计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(12):412−419 [doi: 10.15888/j.cnki.csa.008870]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于改进 UNet 网络的室内运动目标阴影分割①

刘　莹, 杨　硕

(沈阳化工大学 计算机科学与技术学院, 沈阳 110142)
通信作者: 刘　莹, E-mail: lying220601@163.com

摘　要: 针对室内环境下智能监控视频对光照变化产生的阴影难以识别、分割困难等问题, 提出一种结合迁移学习方式和 SENet 通道注意力机制的 UNet 网络. 首先, 针对阴影特征模糊难以有效提取的问题, 在 UNet 模型的上采样部分, 添加 SENet 通道注意力机制, 在不增加网络参数的同时, 提高有效区域的特征权重; 并将预训练好的 VGG16 网络迁移到 UNet 模型中, 实现特征迁移和参数共享, 提高模型的泛化能力, 减少训练成本; 最后通过解码器得到分割结果. 实验结果表明, 改进的 UNet 算法相比于原 UNet 算法在对运动目标的分割精度上达到了 96.09%, 对阴影的分割精度上达到 92.24%, 平均交并比 (*MIOU*) 达到 92.58%, 算法性能指标有显著提升.

关键词: 阴影; 迁移学习; 注意力机制; UNet; 深度学习

引用格式: 刘莹,杨硕.基于改进 UNet 网络的室内运动目标阴影分割.计算机系统应用,2022,31(12):412–419. http://www.c-s-a.org.cn/1003-3254/8870.html

Segmentation of Indoor Moving Object Shadow Based on Improved UNet Network

LIU Ying, YANG Shuo

(College of Computer Science and Technology, Shenyang University of Chemical Technology, Shenyang 110142, China)

Abstract: Considering that shadows caused by changes in lighting are difficult to identify and segment for intelligent surveillance videos in indoor environments, this study proposes a UNet network combining the transfer learning method and the SENet channel attention mechanism. Specifically, because shadow features are blurry and difficult to extract effectively, the SENet channel attention mechanism is added to the upsampling part of the UNet model to improve the feature weight of the effective area without increasing the network parameters. A pre-trained VGG16 network is then migrated into the UNet model to achieve feature migration and parameter sharing, improve the generalization ability of the model, and reduce training costs. Finally, the segmentation result is obtained by a decoder. The experimental results show that compared with the original UNet algorithm, the improved UNet algorithm offers significantly enhanced performance indicators, with its segmentation accuracy on moving objects and shadows respectively reaching 96.09% and 92.24% and a mean intersection-over-union (*MIOU*) of 92.58%.

Key words: shadow; transfer learning; attention mechanism; UNet; deep learning

自然界中的阴影无处不在, 阴影是由于光的传播途径中有遮蔽物而产生的. 阴影的优点是它包含光源和场景的物体信息, 有助于理解场景的目标. 但是阴影的缺点也不容忽视, 由于阴影与运动目标存在粘连, 使检测到的目标轮廓不够精准, 妨碍现有的很多影像、视频处理及分析工作, 也对后续进行目标的跟踪和识别等操作产生干扰. 因此, 对阴影检测及消除方法的研究, 既有理论意义又有现实意义, 被认为是机器视觉的一项富有挑战性的任务.

在对室内运动目标进行检测与跟踪等工作时, 阴

① 收稿时间: 2022-04-21; 修改时间: 2022-05-22; 采用时间: 2022-06-06; csa 在线出版时间: 2022-08-26

影会因室内存在光照变化而出现. 由于产生的阴影与运动目标的运动性质一致, 随着目标的运动, 阴影的形状也会随时发生变化, 并且室内光线越充足, 阴影颜色越明显. 对于视频序列而言, 在对运动目标进行检测时, 极容易将阴影也判断成是目标, 这对室内运动目标的行为识别的研究造成了很大的困难, 最近几年的前景检测方法也还是被这个问题所困扰.

近年来, 阴影的检测与消除已成为智能监控领域的一个热门话题. 早期的传统方法是利用手工设计的特征 (如颜色[1]、光照变化[2] 等), 通过颜色空间 (如 HSV 颜色空间[3]) 或基于物理模型来检测和消除阴影. 武明虎等人[4] 首先通过混合高斯模型获得视频中的前景, 再通过 HSV 空间实现对阴影的检测, 并结合纹理特征的方法消除所携带的阴影.然而, 传统的阴影检测方法 (如多种特征融合[5]、改进 HSV 的阴影检测方法[6] 等), 已经逐渐不能满足对各种形状下阴影检测的需求.

随着深度学习的迅速发展, 越来越多的研究者们使用深度学习方法检测和去除图像与视频中的阴影. Khan 等人[7] 利用卷积网络分别提取图像边缘信息和位置信息, 并将输出特征输入条件随机场来检测阴影, 取得了很好的效果. Shen 等人[8] 提出 structured CNN, 实现了对图像中阴影区域边缘的检测, 但对阴影轮廓的提取后仍不足以准确检测出整体的阴影区域, 阴影轮廓内部的信息也不太明确. Hosseinzadeh 等人[9] 利用 patched-CNN 和传统方法结合实现对阴影的检测, 相比于一些算法来说提高了准确率与效率. 但该方法不仅要在大量的数据集上进行预训练, 而且使用代码的计算机环境配置十分复杂, 使用较不易.

因此, 采用深度学习的方法进行阴影检测不仅可以获得较高的精度, 而且节省复杂的物理建模过程, 大大简化了阴影检测任务的复杂性, 为研究提供了新的思路.另外, 深度学习在图像语义分割工作中也取得了较高的成果, 如 FCN[10]、UNet[11]、DeepLab[12] 等. UNet 是在 FCN 基础上发展而来的卷积神经网络模型, 本文利用语义分割模型 UNet, 提出了一种结合迁移学习和有效的通道注意机制的 UNet 网络来检测阴影.

1 UNet 模型

UNet 是一种十分对称的端到端的分割算法, 它包括主干特征提取部分和加强特征提取部分. 在主干特征提取结构中, 采用连续下采样过程提取更深层次的特征信息. 这个过程主要是通过 3×3 的卷积层和 2×2 的最大池化层来实现. 通过特征图的大小不断压缩, 特征通道的数量逐渐增加, 提取的特征更加抽象和丰富, 对目标的表达能力更强, 这一部分可以得到 5 个初步有效特征层. 由于图像的不断压缩, 会导致丢失很多目标的细节信息, 因此在加强特征提取结构中, 将上半部分获得的有效特征层进行上采样操作, 逐像素的恢复原始图像的精度, 恢复细节信息, 再通过跳跃连接将两部分获取的特征层进行特征融合. 最后使用 1×1 的卷积层来修改通道数, 获得最终分割结果. 在 UNet 语义分割模型中, 左侧部分用于编码, 右侧部分用于解码, 图 1 展示了 UNet 的网络结构.

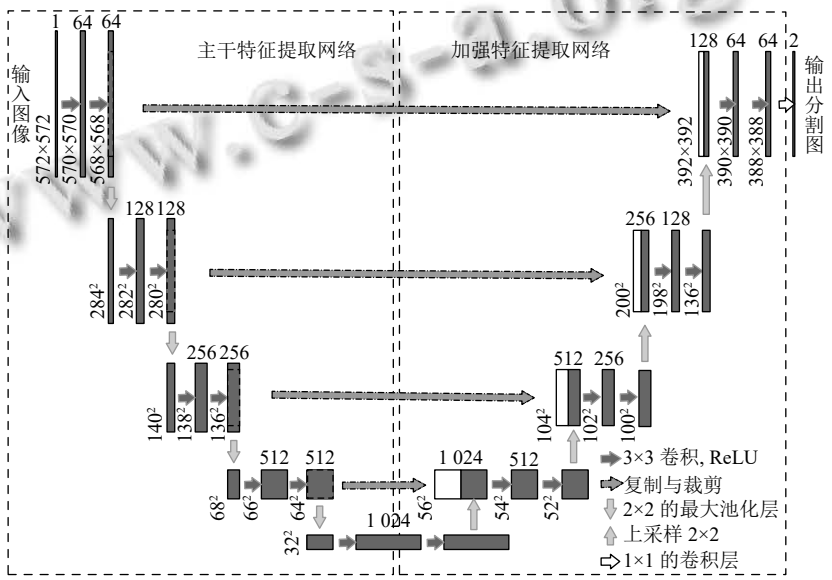


图 1 UNet 网络结构图

2 UNet 网络结构改进

由于在室内环境中, 如开关灯或太阳光线照射等原因, 室内光线强度随时发生变化, 对运动目标的检测变得更加复杂. 为了更加细致且精准的分割运动物体的阴影, 需要高效提取运动目标及其阴影的细节特征, 使阴影轮廓内部的信息更加明确, 因此进行更深层次的卷积运算是十分必要的. 然而随着卷积层的增加, 参数量也会增加. 因此, 改进后的模型需要达到能够高效提取运动目标和阴影的有效特征的同时, 还要减少网络参数.

2.1 VGG 网络

为了高效提取特征, 减少网络参数, 本文采用 VGG16 网络 (不包含全连接层)[13], 作为主干特征提取网络. 与 AlexNet[14] 相比, VGG 使用连续的小卷积核代替大卷积核, 减少了网络训练参数. 而且, 模型的感受野较小的优点会对图像的细节把握更加准确, 保证了良好的学习能力. 此外, 本文利用在数据集 ImageNet 上面预训练的 VGG16 模型, 将其训练参数和训练特征迁移到 UNet 语义分割网络中, 实现特征迁移和参数共享, 提高网络泛化能力, 减少模型对训练样本的依赖, 降低训练成本. 如图 2 所示, 通过 VGG16 网络可以获得 5 个初步有效特征层.

2.2 SENet 通道注意力机制

注意力机制的核心焦点是让网络关注到它最需要关注的地方. 当我们使用卷积神经网络处理图片时, 我们会更希望卷积神经网络关注应该需要注意的地方, 而不是关注于一切, 并且也不可能手动去调整需要注意的地方. 此时, 如何使卷积神经网络去自适应的对重要对象的关注就变得极其重要. 注意力机制就是实现网络自适应注意的一个方式.

对于输入进来的特征层, 利用 SENet 通道注意力机制模块[15] 可以让网络关注它最需要关注的通道. SE (squeeze-and-excitation), 即: 压缩-激活网络. 如图 3 所示, 首先 Ftr 这一步是转换操作 Ftr: $X \to U$, $X \in \mathrm{R}^{H' \times W' \times C'}$, $U \in \mathrm{R}^{H \times W \times C}$. 输入 X 为 $C' \times H' \times W'$的张量, 经过卷积运算, 生成 feature map U 为 $C \times H \times W$.

然后对 feature map U 进行如下 3 个操作:

(1) Squeeze 操作

这一步采用全局平均池化来实现, 输入的特征图为 $C \times H \times W$, 通过对每个通道的空间特征进行特征压缩, 获得一个通道的全局信息, 共有 C 个通道. 最后输出为 $1 \times 1 \times C$, 获取了全局的感受野. 如式 (1):

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{1}$$

其中, F_{sq} 表示 squeeze 操作, H, W 分别表示图中 U 的宽和高, u_c 表示 U 的第 c 个通道的输出.

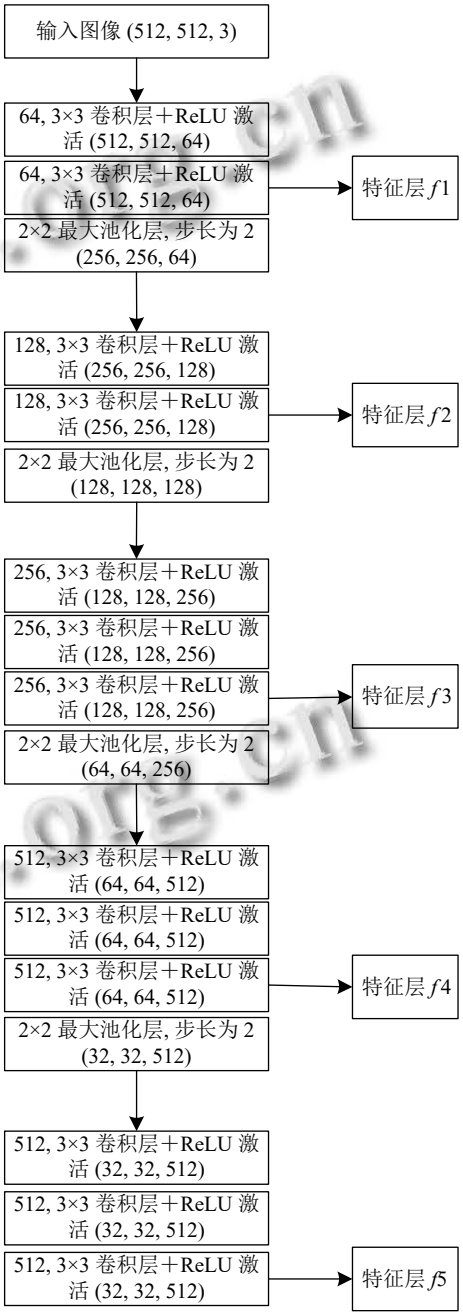


图 2 VGG16 网络结构

(2) Excitation 操作

通过第 1 步的全局平均池化操作会得到一个结

果 z, 然后对结果 z 进行 excitation 操作, 目的是获得输入特征层每一个通道的权值, 学习通道之间的相关性.

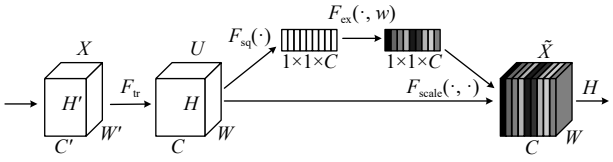


图 3 SENet 注意力机制模块

该过程需要经过两个全连接层和两次激活操作. 首先 z 与 W_1 相乘 (W_1 维度是 $C/r\times C$), 其中, r 是通道压缩倍率, 将 $1\times1\times C$ 压缩为 $1\times1\times C/r$, 完成第 1 个全连接操作, 再经过 ReLU 激活. 输出后进行第 2 个全连接层, 与 W_2 相乘 (W_2 维度是 $C\times C/r$), 然后经过 Sigmoid 函数激活将值固定为 0–1 之间, 输出结果 s, 维度是 $1\times1\times C$. 如式 (2):

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2\delta(W_1 z)) \quad (2)$$

其中, F_{ex} 表示 excitation 操作, δ 表示 ReLU 操作, σ 表示 Sigmoid 操作.

(3) Scale 操作

Scale 操作是运用乘法, 将第 2 步中获得的权重 s, 逐通道加权到先前的特征上. 目的是增强重要的特征信息, 减弱无用的特征信息. 如式 (3):

$$\tilde{x} = F_{scale}(u_c, s_c) = s_c \cdot u_c \quad (3)$$

其中, F_{scale} 表示 scale 操作, s_c 为第 2 步得到的权重, u_c 为提取到的特征, $\tilde{x}$为相乘之后得到的最终结果.

2.3 模型改进

为了加强模型对重要特征的关注, 减少对不必要特征的提取, 本文对 UNet 模型进行了相关改进, 在加强特征提取结构中加入 SENet 通道注意力机制. 在网络解码阶段, 先将编码阶段获得的有效特征层进行上采样操作, 逐像素的恢复原始图像的精度, 恢复细节信息, 再通过跳跃连接将两部分获取的特征层进行特征融合, 此时得到的细节特征更加全面. 将融合后的特征层进行两次卷积之后, 嵌入 SENet 通道注意力模块, 在前 3 次跳跃连接中, 每进行一次特征融合后都添加一个 SENet 通道注意力模块, 使模型在训练过程中能够始终关注重要特征, 从而提高对视频序列中的运动目标及其阴影分割的准确性和鲁棒性. 具体结构如图 4 所示.

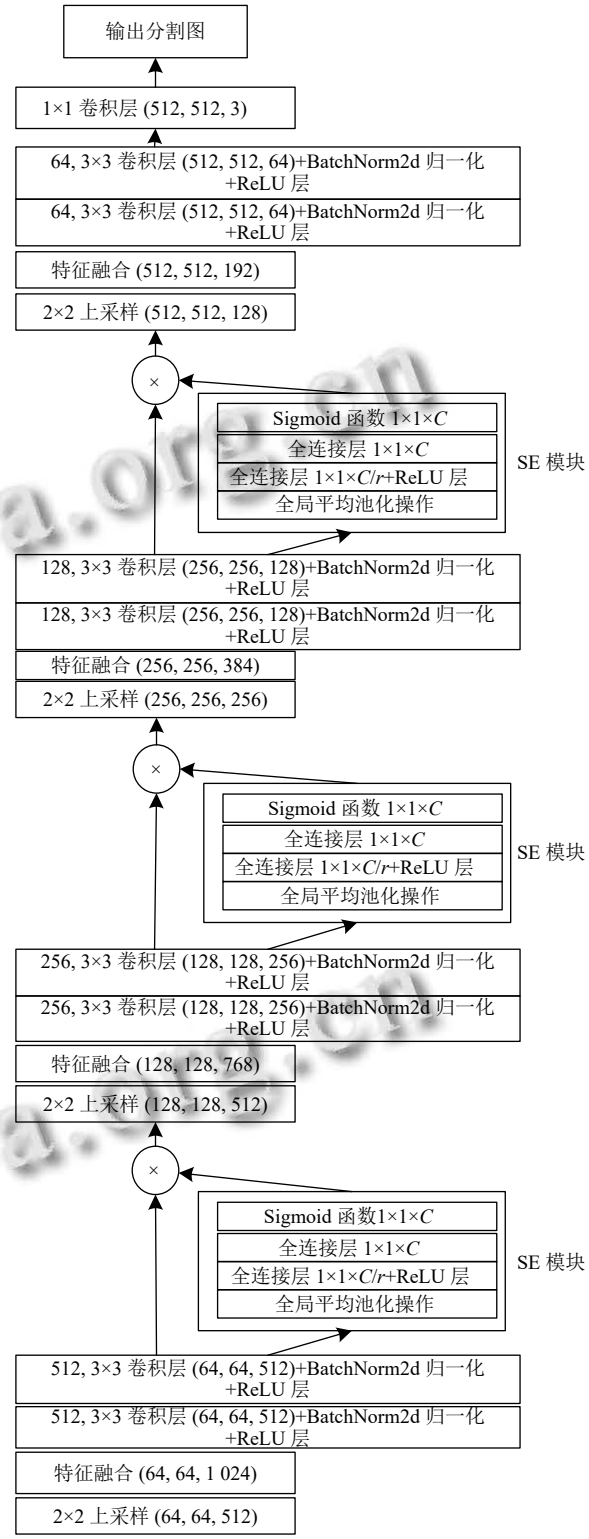


图 4 改进加强特征提取网络的结构图

本文在 UNet 网络结构上进行改进, 使用 VGG16 作为主干特征提取网络, 去掉其全连接层后与加强特征提取网络结构进行有效衔接. 在加强特征提取网络

结构中, 本文选择在特征融合之后嵌入 SENet 通道注意力机制来关注重要特征, 能够增强网络的学习能力, 提高其分割的准确性. 改进后的 UNet 模型网络结构如图 5 所示.

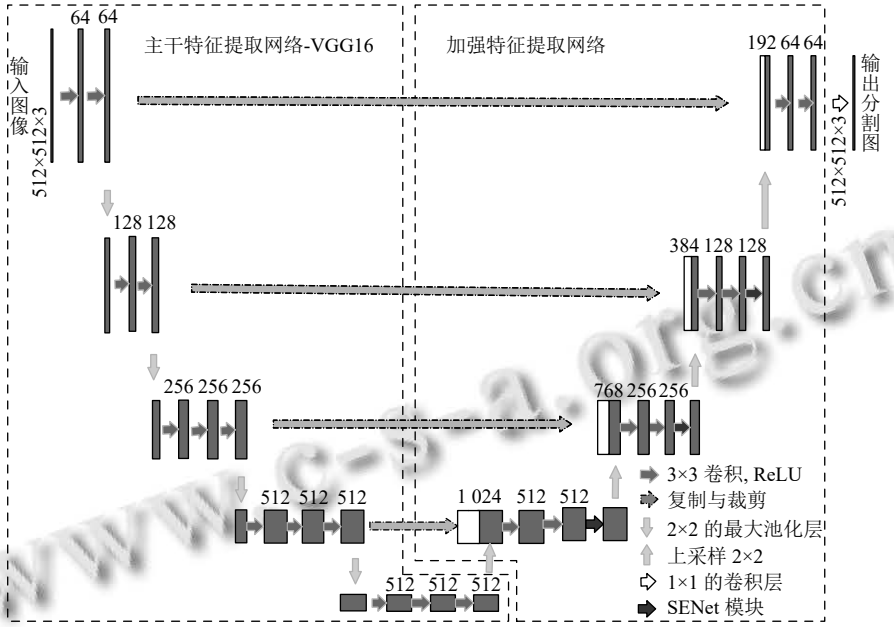


图 5 改进 UNet 的网络结构

3 实验分析

3.1 实验条件配置

本文实验的计算机配置如表 1 所示.

表 1 环境配置

实验条件	配置
处理器	Inter(R) Core(TM)i5-7500 CPU @3.40 GHz
内存	16 GB
显卡	NVIDIA GeForce GTX 1080Ti
深度学习框架	PyTorch 1.2.0
编程语言	Python 3.6

3.2 数据集的构建

本实验采用公共测试数据集 Changedetection 中的 shadow 子集作为数据集, 该数据集是大多数运动检测领域使用的大型且权威的数据集. 其中 shadow 子集里有 6 个类别, 主要场景包含 2 个室内场景, 4 个室外场景. 由于本文研究的是室内环境中运动目标及其阴影的检测与分割任务, 所以选择 shadow 子集中的 cubicle 视频序列作为数据集. 该 cubicle 视频序列的场景为白天室内环境中人物运动存在阴影的场景, 共包含图片 7500 帧, 图像尺寸为 352×240, 选取其中的 2 000 张作为本文实验的数据集.

本文使用 Labelme 图像标注工具对室内环境下的视频图像进行标注, 人工标注每张图像中的运动目标及其阴影, 并输入类别标签为 people 和 shadow 两类. 在完成手工标注后, 生成标签文件. 采用 9:1 的比例划分训练集和验证集, 训练集为 1 800 张, 验证集为 200 张. 测试集选取剩余的 cubicle 视频序列中室内阳光强烈、存在明显阴影的视频序列, 以及选取公开数据集 hallway 和 room 这两个室内运动目标离光源较远、产生的阴影强度非常低的视频序列, 共 80 张. 然后对改进后的 UNet 模型进行训练, 实验流程图如图 6 所示.

3.3 本实验评价标准

本文选择常用的评价分割结果指标, 精确率 (*Precision*)、召回率 (*Recall*)、平均交并比 (mean intersection over union, *MIOU*). 计算公式如下:

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

$$IOU = \frac{TP}{FN+TP+FP} \tag{6}$$

$$MIOU = \frac{1}{k+1} \times \frac{TP}{FN+TP+FP} \tag{7}$$

其中, k+1 表示 k 种类别加一种背景, TP 表示分割正确的正类别数; FP 表示分割错误的正类别数; FN 表示分割错误的负类别数.

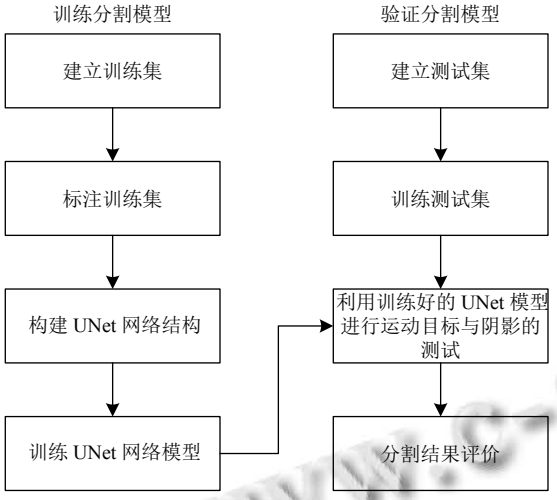


图 6 基于 VGG16 结构的 UNet 阴影分割算法流程图

3.4 实验结果分析

初始学习率设置为 1×10^{-4}, 选取 Adam 优化算法, 学习率的更新因子为 0.96, 间隔设置为 1, 训练迭代次数为 1 000, 批处理图片大小设置为 4. 本文图像分割模型的损失函数为交叉熵损失函数 (cross entropy loss) 与 Dice Loss 函数结合, CE Loss 的计算公式为:

$$Loss=-\sum_{i=1}^{n}p(x_i)\log(q(x_i)) \tag{8}$$

其中, n 为样本数量, $p(x_i)$ 表示第 i 个样本的真实概率分布向量, $q(x_i)$ 表示第 i 个样本模型预测概率分布向量.

Dice 系数, 是一种集合相似度度量函数, 通常用于计算两个样本的相似度 (值范围为 [0, 1]), 计算公式如下:

$$dice=\frac{2|X\cap Y|}{|X|+|Y|} \tag{9}$$

其中, X 和 Y 分别表示预测结果和真实结果, $|X\cap Y|$表示 X 和 Y 之间的交集, 相应的 $Dice\_loss=1-dice$.

从图 7 的训练和验证损失函数变化曲线图得知, 随着迭代次数的逐渐增加, 训练集与验证集的损失值在逐渐减小. 当迭代次数达到 600 次左右时, 损失值基本趋于平稳. 当迭代次数达到 1 000 次, 此时损失值基本收敛, 表示模型获得了良好的训练结果.

3.5 对比实验

本文选择基于 HSV 颜色空间的阴影检测方法和 UNet 网络、DeepLabv3 网络以及本文改进后的 UNet 网络进行对比. 文献 [16] 实现了基于 HSV 的阴影检测方法. 该阴影检测算法选择阴影检测率 (η)、阴影区分率 (ξ)[17] 以及二者的平均值 (Avg) 作为评价指标, 计算公式如下:

$$\eta=\frac{TP_S}{TP_S+FN_S}\times100\% \tag{10}$$

$$\xi=\frac{TP_F}{FN_F+TP_F}\times100\% \tag{11}$$

$$Avg=\frac{\eta+\xi}{2} \tag{12}$$

其中, TP_S 表示正确检测到阴影像素的个数, FN_S 表示把阴影像素误检为前景像素的个数, TP_F 表示正确检测到前景像素点的个数, FN_F 表示将前景像素点误检为阴影像素点的个数.

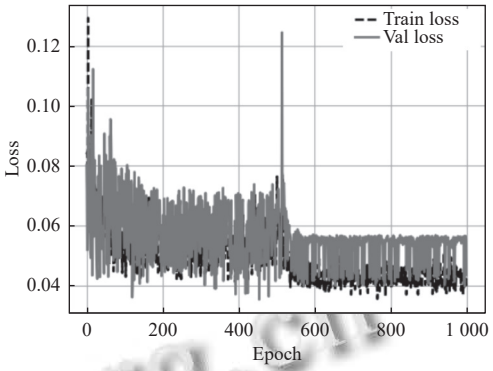


图 7 训练和验证损失函数变化曲线图

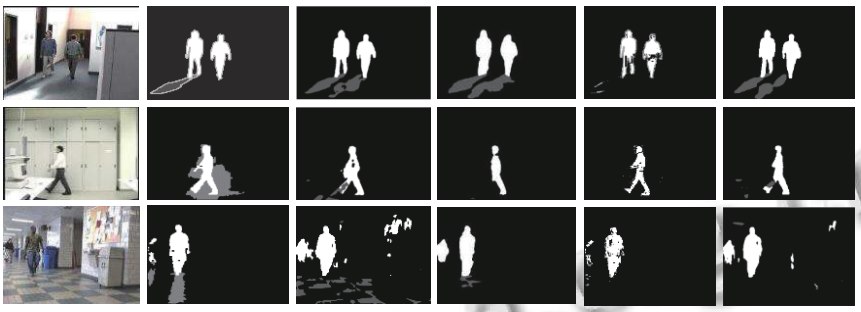
根据实验得出, 基于 HSV 颜色空间的阴影检测方法的阴影检测率为 86.30%, 阴影区分率为 73.28%, 平均值为 79.79%, 平均运行时间为 14.32 ms. 深度学习网络在 cubicle 数据集的实验结果如表 2 所示.

表 2 深度学习网络在 cubicle 数据集的精度结果

算法	类别	*Recall* (%)	*Precision* (%)	*IOU* (%)	*MIOU* (%)	*t* (ms)
UNet	运动目标	96.87	95.9	93.02	92.06	68.96
	阴影	90.3	91.74	83.51		
DeepLabv3	运动目标	94.53	91.11	86.55	87.23	43.27
	阴影	84.55	87.98	75.79		
本文算法	运动目标	97.1	96.09	93.41	92.58	74.48
	阴影	91.17	92.24	84.67		

从以上实验结果可以得知, 使用传统方法即基于 HSV 的阴影检测方法容易将部分运动目标误判为阴

影, 对阴影的检测效果不高. 对比深度学习方法, 改进后的 UNet 分割网络相较于原 UNet 网络, *MIOU* 提高 0.52%, 检测阴影的精准度提高 0.5%, 比 DeepLabv3 检测阴影的精确度提高 4.26%, *MIOU* 提高 5.35%. 本文的运行时间相较于其他方法略长, 但基本上满足实时性的要求. 各个方法进行检测后获得的效果图, 如图 8 所示, 改进后的模型具有更好的分割效果, 分割结果的细节信息更加明显.



(a) 原始图像 (b) Ground truth (c) UNet (d) DeepLabv3 (e) HSV (f) 本文方法

图 8 本文方法与其他方法的对比图

4 结束语

本文对 UNet 语义分割网络进行改进, 采用迁移学习方法并且加入 SENet 通道注意力机制, 能够抓取有效的重要特征, 提高良好的学习能力. 训练和测试改进后的 UNet 模型, 比传统阴影检测的方法效果要好. 相比于原 UNet 模型、DeepLabv3 模型, 在分割精确度上有些许提升. 本文的运行时间相较于其他方法略长, 但基本上满足实时性的要求. 对于部分数据集来说, 图像中人物及其阴影的分割轮廓更加精准、细节更加完善. 实验结果表明, 提出的一种结合迁移学习方式和嵌入 SENet 通道注意机制的 UNet 网络模型的分割算法具有良好的鲁棒性能, 能够基本解决智能监控视频对室内环境下因光照变化产生的阴影难以识别、分割困难等问题. 未来可进一步研究将运动检测与阴影去除结合起来, 便于进行运动目标跟踪及异常行为分析等研究.

参考文献

1 Salvador E, Cavallaro A, Ebrahimi T. Cast shadow segmentation using invariant color features. Computer Vision and Image Understanding, 2004, 95(2): 238–259. [doi: 10.1016/j.cviu.2004.03.008]

2 Guo RQ, Dai QY, Hoiem D. Paired regions for shadow detection and removal. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2956–2967. [doi: 10.1109/TPAMI.2012.214]

3 Cucchiara R, Grana C, Piccardi M, *et al*. Detecting moving objects, ghosts, and shadows in video streams. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003,25(10):1337–1342.[doi:10.1109/TPAMI.2003.1233909]

4 武明虎, 宋冉冉, 刘敏. 结合 HSV 与纹理特征的视频阴影消除算法. 中国图象图形学报, 2017, 22(10): 1373–1380. [doi: 10.11834/jig.170151]

5 张德干, 陈晨, 董悦, 等. 一种基于机器学习的运动目标阴影检测新方法. 光电子·激光, 2018, 29(12): 1317–1324. [doi: 10.16136/j.joel.2018.12.0075]

6 杨春德, 郭帅. 改进基于 HSV 空间的阴影检测算法. 计算机工程与设计, 2018, 39(1): 255–259. [doi: 10.16208/j.issn1000-7024.2018.01.044]

7 Khan SH, Bennamoun M, Sohel F, *et al*. Automatic shadow detection and removal from a single image. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(3): 431–446. [doi: 10.1109/TPAMI.2015.2462355]

8 Shen L, Chua TW, Leman K. Shadow optimization from structured deep edge detection. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 2067–2074.

9 Hosseinzadeh S, Shakeri M, Zhang H. Fast shadow detection from a single image using a patched convolutional neural network. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid: IEEE, 2018. 3124–3129.

10 Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640–651. [doi: 10.1109/TPAMI.2016.2572683]

11 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional

networks for biomedical image segmentation. 18th International Conference on Medical Image Computing and Computer-assisted Intervention. Munich: Springer, 2015. 234–241.

12 Chen LC, Papandreou G, Kokkinos I, *et al*. DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834–848. [doi: 10.1109/TPAMI.2017.2699184]

13 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations. San Diego: ICLR, 2015. 1–14.

14 Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe: ACM, 2012. 1097–1105. [doi: 10.5555/2999134.2999257]

15 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141. [doi: 10.1109/CVPR.2018.00745]

16 Sanin A, Sanderson C, Lovell BC. Shadow detection: A survey and comparative evaluation of recent methods. Pattern Recognition, 2012, 45(4): 1684–1695. [doi: 10.1016/j.patcog.2011.10.001]

17 Prati A, Mikic I, Trivedi MM, *et al*. Detecting moving shadows: Algorithms and evaluation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25(7): 918–923. [doi: 10.1109/TPAMI.2003.1206520]

(校对责编: 牛欣悦)