文章编号：2095-2163(2023)12-0075-05 中图分类号：TP391 文献标志码：A

基于金字塔池化与注意力机制的视频去模糊算法

邹世奇，刘 洪
（贵州大学 大数据与信息工程学院，贵阳 550025）

摘 要：视频拍摄中，常因相机抖动或拍摄对象移动造成视频模糊，给后续的信息获取及视频处理带来干扰。为了更好地利用视频帧的全局上下文信息，本文提出了一种基于金字塔池化和注意力机制的视频去模糊算法。在视频去模糊的复原网络中引入金字塔池化，利用不同尺度的池化获得更加全面的全局上下文信息；使用注意力机制加强对全局上下文信息的利用，以达到提升视频去模糊的效果。在 DVD 数据集上的实验结果表明，该算法能够有效地提升视频复原效果。

关键词：视频去模糊；金字塔池化；注意力机制

Video deblurring algorithm based on pyramid pooling and attention mechanism

ZOU Shiqi, LIU Hong
(College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China)

Abstract: The video is often blurred caused by camera shake or object motions during the exposure time, which interferes with subsequent information acquisition and video processing. In order to make better use of the global context information of video frame, a video deblurring algorithm based on pyramid pooling and attention mechanism is proposed. Pyramid pool is introduced in the video deblurring reconstruction networks to obtain more comprehensive global context information by different scales pooling. Then the attention mechanism is used to enhance the use of global context information to improve the effect of video deblurring. The experimental results on DVD dataset show that the algorithm can effectively improve the video restoration effect.

Key words: video deblurring; pyramid pooling; attention mechanism

0 引 言

视频拍摄设备，如手机、相机和监控器等，在视频拍摄期间，因设备抖动或者拍摄物体快速移动，会使得视频出现运动模糊。模糊的视频不利于对信息的获取以及对视频进行后续的处理。因此，如何有效去除视频中的模糊，得到清晰视频，是计算机视觉领域中一个重要的研究问题。

早期的视频去模糊方法通过搜寻相邻帧的清晰像素来替换目标帧的模糊像素，达到视频去模糊的效果。此类方法主要是依赖于一个观察现象，即一个视频中多帧模糊的视频帧，每一帧都具有不同程度的模糊，都会存在着相对清晰的部分。这类算法对视频去模糊具有一定的效果，但过于依赖于原视频帧的模糊程度。当每一帧的模糊程度都较大时，则恢复效果不佳。

受传统单帧图像去模糊方法的启发，一些学者通过设置先验约束，求解能量函数，从而得到清晰的视频[1]。这类算法将运动信息表示为光流，通过先验信息来对复原帧和光流进行约束，然而对复原帧和光流的约束通常导致能量函数难以求解。

随着神经网络和深度学习的不断发展，学者们引入卷积神经网络（Convolutional Neural Networks，CNN），设计深度神经网络在大量的数据上学习模糊视频和清晰视频之间的映射关系。Kim 等[2]提出了一种具有动态时间混合层的时空递归深度神经网络，用于视频帧的恢复；Zhang 等[3]使用时空 3D 卷积构建深度神经网络以获取更多的时空信息，求解出清晰的视频。该类方法直接利用相邻帧进行复原，未对相邻帧进行对齐，导致复原视频中可能存在伪影。为此，Chen 等[4]使用相邻帧的运动信息实现帧对齐，并以此设计视频去模糊网络；Wang 等[5]提出了金字塔、级联和可变形对齐模块来进行运动估计，以帮助去除视频中的运动模糊。然而，这类算法

基金项目：贵州省科学技术基金（黔科合基础[2019]1063 号）；贵州大学引进人才科研项目（贵大人基合同字(2017)14 号）。

作者简介：邹世奇(1999-)，男，硕士研究生，主要研究方向：图像处理。

通讯作者：刘 洪(1978-)，女，博士，副教授，硕士生导师，主要研究方向：信号处理、图像处理。Email：laniliu@sina.com

收稿日期：2023-09-30

并没有充分地利用视频帧的上下文信息。

金字塔池化通过多个尺度的池化获得更大的感受野,从而提取到更多的上下文信息来帮助视频去模糊,注意力机制能够引导网络关注学习视频去模糊中的重要信息。本文将金字塔池化与注意力机制相结合,利用金字塔池化对提取得到的特征进行 4 个尺度的池化,将池化得到的特征进行卷积、上采样,输出为 4 个具有上下文信息的全局特征,并与原来的特征进行特征融合;再引入注意力机制,对融合的特征赋予权重;最终经过重构网络恢复得到清晰的视频。

1 算法模型

本文通过一个基于金字塔池化与注意力机制的深度网络学习模糊帧与清晰帧之间端到端的映射关系,利用当前模糊帧与相邻模糊帧之间的有用信息来复原当前模糊帧。本文视频去模糊算法框架如图 1 所示。输入连续的三帧模糊视频帧,通过光流估计网络提取中间帧与相邻两帧之间的光流信息,利用所提取到的光流对相邻视频帧进行变换,实现帧对齐;将帧对齐后的三帧视频帧输入进复原网络中,复原得到中间视频帧的清晰帧。

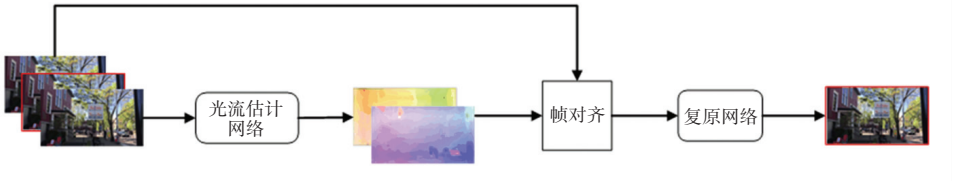

图 1 本文视频去模糊算法框架

Fig. 1 The framework of the proposed video deblurring algorithm

1.1 光流估计网络

光流反映了相邻图像帧之间像素间的相对运动信息。光流估计是指利用图像序列中像素在时间域上的变化及相邻帧之间的相关性来找到两相邻帧像素间的对应关系。本文通过光流估计网络来估计输入相邻视频帧的光流信息,以提供视频帧复原所需的运动信息。光流估计网络采用的是 PWC 网络(CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume)[6]。给定任意两帧相邻视频帧 I_i 和 I_{i+1},I_i 到 I_{i+1} 的光流 $f_{i\to i+1}$ 的求解公式(1):

$$f_{i\to i+1} = P(I_i, I_{i+1}) \tag{1}$$

其中,P 表示 PWC 光流估计网络。

1.2 复原网络

复原网络结构如图 2 所示。复原网络由特征提取网络、金字塔池化模块[7](Pyramid Pooling Module, PPM)、卷积块注意力模块[8](Convolutional Block Attention Module, CBAM)以及重构网络构成。

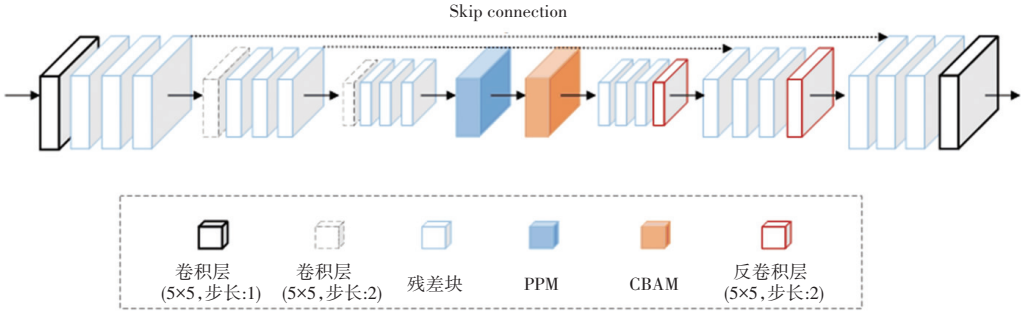

图 2 复原网络结构

Fig. 2 The structure of restoration network.

1.2.1 特征提取与重构网络

尺度递归网络(Scale-Recurrent Network, SRN)是基于多尺度递归的图像去模糊网络。该网络对图像进行降采样,将不同尺度的图像作为输入,“从粗到细”逐步递归恢复出清晰的图像,SRN 中的编码器解码器残差网络是恢复出清晰图像的关键[9]。在复原网络中引入了 SRN 中的编码器解码器残差网络作为本文算法中的特征提取网络和重构网络,对视频帧进行特征提取和特征重构。由于本文算法是单尺度的视频去模糊,所以没有选用原 SRN 网络中 ConvLSTM 模块。

特征提取网络负责将视频帧转换为具有较小空间大小和更多通道的特征图 F。特征提取网络由 3 个卷积块构成,第一个卷积块由一个步长为 1 的卷积层和 3 个残差块构成,并以 LeakyReLU 作为激活函数;另外两个卷积块由步长为 2 的卷积层和 3 个残差

块构成，以 LeakyReLU 作为激活函数。残差块由两个步长为 1 的卷积层和 LeakyReLU 激活函数构成。

重构网络负责对经由特征提取、金字塔池化以及注意力加权得到的特征 F_s 进行复原，得到清晰帧。重构网络由 3 个卷积块组成，其中两个卷积块是由 3 个残差块和一个步长为 2 的反卷积层以及 LeakyReLU 激活函数构成，最后一个卷积块是由 3 个残差块和一个步长为 1 的卷积层以及 LeakyReLU 激活函数构成。

1.2.2　金字塔池化模块

受感受野大小的限制，卷积层对特征实现的是局部感知处理，得到的是局部信息，缺少了全局上下文信息。为了提取到更多的全局上下文信息，本文在编码器与解码器之间引入了金字塔池化，以获得更大的感受野来提高对视频帧上下文全局信息的获取。

金字塔池化模块结构如图 3 所示，金字塔池化模块的输入为特征提取网络提取到的特征 F，对 F 进行 4 个尺度的平均池化，分别得到 1×1、2×2、3×3、6×6 4 个尺度大小的特征。将池化后的 4 个特征使用 1×1 卷积将其通道数减少到 1/4，再通过双线性插值进行上采样，将特征上采样至特征 F 大小，最后将 4 个上采样得到的特征和特征 F 在通道方向上拼接起来，得到一个新的特征 F_P，公式(2)：

$$F_p = N_p(F) \tag{2}$$

其中，N_P 表示金字塔池化模块。

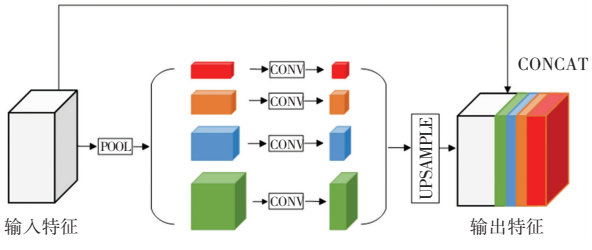

图 3　金字塔池化模块结构

Fig. 3　The structure of pyramid pooling module

1.2.3　卷积块注意力模块

本文采用 CBAM 引导对特征的通道和空间的学习，CBAM 由通道注意力模块和空间注意力模块构成，通道注意力模块引导网络关注学习“什么”更重要，空间注意力模块引导网络关注学习“哪里”更重要。

卷积注意力模块结构如图 4 所示。在注意力模块中，由金字塔池化模块得到的特征 F_P（长、宽以及通道数分别为 H、W、C）输入到通道注意力模块，对 F_P 的每个通道进行最大池化和平均池化，得到两个 $C \times 1 \times 1$ 大小的特征；再通过多层感知机，实现非线性变换；将两个特征对应相加，再经由激活函数输出得到每个通道的权重 M_c。将 M_c 与 F_P 相乘得到特征 F_c；将特征 F_c 输入进空间注意力模块中，在通道方向上求平均池化和最大池化，得到两个 $1 \times H \times W$ 大小的特征；将两个特征在通道方向上进行拼接，再经由激活函数输出得到每个像素的权重 M_s，将 M_s 与 F_c 相乘得到最终的特征 F_s。过程如式(3) ~ 式(6)：

$$M_c = N_c(F_p) \tag{3}$$

$$F_c = M_c * F_p \tag{4}$$

$$M_s = N_s(F_c) \tag{5}$$

$$F_s = M_s * F_c \tag{6}$$

其中，N_c 表示通道注意力模块，N_s 表示空间注意力模块。

1.3　损失函数

损失函数是机器学习和深度学习中用来衡量模型预测值和真实值之间误差的数值评估指标，本文使用 L1loss 来度量复原帧和真实视频帧之间的误差，损失函数定义如式(7)所示：

$$L = \sum_{n=1}^{N_V} \sum_{i=1}^{M} \| I_i^n - I^n{}_{gt,i} \|_1 \tag{7}$$

其中，N_V 表示视频的数量；M 代表视频帧的数量；I 表示复原帧；I_{gt} 表示真实视频帧。

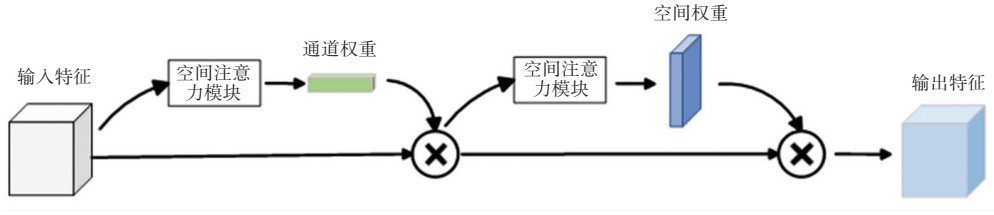

图 4　卷积块注意力模块结构

Fig. 4　The structure of convolutional block attention module

2 实验结果及分析

2.1 实验准备

2.1.1 数据集

本文采用 DVD 数据集来进行训练和评估。该数据集有 71 个视频,每个视频的平均运行时间为 3-5 s,61 个训练视频和 10 个测试视频。本文通过对训练视频帧随机裁剪以扩充数据集,随机裁剪的区域大小为 256×256。

2.1.2 实验环境及参数设置

本文使用预训练模型对 PWC 网络初始化,使用 Adam 优化器,设置参数$\beta_1 = 0.9$,$\beta_2 = 0.999$,训练批次为 8,初始学习率为0.000 001,复原网络的初始学习率为0.000 1,训练的总周期为 300,且在前 200 个训练周期结束之后,学习率呈线性衰减,设置为 0.000 001。硬件环境为:PG500-216 GPU 和 E5-2678 v3 CPU。

2.2 实验结果分析

本文采用峰值信噪比(Peak Signal-to-Noise Ratio, PSNR)和结构相似度(Structural Similarity Index Measurement, SSIM)来对复原结果进行客观评价。峰值信噪比是基于对应像素点之间的误差,即基于误差敏感的图像质量评价,给定复原视频帧 X 和真实视频帧 Y,其峰值信噪比如式(8)所示:

$$PSNR(X,Y) = 10\lg\left(\frac{255^2 N}{\sum_{n=1}^{N}(x_n - y_n)^2}\right) \tag{8}$$

其中,N 表示图像的总像素数;x_n 表示 X 的第 n 个像素值;y_n 表示 Y 的第 n 个像素值。

结构相似性分别从亮度、对比度、结构三方面度量图像相似性,式(9):

$$SSIM(X,Y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{9}$$

其中,μ_x,μ_y,σ_x,σ_y,σ_{xy} 分别是 X 的平均值,Y 的平均值,X 的标准差,Y 的标准差,以及 X 和 Y 的协方差;C_1,C_2 是用于保持稳定的常数。

本文方法参考了文献[9]的编码器解码器残差网络,所以将文献[9]中的 SRN 与本文方法进行了定量对比。为了验证加入的金字塔池化模块和卷积块注意力模块的有效性,同时做了消融实验,实验结果见表 1。

表 1 算法在 DVD 数据集上有效性对比

Table 1 Validity comparison of the algorithm on DVD dataset

方法	SRN	无 PPM、CBAM	只加入 PPM	只加入 CBAM	本文方法
PSNR	29.42	30.42	30.51	30.54	**30.61**
SSIM	0.875 6	0.898 0	0.899 5	0.899 9	**0.900 5**

根据表 1 所示,本文的方法 PSNR 和 SSIM 更高,分别加入 PPM 和 CBAM 的视频去模糊算法,在 PSNR 和 SSIM 上均有提高,证明了 PPM 和 CABM 的有效性。

此外,还将本文方法与 SRN 以及未加入 PPM 和 CBAM 的视频去模糊方法的去模糊效果进行了对比(如图 5 所示),为了便于观察细节,将原始模糊帧以及对比方法的去模糊结果的局部细节进行了放大。从图 5 可以看出使用文中方法复原后的视频帧,其纹理细节以及轮廓更加清晰,并且能够有效地抑制伪影。

将本文算法同当前主流的视频去模糊算法:结合动态时间混合层的时空递归网络(STRCNN+DTB)[2]、增强型可变形卷积视频恢复网络(EDVR)[5]、结合光流的去模糊网络(DBN+FLOW)[10]进行定量比较,对比结果见表 2,实验结果表明,文中方法拥有更高的 PSNR 和 SSIM,证明了本文算法的有效性。

表 2 不同算法在 DVD 数据集上的定量评估

Table 2 Quantitative evaluations on the DVD dataset for different algorithms

方法	原始视频	STRCNN+DTB	EDVR	DBN+FLOW	本文方法
PSNR	27.20	29.95	28.51	30.01	**30.61**
SSIM	0.812 3	0.869 2	0.863 7	0.887 7	**0.900 5**

（a）模糊帧　（b）SRN 方法结果　（c）无 PPM 和 CBAM 方法结果　（d）文本方法结果

图 5　在 DVD 数据集上的去模糊结果

Fig. 5　Deblurred results on the DVD dataset

3　结束语

本文提出了一种有效的视频去模糊算法，在较好地复原出视频帧的轮廓和细节的同时，也有效地抑制了伪影。使用金字塔池化为视频复原提供了更多的全局上下文信息；还加入了注意力机制引导网络关注学习视频复原所需的重要信息，加强了对视频复原有用信息的利用。将金字塔池化同注意力机制相结合，进一步提高模糊视频复原的效果。

参考文献

[1] PAN J, BAI H, TANG J. Cascaded deep video deblurring using temporal sharpness prior [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 3043-3051.

[2] HYUN KIM T, MU LEE K, SCHOLKOPF B, et al. Online video deblurring via dynamic temporal blending network [C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 4038-4047.

[3] ZHANG K, LUO W, ZHONG Y, et al. Adversarial spatio-temporal learning for video deblurring[J]. IEEE Transactions on Image Processing, 2018, 28(1): 291-301.

[4] CHEN H, GU J, GALLO O, et al. Reblur2deblur: Deblurring videos via self-supervised learning[C]//Proceedings of the 2018 IEEE International Conference on Computational Photography (ICCP). IEEE, 2018: 1-9.

[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: 1954-1963.

[6] SUN D, YANG X, LIU M Y, et al. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8934-8943.

[7] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2881-2890.

[8] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 3-19.

[9] TAO X, GAO H, SHEN X, et al. Scale-recurrent network for deep image deblurring[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8174-8182.

[10] SU S, DELBRACIO M, WANG J, et al. Deep video deblurring for hand-held cameras[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1279-1288.