# 基于帧循环网络的视频超分辨率技术

刘 佳,安鹤男,李 蔚,张昌林,涂志伟
(深圳大学 电子与信息工程学院,广东 深圳 518061)

**摘 要**:对比于单张图像超分辨,视频图像超分辨率技术需要对输入的连续时间序列图像进行融合、对齐等处理。基于帧循环的视频超分辨率网络共分为三部分:(1)帧序列对齐网络提取图像特征,并将邻居帧对齐到中心帧;(2)帧融合网络将对齐完成的帧进行融合,使用邻居帧的信息补充中心帧信息;(3)超分辨网络将融合完成的图像放大,得到最终的高清图像。实验表明,与现有算法相比,基于帧循环网络的视频超分辨率技术产生图像更为锐利,质量更高。

**关键词**:视频;超分辨;深度学习
**中图分类号**:TN919.8;TP183 **文献标识码**:A DOI:10.16157/j.issn.0258-7998.200051

**中文引用格式**:刘佳,安鹤男,李蔚,等. 基于帧循环网络的视频超分辨率技术[J].电子技术应用,2020,46(9):43-46.
**英文引用格式**:Liu Jia,An Henan,Li Wei,et al. Video super-resolution based on frame recurrent network[J]. Application of Electronic Technique,2020,46(9):43-46.

# Video super-resolution based on frame recurrent network

Liu Jia,An Henan,Li Wei,Zhang Changlin,Tu Zhiwei
(College of Electronics and Information Engineering,Shenzhen University,Shenzhen 518061,China)

**Abstract**: Compared with single image super-resolution, video super-resolution needs to align and fuse time series images. This frame-recurrent-based video super-resolution network consists of three parts:(1)The frame sequence alignment network extracts the image features and aligns the neighbor frames to the center frame;(2)The frame fusion network fuses the aligned frames and supplements the center frame information with the neighbor frame information;(3)The super-resolution network enlarges the fused image to obtain the final high-definition image. Experiments show that, compared with existing algorithms, video super-resolution technology based on frame loop network produces sharper images and higher quality.

**Key words**: video;super-resolution;deep learning

## 0 引言

在现存硬件技术的基础上,通过现存图像序列或视频相邻进帧之间的时空信息互补,将低分辨率的图像序列或者视频重构为高分辨率的图像序列或视频,一直是数字图像处理领域内的一个重要分支。最初的视频超分辨被认为是图像超分辨领域的简单扩展,但是这些基于单张图片的超分辨技术不能提取视频相邻帧之间的互补信息和存在视频中的动作位移。由于评价函数的关系,这些技术处理完成的视频会导致伪影,观看感觉不连续。基于帧循环网络的视频超分辨方法正是针对上述问题提出,并在公开数据集上验证了模型的有效性。

图像超分辨不仅可以生成高质量的图像,还可以用作目标检测[1]、人脸识别[2]等任务的预处理步骤。深度学习方法的引入为图像超分辨领域带来新的发展[3]。

相比于单幅图像超分辨,视频超分辨可分为对齐、融合、重建3个步骤。对齐网络的结果会直接影响融合网络与重建网络的效果。早期,基于深度学习的视频超分辨方法[4]参考相邻视频帧之间的光流场扭曲邻居帧从而达到对齐的目的。然而,Xue Tianfan 等人[5]指出基于光流场的对齐方法并非视频超分辨的最优解,提出基于任务流的视频超分辨率方法;JO Y H 等人[6]提出了隐式运动补偿的方法规避流场的计算。

对于视频帧的融合,现有的方法大致可以分为两种,第一种是使用卷积层对所有帧进行早期融合[5],第二种是基于循环网络[7]逐帧进行融合。本文对长短期记忆网络结构进行改进并应用于视频帧融合,整体结构图如图1所示。

## 1 帧循环视频超分辨网络

视频图像超分辨经典模型可以表示为:

$$\boldsymbol{I}_i^{\mathrm{L}}=SKW_{\mathrm{o}\to i}\boldsymbol{I}_{\mathrm{o}}^{\mathrm{H}}+\boldsymbol{n}_i \tag{1}$$

其中,$\boldsymbol{I}_{\mathrm{o}}^{\mathrm{H}}$ 表示高清图像,$K$ 和 $S$ 分别是下采样模糊和抽取影响因子,$W_{\mathrm{o}\to i}$ 是从第0帧到第 $i$ 帧扭曲的变形运算符,$\boldsymbol{I}_i^{\mathrm{L}}$ 表示输入图像,$\boldsymbol{n}_i$ 是第 $i$ 帧的加性噪声。

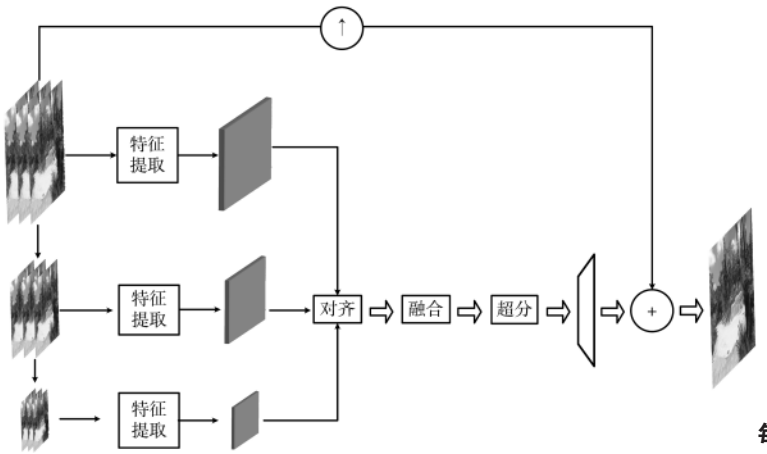



图 1 整体结构图

### 1.1 视频帧对齐网络

根据邻居帧与中心帧的运动状况，将邻居帧对齐到中心帧，是利用邻居帧信息提升中心帧图像超分辨质量的关键步骤。对齐的效果会直接影响到后续的处理过程。

本文提出了一种基于 STN 网络(Spatial Transformer Networks)[8]的金字塔结构对齐网络，具体结构如图 2 所示。

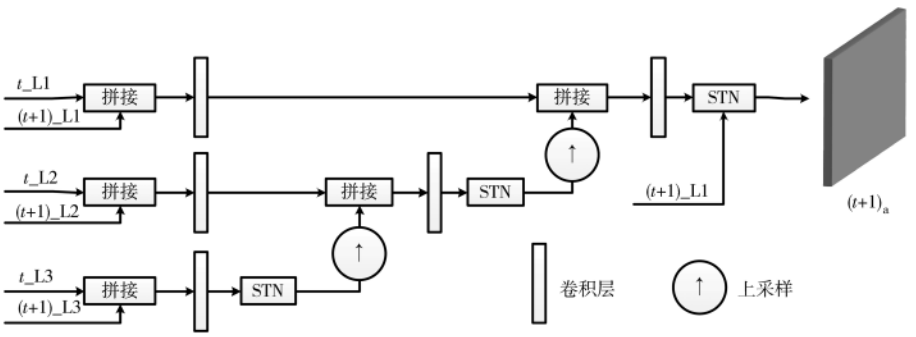



图 2 视频帧对齐网络

对齐网络采用中心帧与某一邻居帧作为输入。所有输入图像会经历两次下采样过程，中间结果保留作为特征提取通道的输入，提取大(特征 L1)、中(特征 L2)、小(特征 L3)3 种尺寸的特征。

相邻帧的特征数据会基于不同的尺寸分别拼接在一起用于提取特征流指导 STN 网络对齐邻居帧。对齐后的特征依次向上拼接，最终在 L1 特征层完成对齐。

### 1.2 视频帧融合网络

LSTM(Long Short Term Memory networks)网络[9]在自然语言处理领域拥有不错的表现，现阶段的方法证明了 LSTM 网络在处理时间连续信息时的有效性。

但是自然语言处理领域关注上下文信息与当前节点的联系，并由上下文信息推导当前节点的信息。

然而，在视频超分辨领域，更加关注上下文对当前节点的补充信息，而非每个节点之间的联系。为此有必要对 LSTM 网络进行改造，使用新的结构应对视频超分辨任务，改进后的 LSTM 网络如图 3 所示。

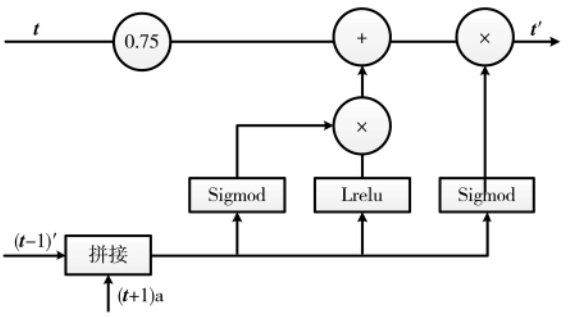



图 3 视频帧融合网络

每次融合将中心帧 $t$ 作为输入可以确保重建网络获取的主体信息来自中心帧，整体结构分为 3 个支路。

$$C_t=t\times 0.75 \tag{2}$$

式(2)可获取用于重建网络的主体信息，$C_t$ 表示在中心帧 $t$ 中获取的主要留存信息。

$$i_t=\sigma(w_i * [(t-1)':(t+1)_a]+b_i) \tag{3}$$

式(3)决定在邻居帧提取的信息在整个重构网络信息中的比重 $i_t$，$(t+1)_a$ 为对其完成的邻居帧，$(t-1)'$ 表示上一时刻网络输出重构信息，$w_i$ 为参数矩阵，$b_i$ 为偏差，[:] 为拼接操作，$*$ 为卷积操作，$\sigma$ 为激活函数 sigmod 函数。

$$C_t'=\text{lrelu}(w_c * [(t-1)':(t+1)_a]+b_c) \tag{4}$$

式(4)决定在邻居帧 $(t+1)_a$ 获取对当前帧 $t$ 的补充信息 $C_t'$，$w_c$ 为参数矩阵，$b_c$ 为偏差，lrelu 为激活函数 LReLu 函数。

$$O_t=\sigma(w_o * [(t-1)':(t+1)_a]+b_o) \tag{5}$$

式(5)对融合完成的信息进行筛选，这个比重参数为 $O_t$，$w_o$ 为参数矩阵，$b_o$ 为偏差。

针对邻居帧 $(t+1)_a$ 最终融合完成后输出信息为 $t'$，计算公式为：

$$t'=O_t\cdot(C_t+i_t\cdot C_t') \tag{6}$$

### 1.3 图像重建网络

图像重建网络主体是密集网络，分两个阶段将图像尺寸提升 4 倍。图像重建网络如图 4 所示。

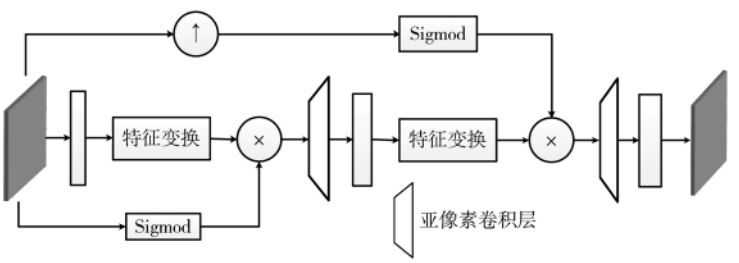



图 4 图像重建网络

特征变换块主体是密集块，最后一层为可变形卷积层，图像重建网络中两个阶段的模型参数共享，这个过程中引入注意力机制帮助网络提取图像数据中的高频细节，弥补生成图像平滑的缺点。

### 1.4 损失函数

损失函数共分为两部分：

(1)为引导对齐网络，保证处理完成的邻居帧与中心帧在特征层面上对齐，仅在训练过程中将对齐网络生成$(t+1)_a$与特征提取网络提取中心帧特征$t_a$做均方差损失，如式(7)所示：

$$\text{loss}_1=L_{\text{mse}}(t_a,(t+1)_a) \tag{7}$$

式中，$L_{\text{mse}}$表示均方差损失函数。

(2)将生成图像与真实高清图像对比，优化网络输出结果不断接近真实图像，第二部分损失函数为式(8)，其中$f(x,\theta)$表示生成图像，$\theta$代表网络参数，$y$代表真实高清图像。

$$\text{loss}_2=L_{\text{mse}}(f(x,\theta),y) \tag{8}$$

整体损失函数为：

$$\text{loss}=0.4\text{loss}_1+\text{loss}_2 \tag{9}$$

## 2 实验结果与分析

为评估网络性能，本文与其他3种方法在同一数据集训练后作对比，评价指标包括峰值信噪比(PSNR)与结构相似性指数(SSIM)。3种方法分别为：双三次插值算法(BICUBIC)、基于细节的深度视频超分辨算法(DRVSR)[10]、基于卷积神经网络的视频超分辨算法(VSRnet)[11]。

### 2.1 训练数据集

REDS数据集(REalistic and Diverse Scenes dataset)是NTIRE19竞赛中新提出的高质量(720P)视频数据集，包含240个训练片段(每个片段含有100个连续帧)。与现有数据集相比，REDS中的视频运动更为复杂，恢复难度更大。训练过程中将高质量720P视频下采样两次退化至180P作为网络的输入数据。

### 2.2 训练配置

训练使用服务器含有4张1080ti显卡，系统为CentOS7，设置最大迭代次数为540 000，小批量训练数据为2，采用Pytorch框架，学习率初始化为0.000 4，$\beta_1=0.9$，$\beta_2=0.99$梯度下降优化算法为ADAM[12]。

在每次迭代中，从一个序列中随机的对连续3帧进行采样，并随机裁剪140×140图像区域作为训练输入。因此，相应的真实图像裁剪对应的480×480区域。

为防止因LSTM网络而产生的梯度爆炸，本文在训练过程中引入梯度裁剪技术。网络中的上采样与下采样操作全部使用双三次插值算法。

### 2.3 测试结果

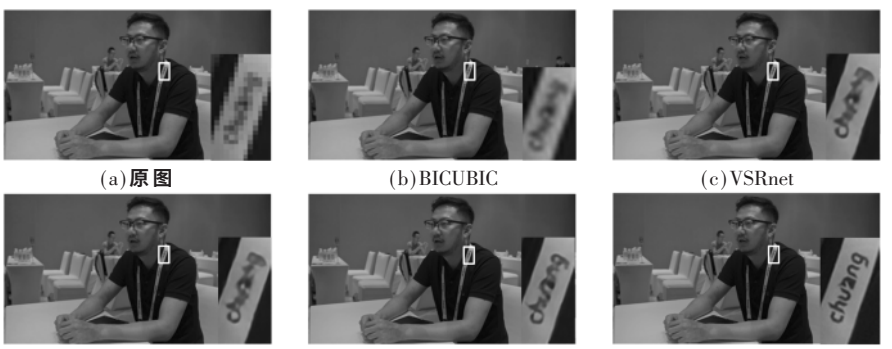
在计算PNSR与SSIM时，本文使用作者提供的网络或者复现的网络在REDS数据集上训练后测试验证集的指标，采取OpenCV保存的图片作为评估指标时的输入，而不采用网络的输出结果。

表1比较了PSNR与SSIM在REDS数据集上本文方法与3种比较方案的数值差异。从表1的数据可以看出，本文方法达到了更为先进的性能。

表1 验证集测试指标

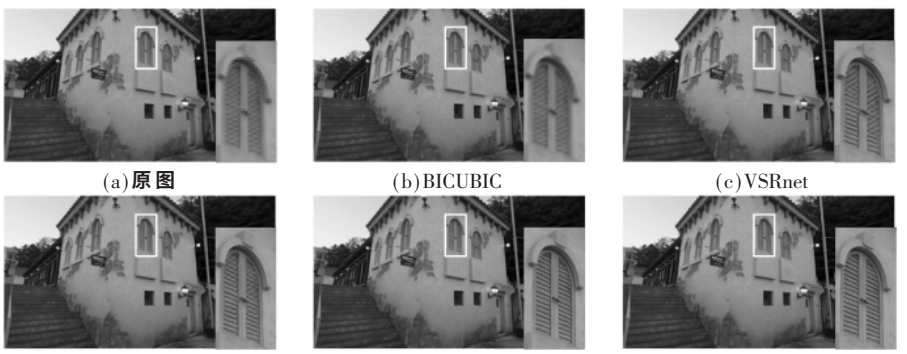
| 方法 | BICUBIC | VSRnet | DRVSR | 本文 |
|---|---|---|---|---|
| PSNR/SSIM | 28.39/0.798 | 29.67/0.838 | 29.86/0.843 | 30.02/0.845 |

图5、图6展示了不同方法的视觉比较结果(人像源自阿里巴巴优酷视频增强和超分辨率挑战赛数据集)。从图中可以看出基于深度学习的方法DRVSR与FRVSR恢复效果全部超越了基于插值的方式BICUBIC。但这两种方法恢复的图像过于平滑。本文相对其他3种方法产生的图像更加锐利，更加接近真实图像。



(a)原图 (b)BICUBIC (c)VSRnet (d)DRVSR (e)本文 (f)真实图像

图5 不同方法恢复结果对比1



(a)原图 (b)BICUBIC (c)VSRnet (d)DRVSR (e)本文 (f)真实图像

图6 不同方法恢复结果对比2

## 3 结论

本文提出了一种可以端到端训练的深度学习视频超分辨网络。视频帧对齐网络能够根据大中小不同尺度特征将邻居帧对齐到中心帧。改进的LSTM网络能够识

别邻居帧中对超分辨有用的信息拟合到中心帧。与现有的方法相比，本文的框架达到更为先进的性能。

**参考文献**


[1] KRISHNA H, JAWAHAR C V. Improving small object detection[C]. Asian Conference on Pattern Recognition, 2017: 340-345.

[2] FOOKES C, LIN F, CHANDRAN V, et al. Evaluation of image resolution and super resolution on face recognition performance[J]. Journal of Visual Communication and Image Representation, 2012, 23(1): 75-93.

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

[4] CABALLERO J, LEDIG C, ANDREW A, et al. Real-time video super-resolution with spatio-temporal networks and motion compensation[C]. 2017 IEEE Conference on CVPR, 2017.

[5] Xue Tianfan, Chen Baian, Wu Jiajun, et al. Video enhancement with task-oriented flow[J]. International Journal of Computer Vision, 2019, 127: 1106-1125.

[6] JO Y H, OH S W, KANG J, et al. Deep video super-resolution network using dynamic upsampling lters without explicit motion compensation[C]. 2018 IEEE/CVF Conference on CVPR. IEEE, 2018.

[7] HARIS M, SHAKHNAROVICH G, UKITA N. Recurrent back-projection network for video super-resolution[C]. 2019 IEEE/CVF Conference on CVPR, 2019.

[8] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. Spatial transformer networks[C]. Advance in NIPS, 2015.

[9] XINGJIAN S, CHEN Z, WANG H, et al. Convolutional lstm network: a machine learning approach for precipitation nowcasting[C]. Advance in NIPS, 2015: 802-810.

[10] TAO X, GAO H, LIAO R, et al. Detail revealing deep video super-resolution[C]. 2017 IEEE ICCV, 2017.

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

[12] KINGMA D P, BA J. Adam: a method for stochastic optimization[J]. arXiv preprint arXiv: 1412.6980, 2014.


（收稿日期：2020-01-17）


**作者简介：**

刘佳(1996-)，男，硕士研究生，主要研究方向：计算机图像处理。

安鹤男(1963-)，通信作者，男，副教授，硕士研究生导师，主要研究方向：计算机视觉、图形处理，E-mail：anhenan@szu.edu.cn。

李蔚(1996-)，男，硕士研究生，主要研究方向：计算机图像处理。


（上接第 42 页）


with deep learning: a review[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(11): 3212-3232.

[6] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[7] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016: 779-788.

[8] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]. IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017: 7263-7271.

[9] REDMON J, FARHADI A. YOLOv3: an incremental improvement[J]. arXiv preprint arXiv: 1804.02767, 2018.

[10] 冯国臣，陈艳艳，陈宁，等.基于机器视觉的安全帽自动识别技术研究[J].机械设计与制造工程，2015，44(10)：39-42.

[11] 刘晓慧，叶西宁.肤色检测和 Hu 矩在安全帽识别中的应用[J].华东理工大学学报(自然科学版)，2014，40(3)：365-370.

[12] 施辉，陈先桥，杨英.改进 YOLO v3 的安全帽佩戴检测方法[J].计算机工程与应用，2019，55(11)：213-220.

[13] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]. IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018: 7132-7141.

[14] PENG D, SUN Z, CHEN Z, et al. Detecting heads using feature refine net and cascaded multi-scale architecture[C]. International Conference on Pattern Recognition, Beijing, 2018: 2528-2533.

[15] PASZKE A, GROSS S, MASSA F, et al. PyTorch: an imperative style, high-performance deep learning library[C]. Advances in Neural Information Processing Systems, Vancouver, 2019: 8024-8035.


（收稿日期：2020-02-19）


**作者简介：**

刘欣(1985-)，男，硕士，助理研究员，国家安全评价师，主要研究方向：计算机技术、煤矿自动化及安全。

张灿明(1984-)，通信作者，男，硕士，助理研究员，主要研究方向：深度学习、煤矿自动化及安全，E-mail：zhangcm0103@126.com。

# 版权声明