\frac{w}{2}\right],j-\left[\frac{w}{2}\right]\right) \tag{10}$$

$\mathrm{F}(x,y)$ 将 2D 连续坐标投影到高维空间中的 2D 位置编码函数：

$$f(x,y):=\begin{cases}\begin{cases}\sin(w_k.x), & \text{if } c=2k \text{ and } c\leqslant C/2\\ \cos(w_k.x), & \text{if } c=2k+1 \text{ and } c\leqslant C/2\end{cases}\\ \begin{cases}\sin(w_k.y), & \text{if } c=2k \text{ and } c>C/2\\ \cos(w_k.y), & \text{if } c=2k+1 \text{ and } c>C/2\end{cases}\end{cases} \tag{11}$$

PE_Q 和 PE_M 是以 $([x+dx],[y+dy])$ 为基准并计算相对偏移，PE_Q 表示 $I'(x+dx,y+dy)$ 的相对子像素偏移；假设 query 点 $I'(x+dx,y+dy)$ 与 $I(x,y)$ 具有相似的值，将此位置编码添加到 $I(x,y)$ 中。

3.2 隐式对齐模块

使用光流方法[8]对齐信息，准确性会受限，而使用可变形卷积[5]进行对齐，训练过程中偏移量过大，会导致最终训练崩溃。本文基于隐式对齐方法和光流指导的可变形卷积[10]提出新的隐式对齐模块，如图 5 所示，该方法利用隐式对齐方法和可变形卷积的优势，能够有效利用相邻多帧的时空信息，进一步提升对齐信息的准确性。

首先使用预训练好的光流网络估计相邻帧之间的光流 S，F_{t-1} 是上一个时间步计算的特征，然后使用光流进行隐式对齐：

$$F_{t-1}^{f}=I(F_{t-1},S) \tag{12}$$

I 是隐式对齐方法，我们使用预对齐的特征计算偏移量和 $mask$ 来计算残差，而不是直接计算偏移量，这样可以更好地利用特征之间的相关性：

$$O=S+C(c(F_t,F_{t-1}^{f})) \tag{13}$$

$$M=\sigma(C(c(F_t,F_{t-1}^{f}))) \tag{14}$$

其中，C 表示卷积操作，σ 表示 sigmoid 函数，D 表示可变形卷积，使用可变形卷积对齐特征：

$$F_t^{\wedge}=D(F_{t-1},O,M) \tag{15}$$

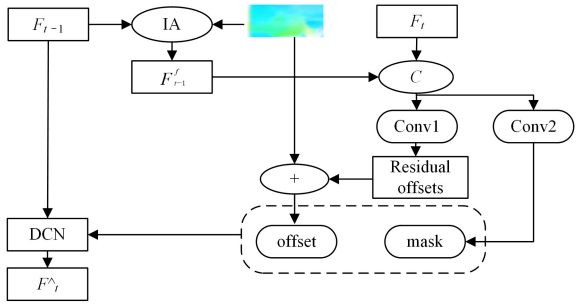

图 5 隐式对齐模块

Fig. 5 Implicit alignment module

3.3 传播和上采样模块

对于传播模块，IAVSR 采用双向传播方法[8]来获取前后两个方向的多个时间步的信息。它利用前两帧传播过来的信息对齐当前帧的信息，具体来说，其扩展了隐式对齐模块，将前两个时间步的信息进行 concat 操作拼接，并将结果与当前帧的信息一起送入隐式对齐模块进行对齐。这样做可以更好地利用前两帧传播过来的信息，增强对当前帧的特征提取和对齐效果，并通过残差网络连接来进一步优化结果。

对于上采样模块，利用正向和反向传播的特征进行恢复。为了更好地指导当前帧的恢复，使用卷积提取当前特征，将多个特征进行拼接，通过多个残差块细化特征，并使用像素重排[22]方法扩大 4 倍。最后将低分辨率图像使用双线性插值放大 4 倍，两者相加得到高分辨率图像，如图 6 所示。

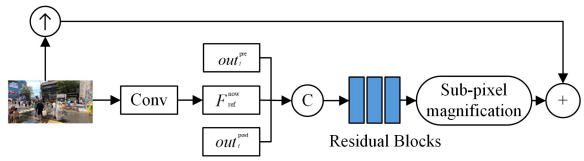

图 6 重建模块

Fig. 6 Reconstruction module

3.4 损失函数

Charbonnier loss[23]：在训练过程中，使用 Charbonnier 损失函数来评估模型的性能。该函数相比于传统的 L1 loss 和 L2 loss，可有效地提升视频超分辨率的性能，因此成为视频超分辨率领域的常用损失函数。其具体表达式如下：

$$Loss_{\mathrm{charbonnier}}=\sum_{i=1}^{S}\sqrt{(E(I_{\mathrm{LR}}^{i})-I_{\mathrm{HR}}^{i})^2+\varepsilon^2} \tag{16}$$

其中，$E(I_{\mathrm{LR}}^{i})$ 表示模型对第 i 个低分辨率图像的预测值，E 是超分模型，I_{HR}^{i} 表示第 i 个高分辨率图像的真实值或目标图像；i 表示帧的索引，从 1 到 S；ε 是损失函数系数，通常设定为一个很小的值(如 0.001)，用于避免当 I_{LR}^{i} 和 I_{HR}^{i} 非常接近时出现数值不稳定的情况。

4 实验

4.1 训练数据集和参数细节

1)数据集设置

在视频超分领域,数据集的规模和分辨率至关重要,因此使用 3 个常用的公开数据集 Vimeo-90k[24], REDS[25] 和 Vid4[26]。Vimeo-90k 包含超过 90 000 个视频序列,每个序列由 7 个连续的帧组成,分辨率为 448×256。REDS 数据集由高帧率(120 FPS)的视频组成,视频分辨率为 1 080×1 920,每个视频片段包含 100 个连续帧。Vid4 数据集是常用的视频超分测试数据集,包含了 4 个视频序列,每个视频片段平均有 41 个连续帧。REDS 是在 NTIRE19 竞赛中提出的数据集。

2)实现细节

所有模型使用 BI(Bicubic)退化进行 4 倍下采样测试,采用 REDS 和 Vimeo-90k 数据集进行训练,同时将 REDS4, Vid4 和 Vimeo90k-T 作为 IAVSR 模型的测试集。将峰值信噪比(PSNR)和结构相似性(SSIM)作为评估指标,并在 YCbCr 空间中的 Y 通道上进行测试。光流估计网络模块使用具有恒定学习率的 Adam[27] 优化器和预训练的 SpyNet[28],输入的 LR 大小为 64×64。光流估计网络模块的初始学习率设为 2.5×10^{-5},其他模块的学习率为 1×10^{-4},总迭代次数为 300 000 次,使用 Charbonnier loss 作为损失函数,batchsize 设置为 8,模型使用 4 张 NVIDIA Tesla V100 进行训练。

4.2 对比实验和结果分析

将 IAVSR 与 RBPN[29], BasicVSR[6], TTVSR[30] 等模型进行了比较。如表 1 所列,在 REDS4 和 Vid4 上,IAVSR 的 PSNR 指标超过所有对比模型。特别是在 IAVSR 的参数量和运行时间与 BasicVSR 和 IconVSR 相当的情况下,其相比 BasicVSR 在 Vid4 数据集上提升 0.38 dB,且在 REDS4 数据集上提升 0.6 dB。在 Vimeo-90k-T 上,IAVSR 的表现略次于 TCNet,这可能是因为 Vimeo 数据集中每个片段仅包含 7 个帧序列,而 REDS4 和 Vid4 的每个片段具有较长的连续帧(如 REDS4 中每个视频片段包含 100 个连续帧)。说明 IAVSR 在长帧序列数据集上效果更佳,适用于聚合长期信息。

表 1 定量比较(PSNR)

Table 1 Quantitative comparison (PSNR)

模型	参数量	运行时间/ms	Bl 退化(PSNR/SSIM)		
			REDS4	Vimeo-90k-T	Vid4
Bicubic	—	—	26.14/0.7292	31.32/0.8684	23.78/0.6347
RBPN[29]	12.2×10^6	1507	30.09/0.8590	37.07/0.9435	27.12/0.8180
EDVR[5]	20.6×10^6	378	31.09/0.8800	37.61/0.9489	27.35/0.8264
BasicVSR[6]	6.3×10^6	63	31.38/0.8893	37.14/0.9446	27.21/0.8229
IconVSR[6]	8.7×10^6	70	31.58/0.8932	37.44/0.9473	27.36/0.8255
TTVSR[30]	6.8×10^6	150	<u>31.92/0.9908</u>	37.58/0.9504	27.39/0.8342
PFDVR[31]	5.7×10^6	68	—	37.71/0.9501	<u>27.56/0.8385</u>
TCNet[14]	9.6×10^6	94	31.82/0.9004	**37.84/0.9514**	27.48/0.8380
Ours	9.3×10^6	75	**31.98/0.9012**	<u>37.75/0.9507</u>	**27.59/0.8391**

注:除 REDS4 外,其余结果在 Y 通道计算并取多帧平均;加粗表示最优结果,下划线表示次优结果。

此外还比较了模型的运行时间和参数量。如图 7 所示,比较了在 Vid4 上各模型的性能增益。从参数量来看,EDVR 和 RBPN 的参数最多,而 BasicVSR 和 PFDVR 相对较少,但 PSNR 指标略低。与这些模型相比,IAVSR 在增加了一定参数量的情况下,PSNR 指标更加平衡。从运行时间上看,尽管 IAVSR 的运行时间相对于 BasicVSR, IconVSR 和 PFDVR 模型较长,但相对于最短运行时间的 BasicVSR 模型,IAVSR 的 PSNR 实现了 0.38 dB 的增益。

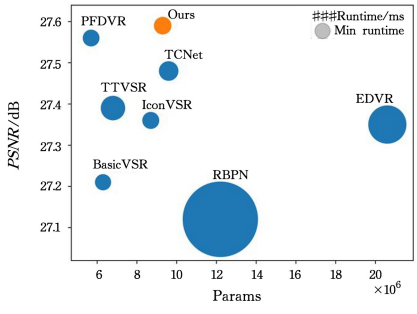

图 7 参数量和运行时间对比

Fig. 7 Comparison of parameters and running time

图 8 展示了 IAVSR 模型相对于基准模型 BasicVSR 和改进模型 IconVSR 在训练过程中 PSNR 值随训练次数变化的情况,每个模型训练 30 万次。

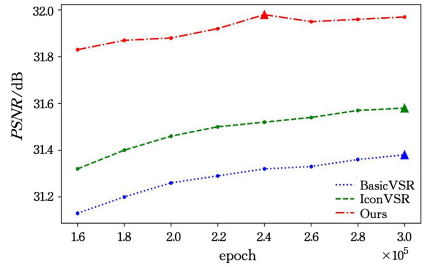

图 8 训练次数对比

Fig. 8 Comparison of training time

结果显示,IAVSR 训练到 24 万次时已经收敛,PSNR 达到了最高值,而 BasicVSR 和 IconVSR 在 30 万次训练后仍未完全收敛,且提升很小。这是因为 IAVSR 在传播模块中引入残差网络结构,在多个残差块中提取特征后传播到下一帧进行对齐,提高对齐信息准确性的同时加快了训练过程。

图 9 展示了 IAVSR 在 REDS 数据集上的效果图。在测试目标为公路牌的情况下,IAVSR 较具有最佳指标的 TTVSR 模型提升了 0.09 dB,公路牌中人行道处呈现出更加明显的轮廓和纹路。相比之下,在其他模型的恢复效果图中无法看清斑马线的黑白分界,特别是 BasicVSR。同时,IAVSR 是唯一一个可以清晰恢复图 10 指示牌中标记处的韩文

字母的方法，较 TCNet 模型提升了 0.08 dB。因为本文的隐式对齐模块利用隐式对齐方法捕捉高频信息和可变形卷积学习偏移多样性的优势，有效利用相邻多帧的时空信息，进一步提升了对齐信息的准确性。

图 9 公路牌效果图

Fig. 9 Highway sign effect

图 10 指示牌效果图

Fig. 10 Signage effect

4.3 消融实验

为了阐明所提出组件的贡献，从基准模型中逐步插入各个组件，并记录了每个组件引入后的改进以及 PSNR 值的变化。表 2 中列出了每个组件的名称和引入后的效果，包括 IA（隐式对齐方法）、IAM（隐式对齐模块）、Two（两时间步对齐）和 Rec（重建模块）。

表 2 不同组件的提升

Table 2 Different component improvements

Method	A	B	C	D	Ours
IA		√	√	√	√
IAM			√	√	√
Two				√	√
Rec					√
PSNR	31.38	31.62	31.80	31.93	31.98

隐式对齐方法：在神经网络的低频信号处，显示出的差异较小，在这种情况下，使用基于最邻近插值的对齐会导致性能差异，我们所使用的隐式对齐避免了该缺点，没有对插值施加平滑度约束，从而避免对一些高频信息产生平滑效应。从图 11 可以看出，使用隐式对齐后，高频信息更加真实，字母整体轮廓更加完整。

图 11 隐式对齐方法效果图

Fig. 11 Implicit alignment method renderings

隐式对齐模块：隐式对齐模块相比光流带来了更大的增益，可视化了光流以及隐式对齐模块对齐特征后的特征图，如图 12 所示。可以看到，网络学习到的偏移量与光流偏移量高度相似，但存在可观察到的差异。因为光流法仅从一个空间位置聚合信息，而提出的隐式对齐模块利用了可变形卷积，可以学习到多个方向的特征，能够从周围多个空间位置检索信息。

图 12 光流可视化

Fig. 12 Optical flow visualization

如图 13 所示，在隐式对齐模块加持下，门窗上恢复的纹路更加完整，避免了过度平滑的情况发生。特别是在处理细节复杂的区域时，隐式对齐模块能够更好地聚合相邻帧的信息，减少了对光流网络准确性的依赖，从而提高了对齐信息的准确性，有助于保留图像的细节，并提升了整体的视觉质量。

图 13 隐式对齐模块效果图

Fig. 13 Implicit alignment module renderings

两时间步对齐：正向传播和反向传播只能利用前一个方向上的信息，采用越多相邻连续帧传播过来的信息进行对齐，模型所需的训练成本越大。为了使 IAVSR 在具有与基准模型相当的参数量和运行时间下取得最佳 PSNR 值，利用前两

帧传播过来的信息进行对齐。通过利用两个时间步对齐的信息，模型能够更准确地捕捉到细微的变化和细节，从而产生清晰、更具纹理的图像边缘，尤其在处理具有复杂纹理或细微细节的区域时的效果更为显著。如图 14 所示，汽车车牌上有多位数字，IAVSR 模型恢复的数字边缘更加清晰，纹路比较明显，而只有单个时间步对齐的信息不够准确，导致图像的细节信息缺失，难以恢复高质量图像。

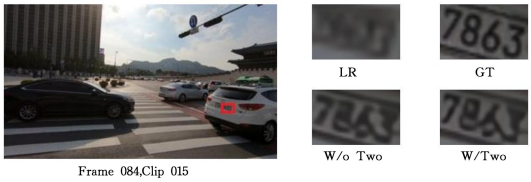

图 14 两时间步对齐效果图

Fig. 14 Two time step alignment renderings

重建模块：重建模块利用了前后两个方向的特征，并且融合了当前低分辨率图像特征，能够更好地理解图像内容并准确地恢复细节，重建出高分辨率图像。

结束语 本文提出了一个基于隐式对齐的视频超分辨率模型 IAVSR，使用隐式对齐模块代替光流，提高了对齐特征的准确性。为了充分利用双向传播的优势，使用两时间步对齐来指导当前帧恢复，同时使用残差结构[32]减轻网络负载并加快训练过程。最后重建模块将两个方向特征以及当前帧特征进行融合，以此获得高质量的高分辨率图像。实验表明，在保证一定效率的情况下，IAVSR 取得了优异的性能，并且可以推广到其他视频恢复的下游任务中[33-34]。现有的方法通过不断增加模型的参数量来提升性能，但训练的成本阻碍了视频超分的发展，因此未来的研究将重点放在优化模型参数量上。

参 考 文 献

[1] HUANG Z, HUANG A, HU X, et al. Scale-Adaptive Feature Aggregation for Efficient Space-Time Video Super-Resolution [C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024: 4228-4239.

[2] ZHU Y, LI G. A Lightweight Recurrent Grouping Attention Network for Video Super-Resolution[J]. Sensors, 2023, 23(20): 8574.

[3] CHEN Y H, CHEN S C, LIN Y Y, et al. MoTIF: Learning Motion Trajectories with Local Implicit Neural Functions for Continuous Space-Time Video Super-Resolution [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 23131-23141.

[4] TUO Z, YANG H, FU J, et al. Learning data-driven vector-quantized degradation model for animation video super-resolution[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 13179-13189.

[5] WANG X, CHAN K C K, YU K, et al. Edvr: Video restoration with enhanced deformable convolutional networks [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019.

[6] CHAN K C K, WANG X, YU K, et al. Basicvsr: The search for essential components in video super-resolution and beyond[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 4947-4956.

[7] ISOBE T, JIA X, GU S, et al. Video super-resolution with recurrent structure-detail network [C] // European Conference on Computer Vision. Cham: Springer, 2020: 645-660.

[8] CHAN K C K, ZHOU S, XU X, et al. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 5972-5981.

[9] TIAN Y, ZHANG Y, FU Y, et al. Tdan: Temporally-deformable alignment network for video super-resolution[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 3360-3369.

[10] KIM S Y, LIM J, NA T, et al. 3dsrnet: Video super-resolution using 3d convolutional neural networks[J]. arXiv: 1812. 09079, 2018.

[11] LIN J, HUANG Y, WANG L, et al. FDAN: Flow-guided deformable alignment network for video super-resolution[J]. arXiv: 2105. 05640, 2021.

[12] ROTA C, BUZZELLI M, VAN DE WEIJER J. Enhancing Perceptual Quality in Video Super-Resolution through Temporally-Consistent Detail Synthesis using Diffusion Models[J]. arXiv: 2311. 15908, 2023.

[13] XU K, YU Z, WANG X, et al. An Implicit Alignment for Video Super-Resolution[J]. arXiv: 2305. 00163, 2023.

[14] LIU M, JIN S, YAO C, et al. Temporal Consistency Learning of Inter-Frames for Video Super-Resolution[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 33(4): 1507-1520.

[15] PRIESSNER M, GABORIAU D C A, SHERIDAN A, et al. Content-aware frame interpolation(CAFI): Deep Learning-based temporal super-resolution for fast bioimaging[J]. Nature Methods, 2024, 21(2): 322-330.

[16] LI A, ZHANG L, LIU Y, et al. Feature Modulation Transformer: Cross-Refinement of Global Representation via High-Frequency Prior for Image Super-Resolution[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 12514-12524.

[17] YIN Z, LIU M, LI X, et al. MetaF2N: Blind Image Super-Resolution by Learning Efficient Model Adaptation from Faces[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 13033-13044

[18] WEI P, SUN Y, GUO X, et al. Towards Real-World Burst Image Super-Resolution: Benchmark and Method[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 13233-13242.

[19] WANG X, YU K, WU S, et al. Esrgan: Enhanced super-resolution generative adversarial networks[C]// Proceedings of the European Conference on Computer Vision(ECCV) Workshops.

[20] LIANG J, FAN Y, XIANG X, et al. Recurrent video restoration transformer with guided deformable attention[J]. Advances in Neural Information Processing Systems, 2022, 35: 378-393.

[21] SHI S, GU J, XIE L, et al. Rethinking alignment in video super-

resolution transformers[J]. Advances in Neural Information Processing Systems，2022，35：36081-36093.

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

[23] CHARBONNIER P，BLANC-FERAUD L，AUBERT G，et al. Two deterministic half-quadratic regularization algorithms for computed imaging[C]// Proceedings of 1st International Conference on Image Processing. IEEE，1994：168-172.

[24] XUE T，CHEN B，WU J，et al. Video enhancement with task-oriented flow[J]. International Journal of Computer Vision，2019，127：1106-1125.

[25] NAH S，BAIK S，HONG S，et al. Ntire 2019 challenge on video deblurring and super-resolution：Dataset and study[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019.

[26] LIU C，SUN D. On Bayesian adaptive video super resolution[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2013，36(2)：346-360.

[27] KINGMA D P，BA J. Adam：A method for stochastic optimization[J]. arXiv：1412. 6980，2014.

[28] RANJAN A，BLACK M J. Optical flow estimation using a spatial pyramid network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017：4161-4170.

[29] HARIS M，SHAKHNAROVICH G，UKITA N. Recurrent back-projection network for video super-resolution[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019：3897-3906.

[30] LIU C，YANG H，FU J，et al. Learning trajectory-aware transformer for video super-resolution[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022：5687-5696.

[31] QING T，YING X，SHA Z，et al. Video Super-Resolution with Pyramid Flow-Guided Deformable Alignment Network[C]// 2023 3rd International Conference on Electrical Engineering and Mechatronics Technology. IEEE，2023：758-764.

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

[33] ZHONG Z，CAO M，JI X，et al. Blur Interpolation Transformer for Real-World Motion from Blur[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023：5713-5723.

[34] YUE H，CAO C，LIAO L，et al. RViDeformer：Efficient Raw Video Denoising Transformer with a Larger Benchmark Dataset[J]. arXiv：2305. 00767，2023.

WANG Fengling，born in 2000，postgraduate. Her main research interests include video super-resolution and time series.

XIE Jingming，born in 1977，Ph.D，professor. His main research interests include artificial intelligence technology application and so on.

（责任编辑：喻藜）