多通道时空融合网络双人交互行为识别

裴晓敏[1], 范慧杰[2], 唐延东[2]

(1. 辽宁石油化工大学 信息与控制工程学院, 辽宁 抚顺 113001;
2. 中国科学院沈阳自动化研究所 机器人学国家重点实验室, 辽宁 沈阳 110016)

摘 要: 提出一种基于多通道时空融合网络的双人交互行为识别方法,对双人骨架序列行为进行识别。首先,采用视角不变性特征提取方法提取双人骨架特征,然后,设计两层级联的时空融合网络模型,第一层基于一维卷积神经网络 (1DCNN) 和双向长短时记忆网络 (BiLSTM) 学习空间特征,第二层基于长短时记忆网络 (LSTM) 学习时间特征,得到双人骨架的时空融合特征。最后,采用多通道时空融合网络分别学习多组双人骨架特征得到多通道融合特征,利用融合特征识别交互行为,各通道之间权值共享。将文中算法应用于 NTU-RGBD 人体交互行为骨架库,双人交叉对象实验准确率可达 96.42%,交叉视角实验准确率可达 97.46%。文中方法与该领域的典型方法相比,在双人交互行为识别中表现出更好的性能。

关键词: 双人交互行为; 卷积神经网络; 长短时记忆网络; 时空融合网络; 多通道

中图分类号: TP183 **文献标志码:** A **DOI:** 10.3788/IRLA20190552

Two-person interaction recognition based on multi-stream spatio-temporal fusion network

Pei Xiaomin[1], Fan Huijie[2], Tang Yandong[2]

(1. School of Information and Control Engineering, Liaoning Shihua University, Fushun 113001, China;
2. State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China)

Abstract: Two-person interaction recognition based on multi-stream spatio-temporal fusion was proposed. Firstly, a method to describe two-person's skeleton which invariable with angle of view was proposed. Then a two-layer spatio-temporal fusion network model was designed. In the first layer, the spatial correlation features were obtained based on one-dimensional convolutional neural network (1DCNN) and bi-directional long short term memory(BiLSTM). In the second layer, the spatio-temporal fusion features were obtained based on LSTM. Finally, the multi-stream spatio-temporal fusion network was used to obtain the multi-stream fusion features, which learned one kind of feature by one stream and fusion features for all streams together at last. The weights for each stream was shared, and every stream had the same structure. After features were fusion for all streams, it could be used for interaction recognition. By applying this algorithm to NTU-rgbd datasets, the accuracy for two person interaction recognition for cross-subject could reach 96.42%, and the accuracy of two person interaction recognition for cross-view could reach 97.46%. Compared with the state of art methods in this field, this method performed best in two person interaction recognition.

Key words: two-person interaction; CNN; LSTM; spatio-temporal fusion network; multi-stream

收稿日期: 2020−01−09; **修订日期:** 2020−02−15

基金项目: 国家自然科学基金 (61401455); 辽宁省自然科学基金 (2019ZD0066)

作者简介: 裴晓敏 (1981−), 女, 讲师, 博士后, 主要从事图像处理与机器视觉方面的研究。Email: pxm_neu@126.com

0 引 言

人体行为识别在视频大数据分析、公共安全、人机交互等领域具有广泛应用。根据人体行为活动中执行动作的个体数量不同,可分为单个行为人的单人行为识别、两个行为人的交互行为识别、多个行为人的群体行为识别。双人交互行为中两行为人之间相互关联,如“握手”、“传球”、“互相拍打”等,这些行为经常出现在日常生活中。因而,对交互行为识别具有较高的实用价值和广阔的应用前景。

目前,针对于双人交互行为的识别方法主要有两种。一种是先分别提取每个参与人的运动信息,然后计算双人之间的相关性来预测交互双人的行为,该方法将交互行为看作是两个独立个体的个人行为,忽略了双人间的关联性,且产生大量无用信息。另一种是在每个时间点上提取双人间的运动关联信息,然后建立相关模型,由于该方法多采用自然场景图像,双人紧密接触时交互区域很难定位[1−3]。

近年来,基于深度学习网络的人体骨架行为识别方法受到广泛关注。深度相机可实时采集人体骨架,基于骨架的行为识别方法可避免因背景遮挡和其他环境因素,如光照、物体遮挡等影响识别结果。Yong Du[4] 等提出基于 HBRNN 网络的骨架行为识别方法,将人体行为骨架依照解剖学特征分成若干部分,依次输入到多个级联的循环神经网络 (Recurrent neural network , RNN) 中。

Amir Shahroudy[5] 等提出基于 Part-aware 长短时记忆 (Long short term memory, LSTM) 的骨架行为识别方法,将骨架分成若干部分后,输入到 LSTM 网络中。Pengfei Zhang 等[6] 提出基于 View adaptive LSTM 视角自适应循环神经网络的行为识别方法。Chenyang Si 等[7] 提出基于 Graph LSTM 的骨架行为识别方法,将骨架视为图,骨架点为节点,骨架之间的关联性组成边,利用时空图网络学习骨架的行为特征。

上述方法对单人行为识别的准确率较高,但未考虑双人间的相关性,在双人交互行为识别中准确率较低。为解决这一问题,文中提出针对双人交互行为的多通道时空融合网络双人交互行为识别方法,主要创新点如下:

(1) 提出一种视角不变的双人交互行为描述方法。考虑双人骨架节点间距离不受视角变化影响的特点,定义四组包含骨架自身特征和双人关联特征的骨架距离特征。

(2) 设计两层时空融合网络模型学习序列的时空融合特征。第一层采用基于一维卷积神经网络和双向长短时记忆网络学习序列的空间关联特征 (CNN-BiLSTM); 第二层采用 LSTM 网络学习序列的时间特征。

(3) 提出一种多通道网络结构。在不增加网络参数的前提下,得到多组时空特征,利用四个通道分别处理四组骨架特征,四通道结构相同、权值共享。

1 双人交互特征表示

在交互行为识别中,单人行为特征、双人交互特征同等重要。骨架采集阶段以摄像头为坐标中心,随着摄像头位置的变化,即使同一个人的同一行为骨架坐标也存在较大差异,图 1(a) 为在骨架采集时,以摄像头 P_0 点为坐标原点,得到各骨架点的坐标值,图 1(b) 为各骨架节点的序号。当以两行为人的中心点 o_c 为坐标原点时,行为人本身骨架各节点之间的距离及两参与人之间的距离不随摄像头位置变化,具有视角不变性,如图 1(c) 所示。为此,文中提出基于骨架关节点距离的单人骨架特征和双人交互特征表示方法。

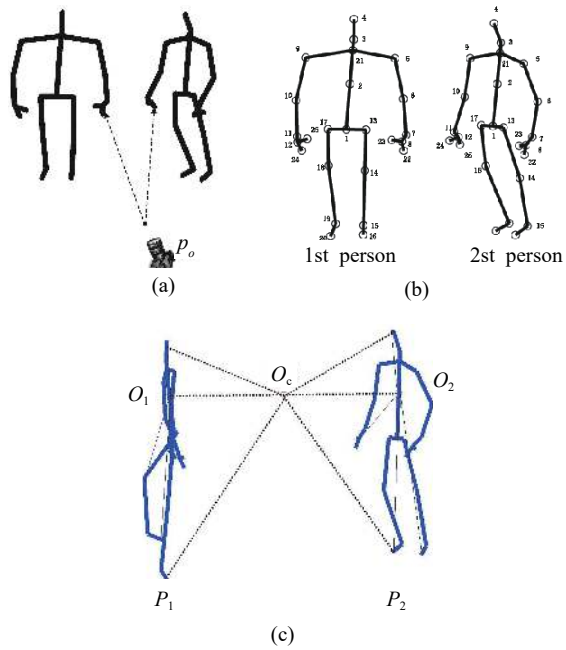

图 1 双人行为骨架

Fig.1 Two person action skeletons

图 1(c) 中，设行为参与人P_1, P_2骨架脊柱中心点O_1,O_2为单人中心点，参与人P_1中心点O_1到该人所有骨架点的距离为单人骨架距离特征d_{i1}，参与人P_2中心点O_2到该人所有骨架点的距离为单人骨架距离特征d_{i2}，采用欧式距离计算d_{i1}, d_{i2}，见公式 (1)：

$$d_{ij}=\sqrt{(x_i-x_j)^2+(y_i-y_j)^2+(z_i-z_j)^2}$$
$$i\in(1,2,\ldots,25),j\in(1,2) \tag{1}$$

设O_1, O_2连线中点为双人中心点O_c，分别计算P_1, P_2所有骨架点到双人中心点O_c的距离，见公式 (2)，得到双人关联的距离特征dc_{i1}, dc_{i2}。

$$dc_{ij}=\sqrt{(x_{ij}-x_{12})^2+(y_{ij}-y_{12})^2+(z_{ij}-z_{12})^2}$$
$$i\in(1,2,\ldots,25),j\in(1,2) \tag{2}$$

计算行为序列的单人骨架特征、双人关联骨架特征，得到四组骨架距离特征序列D_1, D_2, Dc_1, Dc_2，见公式 (3)：

$$\left.\begin{aligned}D_1&=\{d_{1j}(t),t\in(1,2,\ldots,T),j\in(1,2\ldots,25)\}\\D_2&=\{d_{2j}(t),t\in(1,2,\ldots,T),j\in(1,2,\ldots,25)\}\\Dc_1&=\{dc_{1j}(t),t\in(1,2,\ldots,T),j\in(1,2,\ldots,25)\}\\Dc_2&=\{dc_{2j}(t),t\in(1,2,\ldots,T),j\in(1,2,\ldots,25)\}\end{aligned}\right\} \tag{3}$$

2 时空融合特征学习

时空特征融合网络采用两层级联结构，第一层空间特征学习，第二层时间特征学习，最后输出时空融合特征。

空间特征学习网络层如图 2 所示，该层学习$t,t\in(1,\ldots,T)$时刻的骨架空间关系特征。为保持序列的时序性，采用M个一维 CNN 滤波器ω 对长度为T，维度为N的序列F滤波，一维最大池化层提取邻域内的最大特征，得到尺度为(T,P)的M个特征图f_m，见公式 (4)：

$$f_m=\sigma(\omega_{(1,l)}*F+b),m=(1,2,\ldots,M) \tag{4}$$

采用双向长短时记忆网络[5](Bi-LSTM) 学习特征图f_m空间各节点的关联性。LSTM 网络定义如公式 (5) 所示，包含输入门i_t、输出门o_t、遗忘门f_t和记忆门c_t，可避免循环神经网络 (RNN) 引起的梯度消失。Bi-LSTM 网络由前向 LSTM 和后向 LSTM 组合而成。点i, $i\in(1,\ldots,P)$在特征图$(1,\ldots,M)$上的值组成M维特征向量$fs(i,t)=[f_1(i,t),f_2(i,t),\ldots,f_M(i,t)]$。将$fs(i,t)$, $i\in(1,\ldots,P)$输入 Bi-LSTM 网络，输出为t时刻节点关联性的空间特征表示$fsr(t)$，见公式 (6)：

$$\left.\begin{aligned}i_t&=\sigma(W_{iy}y_t+W_{ih}h_{t-1}+b_i)\\f_t&=\sigma(W_{fy}y_t+W_{fh}h_{t-1}+b_f)\\o_t&=\sigma(W_{oy}y_t+W_{oh}h_{t-1}+b_o)\\c_t&=f_t\cdot c_{t-1}+i_t\cdot\tanh(W_{cy}y_t+W_{ch}h_{t-1}+b_c)\\h_t&=o_t\cdot\tanh(c_t)\end{aligned}\right\} \tag{5}$$

$$\left.\begin{aligned}hs_i^{forward}(t)&=LSTM^{forward}(h_{i-1},fs(i,t),c_{i-1})\\hs_i^{backward}(t)&=LSTM^{backward}(h_{i-1},fs(i,t),c_{i-1})\\fsr(t)&=\{hs_i^{forward}(t),hs_i^{backward}(t)\}\end{aligned}\right\} \tag{6}$$

时域 LSTM 网络层学习序列的时序特征，将t时刻的空间特征$fsr(t)$输入到时域 LSTM 网络，学习 1~T时刻序列的时间关联性，得到向量fsr，表示该行为序列的时空融合特征，见公式 (7)：

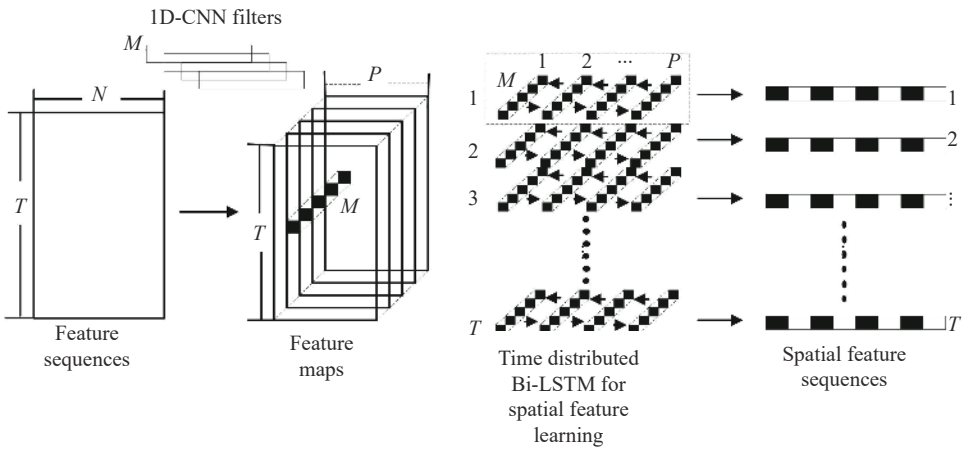

图 2 空间特征学习

Fig.2 Spatial feature learning

$$fsr = LSTM(fsr(t)) \tag{7}$$

3 多通道权值共享时空特征融合网络

设计基于距离特征的多通道特征融合网络分别学习单人特征和双人关联特征，如图 3 所示。输入到该网络各通道的特征都属于骨架的距离特征，有相同的物理意义，因此，采用多通道权值共享时空特征融合网络结构。在单通道中采用多个卷积核提取特征，其余通道采用相同结构，卷积核权值共享。多通道权值共享时空融合网络结构具有两个优点，一是降低网络参数，二是避免多通道结构在训练过程中梯度弥散。

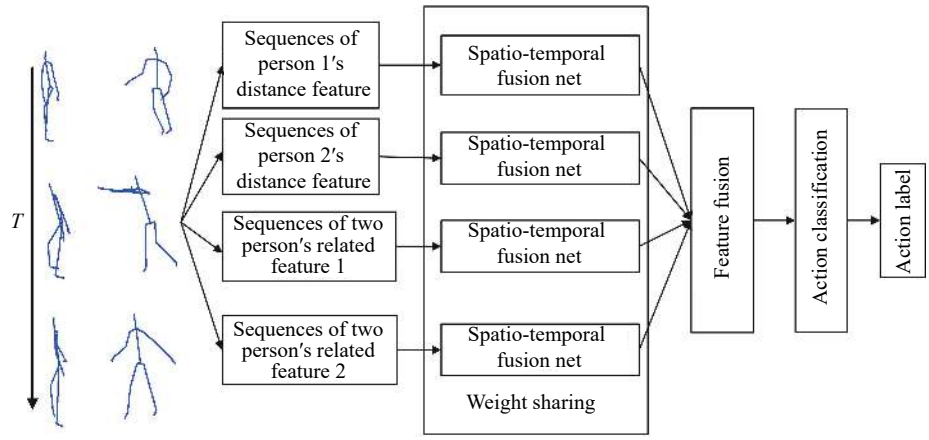

图 3 多通道时空融合网络

Fig.3 Multi-stream spatio-temporal fusion network

将D_1, D_2, D_{c1}, D_{c2}四组序列分别输入到结构为 1DCNN-LSTM LSTM 时空特征融合模型，设该模型实现的函数功能为F_{SPT}，四分支结构相同，权值共享，见公式 (8)：

$$\left.\begin{aligned} fsr_{d1} &= F_{SPT}(D_1) \\ fsr_{d2} &= F_{SPT}(D_2) \\ fsr_{dc1} &= F_{SPT}(D_{c1}) \\ fsr_{dc2} &= F_{SPT}(D_{c2}) \end{aligned}\right\} \tag{8}$$

输出结果融合到一起形成多分支融合特征f_{fusion}，见公式 (9)：

$$f_{fusion} = [fsr_{d1}, fsr_{d2}, fsr_{dc1}, fsr_{dc2}] \tag{9}$$

利用全连接网络学习融合特征得到交互行为标签，见公式 (10)：

$$L = soft\max(W * f_{fusion}) \tag{10}$$

4 实验及结果分析

4.1 网络参数设置

实验中采用 tensorflow+keras 框架设计网络，学习率初始化为 0.001，每 100 次迭代学习率下降 0.1，处理批次为 32，网络采用 rmsprop 优化，损失函数为交叉熵。为防止过拟合，Dropout 参数设置为 0.5。分别进行交叉对象和交叉视角实验[5]，训练集中每种行为包含 10 000 组序列，测试集中每种行为包含 1 000 组序列。

4.2 典型算法对比

采用该领域典型算法对 NTU RGB+D 数据库中交互行为进行识别，其分类准确率见表 1。表中单通道时空融合网络是将上述四组特征全部输入到只含有单个通道的时空融合网络，该网络采用 1D CNN-LSTM LSTM 结构。

表 1 NTU-RGBD 数据库行为识别准确率

Tab.1 Accuracy for human action recognition for NTU-RGBD dataset

Method	Cross-subject	Cross view
HBRNN[4]	59.1%	64.0%
Part-aware LSTM[5]	62.9%	70.3%
VA LSTM[6]	79.4%	87.6%
Trust Gate ST-LSTM[7]	69.2%	77.7%
AGC-LSTM[8]	95.0%	89.2%
ST-GCN[9]	81.5%	88.3%
Single stream (SST)	85.61%	92.42%
Our method (MST)	**96.42%**	**97.46%**

由实验结果可见，多通道时空融合网络交互行为识别交叉对象实验准确率可达 96.42%，交叉视角实验准确率可达 97.46%，在所有方法中行为识别准确率最高。

将多通道时空融合网络应用于 SBU-Kinect interaction dataset 交互行为识别，其准确率如表 2 所示。文中算法对 SBU-kinect dataset 中交互行为识别准确率可达 98.92%，在五种方法中行为识别准确率最高。

表 2 SBU 数据库行为识别结果

Tab.2 Human action recognition accuracy for SBU dataset

Method	Accuracy
Co-occurrence RNN[10]	90.4%
STA-LSTM[5]	91.5%
Trust Gate ST-LSTM[7]	93.3%
VA-LSTM[6]	97.6%
Our method(weighted multi-stream)	**98.92%**

4.3 交互行为识别实验

采用多通道时空融合网络对 NTU-RGBD120 数据库[11−13] 中的 26 种行为分类识别准确率如图 4，5 所示。由图 4 交叉对象实验结果可见，在交叉对象实验中对交互行为识别准确率在 63%~100% 之间，最高可达 100%。由图 5 交叉视角实验结果可见，在交叉视角实验中仅对 hugging, touch pocket 两种行为的识别准确率低于 90%，而对其他行为的识别准确率都在 90% 以上。多通道时空融合网络在交叉视角实验中准确率提升更为明显。

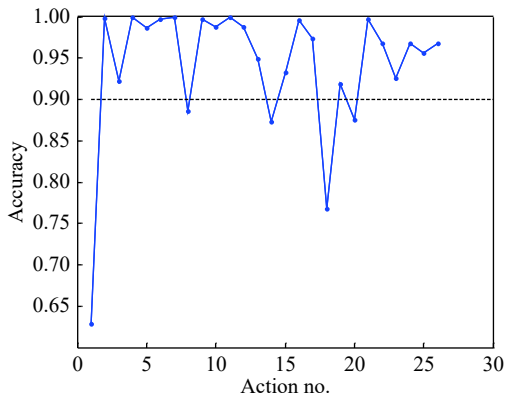

图 4 多通道时空融合网络交叉对象实验行为识别准确率

Fig.4 Multi-stream spatio-temporal fusion model classification accuracy for cross subject

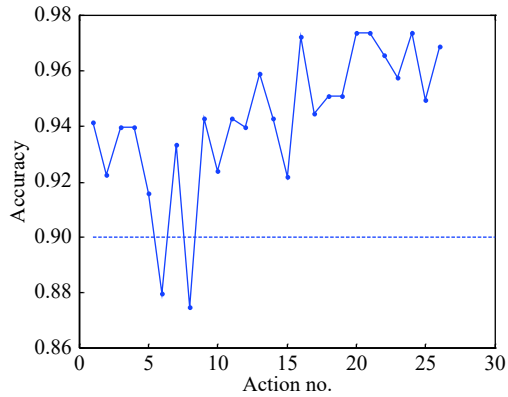

图 5 多通道时空融合网络交叉视角实验行为识别准确率

Fig.5 Multi-stream spatio-temporal fusion model classification

4.4 网络结构对比实验

对多通道时空融合网络和单通道时空融合网络在交互行为识别上进行对比实验，结果如表 3 所示。

表 3 网络结构对比

Tab.3 Comparison of the network structure

Net structural	Parameter	Convergence	Accuracy
Multi-stream spatio-temporal model	*N*	100epochs	96.42%
Single-stream spatio-temporal model	*N*	100epochs	85.61%
Single-stream spatio-temporal model	3*N*	200epochs	86.74%

当多通道网络与单通道网络结构参数相同时，多通道网络行为识别准确率高于单通道网络。继续增加单通道网络的层数和各层参数，当网络参数达到多通道网络的三倍时，行为识别准确率仍低于多通道网络，继续增加参数，行为识别准确率也未见提升，说明多通道时空融合结构可有效提升双人交互行为识别的准确率。

5 结 论

文中提出一种基于深度学习的多通道时空融合网络双人交互行为识别方法，实验结果验证 1DCNN-LSTM LSTM 结构可以提取行为序列的时空融合特征。设计了一种新颖的双人行为表示方法，利用四组距离特征表示原始骨架，得到双人视角不变性特征表示，提升了行为特征的视角不变性。设计多通道权值共享时空特征融合网络模型，采用多个通道分别处理各组特征，多通道间权值共享，在不增加网络参数的

前提下，提取多组时空融合特征。该方法在双人行为识别中具有较高的准确率，与该领域的典型算法对比实验结果表明，文中所提出的方法在双人交互行为识别上具有明显优势。后续将引入多模态行为特征，实现更复杂场景的行为分析。

参考文献：

[1] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C]//Adavances in Neural Information Processing Systems, 2014: 568-576.

[2] Tran D, Bourdev L, Fergus R, et al. Learning spatio temporal features with 3D convolutional networks[C]// 2015 IEEE International Conference on Computer Vision (ICCV), 2015: 4489-4497.

[3] Pei Xiaomin, Fan Huijie, Tang Yandong. Action recognition method of spatio-temporal feature fusion deep learning network [J]. ***Infrared and Laser Engineering***, 2018, 47(2): 0203007. (in Chinese)

[4] Du Yong, Wang W, Wang L. Hierarchical recurrent neural network for skeleton based action recognition[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 1110-1118.

[5] Shahroudy A, Liu J, Ng T -T, et al. NTU RGB+D: A large scale dataset for 3D human activity analysis[C]//Proc CVPR, 2016: 1010-1019.

[6] Zhang Cuiling Lan, Junliang Xing, et al. View adaptive recurrent neural networks for high performance human action recognition from skeleton data[C]//ICCV, 2017.

[7] Li J, Shahroudy A, Xu D, et al. Spatio-temporal lstm with trust gates for 3d human action recognition[C]//ECCV, Springer, 2016: 816-833.

[8] Si Chenyang, Chen Wentao, Wang W. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition[C/OL]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).[2019-06-15]. https://arxiv.org/abs/1902.09130.

[9] Yan Sijie, Xiong Yuanjun, Lin Dahua. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//AAAI, 2018.

[10] Li Chao, Zhong Qiaoyong, Xie Di, et al. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation[C]//IJCAI, 2018.

[11] Zhang Xiangyue, Ding Qinghai, Luo Haibo. Infrared dim target detection algorithm based on improved LCM [J]. ***Infrared and Laser Engineering***, 2017, 46(7): 0726002. (in Chinese)

[12] Wang Zhongyu, Ni Xianyang, Shang Zhendong. Autonomous driving semantic segmentation with convolution neural networks [J]. ***Optics and Precision Engineering***, 2019, 27(11): 0726002. (in Chinese)

[13] Wu Yanfeng, Wang Yanjie, Sun Haijiang, et al. LSS-target detection in complex sky backgrounds [J]. ***Chinese Optics***, 2019, 12(4): 854−866. (in Chinese)