# 时域注意力特征对齐的视频压缩感知重构网络

魏志超,杨春玲
(华南理工大学电子与信息学院,广东广州 510640)

**摘 要:** 现有视频压缩感知神经网络重构算法采用的光流对齐和可变形卷积对齐的运动补偿方式存在误差积聚、信息感知范围有限等问题,极大地限制了其有效性和实用性. 为了在不引入额外参数的条件下自适应提取参考帧的全局信息,本文提出了利用注意力机制实现视频压缩感知重构过程中运动估计/运动补偿的创新思想,并设计了时域注意力特征对齐网络(Temporal-Attention Feature Alignment Network, TAFA-Net)进行实现. 在此基础上,提出了联合深度重构网络(Joint Deep Reconstruction Network Based on TAFA-Net, JDR-TAFA-Net),实现非关键帧的高性能重构. 先利用本文所提的TAFA-Net获得参考帧到当前帧的对齐帧;然后,利用基于自编码器架构的融合网络充分提取已有帧信息,增强非关键帧的重构质量. 仿真结果表明,与最优的迭代优化算法SSIM-InterF-GSR相比,所提算法重构帧的峰值信噪比(Peak Signal to Noise Ratio, PSNR)最高提升了4.74 dB;与最优的深度学习算法STM-Net相比,所提算法重构帧的PSNR最高提升了0.64 dB.

**关键词:** 视频压缩感知; 神经网络; 时域注意力; 特征对齐; 运动补偿; 深度重构
**中图分类号:** TN919.8 **文献标识码:** A **文章编号:** 0372-2112(2022)11-2584-09
**电子学报URL:**http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220041

# Video Compressed Sensing Reconstruction Network Based on Temporal-Attention Feature Alignment

WEI Zhi-chao, YANG Chun-ling
(*School of Electronic and Information Engineering, South China University of Technology, Guangzhou, Guangdong* 510640, *China*)

**Abstract:** The motion compensation methods of optical flow alignment and deformable convolution alignment adopted by the existing video compressed sensing reconstruction algorithms have problems such as error accumulation and limited information perception range, which greatly limit their effectiveness and practicability. In order to adaptively extract the global information of the reference frame without introducing extra parameters, this paper first proposes an innovative idea of using the attention mechanism to realize motion estimation and motion compensation in video compressed sensing reconstruction, and then designs the temporal-attention feature alignment network(TAFA-Net) for implementation. On this basis, a joint deep reconstruction network(JDR-TAFA-Net) is proposed to achieve high-performance reconstruction for non-key frames. First, the reference frames are adaptively aligned to the current non-key frame through TAFA-Net, and then a fusion network based on the auto-encoder is introduced to fully extract the relevant information from existing frames to further enhance the reconstruction quality of the non-key frames. Experimental results show that, compared with the state-of-the-art iterative optimization-based method SSIM-InterF-GSR, the proposed method can improve PSNR(Peak Signal to Noise Ratio) by 4.74 dB, and compared with the state-of-the-art deep learning-based method STM-Net, the proposed method can improve PSNR by 0.64 dB.

**Key words:** video compressed sensing; neural network; temporal attention; feature alignment; motion compensation; deep reconstruction

## 1 引言

压缩感知(Compressed Sensing, CS)[1]是一种强大的信号采样技术,它突破了奈奎斯特采样定理的限制,能够在对信号采样的同时完成压缩,极大地减轻了编

收稿日期:2022-01-05;修回日期:2022-04-01;责任编辑:覃怀银
基金项目:广东省自然科学基金(No.2019A1515011949)

码端的计算和存储负担. CS理论表明,如果信号能够在某个变换域稀疏表示,则能够通过重构算法从少量的观测值中准确恢复原始信号,这本质上是一个不适定问题. 对于视频压缩感知(Video Compressed Sensing, VCS)来说,如何有效地利用视频信号中存在的大量帧间冗余信息提升重构性能是研究关键.

现有VCS重构算法大都基于分布式视频压缩感知框架而设计. 在该框架下,视频帧在编码端以不同采样率进行独立采样,在解码端则通过建立帧间对应关系进行联合重构. 多假设算法(Multi-Hypothesis, MH)[2]基于最小均方误差准则首先从参考帧中找到当前图像块的假设块集,然后通过残差重构恢复原始视频帧. 由于残差比原始信号更加稀疏,MH算法及其改进算法[3,4]取得了不错的重构性能. 另一类算法[5]在帧间运动估计中采用更优的匹配准则,通过构造稀疏表示能力更强的自适应字典进一步提升视频帧的重构质量. 然而,上述方法均通过迭代优化的方式重构视频信号,计算复杂度较高,且泛化能力不理想,限制了其实际应用能力.

近年来,随着深度学习技术在图像压缩感知(Image Compressed Sensing, ICS)领域取得成功,一些基于深度学习的VCS重构算法[6~9]也陆续被提出,并取得了相较于传统算法更优的重构性能. 这些算法通常包含四个步骤:逐帧独立重构、特征提取、时域对齐以及信息融合. 由于物体或相机的运动,视频帧不是完全对齐的,如何在遮挡以及复杂运动条件下准确建立帧间对应关系以进行时域对齐是提升重构性能的关键. 现有基于深度学习的VCS重构算法通常采用光流网络或可变形卷积网络进行时域对齐. 基于光流对齐的算法[7,9]首先估计参考帧和当前帧之间的运动信息,然后基于光流场把参考帧对齐到当前帧. 尽管光流网络的相关研究较为成熟,但精确的光流估计仍旧难以实现,光流计算及对齐过程中产生的误差会逐渐积聚,最终在对齐帧中产生伪影. 与基于光流的两阶段显式对齐方法不同,另一类算法[8]采用可变形卷积[10]进行一阶段隐式对齐. 由于可变形卷积中卷积核的采样位置具有自适应的偏移量,这类方法在处理一些复杂数据时会表现出更好的性能. 然而,估计偏移量所引入的额外参数以及较大的网络训练难度限制了这类方法的灵活性.

在过去几年中,卷积神经网络(Convolutional Neural Network, CNN)凭借参数共享、局部连接的优势被广泛用于图像和视频数据的特征提取. 然而,CNN只能捕获局部依赖关系,这在一定程度上限制了其在视频任务上的表现. 近来,一批视觉Tranformer[11~13]被提出,并在多种任务上取得了相较于CNN更优的性能. Transformer架构的核心在于注意力机制的使用,通过建立全局依赖关系获取远距离的上下文信息. 需要指出的是,尽管该过程并未引入额外参数,但高维矩阵相乘却带来了较大的计算代价.

基于以上分析,本文将计算高效的单输入criss-cross注意力[14]扩展到多输入的视频任务中进行运动补偿,旨在减小注意力技术计算复杂度的同时利用其优势解决现有时域对齐方法存在的问题. 首先提出了时域注意力特征对齐网络(Temporal-Attention Feature Alignment Network, TAFA-Net),能够在复杂运动及较远时域距离的条件下,通过多尺度的特征迁移和特征融合生成高质量的对齐帧. 在此基础上,提出了一种联合深度重构网络(Joint Deep Reconstruction Network Based on TAFA-Net, JDR-TAFA-Net)充分利用已有信息,增强非关键帧的重构质量. 接下来将详细描述本文所提思想及神经网络实现.

## 2　时域注意力特征对齐的视频压缩感知重构网络

### 2.1　端到端的视频压缩感知框架

本文设计的基于所提网络TAFA-Net和JDR-TAFA-Net的分布式视频压缩感知框架如图1所示,其对输入视频帧尺寸不做固定要求,因而具有较高的灵活性,适用于任意分辨率的视频数据. 训练时,整个框架作为一个整体进行联合训练;实际应用时,可以方便地拆分为采样和重构两个部分分别置于编码端和解码端.

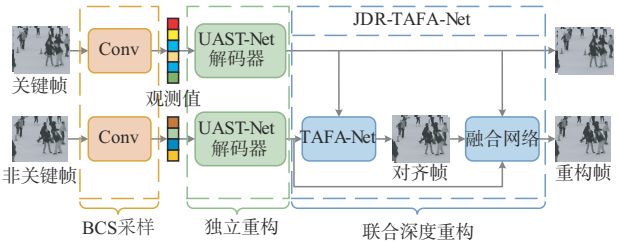



图1　本文设计的分布式视频压缩感知框架

在编码端,视频序列被划分为固定长度的图像组(Group Of Pictures, GOP),每个GOP的第一帧为关键帧,采样率高于其他非关键帧. 与现有大多数算法采用随机高斯矩阵进行采样的方式不同,本文使用一个可学习的卷积层$s(\cdot)$把信号采样转化为特征提取过程,使得观测值中保留更多原始信号的结构信息[15,16]. 为了减小编码端的计算和存储负担,对视频帧进行分块压缩感知(Block-based Compressed Sensing, BCS)[17]采样,$s(\cdot)$中卷积核大小及滑动步长均设为非重叠图像块的大小$B$. 在此条件下,对于采样率$r$,观测值数量为$rB^2$.

解码端包含两个部分:独立重构和联合深度重构. 为充分利用帧内相关性,首先采用具有良好可解释性的ICS网络UAST-Net[9]进行逐帧独立重构. 由于关键

帧的采样率相对较高，其独立重构帧也相应地具有更高的质量，因而直接作为最终重构结果．对于非关键帧来说，本文设计了联合深度重构网络进一步提升其重构质量．

## 2.2 联合深度重构网络

经过独立重构后，关键帧具有较高的重构质量．本文提出了联合深度重构网络JDR-TAFA-Net，把离当前非关键帧最近的两个关键帧（来自当前GOP和下一个GOP）的独立重构帧视为参考帧，对当前帧的独立重构帧进行补偿．该网络的实现包含两个过程：时域注意力特征对齐和多帧信息融合．在利用所提时域注意力技术获得高质量对齐帧后，通过具有较大感受野的融合网络充分提取已有帧信息，进一步增强非关键帧的重构质量．

### 2.2.1 时域注意力特征对齐思想及神经网络实现

如图2所示，TAFA-Net在结构流程上包含三个核心操作：多尺度特征提取、基于注意力机制的特征对齐，以及基于混合空洞卷积后处理．为了在复杂运动或较远时域距离条件下保证高质量的运动补偿，TAFA-Net采用由粗到细的方式进行多尺度的时域对齐，即从特征金字塔最高层的特征开始，逐层进行特征迁移及融合．最终得到的最大尺寸的对齐特征将通过上采样和后处理生成高质量的对齐帧．

值得强调的是，在时域特征对齐操作中，本文提出使用参考帧$\boldsymbol{x}^{\mathrm{r}}$的低采样率重构帧$\boldsymbol{x}^{\mathrm{r_ls}}$作为替代与当前帧

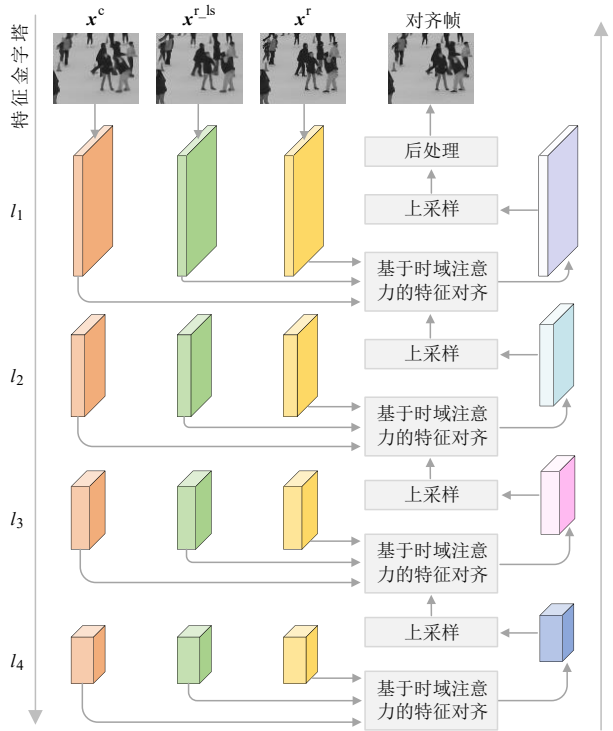



图2 TAFA-Net网络结构

$\boldsymbol{x}^{\mathrm{c}}$进行特征匹配的思想，以实现更加有效的运动估计．原因为，虽然时域对齐的目的是用$\boldsymbol{x}^{\mathrm{r}}$中的信息来补偿$\boldsymbol{x}^{\mathrm{c}}$，但它们的特征分布具有非常不同的模式，直接进行特征匹配会严重影响匹配的准确度，进而影响对齐特征的质量．

$\boldsymbol{x}^{\mathrm{r_ls}}$的生成方法为：在解码端以$\boldsymbol{x}^{\mathrm{r}}$作为非关键帧采样网络（采样矩阵需要传输到解码端）和重构网络的输入，输出结果就是$\boldsymbol{x}^{\mathrm{r_ls}}$．由于$\boldsymbol{x}^{\mathrm{r_ls}}$与$\boldsymbol{x}^{\mathrm{c}}$特征分布相匹配，同时采用共享参数的特征提取器，本文把其定义为相同质量参考帧．而在进行后续对齐操作时，TAFA-Net仍采用质量更高的原始参考帧$\boldsymbol{x}^{\mathrm{r}}$作为补偿信息来源．接下来将对TAFA-Net的三个核心操作进行详细介绍．如无特别指定，TAFA-Net中所有的卷积层均使用大小为$3\times3$的卷积核．

（1）多尺度特征提取

TAFA-Net采用由粗到细的方式进行特征匹配与对齐，即生成$l$层的特征金字塔以同时获取低层细节信息和高层语义信息，从而实现复杂运动场景下相关信息的提取和不相关信息的抑制．给定$\boldsymbol{x}^{\mathrm{c}}$，$\boldsymbol{x}^{\mathrm{r_ls}}$和$\boldsymbol{x}^{\mathrm{r}}$，特征提取过程表示为

$$\begin{aligned}\boldsymbol{F}^{\mathrm{c},\,l}&=\mathrm{MFE}^{l}\left(\boldsymbol{x}^{\mathrm{c}}\right)\\ \boldsymbol{F}^{\mathrm{r\_ls},\,l}&=\mathrm{MFE}^{l}\left(\boldsymbol{x}^{\mathrm{r\_ls}}\right)\\ \boldsymbol{F}^{\mathrm{r},\,l}&=\mathrm{MFE}^{l}\left(\boldsymbol{x}^{\mathrm{r}}\right)\end{aligned}\tag{1}$$

其中，$\mathrm{MFE}^{l}(\cdot)$表示特征金字塔第$l$层的输出．TAFA-Net采用步长为2的卷积得到深度为5层的特征金字塔，其中第0层为输入的独立重构帧，而1到4层每层由3个级联卷积组成，不同层特征通道数分别为32，64，96和128．

（2）基于注意力机制的特征对齐

如图3所示，TAFA-Net中基于注意力机制的特征对齐包含两个过程：特征迁移与特征融合．

为高效地实现特征迁移，本文提出的TAFA-Net开创性地把运动估计和运动补偿建模为注意力机制中的相关性计算与特征聚合步骤．因此，TAFA-Net不但利用了注意力技术强大的建模能力，同时还具有良好的可解释性．与光流单假设或可变形卷积多假设的思想不同，TAFA-Net中运动估计和运动补偿是基于全局假设的，即需要求出当前帧与参考帧任意位置特征向量之间的相关性，并以此作为权重对目标位置补偿全局信息．需要强调的是，尽管对参考帧在特征空间进行全局搜索的方式具有更高的灵活性，可以更加有效地处理不同类型的视频序列，但却引入了较大的计算量．TAFA-Net通过以下两种设计来缓解该问题：（$a$）时域对齐只在特征金字塔较小尺寸上进行（不包括原始输入尺寸）；（$b$）把图像任务中计算高效的criss-cross注意力[14]扩展到视频任务上．

特征融合包括迁移特征与当前帧原始特征的融合,以及特征金字塔不同层对齐特征的融合. 前者用来自适应提取已有特征中的信息,避免不相关的运动补偿影响对齐效果;后者用来融合粗粒度的高层对齐特征与细粒度的低层对齐特征,最终得到结构与纹理信息更加丰富的对齐帧.

(*a*) 基于时域相关性的两阶段特征迁移

为了减小计算和存储代价,本文采用criss-cross注意力[14]实现特征迁移,迁移过程的时间复杂度和空间复杂度均为 $\mathcal{O}(N\sqrt{N})$($N$ 为特征向量个数). 与文献[14]使用单一输入特征得到query、key和value的方式不同,本文提出对当前帧特征、相同质量参考帧特征与原始参考帧特征分别进行线性变换得到query、key和value,从而赋予后续相关性计算和特征聚合操作明确的意义.

如图3所示,TAFA-Net中计算高效的特征迁移过程包含紧密相关的两个阶段,通过连续提取局部信息的方式实现参考帧全局信息的提取. 每个阶段包含相似的三个实现步骤,下面以第一阶段局部特征迁移为例详细描述.

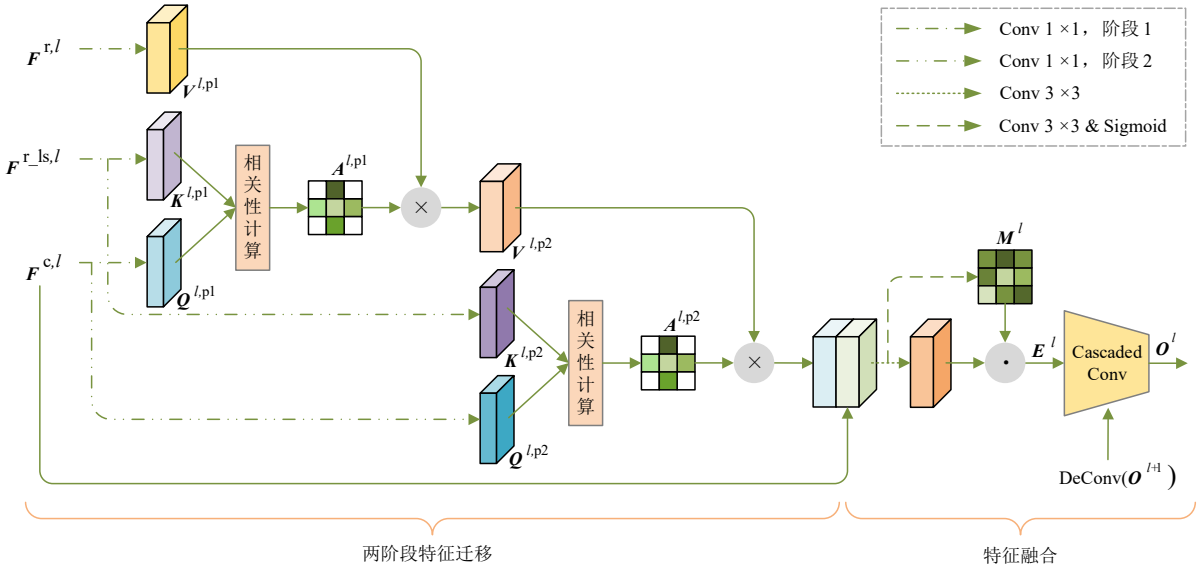



图3　基于时域注意力的特征对齐

**步骤1**　线性变换:将 $\{\boldsymbol{F}^{\mathrm{c},l},\boldsymbol{F}^{\mathrm{r\_ls},l},\boldsymbol{F}^{\mathrm{r},l}\}\in\mathbb{R}^{C\times H\times W}$ 分别通过三个卷积核大小为 $1\times1$ 的卷积层生成注意力机制中的三个基本元素 $\{\boldsymbol{Q}^{l,\mathrm{p1}}(\mathrm{query}),\boldsymbol{K}^{l,\mathrm{p1}}(\mathrm{key})\}\in\mathbb{R}^{C'\times H\times W}$、$\boldsymbol{V}^{l,\mathrm{p1}}(\mathrm{value})\in\mathbb{R}^{C\times H\times W}$,其中p1表示第一阶段特征迁移,$C'=\lfloor C/2\rfloor$ 用于减小计算量;

**步骤2**　相关性计算:对于 $\boldsymbol{Q}^{l,\mathrm{p1}}$ 中空间位置 $i$(按行展开的索引)处的特征向量 $\boldsymbol{q}_i^{l,\mathrm{p1}}\in\mathbb{R}^{C'}$,可以得到其在 $\boldsymbol{K}^{l,\mathrm{p1}}$ 中对应位置criss-cross路径(同一行以及同一列的所有空间位置)上的特征向量集 $\boldsymbol{\Omega}_i^{l,\mathrm{p1}}\in\mathbb{R}^{(H+W-1)\times C'}$. $\boldsymbol{q}_i^{l,\mathrm{p1}}$ 与 $\boldsymbol{\Omega}_i^{l,\mathrm{p1}}$ 中第 $j$ 个元素 $\boldsymbol{\Omega}_{j,i}^{l,\mathrm{p1}}$ 之间的相关运算定义为

$$a_{j,i}^{l,\mathrm{p1}}=\frac{\exp\left(\boldsymbol{q}_i^{l,\mathrm{p1}}\cdot\boldsymbol{\Omega}_{j,i}^{l,\mathrm{p1}}\right)}{\sum_{k=1}^{H+W-1}\exp\left(\boldsymbol{q}_i^{l,\mathrm{p1}}\cdot\boldsymbol{\Omega}_{k,i}^{l,\mathrm{p1}}\right)}\tag{2}$$

其中,$a_{j,i}^{l,\mathrm{p1}}$ 为局部注意力图 $\boldsymbol{A}^{l,\mathrm{p1}}\in\mathbb{R}^{(H+W-1)\times(H\times W)}$ 空间位置 $(j,i)$ 处的值,表示 $\boldsymbol{\Omega}_{j,i}^{l,\mathrm{p1}}$ 与 $\boldsymbol{q}_i^{l,\mathrm{p1}}$ 的相似性;

**步骤3**　特征聚合:在 $\boldsymbol{V}^{l,\mathrm{p1}}$ 中空间位置 $i$ 处,可以得到其criss-cross路径上的向量集 $\boldsymbol{\Gamma}_i^{l,\mathrm{p1}}\in\mathbb{R}^{(H+W-1)\times C}$,根据局部注意力图 $\boldsymbol{A}^{l,\mathrm{p1}}$ 进行特征加权求和的操作定义为

$$\boldsymbol{t}_i^{l,\mathrm{p1}}=\sum_{j=1}^{H+W-1}a_{j,i}^{l,\mathrm{p1}}\boldsymbol{\Gamma}_{j,i}^{l,\mathrm{p1}}\tag{3}$$

其中,$\boldsymbol{\Gamma}_{j,i}^{l,\mathrm{p1}}$ 为 $\boldsymbol{\Gamma}_i^{l,\mathrm{p1}}$ 的第 $j$ 个元素,$\boldsymbol{t}_i^{l,\mathrm{p1}}$ 为局部特征迁移图 $\boldsymbol{T}^{l,\mathrm{p1}}\in\mathbb{R}^{C\times H\times W}$ 中位置 $i$ 处的特征向量.

第一阶段的特征迁移只建立了参考帧特征和当前帧特征在水平和垂直两个方向上的局部依赖关系,当视频序列运动较为复杂时,该方式便无法有效地从参考帧中提取信息. TAFA-Net通过引入第二阶段特征迁移操作来解决这一问题,两次特征迁移步骤相同,不同之处在于:为间接获取参考帧特征的全局信息,第二阶段特征迁移操作中使用的value(即 $\boldsymbol{V}^{l,\mathrm{p2}}$)为第一阶段特征迁移操作得到的特征 $\boldsymbol{T}^{l,\mathrm{p1}}$;为了让网络适应 $\boldsymbol{V}^{l,\mathrm{p1}}$ 和 $\boldsymbol{V}^{l,\mathrm{p2}}$ 的不同特征分布以进行自适应特征提取,第二阶段特征迁移操作用来进行相关匹配的 $\boldsymbol{Q}^{l,\mathrm{p2}}$ 和 $\boldsymbol{K}^{l,\mathrm{p2}}$ 通过新的线性投影得到.

(*b*) 基于软注意力的特征融合

得到包含补偿信息的迁移特征 $\boldsymbol{T}^{l,\mathrm{p2}}$ 后,令其与当前帧原始特征 $\boldsymbol{F}^{\mathrm{c},l}$ 进行融合以充分利用已有信息. 为避免网络对迁移特征产生过度依赖而引入不相关的补

偿信息，融合特征将通过一个软注意力模块得到更加判别的增强特征 $\boldsymbol{E}^l$. 最终，$\boldsymbol{E}^l$ 将与特征金字塔 $l$+1 层的对齐特征 $\boldsymbol{O}^{l+1}$ 进行跨层融合，得到当前层的对齐特征，上述操作可以表示为

$$\boldsymbol{M}^l=\mathrm{Sigmoid}(f_1([\boldsymbol{F}^{\mathrm{c},l},\boldsymbol{T}^{l,\mathrm{p2}}])) \tag{4}$$

$$\boldsymbol{E}^l=f_2([\boldsymbol{F}^{\mathrm{c},l},\boldsymbol{T}^{l,\mathrm{p2}}])\odot\boldsymbol{M}^l \tag{5}$$

$$\boldsymbol{O}^l=g([\boldsymbol{E}^l,\mathrm{DeConv}(\boldsymbol{O}^{l+1})]) \tag{6}$$

其中，[·, ·]和⊙分别表示拼接操作和对应位置元素相乘操作，$f_i(\cdot)(i\in[1,2])$、DeConv(·)及 $g(\cdot)$ 分别表示一个常规卷积层、一个转置卷积层以及 5 个级联卷积层，不同金字塔层中 5 个级联卷积层对应的输出通道数如表 1 所示.

**表 1 不同金字塔层 $g(\cdot)$ 所表示级联卷积层输出通道数**

| 层序 | 5 个级联卷积层输出通道数 | | | | |
|---|---|---|---|---|---|
| $l_1$ | 128 | 96 | 64 | 32 | 16 |
| $l_2$ | 128 | 128 | 96 | 64 | 32 |
| $l_3$ | 128 | 128 | 128 | 96 | 64 |
| $l_4$ | 128 | 128 | 128 | 128 | 96 |

总的来说，上述基于注意力机制的特征迁移与特征融合操作使用了较为高效的方式进行运动补偿，有效解决了现有方法参数量较大、误差积聚、不能提取全局信息等问题.

（3）基于混合空洞卷积的后处理

经过多尺度的特征迁移和特征融合后，得到的高质量对齐特征的宽和高均为输入视频帧的一半. 为了在恢复原始空间尺寸同时生成具有更加精细结构和纹理的对齐帧，TAFA-Net 首先利用转置卷积进行上采样，然后通过空洞卷积提取更多的上下文信息来后处理对齐特征.

本文设计的后处理模块由 8 个空洞卷积组成. 由于单一膨胀率的空洞卷积会产生栅格效应，本文引入具有锯齿状膨胀率的混合空洞卷积[18]以及残差学习来解决这一问题. 这 8 个卷积层的输出通道数及膨胀率分别设为(128,1)、(128,2)、(128,5)、(64,1)、(64,2)、(64,5)、(32,1)、(1,1). 基于此，后处理操作便可以在不增加额外参数和计算量的条件下有效扩大感受野，从而自适应提取更多有用信息，生成高质量的对齐帧.

### 2.2.2 多帧信息融合

当物体或相机具有较为复杂的运动时，TAFA-Net 输出的对齐帧可能达不到理想的效果. 为了避免不相关的运动补偿对重构质量造成影响，本文设计了多帧信息融合网络来自适应地从已有帧中提取信息，生成最终重构帧.

如图 4 所示，本文设计的多帧信息融合网络基于自编码器架构，其输入为当前帧的独立重构帧 $\boldsymbol{x}_i^{\mathrm{c}}$、两个参考帧 $\{\boldsymbol{x}_i^{\mathrm{r},1},\boldsymbol{x}_i^{\mathrm{r},2}\}$ 以及两个对齐帧 $\{\boldsymbol{x}_i^{\mathrm{a},1},\boldsymbol{x}_i^{\mathrm{a},2}\}$，输出为当前帧的最终重构帧 $\hat{\boldsymbol{x}}_i$. 该网络的设计具有以下两方面的优势：编码器进行逐级下采样在减小计算量的同时有效扩大感受野；解码器融合包含细节信息的低层特征与包含语义信息的高层特征生成高质量的重构帧.

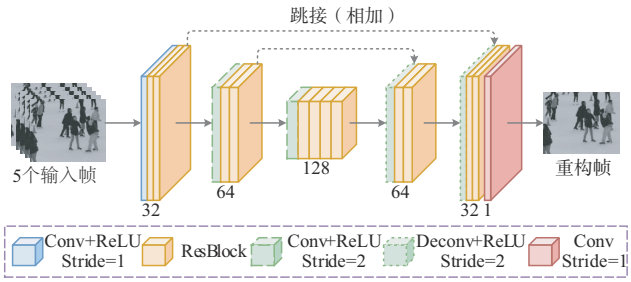



图 4 多帧信息融合网络

## 2.3 损失函数

本文采用两种损失函数进行网络训练，包括像素损失 $L_{\mathrm{pixel}}$ 和观测损失[9] $L_{\mathrm{measurement}}$. 总的损失函数表示为

$$L=L_{\mathrm{pixel}}+\lambda L_{\mathrm{measurement}} \tag{7}$$

（1）本文采用均方误差作为像素损失. 给定输入视频帧 $\{\boldsymbol{x}_i\}$，像素损失定义为

$$L_{\mathrm{pixel}}=\frac{1}{THW}\sum_{i=1}^{T}\left(\left\|\hat{\boldsymbol{x}}_i-\boldsymbol{x}_i\right\|_2^2+\sum_{j=1}^{R}\left\|\boldsymbol{x}_i^{\mathrm{a},j}-\boldsymbol{x}_i\right\|_2^2\right) \tag{8}$$

其中，($T$, $H$, $W$)为 GOP 的尺寸，$R$ 为参考帧的数量. $\boldsymbol{x}_i^{\mathrm{a},j}$ 表示第 $j$ 个参考帧对应的对齐帧.

（2）为进一步利用观测值信息提升网络重构性能，本文引入观测损失[9]来辅助网络训练，基本思想为增强重构帧和原始视频帧在观测空间的相似性. 本文使用的观测损失定义为

$$L_{\mathrm{measurement}}=\frac{1}{rTHW}\sum_{i=1}^{T}\left(\left\|s\left(\hat{\boldsymbol{x}}_i\right)-\boldsymbol{y}_i\right\|_2^2+\sum_{j=1}^{R}\left\|s\left(\boldsymbol{x}_i^{a,j}\right)-\boldsymbol{y}_i\right\|_2^2\right) \tag{9}$$

其中，$r$ 和 $\boldsymbol{y}_i$ 分别表示采样率和视频帧 $\boldsymbol{x}_i$ 的观测值.

# 3 仿真实验结果及分析

## 3.1 网络训练细节

为了使网络各部分关注不同的时域和空域信息，整个训练过程被划分为两个阶段. 在第一个训练阶段，使用 BSDS500[19]数据集训练 UAST-Net. 训练过程中，输入图像被随机裁剪为 256 pixel×256 pixel 大小，并采用随机翻转操作进行数据增强. 该阶段的训练完成后，UAST-Net 的参数将保持固定. 第二个训练阶段使用 UCF-101[20]数据集来训练联合深度重构网络. 其中，训练集、验证集和测试集的划分比例为 8∶1∶1. 训练过程中，视频帧被中心裁剪为 160 pixel×160 pixel 大小，且只使用亮度分量（YCbCr 色彩空间）.

实验过程中,图像块大小$B$设为32,正则项权重$\lambda$设为0.001.网络优化算法采用Adam[21],两个训练阶段的学习率均为0.0001. 仿真硬件配置为一块11 GB显存的NVIDIA RTX2080Ti GPU.

### 3.2　与现有算法重构性能对比

为了评估所提算法的有效性,本节将其与几种具有代表性的迭代优化算法和深度学习算法进行重构性能对比,采用的评价指标包括峰值信噪比(Peak Signal to Noise Ratio, PSNR)、结构相似性(Structural Similarity Index Measure, SSIM)以及运行时间. 凭借GPU加速的巨大优势,深度学习算法在重构速度上要远快于迭代优化算法. 因此,本节只对所提算法与深度学习算法进行运行时间的对比. 在所有实验中,关键帧及非关键帧的采样率分别表示为$SR_K$和$SR_N$.

#### 3.2.1　与基于迭代优化的VCS算法重构性能对比

本小节用来跟所提算法进行对比的算法包括Video-MH[2]、2sMHR[3]以及SSIM-InterF-GSR[5],其中SSIM-InterF-GSR为现有最优的迭代优化VCS重构算法. 为了验证所提算法的有效性,选取5个具有不同运动类型的QCIF视频序列作为测试集. 实验过程中,GOP大小设为8,$SR_K$设为0.5,$SR_N$分别设为0.1、0.05以及0.01.

表2给出了所提算法与对比算法在不同测试序列前12个GOP上的评估结果. 从表中可以看出,所提算法重构帧的PSNR/SSIM显著超过了其他几种算法. 例如,在$SR_N$为0.01时,与Video-MH, 2sMHR和SSIM-InterF-GSR相比,所提算法在PSNR/SSIM指标上平均分别提升了8.09 dB/0.26, 7.55 dB/0.23和4.74 dB/0.12. 为了进一步分析所提算法的重构效果,图5对不同算法的重构帧进行了视觉质量对比. 从图中可以看到:Video-MH算法的重构帧存在明显的块效应,丢失了大量结构和细节信息;2sMHR算法的重构帧相对于Video-MH略微平滑,但仍旧没有完整恢复出运动目标的基本轮廓;SSIM-InterF-GSR算法则恢复出了部分前景目标的基本结构,但却丢失了部分细节信息,甚至产生了虚假的纹理;而本文算法在平滑区域和细节部分均有较好的重构效果. 实验结果表明本文所提出的TAFA-Net能够有效地利用时域相关性提供高质量的对齐帧,而多帧信息融合则能够对重构质量进行进一步增强.

**表2**　JDR-TAFA-Net与迭代优化VCS算法重构PSNR(dB)/SSIM对比

| 算法 | Coastguard | Football | Hall | Ice | Soccer |
|---|---|---|---|---|---|
| $SR_K$=0.5, $SR_N$=0.1 | | | | | |
| Video-MH[2] | 29.83 / 0.85 | 26.25 / 0.72 | 31.85 / 0.95 | 30.07 / 0.92 | 28.75 / 0.84 |
| 2sMHR[3] | 30.17 / 0.86 | 26.77 / 0.74 | 32.24 / 0.95 | 30.92 / 0.94 | 29.71 / 0.86 |
| SSIM-InterF-GSR[5] | 30.25 / 0.87 | 27.22 / 0.76 | 34.46 / 0.97 | 31.74 / 0.95 | 30.34 / 0.87 |
| JDR-TAFA-Net | **32.98 / 0.93** | **29.68 / 0.86** | **38.25 / 0.98** | **36.23 / 0.98** | **34.79 / 0.93** |
| $SR_K$=0.5, $SR_N$=0.05 | | | | | |
| Video-MH[2] | 28.14 / 0.79 | 24.59 / 0.63 | 30.97 / 0.94 | 28.42 / 0.90 | 26.81 / 0.77 |
| 2sMHR[3] | 28.68 / 0.81 | 25.12 / 0.66 | 31.37 / 0.94 | 29.08 / 0.91 | 27.53 / 0.80 |
| SSIM-InterF-GSR[5] | 28.09 / 0.81 | 25.57 / 0.68 | 33.21 / 0.96 | 28.80 / 0.92 | 27.59 / 0.80 |
| JDR-TAFA-Net | **32.23 / 0.91** | **27.78 / 0.79** | **37.26 / 0.98** | **34.83 / 0.97** | **32.25 / 0.90** |
| $SR_K$=0.5, $SR_N$=0.01 | | | | | |
| Video-MH[2] | 21.07 / 0.44 | 20.09 / 0.43 | 23.43 / 0.78 | 21.96 / 0.73 | 21.19 / 0.52 |
| 2sMHR[3] | 21.57 / 0.47 | 20.76 / 0.46 | 24.04 / 0.81 | 22.33 / 0.75 | 21.78 / 0.56 |
| SSIM-InterF-GSR[5] | 25.09 / 0.68 | 22.96 / 0.54 | 28.44 / 0.93 | 24.57 / 0.82 | 23.44 / 0.63 |
| JDR-TAFA-Net | **30.45 / 0.88** | **24.78 / 0.63** | **35.39 / 0.98** | **30.03 / 0.94** | **27.56 / 0.78** |

(*a*) GT　(*b*) Video-MH[2]　(*c*) 2sMHR[3]　(*d*) SSIM-InterF-GSR[5]　(*e*) JDR-TAFA-Net

图5　不同算法重构帧视觉质量对比(Ice序列第34帧,$SR_N$=0.01)

### 3.2.2 与基于深度学习的VCS算法重构性能对比

本小节将JDR-TAFA-Net与4种深度学习VCS重构算法进行对比，包括CSVideoNet[6]、2sRER-VGSR-Net[7]、PRCVSNet[8]和STM-Net[9]. 实验过程中，GOP大小设为10，$SR_K$设为0.2，$SR_N$分别设为0.037、0.018以及0.009.

表3给出了不同算法在UCF-101测试集上的比较结果. 由表可知，在所有采样率下，JDR-TAFA-Net均取得了最高的重构PSNR/SSIM. 例如，在$SR_N$为0.037时，与CSVideoNet、2sRER-VGSR-Net、PRCVSNet以及STM-Net相比，JDR-TAFA-Net重构帧的平均PSNR/SSIM分别提高了6.27 dB/0.13、1.62 dB/0.05、2.05 dB/-以及0.64 dB/0.01.

**表3** JDR-TAFA-Net与深度学习VCS算法重构PSNR(dB)/SSIM对比

| 算法 | $SR_N$=0.037 | $SR_N$=0.018 | $SR_N$=0.009 |
|---|---|---|---|
| CSVideoNet[6] | 26.87 / 0.81 | 25.09 / 0.77 | 24.23 / 0.74 |
| 2sRER-VGSR-Net[7] | 31.52 / 0.89 | 29.87 / 0.86 | 28.60 / 0.83 |
| PRCVSNet[8] | 31.09 / - | 28.93 / - | 26.86 / - |
| STM-Net[9] | 32.50 / 0.93 | 31.14 / **0.91** | 29.98 / **0.89** |
| JDR-TAFA-Net | **33.14 / 0.94** | **31.63 / 0.91** | **30.33 / 0.89** |

为了对比各算法的运行速度，表4给出了它们在UCF-101测试集上平均每帧的重构时间. 从表中可以看出，JDR-TAFA-Net取得了与2sRER-VGSR-Net相当的性能，而略多于CSVideoNet和STM-Net两种算法，但仍可以实现实时重构. 实验结果表明，无论是在低采样率还是在较大时域间隔条件下，所提算法都能够快速重构出高质量的视频帧.

**表4** JDR-TAFA-Net与深度学习VCS算法重构时间对比 单位：s

| 算法 | $SR_N$=0.037 | $SR_N$=0.018 | $SR_N$=0.009 |
|---|---|---|---|
| CSVideoNet[6] | 0.0094 | **0.0085** | **0.0080** |
| 2sRER-VGSR-Net[7] | 0.0152 | 0.0153 | 0.0156 |
| PRCVSNet[8] | - | - | - |
| STM-Net[9] | **0.0087** | 0.0087 | 0.0086 |
| JDR-TAFA-Net | 0.0169 | 0.0169 | 0.0169 |

## 3.3 泛化性能测试

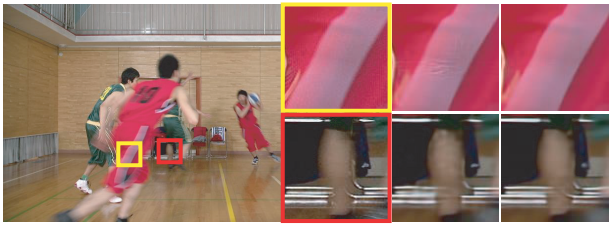
现有VCS文献通常只测试算法在CIF格式或QCIF格式数据集上的性能. 然而在实际应用中，低分辨率的图像已无法满足人们的日常需求. 本小节将通过实验说明，尽管所提算法在单一尺寸的固定数据集(UCF-101)上进行训练，却在不同尺寸、不同场景的测试集上有着较好的泛化性能. 本小节选用3个分辨率为1920×1080的高清视频序列[22]BasketballDrive、Cactus和ParkScene作为测试集. 实验设置与3.2.2节相同.

表5给出了JDR-TAFA-Net与STM-Net在3个高清测试序列前12个GOP上的量化评估结果. 由表可知，

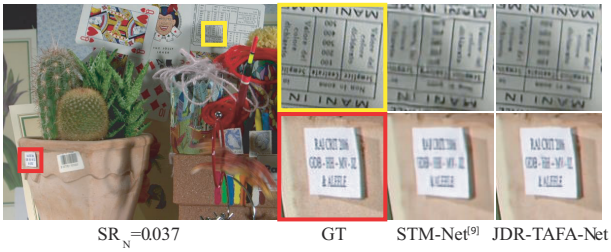
**表5** JDR-TAFA-Net与STM-Net在高清序列上重构PSNR(dB)/SSIM对比

| 算法 | BasketballDrive | Cactus | ParkScene |
|---|---|---|---|
| $SR_K$=0.2, $SR_N$=0.037 | | | |
| STM-Net[9] | 33.48 / 0.87 | 32.01 / **0.88** | 32.57 / 0.88 |
| JDR-TAFA-Net | **33.84 / 0.88** | **32.62 / 0.88** | **32.71 / 0.89** |
| $SR_K$=0.2, $SR_N$=0.018 | | | |
| STM-Net[9] | 31.01 / 0.85 | 31.04 / 0.86 | 31.38 / 0.86 |
| JDR-TAFA-Net | **31.87 / 0.86** | **31.57 / 0.87** | **32.04 / 0.88** |
| $SR_K$=0.2, $SR_N$=0.009 | | | |
| STM-Net[9] | 29.16 / **0.81** | 29.83 / 0.83 | 30.07 / 0.83 |
| JDR-TAFA-Net | **29.46 / 0.81** | **29.95 / 0.84** | **30.33 / 0.84** |

JDR-TAFA-Net与STM-Net在高清数据集上均有不错的重构表现，但JDR-TAFA-Net凭借注意力机制远距离的建模能力取得了相对更优的重构效果. 图6对一些重构帧的视觉质量进行了对比，为方便起见，实验过程中只对YCbCr色彩空间中的Y通道进行重构，展示时则对Y通道进行替换. 从图中可以看出，STM-Net重构帧的部分区域过于平滑，且产生了一些伪影. 相比之下，本文所提算法则能够恢复更多细节信息，得到更加清晰的重构帧. 实验结果表明，本文所提算法具有较好的泛化能力，适用于不同类型的视频数据.




(a) BasketballDrive序列第67帧




(b) Cactus序列第42帧

图6 不同算法高清序列重构帧视觉质量对比

## 3.4 消融实验

本小节将通过消融实验验证所提算法中不同模块的有效性，包括TAFA-Net、多帧信息融合网络以及观测损失. 除$SR_N$设为0.037外，其他实验设置与3.2.2节相同.

（1）TAFA-Net

TAFA-Net主要包含4个操作:多尺度特征提取(Muti-scale Feature Extraction, MFE),特征迁移(Feature Transfer, FT),特征融合(Feature Fusion, FF),以及后处理(Post Processing, PP). 本小节以UAST-Net作为Base网络,在此基础上逐步添加FT-1(单阶段FT)、FT-2(两阶段FT)、FF、PP以及MFE来研究各模块对重构性能的影响. 在没有MFE的情况下,采用相同个数的级联卷积进行特征提取.

表6给出了离当前帧时域距离较近的对齐帧的PSNR/SSIM. 从表中可以看出,在①的算法设置下,尽管只进行了单阶段的特征迁移,重构性能相对于Base网络却有了大幅提升,而在②中进行额外一次特征迁移操作后,性能提升却并不明显. 这一方面验证了注意力机制在时域对齐中的有效性,另一方面则说明了UCF-101测试序列的运动情况并不是很复杂,以致本文所提算法可以较为容易地建立当前帧特征与参考帧特征的依赖关系,进行准确的运动补偿. 在③和④的算法设定下,重构PSNR/SSIM也有一定的增益,这说明软注意力模块能够根据迁移特征的相关程度对其进行加强或抑制,而后处理模块则能够扩大感受野,捕获更多的结构信息. ⑤与④的差异在于⑤通过特征金字塔进行了多尺度的特征对齐,由实验结果可知,这种方式不但能够减小计算量,还具有更优的重构性能,说明不同粒度的特征对齐更容易处理复杂运动序列.

**表6　不同网络设置对TAFA-Net重构PSNR(dB)/SSIM的影响**

| 设置 | FT-1 | FT-2 | FF | PP | MFE | PSNR / SSIM |
|---|---|---|---|---|---|---|
| Base | × | × | × | × | × | 27.75 / 0.82 |
| ① | √ | × | × | × | × | 31.82 / 0.91 |
| ② | × | √ | × | × | × | 32.04 / 0.92 |
| ③ | × | √ | √ | × | × | 32.14 / 0.93 |
| ④ | × | √ | √ | √ | × | 32.26 / 0.93 |
| ⑤ | × | √ | √ | √ | √ | 32.47 / 0.93 |

（2）多帧信息融合网络

如前文所分析,在复杂运动场景下,TAFA-Net输出的对齐帧可能达不到理想的效果,因此本文设计了基于自编码器架构的融合网络来自适应提取已有重构帧的信息,得到更高质量的重构帧. 如表7所示,在TAFA-Net时域对齐的基础上,多帧信息融合网络可以显著提升视频帧的重构质量,这一方面说明了复杂运动场景下建模时域依赖关系的难度,另一方面则验证了本文所设计融合网络的有效性.

（3）观测损失

观测损失用来约束重构帧的观测值与原始视频帧的观测值尽可能地相似. 表7给出了观测损失对重构性能的影响. 可以看到,与单纯采用像素损失相比,加上观测损失后重构帧的PSNR/SSIM可以提升0.14 dB/0.01,这说明观测损失可以恢复更多细节信息,得到视觉质量更高的重构帧.

**表7　融合网络及观测损失对重构PSNR(dB)/SSIM的影响**

| Base | TAFA-Net | 融合网络 | 观测损失 | PSNR / SSIM |
|---|---|---|---|---|
| √ | × | × | × | 27.75 / 0.82 |
| √ | √ | × | × | 32.47 / 0.93 |
| √ | √ | √ | × | 33.00 / 0.93 |
| √ | √ | √ | √ | 33.14 / 0.94 |

## 4　结论

本文首先提出了具有良好可解释性的时域注意力特征对齐网络TAFA-Net,能够在复杂运动及较远时域距离的条件下,通过多尺度的特征迁移和特征融合生成高质量的对齐帧. 在此基础上,提出了联合深度重构网络JDR-TAFA-Net实现视频压缩感知中非关键帧的高质量重构. 仿真实验结果表明,与现有算法相比,本文所提算法可以在保持实时重构的条件下,在重构质量上取得显著提升. 未来工作中,将考虑TAFA-Net在其他视频任务上的应用.

**参考文献**


[1] DONOHO D L. Compressed sensing[J]. IEEE Transactions on Information Theory, 2006, 52(4): 1289-1306.

[2] TRAMEL E W, FOWLER J E. Video compressed sensing with multihypothesis[C]//2011 Data Compression Conference. Snowbird: IEEE, 2011: 193-202.

[3] OU W, YANG C, LI W, et al. A two-stage multihypothesis reconstruction scheme in compressed video sensing[C]//2016 IEEE International Conference on Image Processing. Phoenix: IEEE, 2016: 2494-2498.

[4] LI W, YANG C, MA L. A multihypothesis-based residual reconstruction scheme in compressed video sensing[C]//2017 IEEE International Conference on Image Processing. Beijing: IEEE, 2017: 2766-2770.

[5] 和志杰, 杨春玲, 汤瑞东. 视频压缩感知中基于结构相似的帧间组稀疏表示重构算法研究[J]. 电子学报, 2018, 46(3): 544-553.
HE Zhi-jie, YANG Chun-ling, TANG Rui-dong. Research on structural similarity based inter-frame group sparse representation for compressed video sensing[J]. Acta Electronica Sinica, 2018, 46(3): 544-553. (in Chinese)

[6] XU K, REN F. CSVideoNet: A real-time end-to end learning framework for high-frame-rate video compressive sensing[C]//2018 IEEE Winter Conference on Applications of Computer Vision. Lake Tahoe: IEEE, 2018: 1680-1688.

[7] 禤韵怡, 杨春玲. 基于帧间组稀疏的两阶段递归增强视频压缩感知重构网络[J]. 电子学报, 2021, 49(3): 435-442.
XUAN Yun-yi, YANG Chun-ling. Two-stage recursive enhancement reconstruction based on video inter-frame group sparse representation in compressed video sensing [J]. Acta Electronica Sinica, 2021, 49(3): 435-442. (in Chinese)

[8] LING X, YANG C, PEI H. Compressed video sensing network based on alignment prediction and residual reconstruction[C]//2020 IEEE International Conference on Multimedia and Expo. London: IEEE, 2020: 1-6.

[9] WEI Z, YANG C, XUAN Y. Efficient video compressed sensing reconstruction via exploiting spatial-temporal correlation with measurement constraint[C]//2021 IEEE International Conference on Multimedia and Expo. Shenzhen: IEEE, 2021: 1-6.

[10] DAI J, QI H, XIONG Y, et al. Deformable convolutional networks[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 764-773.

[11] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//European Conference on Computer Vision. Virtual: Springer, 2020: 213-229.

[12] CHEN H, WANG Y, GUO T, et al. Pre-trained image processing transformer[C]//2021 IEEE Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 12299-12310.

[13] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//2021 IEEE International Conference on Computer Vision. Virtual: IEEE, 2021: 10012-10022.

[14] HUANG Z, WANG X, HUANG L, et al. CCNet: Criss-cross attention for semantic segmentation[C]//2019 IEEE International Conference on Computer Vision. Seoul: IEEE, 2019: 603-612.

[15] SHI W, JIANG F, ZHANG S, et al. Deep networks for compressed image sensing[C]//2017 IEEE International Conference on Multimedia and Expo. Hong Kong: IEEE, 2017: 877-882.

[16] 裴翰奇, 杨春玲, 魏志超, 曹燕. 基于SPL迭代思想的图像压缩感知重构神经网络[J]. 电子学报, 2021, 49(6): 1195-1203.
PEI Han-qi, YANG Chun-ling, WEI Zhi-chao, CAO Yan. Image compressive sensing reconstruction network based on iterative SPL theory[J]. Acta Electronica Sinica, 2021, 49(6): 1195-1203. (in Chinese)

[17] GAN L. Block compressed sensing of natural images[C]//2007 International Conference on Digital Signal Processing. Cardiff: IEEE, 2007: 403-406.

[18] WANG P, CHEN P, YUAN Y, et al. Understanding convolution for semantic segmentation[C]//2018 IEEE Winter Conference on Applications of Computer Vision. Lake Tahoe: IEEE, 2018: 1451-1460.

[19] ARBELAEZ P, MAIRE M, FOWLKES C, et al. Contour detection and hierarchical image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 33(5): 898-916.

[20] SOOMRO K, ZAMIR A R, SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild [EB/OL]. [2022-04-01]. https://arxiv.org/abs/1212.0402.

[21] KINGMA D P, BA J. Adam: A method for stochastic optimization[EB/OL]. [2022-04-01]. https://arxiv.org/abs/1412. 6980.

[22] SULLIVAN G J, OHM J R, HAN W J, et al. Overview of the high efficiency video coding(HEVC) standard[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 22(12): 1649-1668.

## 作者简介

**魏志超** 男,1996年出生,河南禹州人. 现为华南理工大学电子与信息学院硕士研究生. 主要研究方向为视频压缩感知.
E-mail: zcwei2306@outlook.com

**杨春玲(通讯作者)** 女,1970年出生,河南新乡人. 现为华南理工大学电子与信息学院博士生导师. 主要研究方向为图像/视频压缩编码、图像质量评价.
E-mail: eeclyang@scut.edu.cn