一种低分辨率视频实例分割算法的研究

达 婷*，杨 靓

(西安微电子技术研究所 西安 710065)

【摘要】由于图像采集设备的限制和采集条件不佳等原因，在现实中很难获得高质量的图像，尤其是视频图像。现有的实例分割算法在低分辨率 (low resolution，LR) 视频中达不到理想效果。另一方面，现有复杂的实例分割模型很难直接应用于移动设备上。该文基于 MobileNet 建立了一种高效的轻量化实例分割模型。同时，针对低分辨率视频，提出一种改进的超分辨算法 (SCN) 作为预处理，并建立一种基于运动矢量预测帧间掩码的后处理算法。通过对比实验说明，该算法应用于低分辨率实时场景时中精度高，内存小，且易于移植到嵌入式平台中。

关 键 词 模型压缩； MobileNet； 运动矢量； 超分辨重建； 实例分割

中图分类号 TP183 **文献标志码** A doi:10.12178/1001-0548.2020075

An Algorithm for Instance Segmentation in Low-Resolution Video Sequences

DA Ting* and YANG Liang

(Xi'an Microelectronics Technology Institute Xi'an 710065)

Abstract Due to the limitations of image acquisition equipment and poor acquisition conditions, it is a challenge to obtain high-quality images in reality, especially for video images. However, the existing instance segmentation algorithms can hardly handle low-resolution (LR) videos. Moreover, since existing complicated instance segmentation models can be barely applied to mobile devices in practical applications. Accordingly, the proposed method develops an efficient and lightweight instance segmentation model built upon MobileNet. At the same time, an improved super-resolution coding-based network (SCN) algorithm for low-resolution video is proposed as preprocessing. In addition, the motion vector is employed as post-proposing to predict the inter-frame mask. Experimental results have demonstrated that the proposed algorithm could be easily transplanted to embedded platforms in LR real-time street view dataset thanks to its remarkably low memory cost and high precision.

Key words model compression; MobileNet; motion vector; super-resolution; words image segmentation

视频图像是生活中最常见的一种采集数据的方式，但是由于图像采集设备的限制和采集条件不佳等原因，很难获得高质量的视频图像。尤其在移动端实际应用中，比如视频监控、目标追踪、行为分析等。同时，实例分割作为人机交互的一个重要研究方向被广泛应用于各种计算机视觉应用场景中。一方面，现实环境中存在背景复杂、物体被遮挡，还有很多小目标的低分辨率的视频数据。另一方面，在实际应用中，我们既希望得到高准确率又同时要求算法轻量化，所以使得实例分割任务难度较大。

随着深度学习在计算机视觉方面的大规模运用，利用深度学习算法进行图像分割成为一种必然趋势。卷积神经网络 (convolution neural network, CNN)[1] 已经普遍应用在计算机视觉领域，并且取得了不错的效果，如深度残差网络 (residual neural network, ResNet)[2]。目标检测模型的提出，从 CNN[1]、R-CNN[3]、Fast R-CNN[4]、Faster R-CNN[5] 发展到 Mask R-CNN[6]。R-CNN(regions with CNN features) 是一种物体检测任务的经典算法，算法思路是从生成的候选区域中生成特征向量，然后使用支持向量机 (SVM)[7] 进行检测分类。Fast R-CNN 的提出基

收稿日期：2020－02－24；修回日期：2020－09－07
基金项目：国家“十三五”技术基础科研项目 (JSZL2017203B023)
作者简介：达婷 (1992−)，女，博士生，主要从事计算机视觉方面的研究. E-mail：dpting1222@163.com

于 R-CNN，同时结合了 SPPNet[8] 的特点，主要解决了模型速度慢和训练占用空间大的问题。Faster R-CNN 是通过 RPN[5] 生成区域建议框，并实现两个网络卷积层的特征共享，大大降低了计算复杂度，真正实现了端到端的目标检测。Mask R-CNN 可以增加不同的分支完成不同的任务，如实例分割、关键点检测等多种任务。由于算法的高精度性，已经成为目前最好的实例分割算法之一。

深度神经网络已经可以解决越来越复杂的难题。面对海量的数据，为了提升测试的精度，大家往往会选择通过增加网络的层数以及更为复杂的网络来提高模型精度。一方面，复杂的模型面临着内存不足的问题；另一方面，高精度、响应速度快、低延迟性，才能达到我们的应用需求。然而真实的应用场景中，如移动端或嵌入式设备 (摄像头、无人机等)，庞大而复杂的模型难以被直接应用。所以，嵌入式的平台由于硬件资源的限制，需要一种内存小、测试速度快、精度高的灵活便捷的模型。因此，研究小而高效的模型至关重要，也是目前的一大热点研究方向。同时，虽然高分辨 (high resolution, HR) 的视频图像已经有很多成熟的算法，但在实际应用时，所采集到的视频数据往往是低分辨率 (low resolution, LR) 的。另一方面，分割算法是一种像素级的分割技术，模型复杂性较高，相比目标检测和识别来说一直是难题中的难题，尤其是实例分割。对于低分辨率视频实时场景的实例分割模型，目前一直没有较为成熟的算法研究。

1 相关算法研究

目前有一些针对高分辨率视频的目标检测算法。文献 [9] 提出了一种类似于 Faster R-CNN 的算法。文献 [10] 提出 P3D，组合了 3 种不同的模块结构，得到 P3D ResNet。文献 [11] 以 C3D 网络为基础，对输入数据提取特征，生成候选时序，最后进行目标检测。针对于移动端检测算法的研究，文献 [12] 提出了一种改进 SSD 的轻量化算法用于小目标检测。文献 [13] 等提出了一种应用于嵌入式平台的 MTYOLO 算法。对于实例分割的研究，文献 [14] 将 Mask R-CNN 应用于无人驾驶汽车的研究。还有很多其他成熟的视频检测和分割算法[15-20]。然而，实例分割算法由于其高复杂性在嵌入式平台的研究还未成熟。

针对复杂模型的压缩优化算法，目前主要有两个方向：1) 对训练好的复杂模型进行压缩优化得到较小模型；2) 直接设计轻量化模型进行训练。模型压缩的目标是在保持模型性能的前提下降低模型大小，同时提升模型的测试速度。针对已存在复杂模型的压缩优化算法有剪枝 (Pruning)[21-23]、量化 (Quantization)[24-26]、Huffman[27] 编码。此类算法可以将模型进行优化，但是压缩程度范围较低。针对第二个方向，目前轻量级模型有 MobileNet V1[28]、MobileNet V2[29]、ShuffleNet V1[30]、ShuffleNet V2[31] 等。基于 MobileNet V1 轻量级模型由于更容易实现及保持模型性能的特点被广泛应用，所以本文所选择基于 MobileNet V1 设计实例分割模型。

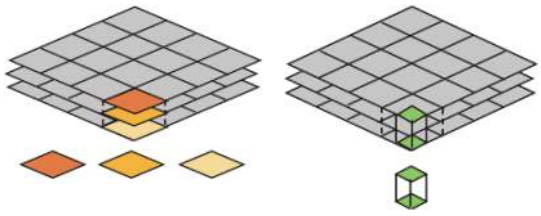
此外，图像超分辨 (super resolution, SR) 是一种由给定的低分辨率 (LR) 图像重建出它相应的高分辨率 (high resolution, HR) 图像的算法。目前图像超分辨重建算法主要有以下 4 个方向：1) 基于经典稀疏编码[32-34]；2) 基于邻域回归的算法[35-38]；3) 基于深度学习方向的重构算法，这也是近年来很流行的方向[39-43]；4) 利用 GAN 网络的算法[44-45]。SCN (sparse coding based network)[46] 算法是一种基于稀疏编码网络的方法，该算法结合稀疏先验和深度学习的优点，且重建图像视觉质量较好。所以本文改进 SCN 模型作为低分辨视频的预处理算法。同时，视频的时序性使得视频检测与图像不同，常常忽略视频序列间的强相关性，运动矢量[47-48] 是一种高效的帧间预测算法，该算法被作为模型的后处理优化算法。

综上所述，本文针对低分辨率视频，提出了一种基于 MobileNet 的新型轻量化的实例分割模型。

2 基于 Mask R-CNN 的模型压缩

2.1 MobileNet 模型

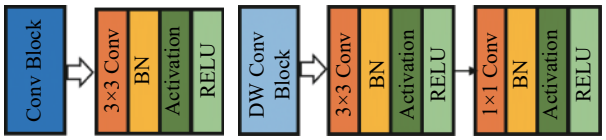
MobileNet 是一种高效且易于实现的轻量级神经网络。其基本单元是深度可分离卷积 (depthwise separable convolution)。深度可分离卷积可以分解为两个更小的操作：depthwise 卷积和 pointwise 卷积，如图 1 所示。



a. depthwise 卷积　　b. pointwise 卷积

图 1　depthwise 卷积和 pointwise 卷积

MobileNet 的设计包含两大结构，如图 2 所示。



a. 标准卷积结构 (Conv_Block)　　b. 深度可分离卷积结构 (DW_Conv_Block)

图 2　标准卷积结构和深度可分离卷积结构的定义

为了减小计算量该模型利用 1×1 卷积降通道。在 MobileNet 中，有 95% 的计算量和 75% 的参数属于 1×1 卷积。将 N 个大小为 D_k、通道数为 M 的卷积核作用于大小为 D_f、通道数同为 M 的特征图上，最后得到大小为 D_p、通道数为 N 的输出，如图 3 所示。

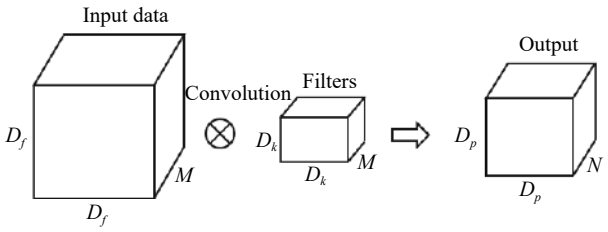


图 3　卷积的过程

标准卷积的计算量为：

$$D_f^2 D_k^2 MN \tag{1}$$

深度可分离卷积总计算量为：

$$D_f^2 D_k^2 M + D_f^2 MN \tag{2}$$

可以看出深度可分离卷积计算量远远小于标准卷积的计算量。

2.2　MobileNet 模型的实现

2.2.1　MobileNet 的模型设计

MobileNet 的模块设计如图 4 所示。其中包含 5 个阶段 (Stage)。输出的通道数分别为：32，64，128，256，512，1 024。具体的细节如表 1 所示，其中 M 代表输入通道，N 代表输出通道。Mask R-CNN 是一种实例分割表现很好的算法，同时 ResNet-FPN 作为特征提取的主干网络 (backbone) 效果较好。所以本文提出的模型优化的算法框架基于 Mask R-CNN，将 MobileNet V1-FPN 做为主干网络进行特征提取，分割算法模型如图 5 所示。

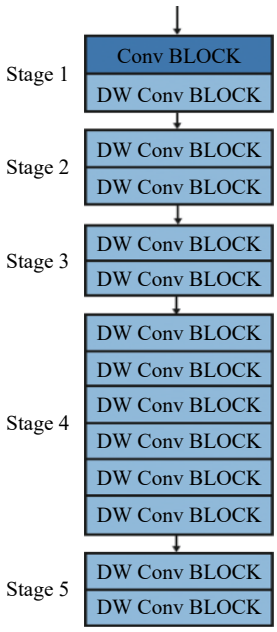


图 4　MobileNet 的模块设计

表 1　MobileNet 的设计

Type	Kernel	M	N	Operator	Stage	Type	Kernel	M	N	Operator	Stage
conv1	(3,3)	3	32	Conv_Block		conv_pw_7	(1,1)	512	512		
conv_dw_1	(3,3)	32	1	DW_Conv_Block	Stage 1	conv_dw_8	(3,3)	512	1	DW_Conv_Block	
conv_pw_1	(1,1)	32	64			conv_pw_8	(1,1)	512	512		
conv_dw_2	(3,3)	64	1	DW_Conv_Block		conv_dw_9	(3,3)	512	1	DW_Conv_Block	
conv_pw_2	(1,1)	64	128		Stage 2	conv_pw_9	(1,1)	512	512		Stage 4
conv_dw_3	(3,3)	128	1	DW_Conv_Block		conv_dw_10	(3,3)	512	1	DW_Conv_Block	
conv_pw_3	(1,1)	128	128			conv_pw_10	(1,1)	512	512		
conv_dw_4	(3,3)	128	1	DW_Conv_Block		conv_dw_11	(3,3)	512	1	DW_Conv_Block	
conv_pw_4	(1,1)	128	256		Stage 3	conv_pw_11	(1,1)	512	512		
conv_dw_5	(3,3)	256	1	DW_Conv_Block		conv_dw_12	(3,3)	512	1	DW_Conv_Block	
conv_pw_5	(1,1)	256	256			conv_pw_12	(1,1)	512	1 024		Stage 5
conv_dw_6	(3,3)	256	1	DW_Conv_Block		conv_dw_13	(3,3)	1 024	1	DW_Conv_Block	
conv_pw_6	(1,1)	256	512		Stage 4	conv_pw_13	(1,1)	1 024	1 024		
conv_dw_7	(3,3)	512	1	DW_Conv_Block							

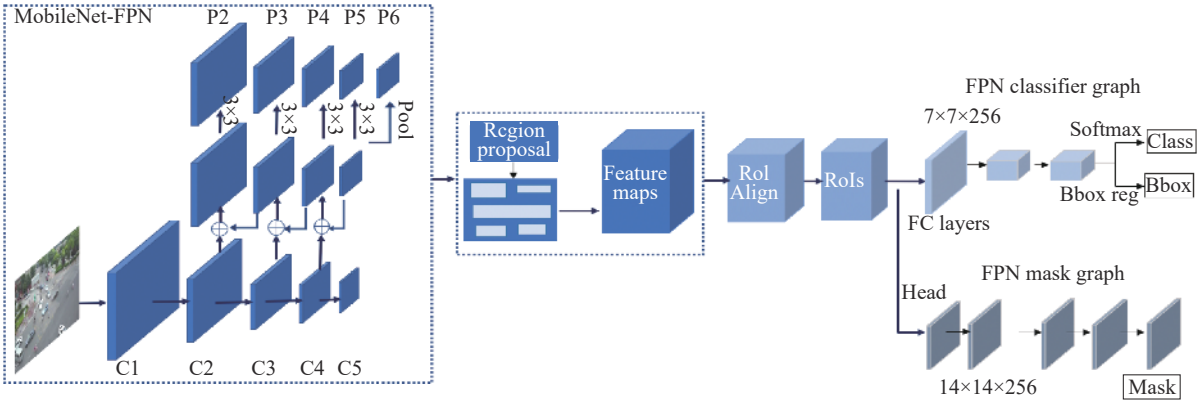


图 5 基于 MobileNet 的实例分割算法框架

2.2.2 SR 重建模型分支设计

由于实例分割模型的像素级特性与目标检测模型不同，对数据的像素级要求很高。同时，视频数据的质量高低对目标检测有很大的影响，所以首先采用超分辨重建算法对视频数据进行预处理。SCN[46] 是一种基于稀疏编码网络的方法，重建图像质量好，模型不需要额外任何数据的支持。该算法结合稀疏先验和深度学习的优点，将原方法中稀疏表示、映射、稀疏重建 3 个独立优化的模块纳入到一个稀疏网络中，得到全局最优解。稀疏编码网络充分利用了图像的先验信息，首先通过特征提取层得到图像的稀疏先验信息，然后通过 LISTA[49] 建立了一个前馈神经网络，该网络可实现图像的稀疏编码。

在网络结构方面，SCN 保留了 SRCNN[50] 的图像块提取、表示层和重建层部分，并在中间层加入 LISTA 网络，如图 6 所示。输入图像I_y首先经过卷积层 H，该卷积层 H 为每个 LR 补丁提取特征。然后，将每个 LR 补丁 y 输入具有有限数量的 k 个循环级的 LISTA 网络中，以获得其稀疏系数α，再重建 HR 图像补丁，最终生成 HR 图像。该算法用均方误差 (mean square error, MSE) 作为代价函数用来训练网络。

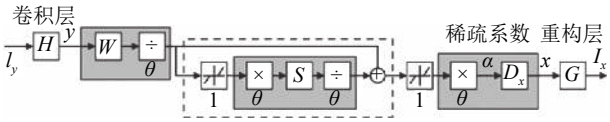


图 6 SCN 算法流程图

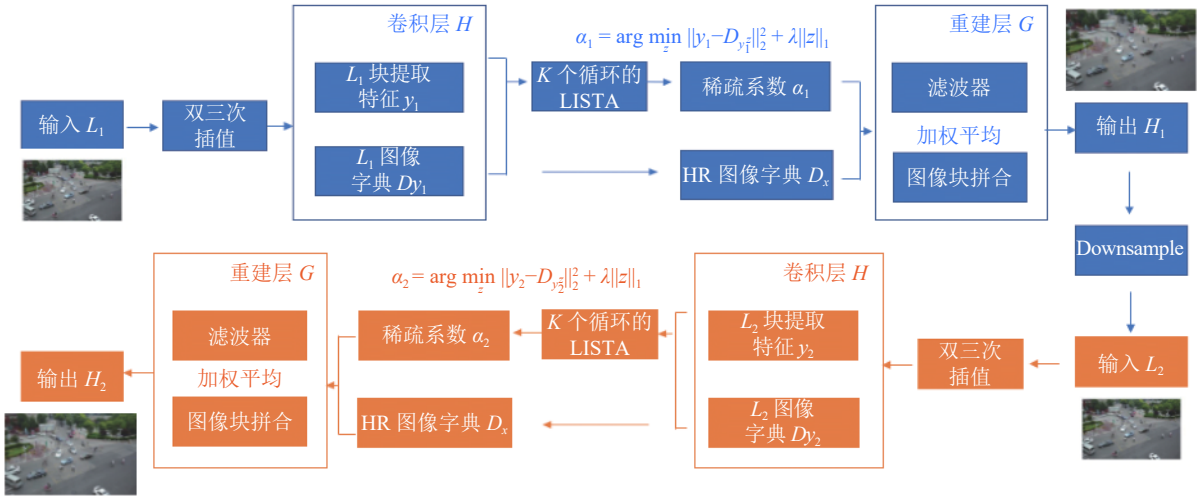
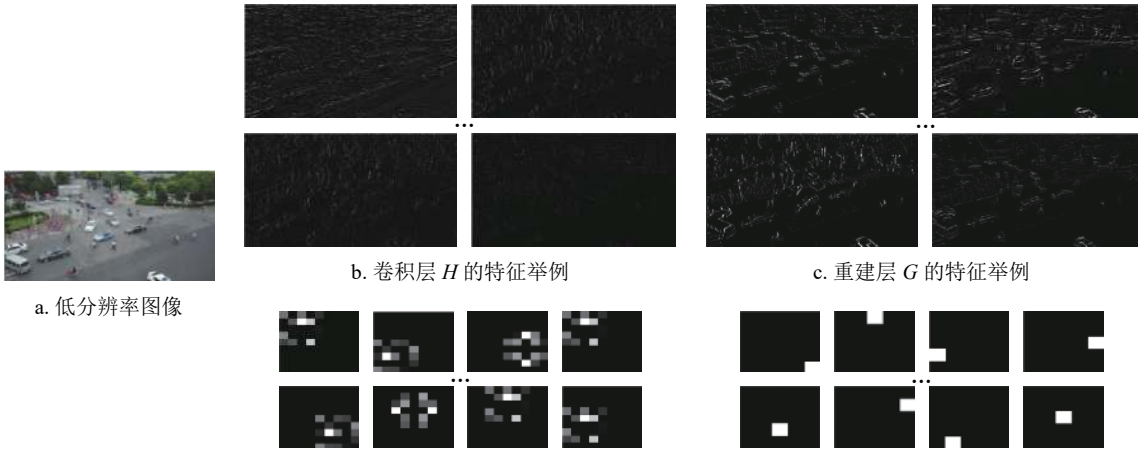


图 7 改进的 SR 重建算法框架

进一步改进该算法，由于图像重建后包含了更多的高频信息，同时下采样 (downsampling) 可以保留图像的边缘、纹理和细节信息，所以采取多次下采样及重构获得更高质量的图像。将输入的

L_1(LR) 图像经过 SCN(1) 生成 H_1(HR) 图像后，经过下采样得到 L_2，再次经过 SCN(2)，得到最终的 H_2(HR)图像，改进的算法流程如图 7 所示。图 8 展示了卷积层和重建层结果举例，图中展示了卷积层 H 和重建层 G 的特征及卷积核。本文用 PSNR[51]、SSIM[52]进行实验验证，实验结果如表 2 所示，从表中可以看出，PSNR 及 SSIM 的结果均有所提高。实验结果如图 9 所示，图中展示了 4 组数据的对比结果，对比算法包括双三次插值 (bicubic interpolation)、SCN 及本文算法。从图中可以看出改进算法重建的图像在边缘部分更加光滑，残差图对比结果如图 10 所示，图中展示了 3 种不同重建算法和 HR 图像的残差图，从图中可以看出本文算法与原图的残差图最接近，因此重建算法效果更好。



a. 低分辨率图像　b. 卷积层 H 的特征举例　c. 重建层 G 的特征举例

d. H 层卷积核举例　e. G 层卷积核举例

图 8　卷积层和重建层结果举例

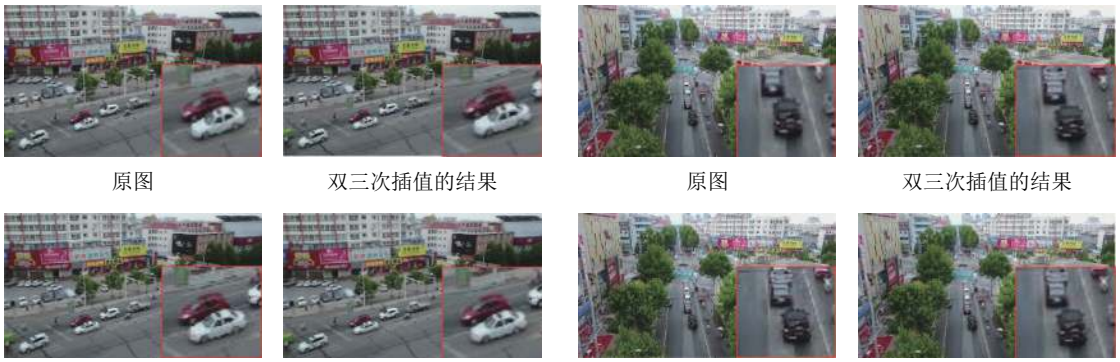
表 2　PSNR 和 SSIM 结果对比

Data	PSNR			SSIM		
	Bic	SCN	本文	Bic	SCN	本文
data 01	29.785 432	29.629 101	30.243 904	0.836 788	0.844 327	0.852 099
data 02	34.276 908	34.038 177	34.398 706	0.891 181	0.886 873	0.891 368
data 03	32.976 318	32.810 872	33.490 763	0.890 747	0.876 704	0.894 202
data 04	33.801 457	33.770 298	34.146 904	0.915 566	0.914 394	0.917 656

原图　双三次插值的结果　原图　双三次插值的结果

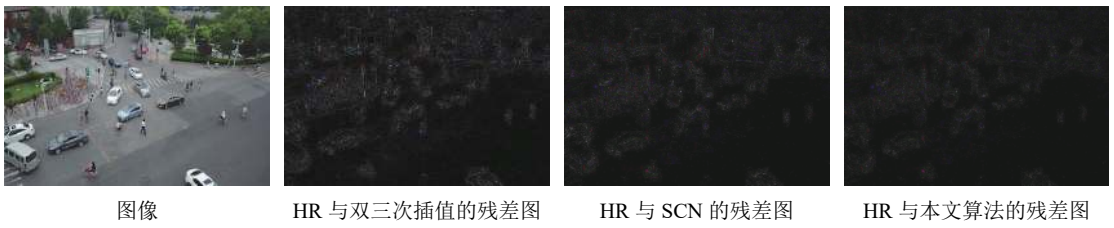
SCN 的结果　本文算法的结果　SCN 的结果　本文算法的结果

a. 实验结果对比 1　b. 实验结果对比 2

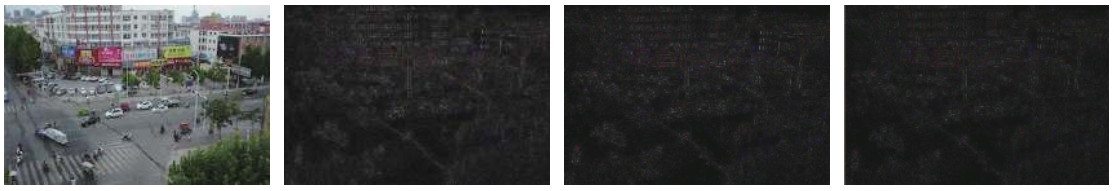


c. 实验结果对比 3　　d. 实验结果对比 4

图 9　改进的 SR 重构算法结果对比



a. 实验结果对比 1



b. 实验结果对比 2

图 10　残差图结果对比

2.2.3　掩码的预测模型分支设计

对于视频帧的进行实例分割时，往往忽略了帧与帧之间的时序性。很多目标在某些帧未被检测到，但在相邻帧检测到。由于帧间的强相关性，对未检测到的掩码 (mask) 进行后处理，得到最终的掩码序列 (mask sequence)。

运动矢量 (motion vector, MV) 是一种高效的帧间预测算法，本文采取效果最好的全局搜索算法。假设，每个块的大小为 $M\times N$(M=2，N=2)。水平方向和垂直方向可搜索的最大位移为 1，把搜索候选区域 3×3 内所有的像素块逐个与当前宏块进行比较，偏移量为 2。块匹配估计准则查找具有最小匹配误差的一个像素块为匹配块，其对应偏移量即为所求运动矢量，最后找到全局最优点，即最佳匹配块。块匹配估计准则是判断块相似程度的依据, 本文采取平均绝对误差函数 (mean absolute deviation, MAD) 作为块匹配估计准则。其函数定义为：

$$\mathrm{MAD}(x,y)=\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}|f_k(m,n)-f_{k-1}(m+x,n+y)| \tag{3}$$

算法具体流程如下，首先找到每个掩码的重心，并用相邻帧的重心最小距离将掩码进行排序，得到每个目标的初始掩码序列。然后，对于未检测到的掩码，用相邻已得到的掩码作为参考图像，从左上角开始，以宏块大小为步长，对于每一个宏块，得到运动矢量，并将参考图像中的该宏块放到所预测的图像中，得到预测的掩码，以此类推，得到最终掩码预测序列。该算法的预测过程如图 11 所示。其中，红色实线框代表已检测到的目标，黄色虚线框代表未检测到的目标。从图中看出，Frame(i) 通过相邻帧 Frame(i−2) 及 Frame(i−1) 得到预测的掩码，Frame(i+1) 通过 Frame(i) 及 Frame(i+2) 进行预测。图 12 展示了运动矢量的实验结果，其中图 12e 和 12f 列举了运动矢量的统计结果。图 13

进一步展示了掩码的预测过程。同时图 14 展示了两组预测结果。两组结果的第二排展示出相邻两帧间的残差图。图 15 列举了最终掩码序列的预测结果。

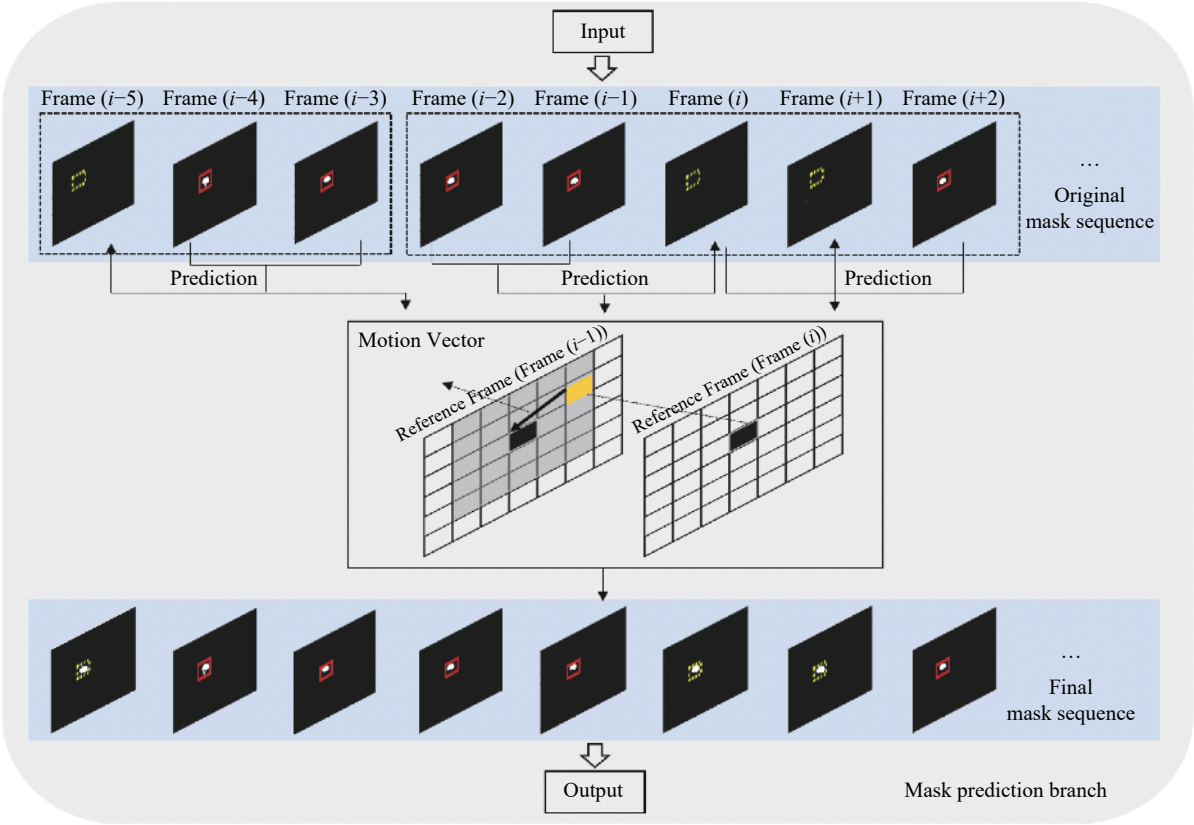
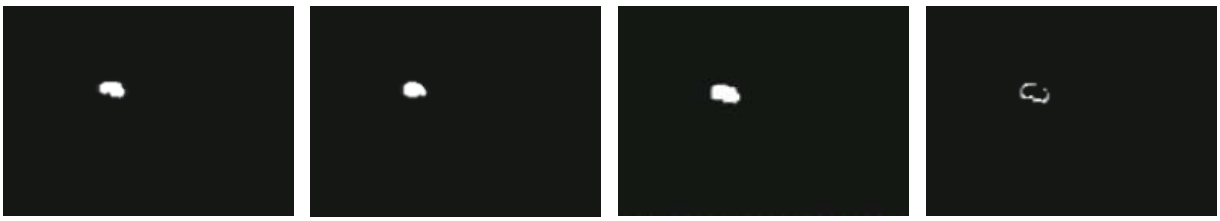
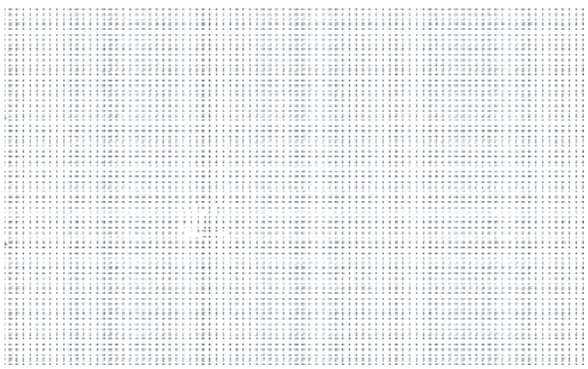
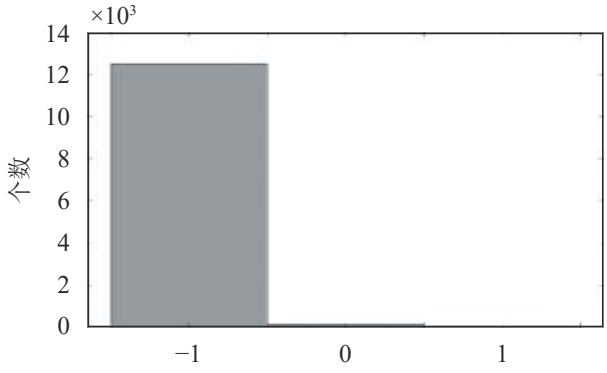


图 11　掩码的预测模型



a. 参考帧　　b. 相邻帧　　c. 预测桢　　d. 预测帧与参考帧的残差





e. 运动矢量　　f. 运动失量

图 12　运动矢量的实验结果

本文算法的整体框架如图 16 所示，该模型包含 SR 重建模型分支 (SR reconstruction branch)、基于 MobileNet 的实例分割模型分支和掩码的预测模型分支 (Mask prediction branch)3 大部分，是一种针对低分辨率视频的新型轻量化实例分割模型。

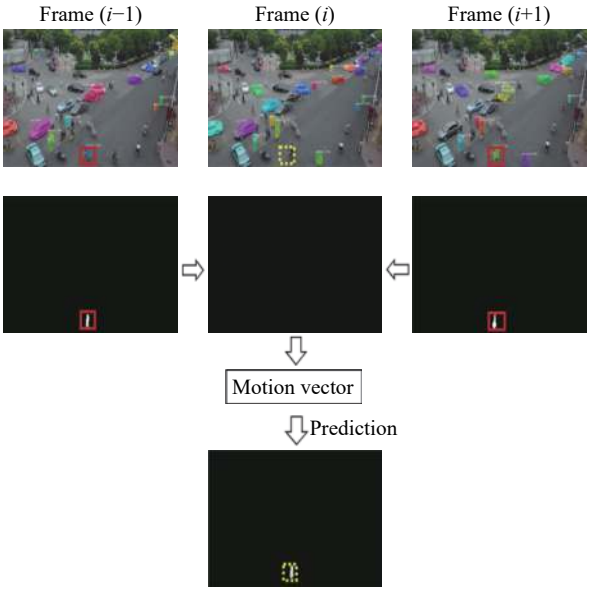
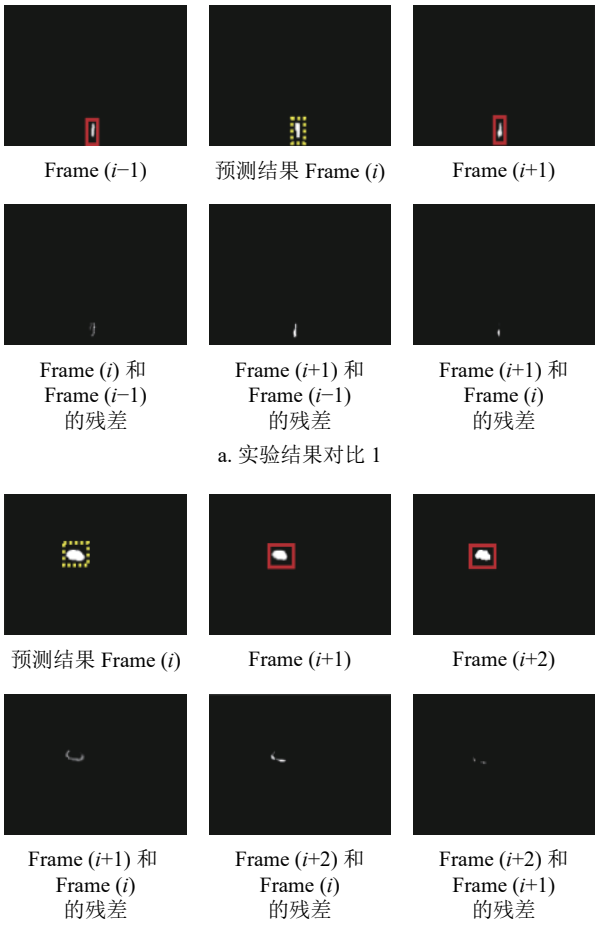


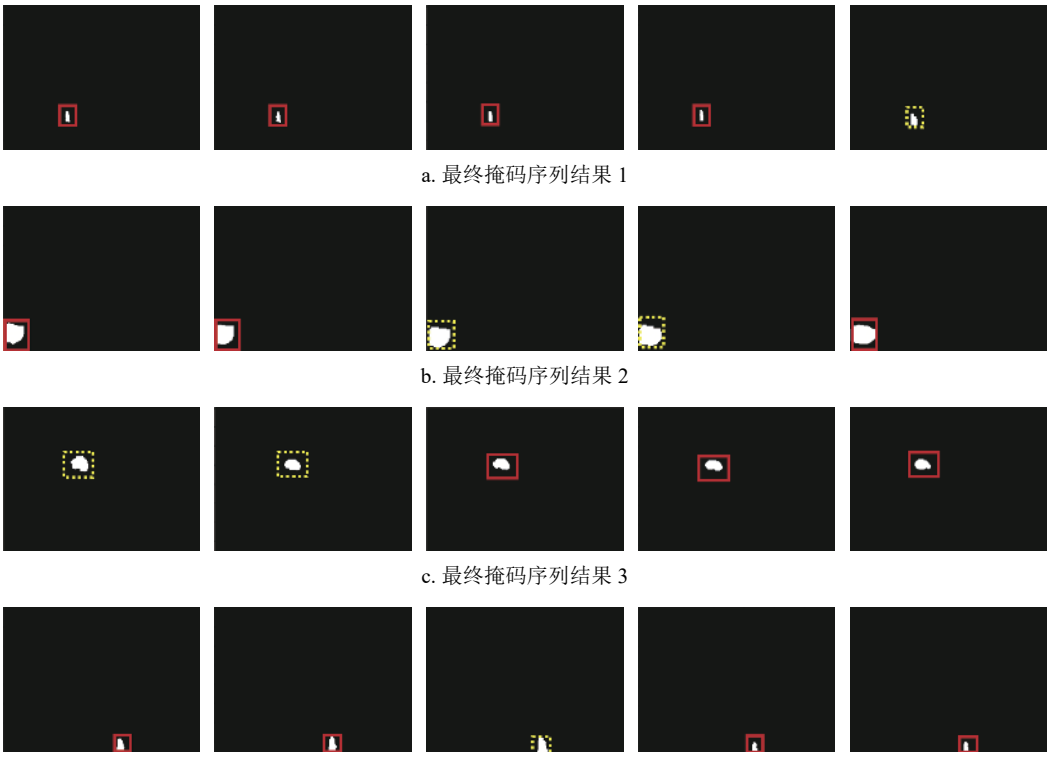
图 13　掩码的预测过程



a. 实验结果对比 1

b. 实验结果对比 2

图 14　两组掩码预测结果



a. 最终掩码序列结果 1

b. 最终掩码序列结果 2

c. 最终掩码序列结果 3

d. 最终掩码序列结果 4

图 15　最终掩码序列的预测结果

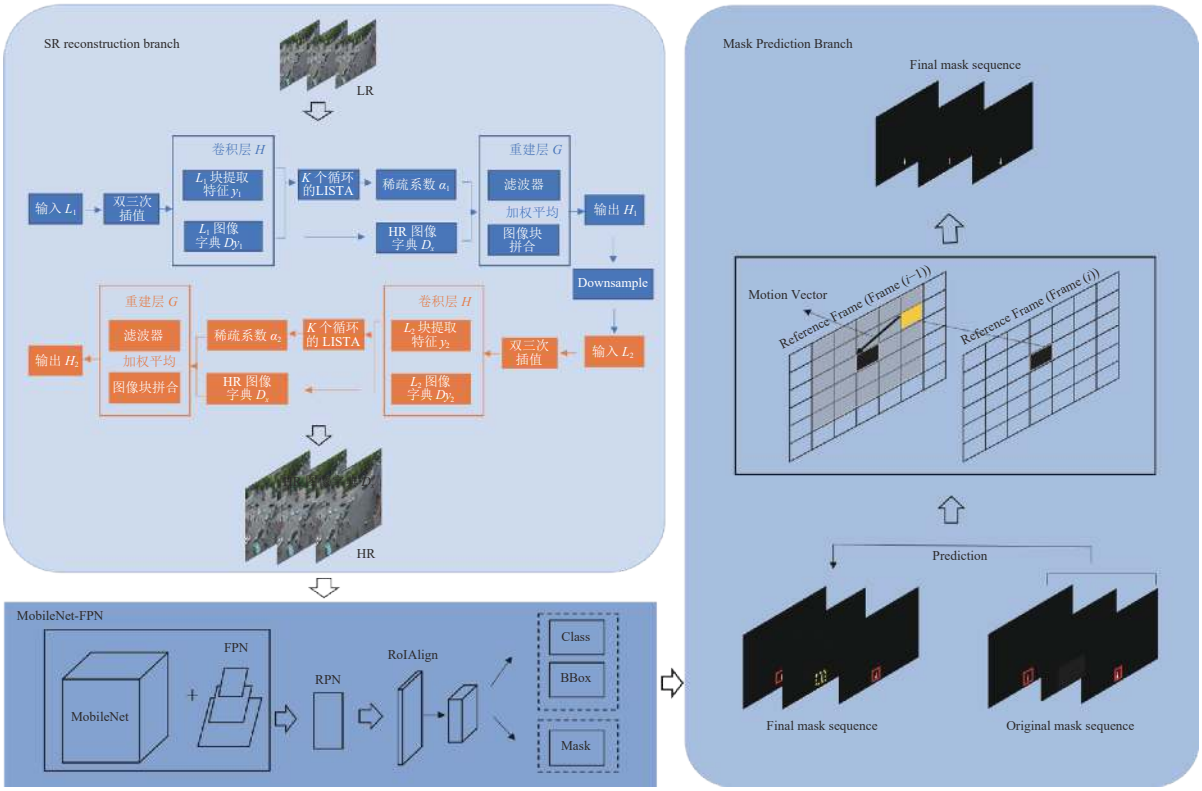


图 16　实例分割模型整体框架

3　实验结果及分析

VisDrone2018 是由无人机拍摄的实时场景视频数据。该数据的特点是每一帧存在很多小尺寸目标，具有低分辨率，而且很多数据由于镜头抖动等造成数据质量较低，能很好展示实时街景场景，具有数据多样性，另一组数据集 Street_View_MB7500 也反映实时街景，符合低分辨率特性。数据集描述如表 3 所示。本文采用该两组数据集作为实验数据。将每个类按 5∶3∶2 随机抽取作为训练数据集、验证集及测试集。

表 3　实验数据集

Data	Names	Frames
Dataset (1)	VisDrone2018 (1)	350
Dataset (2)	VisDrone2018 (2)	270
Dataset (3)	VisDrone2018 (3)	230
Dataset (4)	VisDrone2018 (4)	360
Dataset (5)	Street_View_MB7500	7500

NVIDIA Jetson TX2 采用 256 核 NVIDIA Pascal 架构和 8 GB 内存，如表 4 所示，计算更快，推理能力强，是一种嵌入式 AI 计算设备。本文采用此设备作为嵌入式平台并进行实验分析。训练模型的学习率设为 0.0001，图 17 表示了训练过程的 loss 变化。

表 4　NVIDIA TX2 的技术规格

参数	NVIDIA Jetson TX2
GPU	256个NVIDIA CUDA核心
Memory	8 GB 128位LPDDR4
CPU	双核Denver 2 64位CPU和四核ARM A57 Complex

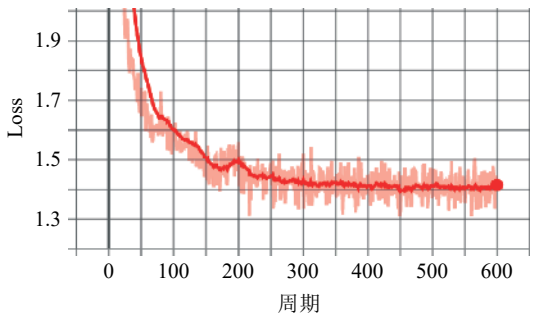


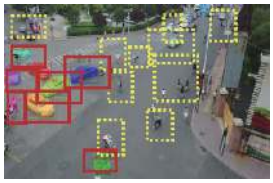
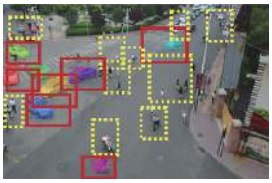
图 17　Loss 的变化结果

本文对 ResNet-FPN 及 MobileNet-FPN 作为 backbone 进行了对比实验。表 5 表示了压缩前后实验结果对比。从表中看出，未压缩的 ResNet-FPN 训练模型大小为 255 M，基于 MobileNet-FPN 的模型大小为 88 M，所以模型降低了近 2/3 大小的空

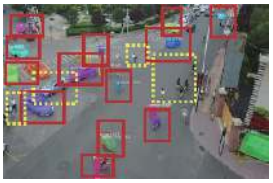
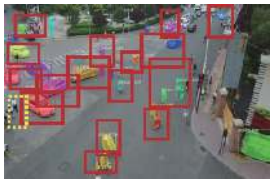
间。同时，SPF (seconds per frame)在模型压缩后大大减少，吞吐量 (Throughput) 也有很大提高。基于 MobileNet-FPN 的准确率高于 ResNet-FPN。图 18 列举了压缩前后基于 ResNet-FPN 与 MobileNet-FPN 的分割结果对比。虚线框表示未检测到的目标，实线框则表示已检测的物体。从图中看出，基于 MobileNet 的模型的结果优于 ResNet，且准确率较高。在两类模型中，本文算法的 HR 图像的检测率高于 LR 图像。实验证明，该算法在目标检测方向获得了较好的结果。

表 5　压缩前后实验对比

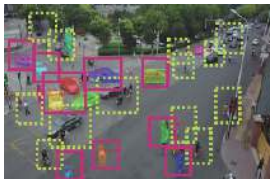
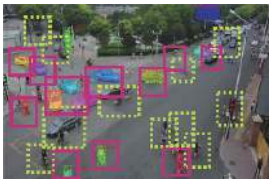
GPU	Data	Backbone	Kernels	SPF	Throughput	mAP/%	model size/M
NVIDIA 1 080 Ti	HR (本文)	ResNet-FPN	125	1.32	45	64.7	255
NVIDIA 1 080 Ti	HR (本文)	MobileNet-FPN (本文)	48	0.62	97	69.1	88
NVIDIA TX2	HR (本文)	MobileNet-FPN (本文)	48	0.88	68	69.1	88



LR 图像　　HR 图像 (本文算法)

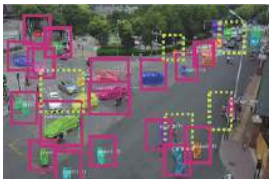
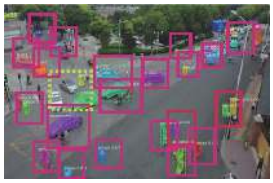
a. 基于 ResNet-FPN 的结果对比 (数据 1)



LR 图像　　HR 图像 (本文算法)

b. 基于 MobileNet-FPN 的结果对比 (数据 1)



LR 图像　　HR 图像 (本文算法)

c. 基于 ResNet-FPN 的结果对比 (数据 2)



LR 图像　　HR 图像 (本文算法)

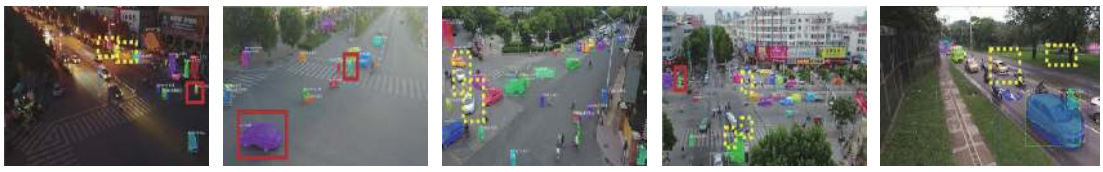
d. 基于 MobileNet-FPN 的结果对比 (数据 2)

图 18　基于 MobileNet 与 ResNet 的分割结果对比

在基于的 MobileNet-FPN 的分割模型下，不同数据的实例分割结果对比如图 19 所示。虚线框表示该目标未被检测到，实线框表示已检测到。为了更好展现对比，图中只展示出部分目标的结果。从图中看出，图像经过 SR 重构后，实线框明显增多，说明图像经过 3 种不同的重建算法后，实例分割效果逐渐提升。所以图像的质量高低对于实例分割算法具有很大的影响。表 6 列举了实验最终结果统计，从 LR、HR(双三次插值)、HR (SCN) 到 HR (本文算法)。表中可以看出 SCN 预处理后， mAP (mean average precision) 提升到 67.9%，改进 SCN 算法后 mAP 达到 69.1%。最后经过掩码预测的算法，模型的 mAP 提高到了 69.8%。因此，本文算法的 mAP 和 mIoU (mean intersection over union) 在一定程度上均有所提高，具体的实验结果统计如图 20 所示。综上所述，本文提出的基于 MobileNet-FPN 的新型轻量化实例分割算法在低分辨率场景中具有内存小、测试速度快、精度高等特性，且易于移植到嵌入式平台中。

a. LR 图像

b. HR 图像 (双三次)



c. HR 图像 (SCN)

d. HR 图像 (本文算法)

图 19　MobileNet-FPN 模型在不同数据的分割结果对比

表 6　实验对比结果

Data	Backbone	mAP/%	mIoU/%
LR	MobileNet-FPN (本文)	65.5	56.5
HR (Bicubic)	MobileNet-FPN (本文)	67.4	60.7
HR (SCN)	MobileNet-FPN (本文)	67.9	61.4
HR (本文)	MobileNet-FPN (本文)	69.1	63.3
HR (本文)	MobileNet-FPN with Mask prediction (本文)	69.8	63.4

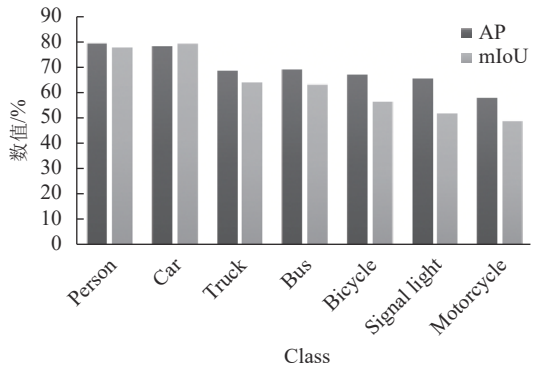


图 20　实验结果统计

4　结 束 语

本文提出了一种基于 MobileNet，结合超分辨及帧间预测的新型轻量化实例分割算法。实验证明该算法在实时场景视频中的准确率较高，内存小，测试速度快且易于移植到嵌入式平台中。

参　考　文　献

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. In Advances in Neural Information Processing Systems, 2012, 25(2): 1097-1105.

[2] 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. [S.l.]: IEEE, 2016: 770-778.

[3] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.]: IEEE, 2014: 580-587.

[4] GIRSHICK R. Fast R-CNN[C]//Proceedings of the IEEE international conference on computer vision. [S.l.]: IEEE, 2015: 1440-1448.

[5] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. In Advances in Neural Information Processing Systems, 2015, DOI: 10.1109/TPAMI.2016.2577031.

[6] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//In Proceedings of the IEEE international conference on computer vision. [S.l.]: IEEE, 2017: 2961-2969.

[7] SUYKENS J A, VANDEWALLE J. Least squares support vector machine classifiers[J]. Neural Processing Letters, 1999, 9(3): 293-300.

[8] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[9] GAO J, YANG Z, CHEN K, et al. Turn tap: Temporal unit regression network for temporal action proposals[C]//

Proceedings of the IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2017: 3628-3636.

[10] QIU Z, YAO T, MEI T. Learning spatio-temporal representation with pseudo-3d residual networks[C]// Proceedings of the IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2017: 5533-5541.

[11] XU H, DAS A, SAENKO K. R-c3d: Region convolutional 3d network for temporal activity detection[C]// Proceedings Of The Ieee International Conference on Computer Vision. [S.l.]: IEEE, 2017: 5783-5792.

[12] 吴天舒, 张志佳, 刘云鹏, 等. 基于改进 SSD 的轻量化小目标检测算法[J]. 红外与激光工程, 2018, 47(7): 47-53.
Wu Tian-shu, Zhang Zhi-jia, Liu Yun-peng, et al. A lightweight small object detection algorithm based on improved SSD[J]. Infrared and Laser Engineering, 2018, 47(7): 47-53.

[13] 崔家华, 张云洲, 王争, 等. 面向嵌入式平台的轻量级目标检测网络[J]. 光学学报, 2019, 39(4): 0415006.
CUI Jia-hua, ZHANG Yun-zhou, WANG Zheng, et al. Light-weight object detection networks for embedded platform[J]. Acta Agronomica Sinica, 2019, 39(4): 0415006.

[14] 邓琉元, 杨明, 王春香, 等. 基于环视相机的无人驾驶汽车实例分割方法[J]. 华中科技大学学报 (自然科学版), 2018, 46(12): 24-29.
DENG Liu-yuan, YANG Ming, WANG Chun-xiang, et al. Surround view cameras based instance segmentation method for autonomous vehicles[J]. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2018, 46(12): 24-29.

[15] XU H, LI B, RAMANISHKA V, et al. Joint event detection and description in continuous video streams[C]//2019 IEEE Winter Conference on Applications of Computer Vision (WACV). [S.l.]: IEEE, 2019: 396-405.

[16] ZHU Y, KARAN S, FITSUM A, et al. Improving semantic segmentation via video propagation and label relaxation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 8856-8865.

[17] XIA Hong-xia, LI Ting, LIU Wen-xuan, et al. Abnormal event detection method in surveillance video based on temporal CNN and sparse optical flow[C]//Proceedings of the 2019 5th International Conference on Computing and Data Engineering. [S.l.]: ACM, 2019: 90-94.

[18] YANG Xi-tong, YANG Xiao-dong, LIU Ming-yu, et al. Step: Spatio-temporal progressive learning for video action detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 264-272.

[19] 彭秋辰, 宋亦旭. 基于 Mask R-CNN 的物体识别和定位[J]. 清华大学学报 (自然科学版), 2019, 59(2): 135-141.
PENG qiu-chen, SONG Yi-xu. Object recognition and localization based on mask R-CNN[J]. Journal of Tsinghua University (Science and Technology), 2019, 59(2): 135-141.

[20] 王子愉, 袁春, 黎健成. 利用可分离卷积和多级特征的实例分割[J]. 软件学报, 2019, 30(4): 954-961.
WANG Zi-yu, YUAN Chun, LI Jian-cheng. Instance segmentation with separable convolutions and multi-level features[J]. Journal of Software, 2019, 30(4): 954-961.

[21] LI Hao, KADAV A, DURDANOVIC I, et al. Pruning filters for efficient convnets[C]//5th International Conference on Learning Representations (ICLR). [S.l.]: ICLR, 2017: .

[22] LUO Jian-hao, WU Jian-xin. An entropy-based pruning method for CNN compression[EB/OL]. [2018-11-10]. https://arxiv.org/pdf/1706.05791.pdf.

[23] LIN Shao-hui, JI Rong-rong, YAN Chen-qian, et al. Towards optimal structured cnn pruning via generative adversarial learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 2790-2799.

[24] COURBARIAUX M, BENGIO Y, DAVID J P. Binaryconnect: Training deep neural networks with binary weights during propagations[C]//Advances in Neural Information Processing Systems. [S.l.]: Neural Information Processing Systems Foundation, 2015: 3123-3131.

[25] CHEN Li-ming, WANG Bo-si, YU Wei-jie, et al. CNN-based fast HEVC quantization parameter mode decision[J]. Journal of New Media 1, 2019(3): 115-126.

[26] GUO Y, YAO A, ZHAO H, et al. Network sketching: Exploiting binary structure in deep CNNs[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017, DOI: 10.119/CVPR.2017.430.

[27] HAN Song, MAO Hui-zi, WILLIAM J D. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding[EB/OL]. [2018-11-12]. https://arxiv.org/pdf/1510.00149.pdf.

[28] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[EB/OL]. [2018-11-14]. https://arxiv.org/pdf/1704.04861.pdf.

[29] SANDLER M, HOWARD A, ZHU M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 4510-4520.

[30] ZHANG X, ZHOU X, LIN M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 6848-6856.

[31] MA N N, ZHANG X Y, ZHENG H T, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design[C]//Proceedings of the European Conference on Computer Vision (ECCV). [S.l.]: Springer Verlag, 2018: 122-138.

[32] YANG J, WRIGHT J, HUANG T, et al. Image super-resolution as sparse representation of raw image patches[C]//IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2008: 1-8.

[33] YANG J, WRIGHT J, HUANG T S, et al. Image super-

resolution via sparse representation[J]. IEEE Transactions on Image Processing, 2010, 19(11): 2861-2873.
[34] YANG J, WANG Z, LIN Z, et al. Coupled dictionary training for image super-resolution[J]. IEEE Transactions on Image Processing, 2012, 21(8): 3467-3478.
[35] TIMOFTE R, DE SMET V, VAN GOOL L. Anchored neighborhood regression for fast example-based super-resolution[C]//Proceedings of the IEEE international Conference on Computer Vision. [S.l.]: IEEE, 2013: 1920-1927.
[36] TIMOFTE R, DE SMET V, VAN GOOL L. A+: Adjusted anchored neighborhood regression for fast super-resolution[C]//Asian Conference on Computer Vision. [S.l.]: Springer, 2014: 111-126.
[37] TIMOFTE R, ROTHE R, VAN GOOL L. Seven ways to improve example-based single image super resolution[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2016: 1865-1873.
[38] HE H, SIU W C. Single image super-resolution using Gaussian process regression[C]//Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2011: 449-456.
[39] YANG Wen-ming, ZHANG Xue-chen, TIAN Ya-peng, et al. Deep learning for single image super-resolution: A brief review[J]. IEEE Transactions on Multimedia, 2019(12): 3106-3121.
[40] KIM J, KWON L J, MU L K. Accurate image super-resolution using very deep convolutional networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2016: 1646-1654.
[41] LEDIG C, THEIS L, HUSZÁR F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017: 4681-4690.
[42] WANG H D, YAIR R, JIN Y Y, et al. Deep learning enables cross-modality super-resolution in fluorescence microscopy[J]. Nature Methods, 2019(1): 103-110.
[43] YANG J X, ZHAO Y Q, JONATHAN C W C, et al. A Multi-scale wavelet 3D-CNN for hyperspectral image super-resolution[J]. Remote Sensing, 2019(13): 1557.
[44] YE Z H, FAN L, LI L Y, et al. SR-GAN: Semantic rectifying generative adversarial network for zero-shot learning[C]//2019 IEEE International Conference on Multimedia and Expo (ICME). [S.l.]: IEEE, 2019: 85-90.
[45] SEFI B K, ASSAF S, MICHAL I. Blind super-resolution kernel estimation using an internal-GAN[C]//33rd Annual Conference on Neural Information Processing Systems (NeurIPS 2019). [S.l.]: Neural information processing systems foundation, 2019: 284-293.
[46] WANG Z, LIU D, YANG J, et al. Deep networks for image super-resolution with sparse prior[C]//Proceedings of the IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2015: 370-378.
[47] CHIEN W J, CHEN P S, MARTA K. Video coding using adaptive motion vector resolution: US Patent 10, 536, 701[P]. 2020-1-14.
[48] VISHWANATH B, KENNETH R. Spherical video coding with motion vector modulation to account for camera motion[C]//2019 IEEE Visual Communications and Image Processing (VCIP). [S.l.]: IEEE, 2019: 1-4.
[49] GREGOR K, YANN L. Learning fast approximations of sparse coding[C]//Proceedings of the 27th International Conference on International Conference on Machine Learning. Haifa, Israel: [s. n.], 2010: 399-406.
[50] DONG C, LOY C C, HE K M, et al. Image super-resolution using deep convolutional networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(2): 295-307.
[51] HUYNH-THU Q, GHANBARI M. Scope of validity of PSNR in image/video quality assessment[J]. Electronics Letters, 2008, 44(13): 800-801.
[52] WANG Z, BOWWVIK A C, SHEIKH H R, et al. Image quality assessment: From error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

编　辑　叶　芳