文章编号:1001-9014(2023)06-0906-10 DOI:10.11972/j.issn.1001-9014.2023.06.024

基于自监督学习的热红外图像景深估计方法

丁 萌[1*], 关 松[2], 李 帅[1], 于快快[2], 徐一鸣[1]
(1. 南京航空航天大学 民航学院,江苏 南京 211106;
2. 光电信息控制和安全技术重点实验室,天津 300308)

摘要:从热红外图像对比度低、细节信息不足等特点出发,提出了一种面向热红外图像的景深估计方法。首先,设计了一种红外特征聚合模块,提高了对目标物边缘和小目标的全方位深度信息获取能力;其次,在特征融合模块中引入了通道注意力机制,进一步融合通道间的交互信息;在此基础上,建立了一种深度估计网络,实现热红外图像的像素级景深估计。消融实验与对比实验的结果表明,该方法在热红外图像像素级景深估计中性能优于其他代表性方法。

关 键 词:红外图像;无监督学习;单目深度估计;特征聚合;通道注意力机制
中图分类号:TP29 **文献标识码**:A

Depth estimation of thermal infrared images based on self-supervised learning

DING Meng[1*], GUAN Song[2], LI Shuai[1], YU Kuai-Kuai[2], XU Yi-Ming[1]
(1. College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China;
2. Science and Technology on Electro-Optical Information Security Control Laboratory, Tianjin 300308, China)

Abstract: Depth estimation based on unsupervised learning is one of the important issues in the field of computer vision. However, existing algorithms of depth estimation are mainly designed based on visible images. Compared with visible images, thermal infrared images have the disadvantages of low contrast and insufficient detailed information. To this end, a depth estimation network is constructed and an unsupervised depth estimation method is proposed for thermal infrared images according to their characteristics. The network consists of three parts: feature extraction module, feature aggregation module, and feature fusion module. Firstly, a feature aggregation module is designed to improve network ability to acquire the edge information of target objects and the small object information of the image. Secondly, the channel attention mechanism is introduced in feature fusion module to effectively capture the interaction relationship between different channels. Finally, a depth estimation network for thermal infrared images is established. In this network, the model parameters are trained by thermal infrared sequence images to achieve the pixel-level depth estimation of a single thermal infrared image. The results of ablation studies and comparative experiments fully demonstrate that the performance of the proposed method in pixel-level depth estimation of thermal infrared image outperforms other representative methods.

Key words: thermal infrared image, self-supervised learning, monocular depth estimation, feature aggregation, channel attention mechanism

引言

从二维图像中估计场景的深度信息是计算机视觉中的一个经典问题[1],传统方法多以多目及多视图匹配的方式获取场景的深度信息,然而这类算

收稿日期:2022-12-13,**修回日期**:2023-07-14 **Received date**:2022-12-13,**Revised date**:2023-07-14
基金项目:光电信息控制和安全技术重点实验室开放基金(JCKY2022210C005),国家自然科学基金(U2033201),航空科学基金(20220058052001)
Foundation items: Supported by the Open Foundation of Science and Technology on Electro-Optical Information Security Control Laboratory (JCKY2022210C005), the National Natural Science Foundation of China (U2033201), and Aeronautical Science Foundation of China (20220058052001)
作者简介(Biography):丁萌(1981—),男,江苏仪征人,教授,博士学位,主要研究领域为红外图像智能感知. E-mail: nuaa_dm@nuaa.edu.cn
*通讯作者(Corresponding author): E-mail: nuaa_dm@nuaa.edu.cn

法首先需要完成从图像中进行特征提取与匹配等一系列预处理[2-3],且难以获得稠密的和输入图像实现像素级匹配的深度数据。近年来,随着深度学习算法的发展,利用卷积神经网络强大的拟合能力,建立基于深度学习的单视图景深估计方法成为计算机视觉领域的一个研究热点[4]。相比于传统方法,这类方法的主要优势体现在[5]:第一、尽管需要大规模的训练,但是在实际使用过程中,仅需提供单幅图像即可恢复出其对应的深度信息;第二、恢复出的深度信息与输入的二维图像可以实现像素级匹配,无需再进行深度图与原始图像之间的配准,降低了后续任务的难度。

根据是否需要深度标签及深度标签的稠密度,现有的基于深度学习的单视图景深估计方法总体上可以分为三类[6]:第一、监督学习的方法,该方法需要稠密的像素级的深度标签,由于现有的景深测量设备很难获取和图像分辨率相匹配的深度标签,所以这类算法监督信息的获取难度和成本极大[7-8];第二、半监督学习的方法,该类方法主要使用稀疏的深度标签对图像中的部分像素进行标注,虽然其对深度传感器的分辨率要求大大降低,但是其存在的最大问题是需要先对深度数据和图像进行高精度配准,其工作量较大,也存在着一定的误差[9-10];第三、自监督方法,相比于前两类方法,该类方法无需深度标签,在训练过程中利用多视图之间的空间几何关系去建立监督信息,实现了自监督的像素级深度估计,这类空间几何关系主要包括[11-12]:基于左/右视图的方法和基于前/后视图的方法。由于自监督方法在训练样本的获取上较为便利,因此其已成为当前基于深度学习的景深估计算法的主要研究方向。

2016年,Garg等人创造性地提出了一种利用左/右视图的自监督景深估计方法,根据预测的视差值和左视图重建右视图,其中使用光度误差作为网络模型的监督信息,但该方法最大的缺陷是网络模型优化迭代困难[13]。在此基础上,Godard等人利用左/右视图间的几何约束关系,提出了利用左右视差之间的一致性损失作为约束项,来优化网络参数,提高准确率和鲁棒性[14]。Tosi等人以左/右视图作为训练输入,通过两个网络的优化来预测最终的视差[15],首先通过一个网络得到初始特征和初始视差图,然后通过一个额外的残差网络对前一个网络得到的初始特征进行修正,获得一个对初始视差图的补偿,从而完成对深度信息的预测。相比于Godard的单一网络,该方法通过增加一个网络对特征进行修正,提高了景深估计的精度。这类基于左/右视图的方法,要求在网络训练时需要标定好位姿关系的双目图像作为输入,训练样本的采集对设备有着较为严苛的要求。为此,从简化训练样本获取方式的角度,构建新的自监督景深信息成为研究的重点。Zhou等人发现多视图之间的相机位姿与深度值存在紧密联系[16],在训练时将深度估计网络和相机位姿估计网络联合起来,使用不同位姿之间的视图重投影关系作为监督信号,就可以使用图像序列完成景深信息的估计。但是这种重投影关系的成立条件十分严苛,其中最重要的一点就是要求场景中都是刚性物体,即和相机之间存在着一致的相对运动关系。为了解决这一问题,学者们分别提出了不同的解决方案,如将光流估计和立体匹配结合起来[17],在联合学习过程中引入几何约束[18]。在上述研究的基础上,为了解决多视图相对位姿估计和非刚性及遮挡物体带来的误差,Godard等人提出了一种全新的基于自监督学习的单目深度估计的方法Monodepth2[19],该方法分别使用了深度预测网络和位姿变换预测网络分别对当前帧的深度图和连续帧之间的相对位姿进行估计,通过重投影重构当前图像并计算误差构建自监督信息,其对基于深度学习的无监督景深估计研究的主要贡献来自于三个方面:首先,采用了逐像素计算最小重投影损失的方法,解决了单目视频序列中存在的帧间遮挡问题;其次,采用了多尺度采样方法,使模型尽量在较高的输入分辨率下计算重投影误差,降低了深度图中不合理的纹理特性;最后,使用固定像素自动掩膜的方式过滤掉图像序列中没有变化的像素点,避免深度图中出现异常空洞。综上所述,当前基于深度学习的景深估计方法主要是采用自监督模式,将不同视图之间的几何约束关系作为监督信息,实现对景深信息的恢复。

近年来,热红外成像技术快速发展,该技术通过捕获目标发出的热辐射,经过光电转换后将其转换成图像,避免了光照条件的限制,在一定程度上弥补了可见光图像在低能见度条件下不能使用的不足,目前已经被广泛用于安防监控、设备故障诊断、人体医学影像检查等民用领域以及夜视观察、精确制导等军用领域[20-22]。但是,相比于可见光图像,热红外图像存在着对比度低、分辨率低、目标细

节信息缺失等一系列的不足。现有的基于深度学习的景深估计方法基本都是针对可见光图像的，将其直接应用于红外图像，很难有效提取红外图像的特征[23]。比如，尽管Monodepth2用于单目可见光图像深度估计时具有良好的表现，但由于红外图像有效特征的不足，导致了直接利用该方法进行红外图像的深度估计误差较大，生成的深度图质量也较低，因此亟待提出一种能够有效克服红外图像局限性的深度估计方法。当前，展开面向红外图像的深度学习景深估计方法研究的成果鲜有报道。为此，本文以Monodepth2为基本框架，针对红外图像的特点展开针对性改进，建立了一种性能明显优于Monodepth2的景深估计方法。本文所提出的方法以Monodepth2为基础框架，面向红外图像的特点，设计了新的编解码器结构间的跳连接方式，并引入了通道注意力机制，针对边缘模糊的场景目标以及小目标特征的获取，提高了红外特征的表达能力和图像深度估计性能，并通过实验证明了该方法对于红外图像的优势。本文的主要贡献如下：(1)设计了新的特征聚合模块，将深度估计网络编解码器间的跳连接方式改成了包含上采样等操作的密集跳连接方式，加入了中间节点对不同尺度的特征图进行融合，提高景深估计网络对场景目标物体边缘信息和小物体信息的获取能力；(2)改进了特征融合模块，引入了通道注意力机制（Efficient Channel Attention Net，ECANet），根据特征通道的重要性对不同通道进行权重分配，进而提高神经网络重要通道对输出结果的影响比重，ECANet能够有效地捕获通道间的交互关系，避免了降维给通道注意力预测带来的负面影响；(3)在新的特征聚合和融合模块的基础上，建立了一种面向热红外图像的深度估计网络。消融实验与对比实验结果充分证明了本文提出的深度估计网络模型针对热红外图像具有更好的性能。

1 本文方法

本文提出的基于红外图像序列的景深估计方法，根据红外图像存在有效特征不足的特点，针对Monodepth2中的景深估计网络进行了改进，在多层特征图提取的基础上，建立了自底向上的特征聚合模块，通过引入通道注意力机制ECANet改进了特征融合模块，提高网络的特征提取能力和红外图像深度估计性能。

1.1 基于序列图像的自监督景深估计基本原理

自监督信息的实质是根据相机成像和立体几何投影原理，即同一视频序列的相邻两帧之间存在严格的约束关系，利用这种约束关系即可构建自监督信息。不失一般地假设世界坐标系为前一帧图像所在相机位置的机体坐标系，空间点P在第一位置的相机机体坐标系的位置为(X_1, Y_1, Z_1)，则后一帧图像所在相机位置的机体坐标系为(X_2, Y_2, Z_2)，根据两个相机坐标系之间的转换关系，可得：

$$\begin{bmatrix} X_2 \\ Y_2 \\ Z_2 \\ 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{T} \\ \boldsymbol{0}_{1\times 3} & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ Y_1 \\ Z_1 \\ 1 \end{bmatrix} \quad , \quad (1)$$

其中，$\boldsymbol{R} \in \Re^{3\times 3}$为两个相机位置之间的姿态转移矩阵，$\boldsymbol{T} \in \Re^3$为其位置转移向量，根据小孔成像原理与摄像机内参数矩阵$\boldsymbol{K} \in \Re^{3\times 3}$，空间点P在前一帧和后一帧像素坐标下的位置分别为$(u_1, v_1)$、$(u_2, v_2)$，则$(u_1, v_1)$和$(X_1, Y_1, Z_1)$，$(u_2, v_2)$和$(X_2, Y_2, Z_2)$的关系可表示为：

$$Z_1 \begin{bmatrix} u_1 \\ v_1 \\ 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{K} & \boldsymbol{0}_{3\times 1} \end{bmatrix} \begin{bmatrix} X_1 \\ Y_1 \\ Z_1 \\ 1 \end{bmatrix}, Z_2 \begin{bmatrix} u_2 \\ v_2 \\ 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{K} & \boldsymbol{0}_{3\times 1} \end{bmatrix} \begin{bmatrix} X_2 \\ Y_2 \\ Z_2 \\ 1 \end{bmatrix} \quad , \quad (2)$$

其中，$\boldsymbol{0}_{3\times 1}=[0, 0, 0]^T$，根据式(1)和(2)，可得：

$$Z_2 \begin{bmatrix} u_2 \\ v_2 \\ 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{K} & \boldsymbol{0}_{3\times 1} \end{bmatrix} \begin{bmatrix} \boldsymbol{R} & \boldsymbol{T} \\ \boldsymbol{0}_{1\times 3} & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ Y_1 \\ Z_1 \\ 1 \end{bmatrix} \quad . \quad (3)$$

且根据式(2)可得，

$$\begin{bmatrix} X_1 \\ Y_1 \\ Z_1 \end{bmatrix} = Z_1 \boldsymbol{K}^{-1} \begin{bmatrix} u_1 \\ v_1 \\ 1 \end{bmatrix} \quad . \quad (4)$$

由式(3)进一步可得，

$$Z_2 \begin{bmatrix} u_2 \\ v_2 \\ 1 \end{bmatrix} = \boldsymbol{KR} \begin{bmatrix} X_1 \\ Y_1 \\ Z_1 \end{bmatrix} + \boldsymbol{KT} \quad . \quad (5)$$

因此，由式(4)和(5)可得，

$$Z_2 \begin{bmatrix} u_2 \\ v_2 \\ 1 \end{bmatrix} = Z_1 \boldsymbol{KRK}^{-1} \begin{bmatrix} u_1 \\ v_1 \\ 1 \end{bmatrix} + \boldsymbol{KT} \quad . \quad (6)$$

式(6)即为重投影公式，从式(6)可知，(u_2, v_2)和Z_2可以表示为K，R，T和Z_1的函数。因此，在已知摄像机内参数矩阵K、$t-1$时刻到t时刻的摄像机位姿转移矩阵(R, T)和前一个时刻的像素点(u_{t-1}, v_{t-1})及其深度值Z_{t-1}，就可以重建当前时刻的像素点

(u_t, v_t)。利用t时刻实际的像素点作为监督信息,和根据前一时刻$t-1$重建的像素点进行对比,即可建立一种自监督学习框架。

1.2　基于序列图像的自监督景深估计网络架构

如上文所述,基于序列图像的自监督学习框架需要解决两个问题,分别为六自由度位姿参数R和T的估计,和像素级深度Z_1的估计。因此,基于序列图像的自监督单目深度估计方法涉及到了多任务联合训练,需要分别训练深度估计网络和位姿估计网络,基本结构图如下图 1 所示。本文方法是建立在Monodepth2 基础上,因此其位姿估计网络与Monodepth2中的相关部分完全相同,因此在此将不再进行赘述,下文主要针对深度估计网络进行研究。

在进行网络训练过程中,本文利用原图$\boldsymbol{I}_t$和由网络重建的图像$\hat{\boldsymbol{I}}_t$间的差异构建损失函数,该损失函数被称为重投影损失,该损失函数分为两个部分:L1损失和用于描述两幅图像的亮度相似性和对比度相似性的结构相似性度量(Structural Similarity Index Measurement, *SSIM*)[24]。损失函数L定义如下:

$$
\begin{aligned}
L &= (1-\alpha)L_1\left(\boldsymbol{I}_t,\hat{\boldsymbol{I}}_t\right) + \alpha\left(1 - SSIM\left(\boldsymbol{I}_t,\hat{\boldsymbol{I}}_t\right)\right)\\
L_1\left(\boldsymbol{I}_t,\hat{\boldsymbol{I}}_t\right) &= \sum_{i=1}^{num}\left|\boldsymbol{I}_t(u_i,v_i) - \hat{\boldsymbol{I}}_t(u_i,v_i)\right| \qquad (7)\\
SSIM\left(\boldsymbol{I}_t,\hat{\boldsymbol{I}}_t\right) &= \frac{2\mu_t\hat{\mu}_t + c_1}{\mu_t^2 + \hat{\mu}_t^2 + c_1}\cdot\frac{2\sigma_t\hat{\sigma}_t + c_2}{\sigma_t^2 + \hat{\sigma}_t^2 + c_2}
\end{aligned}
$$

其中,$\boldsymbol{I}_t(u_i, v_i)$和$\hat{\boldsymbol{I}}_t(u_i, v_i)$分别表示原图和重建图中$(u_i, v_i)$像素的亮度值,$num$表示图像像素数,$\mu_t$和$\hat{\mu}_t$分别表示原图$\boldsymbol{I}_t$和重建图像$\hat{\boldsymbol{I}}_t$间亮度的平均值,$\sigma_t$和$\hat{\sigma}_t$分别表示两张图像亮度的标准差,$c_1$和$c_2$为防止分母为零而设置的常数。

1.3　景深估计网络结构

深度估计网络为了实现像素级景深估计,其基本构型均为编解码器(Encoder-Decoder)结构,编码器就是一个特征提取模块,提取出输入图像的深度特征信息,再由解码器通过对特征信息的融合将其转化为深度图。最早的编解码结构使用全卷积网络(Fully Convolutional Networks, FCN)[25],其通过对输入图像进行多次卷积和下采样操作之后,直接将低维特征反卷积并和前序下采样相同维度的特征进行通道维度上的连接(Concatenation),从而实现特征融合。然而这种融合方式缺少输入图像的浅层细节特征,导致得到的深度图细节信息缺失,如边缘等不清晰,进而影响像素级深度估计的精度。由于本文研究对象为单目红外图像,相比于可见光图像,红外图像本身有着对比度低、色彩单一、特征信息不足等缺点,为了更好地提取、聚合及融合红外图像的不同尺度信息,本文提出了一种深度估计网络,如图 2 所示。该网络主要由三个部分组成:特征提取模块、特征聚合模块和特征融合模块。

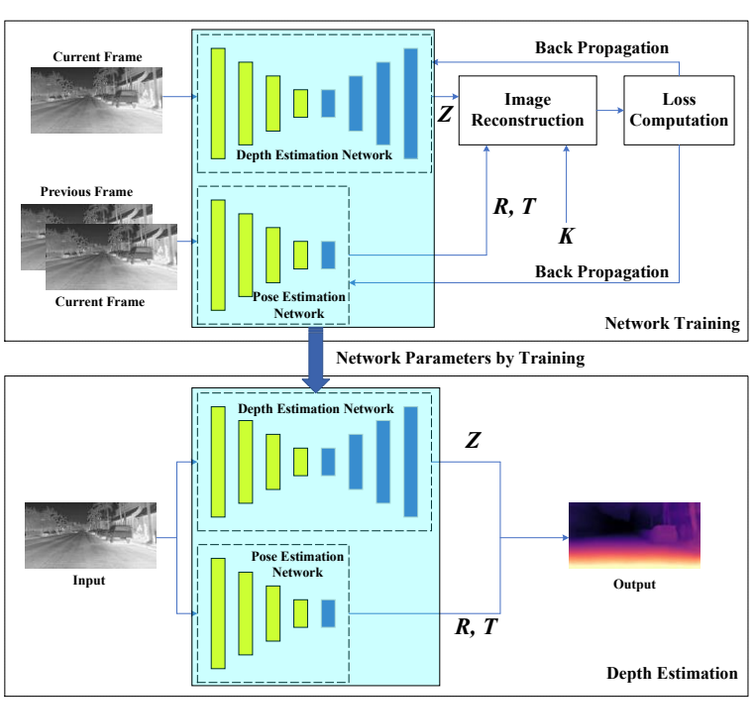

图 1　基于序列红外图像的景深估计网络架构

Fig. 1　The framework of depth estimation using thermal infrared image sequences

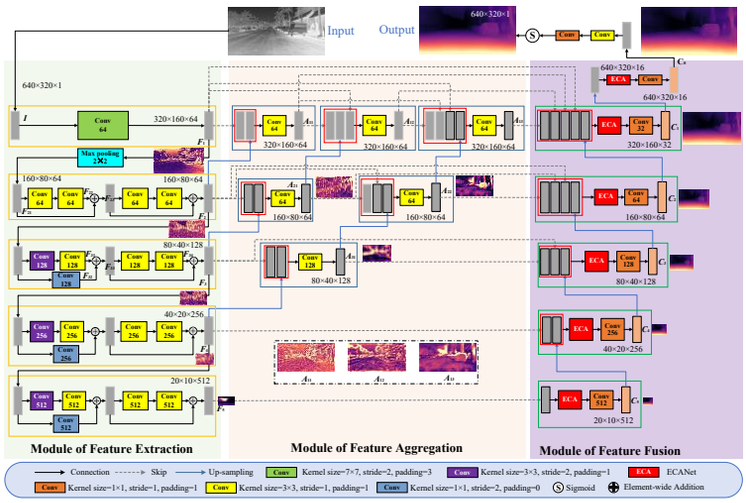

图2 本文的深度估计网络(Conv下方的数字表示卷积核的数量)

Fig. 2 The depth estimation network of this paper (The number below Conv indicates the number of convolutional kernels)

1.3.1 特征提取

在特征提取部分,由于输入的是红外图像,因此设网络的输入为 $\boldsymbol{I} \in \Re^{m \times n \times 1}$(图2中 m=640, n=320),通过特征提取模块,得到五个不同分辨率的特征图$\{F_1, F_2, F_3, F_4, F_5\}$,这五个特征图的维度分别为 $m/2\times n/2\times 64$, $m/2\times n/2\times 64$, $m/4\times n/4\times 128$, $m/8\times n/8\times 256$, $m/16\times n/16\times 512$。高分辨率的浅层特征图表征了图像的细节特征,低分辨率的深层特征图表示了图像的语义特征。

五层特征提取模块(图2中左侧的黄色实线框)共包括三种不同类型的特征提取方法。在第一层特征提取过程中,输入图像通过一个卷积层(图2中的Conv表示卷积操作)得到$m/2\times n/2\times 64$的特征图F_1;在此基础上利用2×2的最大池化(Max pooling)得到第二层特征提取块的输入 $\boldsymbol{F}_{21} \in \Re^{m/4 \times n/4 \times 64}$,第二层特征提取块由两个部分组成,第一部分中特征图首先经过两个卷积层得到 $\boldsymbol{F}_{22} \in \Re^{m/4 \times n/4 \times 64}$,由于$F_{21}$和$F_{22}$维度相同,将这两个张量的对应元素相加(用⊕表示)得到张量F_{23},即 $\boldsymbol{F}_{23} = \boldsymbol{F}_{21} \oplus \boldsymbol{F}_{22}$,这一操作进一步将第一层特征图中的特征进行了强化,上述操作过程重复一次,得到了第二层特征层的输出 $\boldsymbol{F}_2 \in \Re^{m/4 \times n/4 \times 64}$;$F_2$直接作为第三个特征提取块的输入,在第三个特征提取块中F_2首先经过两个卷积层,由于第一个卷积层的步长(Stride)为2,因此经过两个卷积层得到输出 $\boldsymbol{F}_{31} \in \Re^{m/8 \times n/8 \times 128}$,同时将$F_2$经过一个卷积层得到 $\boldsymbol{F}_{32} \in \Re^{m/8 \times n/8 \times 128}$,将$F_{31}$和$F_{32}$的对应元素相加得到$F_{33}$,$F_{33}$再经过两个卷积层得到$F_{34}$,将$F_{33}$和$F_{34}$的对应元素相加得到第三个卷积层的输出 $\boldsymbol{F}_3 \in \Re^{m/8 \times n/8 \times 128}$;第四层与第五层的流程与第三层的方式完全相同,仅仅是卷积核的数量翻倍,分辨率减半,最终得到第四、五层的特征图分别为 $\boldsymbol{F}_4 \in \Re^{m/16 \times n/16 \times 256}$, $\boldsymbol{F}_5 \in \Re^{m/32 \times n/32 \times 512}$。综上,输入红外图像经过五层特征提取模块,共得到五个特征图。

1.3.2 特征聚合

在Monodepth2中,直接使用跳连接(Skip Connect)实现了不同特征图之间的融合[26],保留了浅层特征图中的细节信息。但是,如果能够在特征融合前,将深层的语义信息自底向上聚合到浅层特征图中去,将会更好地提升特征图的融合。为此,本文引入自底向上的特征聚合模块[27],具体过程如下:首先,对第四层特征图F_4进行上采样(Upsampling),将其维度从$m/16\times n/16\times 256$变为$m/8\times n/8\times 256$,将上采样后的特征图与第三层特征图沿着通道维连接(Concatenation),得到一个$m/8\times n/8\times(256+128)$的张量,将这个张量经过一个卷积层得到了聚合后的特征A_{31},其维度等于F_3的维度,从图2中可见A_{31}聚合了特征图F_5, F_4和F_3。针对第二层特征图F_2,得到了两个聚合后的特征图A_{21}和A_{22},它们的维度等于F_2的维度,分别聚合了F_3和F_2,以及F_2,A_{21}和A_{31};针对第一层特征图,设计了三个聚合节点得到三个聚合后的特征A_{11},A_{12}和A_{13},它们的维度等于F_1的维度。六个聚合后的特征图计算过程如下式所示,其中*Conv*是指卷积操作,*Concat*表示特征图通道维连接,*Upsampl*表示双倍上采样操作。

$$\begin{aligned}
\boldsymbol{A}_{i1} &= Conv(Concat(Upsampl(\boldsymbol{F}_{i+1}),\boldsymbol{F}_i)), i = 1,2,3\\
\boldsymbol{A}_{i2} &= Conv(Concat(Upsampl(\boldsymbol{A}_{i+1\ 1}),\boldsymbol{A}_{i1},\boldsymbol{F}_i)), i = 1,2\\
\boldsymbol{A}_{13} &= Conv(Concat(Upsampl(\boldsymbol{A}_{22}),\boldsymbol{A}_{11},\boldsymbol{A}_{12},\boldsymbol{F}_1))
\end{aligned} \quad . \quad (8)$$

1.3.3　特征融合

特征融合模块以特征提取模块得到的五个特征图{F_1, F_2, F_3, F_4, F_5}和特征聚合模块得到的三层聚合特征{A_{11},A_{12}, A_{13}},{A_{21}, A_{22}}和{A_{31}}为输入,自下而上得到六个融合后的特征图{C_0, C_1, C_2, C_3, C_4, C_5},其计算方法如下式(9)所述,其中ECA是指注意力机制ECANet。

$$\begin{aligned}
\boldsymbol{C}_5 &= Conv(ECA(\boldsymbol{F}_5))\\
\boldsymbol{C}_4 &= Conv(ECA(Concat(\boldsymbol{F}_4,Upsampl(\boldsymbol{C}_5)))\\
\boldsymbol{C}_3 &= Conv(ECA(Concat(\boldsymbol{F}_3,\boldsymbol{A}_{31},Upsampl(\boldsymbol{C}_4)))\\
\boldsymbol{C}_2 &= Conv(ECA(Concat(\boldsymbol{F}_2,\boldsymbol{A}_{23},\boldsymbol{A}_{22},Upsampl(\boldsymbol{C}_3)))\\
\boldsymbol{C}_1 &= Conv(ECA(Concat(\boldsymbol{F}_1,\boldsymbol{A}_{13},\boldsymbol{A}_{12},\boldsymbol{A}_{11},Upsampl(\boldsymbol{C}_2)))\\
\boldsymbol{C}_0 &= Conv(ECA(\boldsymbol{C}_1))
\end{aligned} \quad . \quad (9)$$

近年来,人们发现将通道注意力引入卷积块能够明显改善卷积神经网络(Convolutional Neural Network, CNN)性能,具有巨大的潜力。CNN中广泛使用的SENet注意力机制通过学习每个卷积块的通道注意力就能使各种CNN网络模型性能大大提升。SENet主要分为两个部分,压缩(聚合特征)和激励(校准特征)。SENet虽然有着较高的精度,但是会使模型变得十分复杂,从而导致计算负担巨大,计算成本也显著上身;除此之外,由于SENet中采取了降维操作,对通道注意力的预测会产生负面影响,并且效率低下,所以本文采用ECANet注意力机制提高融合特征的表达能力。和SENet相比[28],ECANet摒弃了降维操作[29],并且能够有效地捕获通道间的交互关系,因此避免了降维给通道注意力预测带来的负面影响。ECANet的主要网络结构如图3所示,首先对输入$F=[f_1, f_2, \cdots, f_s]\in \Re^{m\times n\times s}$每一个通道分别进行全局平均池化,得到向量$\boldsymbol{z}=[z_1, z_2, \ldots, z_s]$,

$$z_i = \frac{1}{m\times n}\sum_{h=1}^{m}\sum_{w=1}^{n}\boldsymbol{f}_i(h,w) \quad , \quad (10)$$

其中,$\boldsymbol{f}_i \in \Re^{m\times n}$为第$i$个通道的特征图。在此基础上,ECANet考虑每个通道及其邻近k个通道来获取跨通道交互信息,通过卷积核大小为k的快速一维卷积来实现通道之间的信息交互。卷积核大小k表示有k个相邻通道参与一个通道的注意力预测,即局部跨通道交互的覆盖率,它的数量直接关系到了ECANet模块的计算效率和复杂度。一维卷积的结果在经过Sigmoid函数后输出通道注意力权重$\boldsymbol{w}=[w_1, w_2, \ldots, w_s]$,将权重向量的元素与原始特征图对应通道相乘,最终得到新的特征图$\boldsymbol{c}_i \in \Re^{m\times n}$用于后续深度估计。

$$\boldsymbol{c}_i = w_i\boldsymbol{f}_i \quad , \quad (11)$$

k的值根据特征向量的通道数量s确定:

$$k = \phi(s) = \left|\frac{\log_2(s)}{\gamma} + \frac{b}{\gamma}\right|_{odd} \quad , \quad (12)$$

其中,$|\bullet|_{odd}$表示取离括号内参数最近的奇数,γ=2,b=1。

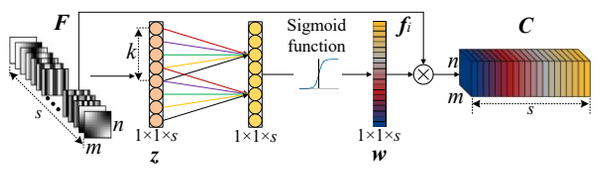

图3　ECANet通道注意力模块结构

Fig. 3　The structure of ECANet

综上所述,本文提出的面向红外图像景深估计的网络,在编解码器之间采取了特征提取能力更强的密集跳连接方式进行级联,实现特征信息的多尺度融合,提高了网路特征提取能力。同时,改进后的方法在解码器部分接入了通道注意力机制ECANet,进一步提升了特征表达与融合能力。

2　实验与分析

2.1　数据集与训练参数设置

实验使用的数据集为FLIR红外数据集(https://www.flir.com/oem/adas/adas-dataset-form,采用FLIR-Tau2热红外相机))和自行拍摄的红外数据集(采用FLIR-A35热红外相机采集),两种数据集中的红外图像数据均是动态背景下的车载图像,由若干个连续的视频序列组成,一共11521张图像。其中,训练集9677张,测试集1844张。表1为热红外相机相关参数。表2为深度估计网络训练所设置的主要参数,三个实验使用的数据集以及训练参数都保持了严格的一致性。

2.2　实验结果分析

在对实验结果进行分析过程中,将从定性和定量两个方面展开对比分析。定性分析主要是通过计算得到的深度图对三种网络模型的结果进行比较;定量分析主要是使用带有真实目标物深度的红外图像进行测试,计算出目标物真实值与利用深度图所预测出的目标物深度值的误差率并进行比较。

2.2.1　定性分析

实验中使用的测试图像包括FLIR数据集中的图像和使用FLIR A35型号热成像仪在实际道路上

表1 热红外相机相关参数

Table 1 Related parameters of thermal infrared cameras

参数	FLIR-Tau2	FLIR-A35
图像分辨率	640×512	320×256
相机参数	HFOV 45° VFOV 37° 13 mm f/1.0	HFOV 48° VFOV 39° 9 mm f/1.0
相机内参数矩阵	$\begin{bmatrix} 0.669 & 0 & 0.5 & 0 \\ 0 & 0.828 & 0.5 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0.6403 & 0 & 0.5 & 0 \\ 0 & 0.8003 & 0.5 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$
图像采样率	30 Hz	30 Hz

表2 训练参数

Table 2 Training parameters

参数	数值
ResNet层数	18
学习率	0.000 1
迭代次数	20

拍摄的红外图像。用于对比的方法主要是:(1)Monodepth2方法,该方法是本文所提出方法的基准方法(Baseline);(2)HR-Depth方法[30],该方法使用SENet注意力机制,其结果与本文方法的结果对比在一定程度上反映ECANet的作用。图4中第一列的五个测试图像来自于FLIR数据集,第二列的五张深度图是利用Monodepth2方法得到的景深估计的结果,第三列是利用HR-Depth得到的结果,第四列是本文方法得到的结果。如图4前三行蓝色框所示,本文提出的方法对类似于柱体和树干这样的细长目标物,在其深度图中边缘细节清晰完整,深度估计的准确性明显高于前两种方法。在五个测试样本中,对于汽车这样的目标物,本文提出的方法得到的深度图中,其边缘及区域深度信息较前两种方法也更加显著。对于第四个测试样本中的自行车,本文提出的方法得到的轮廓较其他两种方法得到的深度图更为清晰。此外,从图4的深度估计结果可见,本文的方法对不同尺度、不同长宽比的目标物均有较好的深度估计能力。

为了验证该方法的泛化能力,我们利用FLIR A35热成像仪获取和训练数据集不相关的场景图像进行测试。由于FLIR A35热成像仪提供的图像分辨率为320×256 pixels,因此在进行深度估计前,利用双线性插值将测试图像的分辨率变为640×512 pixels。图5是FLIR A35热成像仪拍摄的五个测试样本得到的深度估计的结果。从图5最后一列可以看出,本文所设计的景深估计网络针对不同场景具

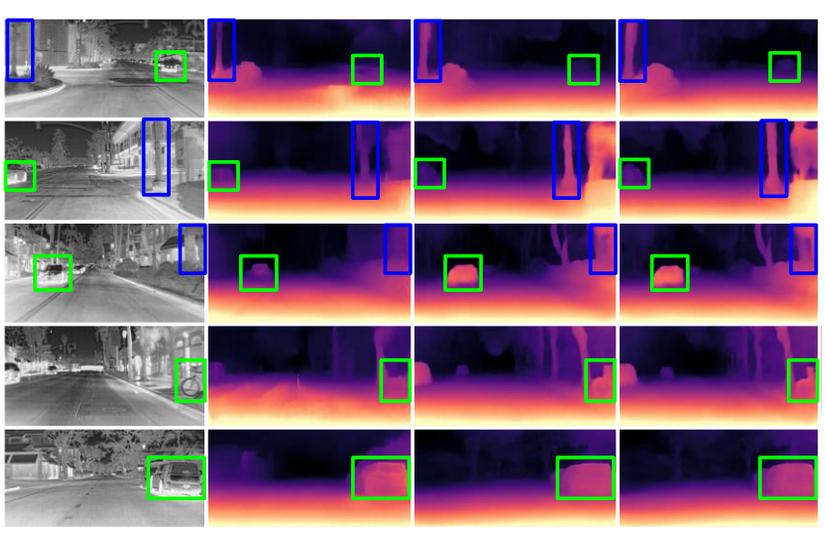

图4 FLIR数据集测试样本与使用不同方法得到的深度图

Fig. 4 Test images from the FLIR dataset and corresponding depth maps

有较强的泛化能力,可以将FLIR数据集上的训练结果泛化到其他场景红外图像中。同时,对比图5的第2、3和4列,可以发现本文提出方法所得到的深度估计的结果边缘清晰,目标物内部颜色均匀,故其得到的深度估计结果优于其他两种方法,这和图4中得到的定性对比的结论具有一致性。

根据图4和图5中的实验结果可以表明,虽然三种方法都是建立在monodepth2框架基础上,但是本文的方法得到的深度估计的结果最好,具体表现在:(1)本文方法生成的深度图内目标轮廓更加明显;(2)本文方法生成的深度图中目标与目标之间的深度区分度较明显;(3)本文方法生成的深度图中各目标的颜色更均匀,说明深度估计结果在同一目标上具有连续性,更加符合目标物各部分之间的实际景深关系。

2.2.2　定量分析

在这一节中,使用的测试图片为FLIR A35拍摄的红外图像,在拍摄前使用激光测距仪测量好场景中的特定目标到相机的真实距离作为基准值,通过比较真实距离和网络估计的距离并引入深度估计的误差率作为评价指标进行量化对比。误差率E定义为:

$$E = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|D_i - D_i^*\right|}{D_i} \times 100\% \quad , \quad (13)$$

其中,N为测试图像总数,D_i为场景目标由激光测距仪得到的真实深度,D_i^*为场景目标的估计深度。作为标准值的场景目标真实深度是通过激光测距仪测量目标平面区域的深度值的平均值所获取。定量对比中共设置了60个目标,并利用激光测距仪确定了其真实深度值,随机分布在区间10~25m内。为了确保标准值的准确性,所选取的目标区域所占整体图像区域的比例均大于1%。在相同测试条件下,使用三种不同深度估计网络对目标进行深度估计并得到估计深度值,进而计算出不同目标距离值,并与各目标真实距离进行比较。图6是一组示例,图6(a)为利用激光测距仪得到的车辆尾部到相机的距离23.02米,根据三种方法得到的深度图,估计出的目标尾部到相机的距离分别为:本文提出的方法得到的结果为24.76米(误差率E=7.56%),HR-depth方法得到的结果为21.17米(误差率E=8.04%),Monodepth2方法得到的结果为20.79米(误差率E=9.69%)。针对全部60个测试目标,得到的距离估计的平均误差率如表3所示。

从表3可以看出,本文提出的改进深度估计网络的误差率最小,HR-Depth次之,Monodepth2最大。由于在计算误差时只考虑了目标中心点的深度值而忽略了其他像素点间的深度误差,导致不同网络间的深度误差率区别不明显,但是依然能看出本文方法的深度估计性能相较于前两者更具有优势。此外,将得到的距离估计的误差率按4个误差区间统计,即误差率分别为<10%、<20%、<30%和>30%,表4统计了三种网络的深度估计误差在不同区间内的占比,即处于该误差区间内测试结果的数量相对于测试样本总数的占比。误差分布区间在

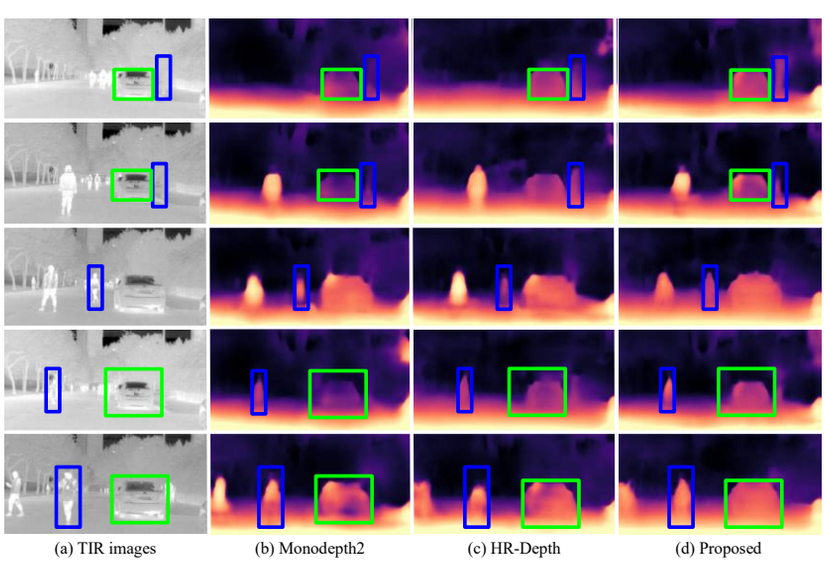

图5　FLIR A35摄像机拍摄的测试样本与使用不同方法得到的深度图

Fig. 5　Test images from the FLIR A35 TIR camera and corresponding depth maps

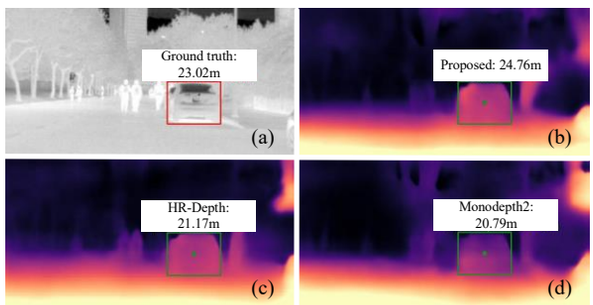

图6 输入图像与距离估计结果,(a) 原图与真实值,(b) 本文方法得到的结果,(c) HR-Depth 得到的结果,(d) Monodepth2得到的结果

Fig. 6 Input image and distance estimation results, (a) the input image and the ground truth, (b) the result of distance estimation by the proposed method, (c) the result of distance estimation by HR-Depth, (d) the result of distance estimation by Monodepth2

表3 不同网络的深度估计误差率

Table 3 Error Rates of depth estimation for different networks

方法	Proposed	HR-Depth	monodepth2
E	19.58%	20.09%	21.68%

一定程度上反映了误差率的分布情况,当多数测试目标的误差率处于较小误差区间时,就可以认为误差率整体偏小,深度估计的精确度更高,反之误差率越大,精确度越差。由表4可以看出,本文方法的测试误差率整体上趋向于更小区间。

表4 不同网络误差分布区间占比(%)

Table 4 Proportions of different network error distribution intervals (%)

方法	*E*			
	<10%	<20%	<30%	>30%
Proposed	41.67%	66.67%	90.00%	10.00%
HR-Depth	36.67%	63.33%	86.67%	13.33%
monodepth2	25.00%	58.33%	85.00%	15.00%

3 结论

由于红外图像本身具有对比度低、分辨率低、目标细节信息不足等缺点,本文构建了一种针对单幅红外图像的自监督深度估计方法。该网络由特征提取模块、特征聚合模块和特征融合模块三个部分组成。首先,设计了一种特征聚合模块,提高景深估计网络对场景目标物体边缘信息和小物体信息的获取能力;其次,在特征融合模块中引入了通道注意力机制,有效获取通道间的交互关系;在此基础上,建立了一种面向热红外图像的深度估计网络。在实验部分,对三种网络模型设置了完全相同的训练集、训练参数和训练环境,在此基础上进行定性和定量两种对比实验。定性结果显示,本文提出的方法生成的深度图像总体质量最好,具体体现在改进网络模型生成的深度图内目标轮廓更加明显、目标与目标之间的区分度较明显。对于定量结果,实验将60个目标的真实深度与三种方法的估计深度进行比较,求出误差率,进而对模型的性能进行比较和判断。最终结果显示,对于整个测试数据集,本文提出的网络模型深度估计的平均误差率最小,整体准确度最高。

References

[1] Huang J, Wang C, Liu Y, *et al*. The progress of monocular depth estimation technology [J]. Journal of Image and Graphics, 2019, **24**(12): 2081-2097.(黄军, 王聪, 刘越, 等. 单目深度估计技术进展综述[J]. **中国图象图形学报**), 2019, **24**(12): 2081-2097.

[2] Jia D, Zhu N D, Yang N H, *et al*. Image matching methods [J]. Journal of Image and Graphics, 2019, **24**(5): 677-699.(贾迪, 朱宁丹, 杨宁华, 等. 图像匹配方法研究综述[J]. **中国图象图形学报**), 2019, **24**(5): 677-699.

[3] Dong X, Garratt A M A, Anavatti G S, *et al*. Towards Real-Time Monocular Depth Estimation for Robotics: A Survey[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2022, **23**(10): 16940-16961.

[4] Liu Y, Jiang J, Sun J, *et al*. A survey of depth estimation based on computer vision[C]//Proceedings of the IEEE 5th international conference on data science in cyberspace, 27-30 July 2020, Hong Kong, China, USA: IEEE, pp. 135-141.

[5] Ming Y, Meng X, Fan C, *et al*. Deep learning for monocular depth estimation: A review[J]. *Neurocomputing*, 2021, **438**: 14-33.

[6] Masoumian A, Rashwan H A, Cristiano J, *et al*. Monocular Depth Estimation Using Deep Learning: A Review[J]. *Sensors*, 2022, **22**(14): 5353.

[7] Qi X, Liao R, Liu Z, *et al*. Geonet: Geometric neural network for joint depth and surface normal estimation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18-23 June 2018, USA: IEEE, pp. 283-291.

[8] Ummenhofer B, Zhou H, Uhrig J, *et al*. Demon: Depth and motion network for learning monocular stereo[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21-26 July 2017, USA: IEEE, pp. 5038-5047.

[9] Luo Y, Ren J, Lin M, *et al*. Single view stereo matching [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18-23 June 2018, USA: IEEE, pp. 155-163.

[10] Xie J, Girshick R, Farhadi A. Deep3d: Fully automatic

2d-to-3d video conversion with deep convolutional neural networks[C]//European Conference on Computer Vision, Amsterdam, The Netherlands, October 11-14, 2016, Germany:Springer, pp. 842-857.

[11] Zhan H, Garg R, Weerasekera C S, *et al*. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18-23 June 2018, USA:IEEE, pp. 340 - 349.

[12] Ding M, Jiang X Y. Scene Depth Estimation Based on Monocular Vision in Advanced Driving Assistance System[J]. Acta Optica Sinica, 2020, **40**(17): 1715001-1-1715001-9.(丁萌, 姜欣言. 先进驾驶辅助系统中基于单目视觉的场景深度估计方法[J]. **光学学报**), 2020, **40**(17):1715001-1-1715001-9.

[13] Garg R, Bg V K, Carneiro G, *et al*. Unsupervised cnn for single view depth estimation: Geometry to the rescue[C]// Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11-14 October 2016, Germany:Springer, pp. 740-756.

[14] Godard C, Mac Aodha O, Brostow G J. Unsupervised monocular depth estimation with left-right consistency[C]// Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, July 21-26 2017, USA:IEEE, pp. 270-279.

[15] Tosi F, Aleotti F, Poggi M, *et al*. Learning monocular depth estimation infusing traditional stereo knowledge[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, June 15-20 2019, USA:IEEE, pp. 9799-9809.

[16] Zhou T, Brown M, Snavely N, *et al*. Unsupervised learning of depth and ego-motion from video[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, July 21-26 2017, USA: IEEE, pp. 1851-1858.

[17] Lai H Y, Tsai Y H, Chiu W C. Bridging stereo matching and optical flow via spatiotemporal correspondence[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, June 15-20 2019, USA:IEEE, pp. 1890-1899.

[18] Zou Y, Luo Z, Huang J B. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency[C]// Proceedings of the European conference on computer vision (ECCV), Munich, Germany, Sep 8-14, 2018, Germany:Springer, pp. 36-53.

[19] Godard C, Mac Aodha O, Firman M, *et al*. Digging into self-supervised monocular depth estimation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, Oct. 27-Nov. 2, 2019, USA:IEEE, pp. 3828-3838.

[20] Li X G, Cao M T, Li B, *et al*. GPNet: Lightweight infrared image target detection algorithm[J]. Journal of Infrared and Millimeter Waves, 2022, **41**(6):1092-1101.(李现国, 曹明腾, 李滨, 等. 2GPNet:轻量型红外图像目标检测算法[J]. **红外与毫米波学报**), 2022, **41**(6): 1092-1101.

[21] Ding M, Chen W-H, Cao Y F. Thermal Infrared Single-Pedestrian Tracking for Advanced Driver Assistance System[J] *IEEE Transactions on Intelligent Vehicles*, online, 2022. DOI: 10.1109/TIV.2022.3140344.

[22] He Y, Deng B, Wang H, *et al*. Infrared machine vision and infrared thermography with deep learning: A review[J]. *Infrared physics & technology*, **116**(103754), 2021.

[23] Li X, Ding M, Wei D H, *et al*. Depth estimation method based on monocular infrared image in VDAS[J]. Systems Engineering and Electronics, 2021, **43**(5): 1210-1217.(李旭, 丁萌, 魏东辉, 等. VDAS中基于单目红外图像的深度估计方法[J]. **系统工程与电子技术**), 2021, **43**(5):1210-1217.

[24] Wang Z, Bovik 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.

[25] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, MA, USA, June 7-12, 2015, USA:IEEE, pp. 3431-3440.

[26] Zhou Z, Rahman Siddiquee M M, Tajbakhsh N, *et al*. Unet++: A nested u-net architecture for medical image segmentation[C]//In Deep learning in medical image analysis and multimodal learning for clinical decision support, 2018, pp. 3-11.

[27] Wang J, Sun K, Cheng T, *et al*. Deep high-resolution representation learning for visual recognition[J]. *IEEE transactions on pattern analysis and machine intelligence*, 2020, **43**(10):3349-3364.

[28] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA, June 18-23 2018, USA:IEEE, pp. 7132-7141.

[29] Wang Q, Wu B, Zhu P, *et al*. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks[C]//Proceedings of the IEEE/CVF international conference on computer vision, Seattle, WA, USA, USA: IEEE, 2020 June 13-19, USA:IEEE, pp. 11534-11542.

[30] Lyu X, Liu L, Wang M, *et al*. HR-depth: High resolution self-supervised monocular depth estimation[C]//Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, British Columbia, Canada, Feb 2-9, 2021, USA:AAAI, vol.**35**, no. 3, pp.2294-2301.