DOI:10.19651/j.cnki.emt.2417573

基于深度学习的多帧瞳孔检测算法*

张国静 李承家 韩敬伟 黄 曼
(杭州电子科技大学信息工程学院 杭州 310000)

摘 要:瞳孔定位在人机交互和生物医学计算应用中起着至关重要的作用。目前,许多复杂的瞳孔定位算法都是通过单幅图像来检测和定位瞳孔位置的。然而,瞳孔运动是一个连续的过程。因此,当无法在当前帧中准确检测和定位瞳孔位置时,可以通过结合前几帧的信息来推断瞳孔位置。这种方法可以更有效地处理困难和具有挑战性的情况,例如反射、睫毛和眨眼遮挡瞳孔,以及偏离中心的瞳孔位置和运动模糊。因此,该方法可以显著提高瞳孔检测的准确性和稳健性,减少定位误差。基于此,一种基于深度学习的多帧瞳孔检测算法被提出。该算法在 Unet 的编码-解码结构的基础上引入连续人眼注视场景的多帧信息进行瞳孔检测。将卷积神经网络 CNN 与卷积长短期记忆网络 ConvLSTM 和注意力机制 CBAM 结构相结合,提出了一种混合语义分割网络。在瞳孔数据集上的实验表明,所提算法相较于其他算法具有更好的性能表现,其中均值交并比 MIoU 得分达 96.78%,均方根误差 RMSE 为 3.83,尤其在处理复杂情况下表现出色。

关键词:卷积神经网络;卷积长短期记忆网络;Unet;瞳孔检测;深度学习
中图分类号:TP391;TN91 **文献标识码**:A **国家标准学科分类代码**:520.6040

Multi-frame pupil detection algorithm based on deep learning

Zhang Guojing Li Chengjia Han Jingwei Huang Man
(Information Engineering College, Hangzhou Dianzi University, Hangzhou 310000, China)

Abstract: Pupil localization plays a crucial role in human-computer interaction and biomedical computing applications. Currently, many sophisticated pupil localization algorithms are designed to detect and locate the pupil position using one single image. However, pupil movement is a continuous process. Therefore, when the pupil position cannot be accurately detected and located in one current frame, the pupil position can be inferred by combining information of previous frames. This approach can more effectively handle difficult and challenging situations such as reflections, pupil occluded by eyelashes and blinks, as well as off-center pupil positions and motion blur. Consequently, it can significantly improve the accuracy and robustness of pupil detection, decreasing localization errors. To address these challenges, propose a pupil detection algorithm based on deep learning using multiple consecutive images. This algorithm enhances the standard Unet encoder-decoder structure by incorporating multi-frame information from continuous eye tracking scenes for improved pupil detection. By combining convolutional neural networks with convolutional long short-term memory networks and a convolutional block attention module, we introduce a hybrid semantic segmentation network. Experiments on a large-scale dataset demonstrate that the proposed method outperforms existing pupil detection algorithms, achieving a mean intersection over union score of 96.78% and a root mean square error value of 3.83, especially in challenging situations.

Keywords: CNN;ConvLSTM;Unet;pupil detection;deep learning

0 引 言

眼睛在人们的日常互动和交流中扮演着至关重要的角色。近年来,对眼睛运动机制的研究日益深入,为本文理解和洞察人类行为认知提供了重要线索。瞳孔检测作为这一研究领域的关键环节,提供了强有力的支持。准确检测瞳孔边界和中心坐标对于多种应用至关重要,包括人机交互[1-3]、驾驶员疲劳检测[4-5]、虚拟现实[6-8]、虹膜识别[9-11]、心

收稿日期:2024-12-06
* 基金项目:杭州电子科技大学信息工程学院科研培育基金(KYP0323015)、浙江省教育厅科研项目(Y202147108)资助

理学和神经科学[12]、神经退行性疾病的诊断[13]等。然而,现有方法大多局限于处理单帧图像,无法有效处理复杂场景下的挑战,如光线条件、眼镜反射、睫毛遮挡等。这些挑战导致算法的精度不足,往往会导致瞳孔位置的不稳定性增加,进而降低了瞳孔检测准确度。这种准确度的下降可能对后续的疾病诊断产生不利影响,特别是在诊断前庭性疾病时,微小眼动的错误识别可能导致误诊。因此,在复杂环境中准确定位瞳孔位置是人机交互领域的重要课题之一。

近年来随着人机交互技术的不断发展,瞳孔检测领域已经涌现出许多高效的算法。其中,一种是基于非监督学习的方法试图通过采用自适应边缘检测或迭代方法来定位瞳孔位置,如 StarBurst[14]、ExCuSe[15]和 ElSe[16]。另一种是基于监督学习方法对瞳孔进行分割,如卷积神经网络(convolutional neural network,CNN)[17-18]、Unet[19-20]等架构。然而,这些方法大多局限于单帧图像的处理,这将导致在处理具有挑战性的复杂场景时性能较低,如光线条件不良、强反射,睫毛及眨眼的遮挡等问题。在这些情况下,瞳孔可能被部分检测到,甚至根本无法被检测到。一个主要原因是,目标帧提供的信息不足以进行准确的瞳孔检测或预测。一般来说,实时瞳孔检测需要考虑到瞳孔的连续性和动态性,而不仅仅局限于单个图像帧。由于瞳孔视频图像场景是连续的,并且在相邻帧之间有很大的重叠,因此相邻帧的瞳孔位置是高度相关的。更准确地说,目标帧中瞳孔位置可以通过使用多个先前的帧来预测,即使瞳孔图像可能受到光线条件不良、强反射、睫毛遮挡、眨眼,以及瞳孔模糊和抖动。这促使本文通过使用连续瞳孔图像研究瞳孔检测。

现今,深度学习作为一种强有力的工具,已经广泛应用于计算机视觉领域,特别是在图像分类、分割和目标跟踪等任务中,取得了显著的成果。现有时序方法通常将 CNN 和递归循环网络(recurrent neural network,RNN)结合起来处理连续时间序列,以提高预测精度[21-24]。由先前的研究可知,长短期记忆(long short-term memory,LSTM)是一种特殊 RNN 结构,LSTM 及其变体具有序列预测的特征,可以存储和学习历史信息[25],这些结构能够捕捉长期依赖信息,有效解决梯度爆炸和梯度消失的问题。

为了捕获时空依赖关系,Donahue 等[26]提出了一种混合 CNN-LSTM 框架(由 FC-LSTM 扩展得到)。CNN 被用来提取特征,然后作为 LSTM 模型的输入,LSTM 被用来学习时间序列信息[27]。Bogaerts 等[27]提出了一种图 CNN-LSTM 神经网络,该网络同时提取交通的空间特征,通过 LSTM 细胞利用图卷积及其时间特征进行短期和长期预测。几项研究发现,混合 CNN-LSTM 模型优于独立 CNN 和 LSTM 模型[28-29]。这些方法在时空预测任务上取得看令人印象深刻的结果。

基于此,本文提出了一种基于多个连续帧的混合深度神经网络,利用了卷积长短期记忆(convolutional LSTM,ConvLSTM)结构进行时序特征学习,并结合卷积块的注意力机制(convolutional block attention module,CBAM)构建了 U 型网络结构。该网络利用对特征序列进行时空建模,实现了对瞳孔位置的连续预测,从而在复杂场景下实现了精确的检测和跟踪。

在这项工作中,本文采用 Unet 作为主干网络,结合 ConvLSTM 块和 CBAM 模块来构建本文的网络。首先,编码器 CNN 生成的特征将被输入到 ConvLSTM 块中,进行序列特征的学习。ConvLSTM 块负责接收并处理编码器传递过来的特征序列,以捕捉时间序列中的动态信息和时空关系。然后,解码器 CNN 网络接收 ConvLSTM 块输出的目标时间帧的特征,同时也接收来自上一级解码器的上采样特征。这些特征经过融合操作,通过通道拼接操作,确保解码器能够同时利用来自编码器和解码器自身的特征信息。最后这样,解码器能够更好地理解目标时间帧的上下文信息,并生成更准确的目标区域和预测结果。

1 提出方法

1.1 系统概述

在本节中,本文提出了一种新的混合神经网络,将 Unet 作为主干网络,利用 ConvLSTM 网络和 CBAM 注意力机制进行时空特征的提取,完成瞳孔检测任务。该方法利用 ConvLSTM 网络进行瞳孔检测,采用多帧连续的注视场景。在连续注视场景中,摄像头拍摄的图像是连续的,目标帧和前一帧的眼睛运动通常是重叠的,这使得瞳孔检测可以在时间序列预测框架内进行。所提出的网络架构,如图 1 所示。编码器 CNN 和解码器 CNN 是两个全卷积网络。以多个连续帧作为输入,每一帧通过编码器 CNN 处理后,生成一个时间序列的特征图,将这些时序特征堆叠在一起,生成时空特征张量。随后,时空特征张量被输入到 ConvLSTM 网络,以提取瞳孔的时序信息。ConvLSTM 的输出被输入到解码器 CNN 中,通过将上一层特征图拼接在一起生成编码层的特征图,逐层上采样最后生成瞳孔预测的概率图。

1.2 网络设计

1)ConvLSTM 网络结构

多个连续帧的视频图像被建模为时间序列,在该网络中,LSTM 块接收编码器 CNN 在每一帧上提取的特征映射作为输入,ConvLSTM 的表达式如下:

$$i_t = \sigma(\boldsymbol{W}_{xi} * \boldsymbol{X}_t + \boldsymbol{W}_{hi} * \boldsymbol{H}_{t-1} + \boldsymbol{b}_i) \tag{1}$$

$$f_t = \sigma(\boldsymbol{W}_{xf} * \boldsymbol{X}_t + \boldsymbol{W}_{hf} * \boldsymbol{H}_{t-1} + \boldsymbol{b}_f) \tag{2}$$

$$o_t = \sigma(\boldsymbol{W}_{xo} * \boldsymbol{X}_t + \boldsymbol{W}_{ho} * \boldsymbol{H}_{t-1} + \boldsymbol{b}_o) \tag{3}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tanh(\boldsymbol{W}_{xc} * \boldsymbol{X}_t + \boldsymbol{W}_{hc} * \boldsymbol{H}_{t-1} + \boldsymbol{b}_c) \tag{4}$$

$$\boldsymbol{H}_t = o_t \odot \tanh(C_t) \tag{5}$$

其中,“*”为卷积算子,“⊙”为 Hadamard 积。$\boldsymbol{X}_t$ 表

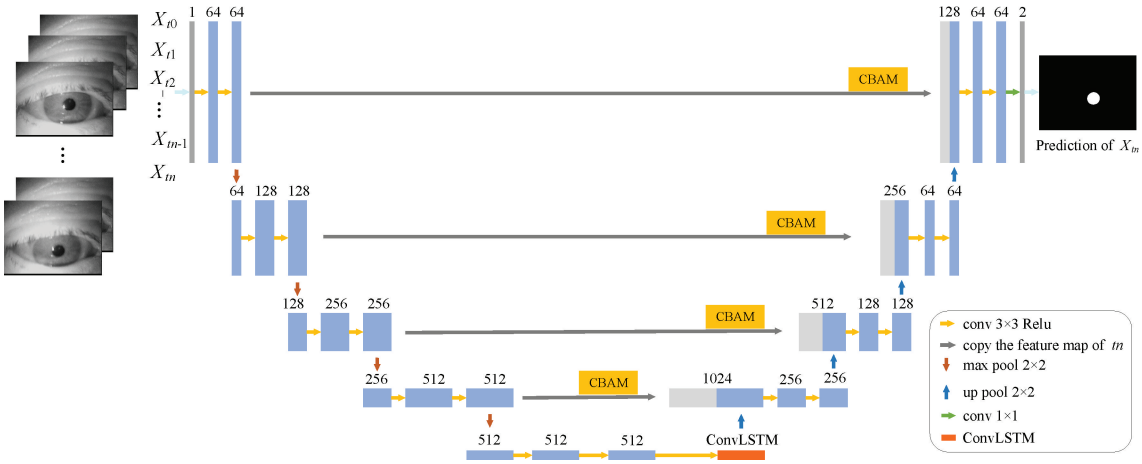


图 1　网络模型结构框图

Fig. 1　Structure diagram of network model

示编码器 CNN 在时刻 t 提取的输入特征映射。i_t、f_t 和 o_t 分别为输入门，遗忘门和输出门。C_t、$\boldsymbol{H}_t$、C_{t-1} 和 $\boldsymbol{H}_{t-1}$ 分别表示时刻 t 和时刻 $t-1$ 的记忆激活和输出激活。$\boldsymbol{W}_{xi}$ 是输入 $\boldsymbol{X}_t$ 对输入门的权值矩阵，$\boldsymbol{b}_i$ 是输入门的偏置。其他 $\boldsymbol{W}$ 和 $\boldsymbol{b}$ 的含义可以从上述规则中推断出来。$\sigma(\cdot)$ 为 Sigmoid 运算，tanh(·)为双曲正切非线性。

2)注意力机制

本文注意力机制 CBAM 是由通道注意力模块和空间注意力模块两部分组成[30]。CBAM 通过将注意力过程划分为两个独立的注意力模块，结构如图 2 所示，不仅能够减少参数量和计算成本，还使其能够作为即插即用的模块融入到现有网络架构中。

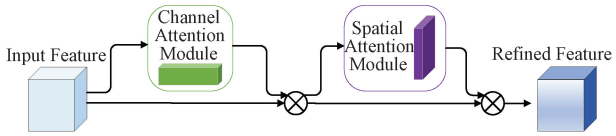


图 2　CBAM 结构

Fig. 2　Structure of CBAM

(1)通道注意力模块

在通道注意力模块中，每个特征通道对应一个特定的检测器，旨在选择具有意义的特征[30]。为了提取空间特征，该模块还结合了全局平均池化和最大池化，从而综合利用不同的信息，其表达式如下：

$$\boldsymbol{F}_{avg}^{C}=AvgPool(\boldsymbol{F}) \tag{6}$$

$$\boldsymbol{F}_{\max}^{C}=MaxPool(\boldsymbol{F}) \tag{7}$$

$$\begin{aligned}\boldsymbol{M}_C(\boldsymbol{F})=&\sigma(MLP(AvgPool(\boldsymbol{F}))+\\&MLP(MaxPool(\boldsymbol{F})))=\\&\sigma(\boldsymbol{W}_1(\boldsymbol{W}_0(\boldsymbol{F}_{avg}^{C}))+\boldsymbol{W}_1(\boldsymbol{W}_0(\boldsymbol{F}_{\max}^{C})))\end{aligned} \tag{8}$$

其中，$\boldsymbol{F}\in \mathrm{R}^{H\times W\times C}$ 为输入特征，它经过全局平均池化和最大池化操作[30]，如图 3 所示，生成两个 $1\times1\times C$ 的特征图，如式(6)和(7)。随后，两个池化操作被送入到两个多层感知机(multi-layer perceptron，MLP)中，输出 $1\times1\times C$ 的特征图。这两个网络构是由一个包含 C/r 个神经元且 ReLU 激活函数和一包含 C 个神经元组成。该网络将两层神经网络的输出相加之后，通过 $\sigma(\cdot)$函数生成通道权重 $\boldsymbol{M}_C\in \boldsymbol{R}^{1\times1\times C}$。MLP 结构采用了 Conv-ReLU-Conv，其中 $\boldsymbol{W}_0$ 和 $\boldsymbol{W}_1$ 为 MLP 权重，Sigmoid 用于生成归一化权重。

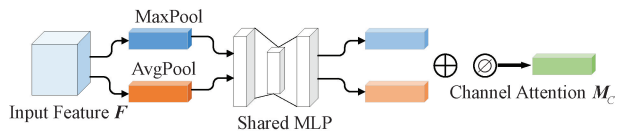


图 3　通道注意力机制

Fig. 3　Channel attention mechanism

(2)空间注意力模块

通道注意力模块之后连接着空间注意力模块，该模块是用于识别具有重要意义的空间位置特征[30]。给定输入特征 $\boldsymbol{F}'\in \mathrm{R}^{H\times W\times C}$，首先对通道维度分别进行平均池化和最大池化操作，得到两个 $H\times W\times1$ 的特征图，如图 4 所示，并将它们在通道维度上拼接[30]。接着，利用 7×7 的卷积层(Sigmoid 激活函数)生成空间权重 $\boldsymbol{M}_S\in \mathrm{R}^{H\times W\times1}$。最终，将权重 $\boldsymbol{M}_S$ 与特征 $\boldsymbol{F}'$ 元素相乘，得到调整后的新特征 $\boldsymbol{F}''$。数学表达式如下：

$$\begin{aligned}\boldsymbol{M}_S(\boldsymbol{F})=&\sigma(\boldsymbol{f}^{7\times7}(Concat[AvgPool(\boldsymbol{F}),\\&MaxPool(\boldsymbol{F})]))\end{aligned} \tag{9}$$

$$\boldsymbol{F}'=\boldsymbol{M}_C(\boldsymbol{F})\otimes\boldsymbol{F} \tag{10}$$

$$\boldsymbol{F}''=\boldsymbol{M}_S(\boldsymbol{F}')\otimes\boldsymbol{F}' \tag{11}$$

其中，$\otimes$ 为逐元素乘法，$\boldsymbol{f}^{7\times7}$ 为 7×7 的卷积核。

1.3　损失函数

在训练模型图像分割时，选择合适的损失函数是非常重要的。本文使用了交叉熵损失函数作为主要的损失函数。同时，考虑到瞳孔近似椭圆形状限制，本文还使用了椭圆拟合误差作为形状的先验损失项，其基本思想是期望分割网络输出为椭圆形状，因此，本文将预测瞳孔图像与真实值之间的偏差添加到损失函数中。本文使用的交叉熵损失

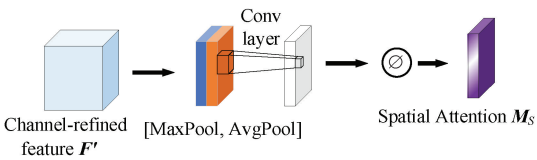


图 4 空间注意力机制

Fig. 4 Spatial attention mechanism

函数[31],损失函数公式表示如下:

$$L_{CE} = -\sum_{c=1}^{M} y_c \log(p_c) \tag{12}$$

其中,M 是类别数,y_c 是样本标签,p_c 是样本属于类别 c 的预测概率。本文属于二分类问题,即 M 设置为 2。

本文在损失函数中添加正则项约束,对预测边界进行椭圆边界限制,公式表示如下[32]:

$$L_{EFE} = \frac{1}{N} \sum_{\substack{p \in e_{GT} \\ q \in e_p}} \| p - q \|^2 \tag{13}$$

其中,$\| \cdot \|^2$ 为最小二乘误差(欧氏距离),p 为真实椭圆边界像素点,q 表示估计的瞳孔边界像素点。e_{GT} 表示真实椭圆边界像素,e_p 表示预测瞳孔边界像素,N 表示预测瞳孔边界的像素个数。L_{EFE} 最小化,因为检测到的瞳孔区域收敛到真实椭圆,因此迫使网络产生椭圆推断。

最终网络模型在训练过程中使用的整个损失函数,可以表示为:

$$L = L_{CE} + \lambda \times L_{EFE} \tag{14}$$

其中,λ 为超参数,设置为 0.1。

2 实验和结果

2.1 数据集

本文从 TEyeD 人眼图像开源数据集[33],以及 ExCuse 和 Else 数据集[15-16],构建一个瞳孔数据集,包含 166 325 个原始图像和人工标记的瞳孔分割图像。每张图像的大小为 384 pixel×288 pixel,每个序列包含了在 1 s 内收集的 25 个连续帧。这些图像是通过头戴式设备拍摄的,采集于驾驶汽车驾驶,模拟驾驶,室内和室外活动,包括 VR 和 AR 等不同的任务场景。数据集包含了多种干扰因素,如光线不足导致的反射、眨眼和睫毛遮挡以及运动模糊等,如图 5 所示。本文将数据集按 7∶3随机分为训练集和测试集。在训练和测试过程中,本文选择 5 个连续图像作为输入,以最后一帧的图像作为目标帧来训练所提出的网络。

2.2 实验环境

本研究是使用了一台配备 Inter(R) Core(TM) i9-10900X CPU@3.70 GHz,RAM 为 128 GB 的 PC 和 11 GB NVIDIA GeForce GTX 2080ti GPU 的计算机上进行训练和性能评估,并使用 Pytorch 库实现。在本研究中,所有模型的输入图像尺寸设为 256×256,需将采集到的瞳孔图像分辨率从 384 pixel × 288 pixel 调整为 256 pixel × 256 pixel。本文选择了 Adam 优化器,并将初始学习率设置为 1×10^{-5},而在训练过程中会自适应调整学习率。训练时,batch size 设置为 6 和 epoch 设置为 50。

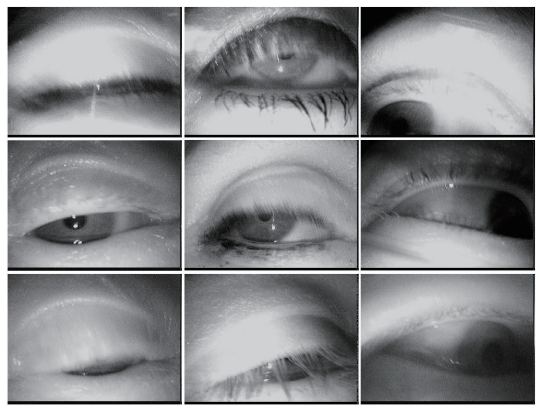


图 5 真实世界场景中受各种因素影响的示例图像:眨眼、睫毛遮挡和眼角遮挡

Fig. 5 Example images in real-world scenarios that are affected by various factors: blinking, eyelash occlusion, and eye-corner occlusion

2.3 评价指标

在本文中,本文引入了一些最流行和广泛使用的图像分割评估指标。其中包括准确度(accuracy,Acc)、精确度(precision,Pre)、召回率(recall,Rec)和 F1 分数,以及交并比(mean intersection-over-union,MIoU)来评估分类器的性能。这些指标的计算方式如下:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{15}$$

$$Pre = \frac{TP}{TP + FP} \tag{16}$$

$$Rec = \frac{TN}{TP + FN} \tag{17}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{18}$$

$$MIoU = \frac{1}{k} \sum_{i=1}^{k} \frac{TP}{TP + FN + FP} \tag{19}$$

其中,TP、FP、FN 和 TN 分别表示真阳性、假阳性、假阴性和真阴性。MIoU 是语义分割任务中的标准度量,通过计算真实样本(ground truth,GT)与预测样本的交集与并集之比来评估,具体公式如式(19)所示。

同时,本文通过计算估计的瞳孔中心坐标与人工标记坐标之间的欧氏距离来评估算法的准确性,即均方根误差(root mean squared error,RMSE)如下:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} ((x'_i - x_i)^2 + (y'_i - y_i)^2)} \tag{20}$$

其中,(x'_i, y'_i) 表示估计的坐标,(x_i, y_i) 表示人工标记的坐标,N 为测试图像数。RMSE 含义是值越小,说明模型能够更好的获得瞳孔中心位置。

此外，为了评估瞳孔定位的准确性，本文计算了误差范围为 Num 个像素内的瞳孔中心数量，即检测率(detection rate，DR)。主要思路是，由于数据标注过程中会存在一定误差，通常将欧氏距离不超过 Num 个像素点视为准确瞳孔定位，其 DR 的计算公式如下：

$$DR = \frac{n}{N} \times 100\% \tag{21}$$

其中，n 代表不超过 Num 个像素的图像数，N 代表数据集的总图像数。其中，本文 Num 设置为 3。DR 越低，模型匹配误差的性能越好。

2.4　实验结果与讨论

本文将提出的 ConvLSTM-CBAM 与现有的一些方法进行了比较，包括单帧语义分割算法和多帧语义分割算法。单帧分割模型有：Unet[19]、TransUnet[34]、SwinUnet[35]、文献[36]和只引入注意力机制 CBAM 的网络。而多帧分割模型有：传统 FCLSTM 网络和只引入 ConvLSTM 的网络。本文在构建的瞳孔数据集上进行了训练和测试，从视觉检测、定量分析和鲁棒性分析三方面对网络模型的性能进行了评估。

1)视觉检测

实验表明，本文的网络在视觉检测方面击败了其他模型框架，如图 6 所示。图 6 表示为不同网络下的瞳孔预测结果，可以看出，由于单帧分割模型缺少连续帧之间的时序信息，分割的结果中存在许多语义碎片，整体分割精度没有多帧分割模型性能精度高。对于单帧分割模型，多帧分割模型减少了分割误差，能够利用连续帧的时序信息准确识别瞳孔目标位置。首先，多帧分割模型能够识别输入连续时间帧图像中的瞳孔目标，而不会遗漏检测或过度检测的情况。单帧分割模型容易将其他区域误识别为目标区域，如眼睫毛。根据本文的预测图，每个目标区域都对应 GT 中的瞳孔区域，显示出多帧分割模型具有较强的边界识别能力，保证了瞳孔位置的准确性。

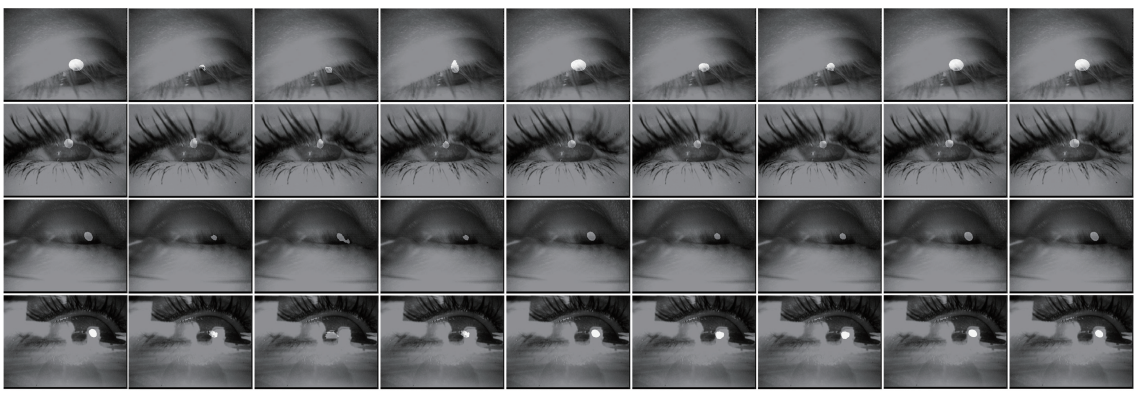


图 6　各种分割算法在瞳孔分割结果的视觉对比

Fig. 6　Visual comparison of the results of pupil segmentation by various segmentation algorithms

然后，在多帧分割模型结果中，瞳孔的位置严格符合背景真实情况，这有助于人机交互系统在真实场景中对瞳孔大小和位置的预测，使其更加可靠和可信。相比之下，单帧分割模型的瞳孔位置更可能偏离正确位置一段距离，预测的瞳孔区域可能不完整的或形状不符合椭圆规则，只能覆盖实际瞳孔的一部分。相比之下，多帧分割模型则能更接近真实值，对瞳孔位置和形状的检测精度较高，如图 6 所示。

最后，在多帧分割模型结果中，瞳孔以近似椭圆的白色形状呈现，较少出现消失点及附近粘连等模糊区域，也较少出现由于瞳孔遮挡或反光等因素导致的模糊预测和未找到目标的情况。此外，多帧分割网络足够强大，当瞳孔有遮挡或形状不规则时，仍可以识别出完整的瞳孔。这些视觉结果显示了与标签高度一致性，表明多帧分割模型在瞳孔检测任务中可以取得良好的性能。

2)定量分析

在训练和测试阶段，单帧网络模型的时间消耗和计算资源需求均小于多帧网络模型，如表 1 所示。具体来说，在训练时，所提网络在完成一代 Epoch 训练所需时间约 80 min 计算资源 GPU 需求约 9.2 GB，而对比 Unet 仅需 28 min 计算资源 GPU 需求为 6.6 GB。在测试时，所提网络确定目标帧瞳孔位置所需推理时间约 33.91 ms，而 Unet 仅需约 8.98 ms。虽然所提网络需要大量时间和计算资源，但是得到更为精确的瞳孔位置，并且预测目标帧所需时间足够的，小于 50 ms。

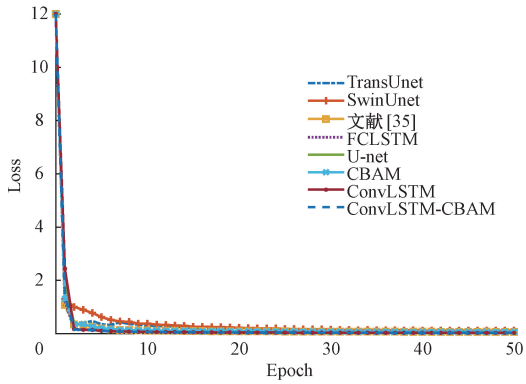
图 7 展示了各种网络模型的 Loss 和 MIoU 曲线，显示了不同模型在训练过程中的性能表现。从图 7 中可以看出，多帧分割模型相比于单帧分割模型具有更快的收敛速度。在训练初期，随着 epoch 增加，Loss 值显著降低，而 MIoU 则不断上升。当训练的 epoch 数大于 20 时，Loss 和 MIoU 趋于稳定，表明多帧分割模型在这个阶段达到了较好的稳定性和鲁棒性。特别值得注意的是，FCLSTM、ConvLSTM 和 ConvLSTM-CBAM 模块在所有模型中表现尤为突出。这两个模块不仅在收敛速度上占据优势，而且在最终的稳定阶段表现出更低的 Loss 值和更高的 MIoU

值,展示了它们在多帧分割任务中的优越性能。结果表明,FCLSTM、ConvLSTM 和 ConvLSTM-CBAM 能够更有效地捕捉时序信息和特征关系,从而显著提升模型的训练速度和效果。

表 1 不同网络模型的性能开销

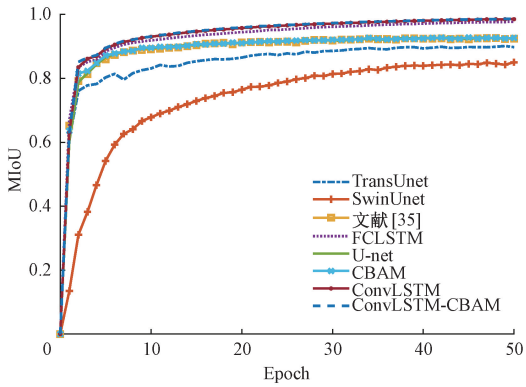
Table 1 Performance overhead of different network models

方法	训练过程		测试过程	
	耗时/(min•周期$^{-1}$)	计算资源 GPU 显存和利用率	耗时/(ms•单次预测$^{-1}$)	计算资源 GPU 显存和利用率
TranUnet	20	6.1 GB, 5%～6%	7.98	1.9 GB, 8%～14%
SwinUnet	20	4.1 GB, 5%～6%	20.94	2.0 GB, 7%～18%
文献[36]	47	6.5 GB, 7%～9%	11.96	1.9 GB, 10%～11%
FCLSTM	135	10.7 GB, 9%～17%	19.95	3.5 GB, 8%～9%
Unet	28	6.6 GB, 6%～7%	8.98	1.9 GB, 10%～12%
CBAM	30	6.6 GB, 6%～7%	10.97	1.9 GB, 10%～11%
ConvLSTM	74	8.7 GB, 5%～7%	27.93	2.0 GB, 7%～8%
本文方法	80	9.2 GB, 6%～7%	33.91	2.2 GB, 6%～7%



(a) 不同迭代次数Loss比较

(a) Comparison of Loss function with different



(b) 不同迭代次数MIoU比较

(b) Comparison of MIoU with different iteration times

图 7 不同网络模型在训练集下的 Loss 值与 MIoU 值曲线

Fig. 7 Loss and MIoU curves of different network models on the training set

表 2 展示了不同网络模型在测试集上进行瞳孔分割的定量评估结果。多帧分割网络在 DR 达到大于 99%的检测水平,在 MIoU 上达到大于 96%,比 Unet 提高了约 3.06%,明显优于单帧分割网络模型。结果表明,多帧预测能够充分利用帧与帧之间的时序信息,不仅提高了瞳孔分割的精度,而且在边界匹配度上表现更好。

在 Acc 指标下,如表 2 所示,在编码器和解码器之间添加多帧时间序列模块进行顺序特征学习后,Unet 的准确率增长了约 0.03%。尽管多帧分割模型已经实现了更高的精度,但本文认为这对瞳孔检测来说这不是一个公平的衡量标准。因为瞳孔检测是一项不平衡的二元分类任务,其中代表瞳孔的像素远小于代表背景的像素,并且它们之间的比率低于 1/60。如果本文将所有像素分类为背景,准确率约为 98%。因此,准确率 Acc 只能被视为一个参考指标。

在 Pre 和 Rec 指标下,如表 2 所示,Unet 加入多帧时间序列模块 FCLSTM、ConvLSTM 和 ConvLSTM-CBAM 后,精确度 Pre 有了显著提高,召回率 Rec 也非常接近最佳结果。结果表明多帧模型的精确度比 Unet 提高了约 1.77%,召回率提升了约 1.55%。

从视觉检测中本文可以直观地看出(如图 6),准确率的提高主要是由于预测更小的瞳孔、模糊粘连区域的减少和误分类的减少。多帧分割模型比单帧分割模型检测的瞳孔边缘更加明确,降低了背景像素分类为瞳孔的可能性,从而降低了误报率。减少消失点周围的模糊区域和瞳孔遮挡区域也导致低假阳性,因为背景像素将不再被识别为瞳孔类。从图 6 中可以看出,加入 FCLSTM、ConvLSTM 和 ConvLSTM-CBAM 后,Unet 导致的其他边界被误分类为瞳孔的情况将得到缓解。

考虑到 Pre 和 Rec 指标仅能反映瞳孔检测性能的部分,本文引入 F1 测度作为综合评价标准。在 F1 测度中,Pre 的权重等于 Rec 的权重。它通过无偏倚地综合 Pre 和 Rec 指标来平衡这种对抗。如表 2 所示,多帧分割模型方

法的 F1 与 Unet 模型相比上升了约 1.72%。结果显示，多帧比单帧更好地预测瞳孔位置，并且在多帧处理框架中，时序网络模块 FCLSTM、ConvLSTM 和 ConvLSTM-CBAM 对序列数据的处理是有效的。

表 2　不同网络模型在测试集下的瞳孔检测性能

Table 2　Pupil detection performance of different segmentation networks on the testing set

方法	DR /%	MIoU /%	Acc /%	Pre /%	Rec /%	F1 /%	RMSE	Params /MB
TranUnet	97.87	92.44±8.82	99.94±0.06	96.73±6.64	95.12±8.30	95.74±7.50	16.82	115
SwinUnet	96.15	90.08±10.20	99.92±0.07	95.62±7.86	93.49±9.27	94.35±8.31	15.89	**68.5**
文献[36]	98.90	93.70±5.80	99.95±0.04	97.35±4.12	96.06±5.14	96.62±4.35	7.13	133
FCLSTM	99.82	96.62±3.16	**99.97±0.02**	98.84±2.43	**97.68±2.75**	98.24±2.53	5.11	694
Unet	98.97	93.72±5.94	99.95±0.04	97.29±4.67	96.11±5.25	96.62±4.66	9.37	133
CBAM	98.95	93.69±6.08	99.95±0.05	97.29±4.71	96.07±5.42	96.59±4.82	9.30	136
ConvLSTM	99.81	96.66±2.95	**99.97±0.02**	99.05±2.05	97.53±2.58	98.27±2.10	4.38	195
本文方法	**99.86**	**96.78±2.89**	**99.97±0.02**	**99.06±1.71**	97.66±2.50	**98.34±2.04**	**3.83**	195

考虑到瞳孔检测的目的是实现瞳孔中心的精确定位，本文引入了 RMSE 测度作为评价指标。通过比较多帧分割模型与单帧分割模型的性能，本文发现多帧分割模型在 RMSE 测度上表现更为优越，仅有 3.83。这表明多帧分割模型在捕捉时间序列信息和减少定位误差方面具有优势。在多帧分割模型中，不同时间序列模块的性能表现也有所差异。尽管 ConvLSTM 模型在瞳孔中心定位任务中表现出较低的 RMSE 值，表 2 显示出其在准确定位上的优势，但其参数量为 195 MB，与 ConvLSTM 有相同的计算复杂度。相比之下，ConvLSTM-CBAM 模块的参数量未变，同时保持了较高的定位精度，增强了模型对时空依赖的捕捉能力，实现了高效且准确的瞳孔中心定位。

尽管 ConvLSTM 在某些指标上表现优异，但 ConvLSTM-CBAM 模块通过增加 CBAM 模块实现了更高效的瞳孔中心定位，并未增加参数量，在整体性能上也表现出色。图 8 显示了 4 张瞳孔图像，所提出的方法完美地找到了瞳孔中心位置（显示为红色叉），与 GT 瞳孔中心坐标（显示为绿色十字）很好的重合。ConvLSTM-CBAM 模块不仅提高了定位精度，还大大降低了计算成本，使得其在实际应用中更具实用性。

3）鲁棒性分析

表 3 展示了不同干扰情况下的瞳孔预测结果对比。多帧网络模型即使在不同干扰情况下，也能完美检测到干扰情况下的瞳孔位置。在某些极端情况下，例如，光线弱、眼镜反光、睫毛或眼睑遮挡等，多帧网络模型也可以准确地识别瞳孔。如表五所示，所提出的模型在不同干扰情况的精度方面都优于其他方法，且有较大的提升，并且在大多数场景中都达到了最高的 F1 得分，这表明了所提出模型的优势。主要原因是由于多帧网络模型利用连续帧之间的时空信息，能够有效应对这些干扰，保持较高的检测精度。结果表明，多帧网络模型在复杂环境下展现了较强的鲁棒性，能够有效应对不同干扰，从而实现准确定位瞳孔位置。

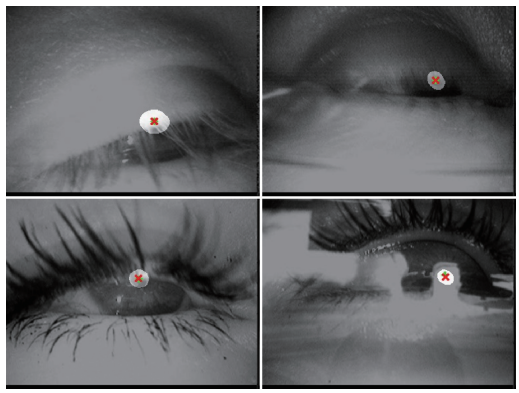


图 8　瞳孔检测结果示例

Fig. 8　Example of pupil detection results

综上所述，尽管本文展示了瞳孔分割算法在实验环境中的有效性，并指出其在人机交互等领域的应用潜力，但尚未在实际应用场景中进行验证。未来的研究应聚焦于在复杂环境中的应用测试，如虚拟现实和自动驾驶等，解决光照、头部姿态变化等因素对算法性能的影响。同时，研究还应关注实时性能优化、个体差异适应以及跨设备验证，进一步提升算法的鲁棒性和应用可行性。

表 3　不同网络模型在不同干扰因素下的 F1 得分

Table 3　F1 scores of different network models under different interference factors

方法 /%	光线变化	眼镜反光	睫毛或眼睑遮挡
TranUnet	93.47±10.77	92.92±10.83	94.81±4.07
SwinUnet	91.92±10.51	89.79±10.45	94.29±2.32
文献[36]	94.60±6.73	95.20±5.55	87.52±4.73
FCLSTM	97.19±3.05	97.19±5.17	98.43±0.33
Unet	94.51±8.25	95.31±5.93	95.88±1.89
CBAM	94.35±8.45	95.32±6.03	96.42±1.49
ConvLSTM	97.28±2.96	97.21±4.34	98.31±0.36
本文方法	**97.32±3.00**	**97.34±4.31**	**98.47±0.28**

3 结 论

本研究提出了一种新的混合神经网络,将 Unet 作为主干网络,将 CNN 与 ConvLSTM 和 CBAM 注意力结构相结合,以进行时空特征的提取,从而完成瞳孔检测任务。该网络结构是由编码器和解码器组成,以多个连续帧作为输入,并通过语义分割的方式预测目标帧的瞳孔。在该框架中,首先通过 CNN 编码器对输入的每一帧的特征进行抽象。然后,对所有输入帧的顺序编码特征进行 ConvLSTM 处理。最后,将 ConvLSTM 的输出送入 CNN 解码器进行信息重构和瞳孔预测。并构建了一个包含连续眼睛图像的数据集用于性能评估。与使用单幅图像作为输入的基准结构相比,该结构取得了明显更好的效果,验证了使用多幅连续帧作为输入的有效性。同时,实验结果表明,在瞳孔检测的背景下,所提混合网络模型(ConvLSTM-CBAM)在序列特征学习和目标信息预测方面优于其他模型。同时,该所提混合网络模型能够稳定地检测各种情况下的瞳孔,并能很好地避免错误识别。

参考文献

[1] CHENG H, LIU Y Q, FU W H, et al, Gazing point dependent eye gaze estimation [J]. Pattern Recognition, 2017,71:36-44.

[2] GALDI C, NAPPI M, RICCIO D, et al. Eye movement analysis for human authentication: a critical survey[J]. Pattern Recognition Letters, 2016,84:272-283.

[3] 孙雨萱,王殊轶.基于眼动追踪的虚拟现实运动评估初步研究[J].电子测量技术,2021,44(18):24-30.
SUN Y X, WANG SH Y. A preliminary study on eye movement evaluation index of cybersickness [J]. Electronic Measurement Technology, 2021, 44(18): 24-30.

[4] ZHUANG Q Y, ZHANG K H, WANG J Y, et al. Driver fatigue detection method based on eye states with pupil and iris segmentation[J]. IEEE Access, 2020,8: 173440-173449.

[5] WENG CH H, LAI Y H, LAI SH H. Driver drowsiness detection via a hierarchical temporal deep belief network [C]. Computer Vision-ACCV 2016 Workshops: ACCV 2016 International Workshops, 2017,13:117-133.

[6] CLAY V, KÖNIG P, KOENIG S. Eye tracking in virtual reality[J]. Journal of Eye Movement Research, 2019,12(1):3-21.

[7] DONUK K, HANBAY D. Pupil center localization based on mini U-net[J]. Computer Science, 2022: 185-191.

[8] LU C, CHAKRAVARTHULA P, LIU K, et al. Neural 3D gaze: 3D pupil localization and gaze tracking based on anatomical eye model and neural refraction correction[C]. 2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE, 2022: 375-383.

[9] 刘笑楠,白雨辰,尹思璐,等.基于类卷积神经网络的可见光虹膜识别方法[J].仪器仪表学报,2017, 38(11): 2651-2658.
LIU X N, BAI Y CH, YIN S L, et al. Iris recognition of visible light based on analogous convolutional neural network[J]. Chinese Journal of Scientific Instrument, 2017,38(11): 2651-2658.

[10] SAAD I A, GEORGE L E, TAYYAR A A. Accurate and fast pupil localization using contrast stretching, seed filling and circular geometrical constraints[J]. Journal of Computer Science, 2014,10(2): 305-315.

[11] ZHOU W J, LU X Y, WAMG Y. A robust pupil localization via a novel parameter optimization strategy[J]. Wireless Communications and Mobile Computing, 2022, 2022(1): 2378911.

[12] CUTUMISU M, TURGEON K L, SAIYERA T, et al. Eye tracking the feedback assigned to undergraduate students in a digital assessment game[J]. Frontiers in Psychology, 2019, 10: 1931.

[13] BRIEN D C, RIEK H C, YEP R, et al. Classification and staging of Parkinson's disease using video-based eye tracking[J]. Parkinsonism & Related Disorders, 2023, 110: 105316.

[14] LI D, WINFIELD D, PARKHURST D J. Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches [C]. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR' 05)-Workshops, 2005: 79.

[15] FUHL W, KÜBLER T, SIPPEL K, et al. Excuse: Robust pupil detection in real-world scenarios[C]. In Computer Analysis of Images and Patterns: 16th International Conference, CAIP 2015. Springer International Publishing, 2015:39-51.

[16] FUHL W, SANTINI T, KÜBLER T, et al. Else: Ellipse selection for robust pupil detection in real-world environments[C]. Ninth Biennial ACM Symposium on Eye Tracking Research & Applications, 2016: 123-130.

[17] FUHL W, SANTINI T, KASNECI G, et al. Pupilnet: Convolutional neural networks for robust pupil detection[J]. Revista De Odontologia Da Unesp, 2016,19(1):806-821.

[18] KONDO N, CHINSATIT W, SAITOH T. Pupil center detection for infrared irradiation eye image using CNN[C]. 2017 56th Annual Conference of the Society

of Instrument and Control Engineers of Japan(SICE). IEEE, 2017: 100-105.

[19] HAN S Y, KWON H J, KIM Y, et al. Noise-robust pupil center detection through CNN-based segmentation with shape-prior loss[J]. IEEE Access, 2020, 8: 64739-64749.

[20] GOWROJU S, KUMAR S, GHIMIRE A, et al. Deep neural network for accurate age group prediction through pupil using the optimized UNet model[J]. Mathematical Problems in Engineering, 2022 (1): 7813701.

[21] GUEN V L, THOME N. Disentangling physical dynamics from unknown factors for unsupervised video prediction[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11474-11484.

[22] LEE S, KIM H G, CHOI D H, et al. Video prediction recalling long-term motion context via memory alignment learning [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 3054-3063.

[23] 陶志勇,曹琦,徐光宪. 基于 FOA-GRNN 模型的太湖水质预测研究[J]. 电子测量与仪器学报, 2021,35(11):83-90.
TAO ZH Y, CAO Q, XU G X. Research on taihu lake water quality prediction based on FOA-GRNN model[J]. Journal of Electronic Measurement and Instrumentation, 2021,35(11): 83-90.

[24] YU W, LU Y CH, EASTERBROOK S, et al. Efficient and information-preserving future frame prediction and beyond[C]. International Conference on Learning Representations, 2020.

[25] WANG Y B, ZHANG J J, ZHU H Y, et al. Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 9154-9162.

[26] DONAHUE J, ANNE H L, GUADARRAMA S, et al. Long-term recurrent convolutional networks for visual recognition and description [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2015: 2625-2634.

[27] BOGAERTS T, MASEGOSA A D, ANGARITA-ZAPATA J S, et al. A graph CNN-LSTM neural network for short and long-term traffic forecasting based on trajectory data[J]. Transportation Research Part C: Emerging Technologies, 2020, 112: 62-77.

[28] KIM T Y, CHO S B. Predicting residential energy consumption using CNN-LSTM neural networks[J]. Energy, 2019, 182: 72-81.

[29] 梁晓龙,李金刚,徐平平,等. 基于 CEEMDAN- CNN-LSTM 的供热异常数据检测与清洗[J]. 电子测量技术,2024,47(11):20-27.
LIANG X L, LI J G, XU P P, et al. Heating data detection and cleaning based on CEEMDAN-CNN-LSTM [J]. Electronic Measurement Technology, 2024, 47(11): 20-27.

[30] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]. European Conference on Computer Vision(ECCV), 2018: 3-19.

[31] ZHANG ZH L, SABUNCU M. Generalized cross entropy loss for training deep neural networks with noisy labels [J]. Advances in Neural Information Processing Systems, 2018, 31, DOI: 10.48550/arXiv.1805.07836.

[32] AKINLAR C, KUCUKKARTAL H K, TOPAL C. Accurate CNN-based pupil segmentation with an ellipse fit error regularization term [J]. Expert Systems with Applications, 2022, 188: 116004.

[33] FUHL W, KASNECI G, KASNECI E. TEyeD: Over 20 million real-world eye images with pupil, eyelid, and iris 2D and 3D segmentations, 2D and 3D landmarks, 3D eyeball, gaze vector, and eye movement types [C]. 2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE, 2021: 367-375.

[34] SHA Y Y, ZHANG Y H, JI X Q, et al. Transformer-unet: Raw image processing with unet[J]. ArXiv preprint arXiv:2109.08417, 2021.

[35] CAO H, WANG Y Y, CHEN J, et al. Swin-unet: Unet-like pure transformer for medical image segmentation[C]. European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 205-218.

[36] 孙语,刘文龙,蒋茂松. 基于注意力机制和空洞卷积的瞳孔定位算法[J]. 电子测量技术,2023, 46(15):126-132.
SUN Y, LIU W L, JIANG M S. Pupil location algorithm based on attention gate and dilated convolution[J]. Electronic Measurement Technology, 2023,46(15):126-132.

作者简介

张国静(通信作者),硕士研究生,助教,主要研究方向为应用数学、模糊数学、信号与信息处理等。
E-mail: zhanggj911109@163.com

李承家,博士,教授,主要研究方向为微分方程的稳定性理论、非线性系统分析与仿真及模糊集理论以及对模糊聚类算法、有效性分析等相关分支的研究。
E-mail:chengjiali@hdu.edu.cn

韩敬伟,博士,副教授,主要研究方向为可积系统等。
E-mail:jingweih@hdu.edu.cn

黄曼,硕士研究生,助教,主要研究方向为应用数学、模糊数学等。
E-mail:20189088@Hziee.edu.cn