面向无人机自主飞行的无监督单目视觉深度估计

赵栓峰，黄涛*，许倩，耿龙龙
西安科技大学机械工程学院，陕西 西安 710054

摘要 针对双目视觉深度估计成本高、体积大以及监督学习需要大量深度图进行训练的问题，为实现无人机在飞行过程中的场景理解，提出一种面向无人机自主飞行的无监督单目深度估计模型。首先，为减小不同尺寸目标对深度估计的影响，将输入的图像进行金字塔化处理；其次，针对图像重构设计一种基于 ResNet-50 进行特征提取的自编码神经网络，该网络基于输入的左视图或右视图以及生成对应的金字塔视差图，采用双线性插值的方法重构出与其对应的金字塔右视图或左视图；最后为提高深度估计的精度，将结构相似性引入到图像重构损失、视差图一致性损失中，并且联合视差图平滑性损失、图像重构损失、视差图一致性损失作为训练的总损失。实验结果表明，经过在 KITTI 数据集上的训练，该模型在 KITTI 和 Make3D 数据集上相比其他单目深度估计方法具有更高的准确性和实时性，基本满足无人机自主飞行对深度估计准确性和实时性的要求。

关键词 图像处理；无监督；自编码神经网络；图像重构；单目深度估计
中图分类号 TN219 **文献标志码** A **doi**：10.3788/LOP57.021012

Unsupervised Monocular Depth Estimation for Autonomous Flight of Drones

Zhao Shuanfeng, Huang Tao*, Xu Qian, Geng Longlong
College of Mechanical Engineering, Xi'an University of Science and Technology, Xi'an, Shaanxi 710054, *China*

Abstract This study proposes an unsupervised monocular depth estimation model for autonomous drone flight to overcome the limitations of high cost and large size in binocular depth estimation and a large number of depth maps required for training in supervised learning. The model first processes the input image into a pyramid shape to reduce the impact of different target sizes on the depth estimation. In addition, the neural network of the automatic encoder used for image reconstruction is designed based on ResNet-50, which is capable of feature extraction. The corresponding right or left pyramid images are subsequently reconstructed by the bilinear sampling method based on the left or right input images, and corresponding pyramid disparity map is generated. Finally, the training loss could be assessed as the combination of the disparity smoothness loss, image reconstruction loss based on the structural similarity, and the loss of disparity consistency. Experimental results indicate that the model is more accurate and timely on KITTI and Make3D compared with other monocular depth estimation methods. When trained on KITTI, the model essentially meets the accuracy requirements and real-time necessities for autonomous drone flight depth estimation.

Key words image processing; non-supervision; neural network of automatic encoder; image reconstruction; monocular depth estimation
OCIS codes 100.4996; 100.5010; 040.3060

1 引 言

随着人工智能、云计算、大数据的迅速发展，无人驾驶汽车、无人机、智能机器人等产品不断出现在人们的生活中。这些产品在工作中需要对周围环境进行实时三维(3D)感知，以便做出正确的决策。深

收稿日期：2019-05-27；**修回日期**：2019-06-14；**录用日期**：2019-07-11
基金项目：陕西省自然科学基金(2017JM5029)、西安市科技计划项目(CXY2017079CG/RC042)
* **E-mail**：775628393@qq.com

度估计是实现 3D 环境感知的基础[1]。目前，深度估计常用的方法有两种：一是利用硬件设备直接获取，比如微软的 Kinect 相机、激光测距仪、双目相机等，但是其价格比较昂贵；二是基于视觉的深度估计[2-3]，视觉深度估计根据拍摄图像所用摄像头的多少，分为单目深度估计和多目深度估计（常用的是双目相机）。双目深度估计通过双目相机获取同一场景的左右视图，然后利用三角形测量法将左右视图之间的匹配信息转换为视差信息，进而根据双目相机的焦距和左右摄像头基线距离推出深度图[4-5]，目前该研究主要集中在提高精度上。单目深度估计则是通过一张图像获取 3D 场景，由于一张图具有无穷多个 3D 场景，因此对于计算机视觉系统通过单目进行深度估计仍然是一个具有挑战性的任务[6-7]，单目深度估计具有很大的研究价值和提升空间。

此前，针对单目深度估计这个问题，国内外已进行大量研究，并取得可观的效果。最早单目深度估计比较经典的算法有，从运动中恢复形状（SFM）[8]、从阴影中恢复形状[9]（SFS）和从对焦或离焦中获取深度[10-11]（DFF/DFD）等，但因其受所需设备比较昂贵、拍摄要求高、结果易受遮挡和对应关系匹配等因素的影响，上述方法均未被大量使用。随着计算机性能的不断提高，机器学习被广泛应用在各个领域，目前，对于单目深度估计这一问题，机器学习主要采用监督学习的方法进行解决，即：以每一张图像和其对应的深度图作为输入；然后构建像素与深度关系的马尔可夫随机场（MRF）或条件随机场（CRF）等模型，将其深度估计问题转换为参数优化问题；紧接着通过大量数据学习输入映射到输出的规律；最后对新输入的图像进行深度预测。但该方法很难找到图像像素与深度之间的模型，导致深度估计精度较低、计算量大、耗时高，因而此方法很难在实际中应用。近几年，随着深度学习理论的不断成熟，深度学习被广泛应用于单目视觉深度估计中，目前基于深度学习的单目视觉深度估计常用的方法如下。1）仅依靠深度学习理论和网络构架进行深度估计，2014 年 Eigen 等[12]提出采用多尺度卷积神经网络（CNN）进行单目深度估计；2016 年许路等[13]提出基于深度卷积神经网络实现单目红外图像的深度估计。2）依靠深度信息本身的性质进行单目深度估计，深度信息的本质就是一个由远到近一层一层的分类，2018 年 Cao 等[14]提出采用全连接残差神经网络通过分类方式实现单目视觉深度估计。3）基于随机场模型与深度学习结合进行单目深度估计，2015 年 Liu 等[15]提出采用 CRF 与 CNN 融合实现单目视觉深度估计，其原理为利用 CRF 的一阶项和二阶项综合训练一阶项 CNN 和二阶项 CNN，两个 CNN 通过 CRF 能量函数统一于一个训练框架中。4）基于无监督学的单目视觉深度估计，2016 年 Chen 等[16]与 Mayer 等[17]提出采用单张图像进行深度估计，其原理为以双目相机拍摄的左右两张图像为约束，基于自编码神经网络重构图像，然后不断优化视差图得到深度图；2017 年 Godard 等[18]针对 Chen 等[16]与 Mayer 等[17]单目深度估计误差大的问题，提出将左右视差图一致性函数作为无监督模型的训练损失函数，来进行单目无监督深度估计；2017 年 Zhou 等[19]基于无监督学习理论提出采用单目视频序列，实现深度与姿态估计；2018 年 Wang 等[20]受直接里程计（DVO）的影响，提出采用直接方法实现无监督单目深度估计。

针对监督学习的单目深度估计需要大量深度数据及无监督学习深度估计准确度低、实时性差等问题，以现有无监督学习深度估计方法为基础，本文提出一种面向无人机自主飞行的无监督单目视觉深度估计模型。本文模型首先为实现多特征提取，将输入图片进行金字塔化处理；其次提出一种ResNet-50自编码神经网络进行特征提取，该网络可以基于预设的视差图重构图像；最后基于双目成像原理得出图像的视差图，同时将结构相似性（SSIM）引入图像重构损失函数和视差图一致性损失函数中，将视差图平滑性损失、图像重构损失、视差图一致性损失作为训练的总损失。

2 单目深度估计原理

借鉴双目深度估计的原理，如图 1 所示，X 表示三维场景映射到二维平面，Z 表示坐标系中沿光轴的方向，z 表示图像的深度，x 表示图像上某一像素点距左相机的铅直距离，x_l 和 x_r 分别表示双目相机左、右两个摄像头成像中的坐标，f 表示双目相机的焦距，b 表示双目相机左右摄像头之间的基线距离。根据以上信息，利用三角形相似原理，可得

$$x-b=\frac{z}{f}\times x_r, \tag{1}$$

$$x=\frac{z}{f}\times x_l, \tag{2}$$

$$d=x_l-x_r=\frac{f\times b}{z}, \tag{3}$$

$$z=\frac{d\times z}{b}, \tag{4}$$

式中：d 表示视差。本模型将单目深度估计问题转换为图像重构问题，然后基于重构的左右视图，依据双目深度估计原理得出图像深度。本模型的具体实现过程为

1）输入双目相机左右摄像头拍摄的一对图像 $\boldsymbol{I}_{\mathrm{l}}$ 和 $\boldsymbol{I}_{\mathrm{r}}$；

2）将 $\boldsymbol{I}_{\mathrm{l}}$ 传入图 1 所示的自编码神经网络，生成预估的视差图 $\boldsymbol{d}_{\mathrm{r}}$，以视差图 $\boldsymbol{d}_{\mathrm{r}}$ 和左视图 $\boldsymbol{I}_{\mathrm{l}}$，基于双线性插值重构出相对应的右视图 $\tilde{\boldsymbol{I}}_{\mathrm{r}}$，同理基于右视图 $\boldsymbol{I}_{\mathrm{r}}$ 重构出相对应的 $\tilde{\boldsymbol{I}}_{\mathrm{l}}$；

3）依据双目成像原理，基于左视图 $\boldsymbol{I}_{\mathrm{l}}$ 和重构的右视图 $\tilde{\boldsymbol{I}}_{\mathrm{r}}$ 推出视差图，根据双目相机的焦距 f 和基线 b 得出深度信息，同理基于右视图 $\boldsymbol{I}_{\mathrm{r}}$ 和重构的左视图 $\tilde{\boldsymbol{I}}_{\mathrm{l}}$ 测出深度信息。

3 无监督单目深度估计模型

针对单个图像输入容易导致小目标丢失而影响深度估计精度的问题，将输入图像进行金字塔处理；针对特征提取网络层数过多易出现退化的问题，引入残差块的 ResNet-50 进行特征提取；为提高单目深度估计的精度，在训练中提出一种新的损失函数，该损失函数主要包括视差平滑度损失、图像重构损失、视差图一致性损失。基于此，设计一种无监督深度估计模型，如图 2 所示。该模型的步骤如下。

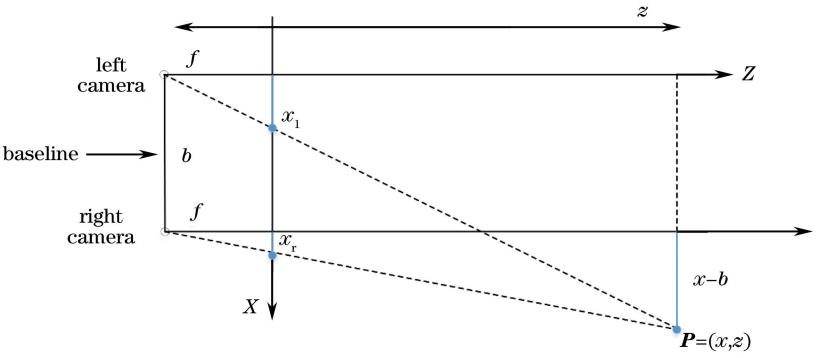

图 1 双目深度估计原理

Fig. 1 Principle of binocular depth estimation

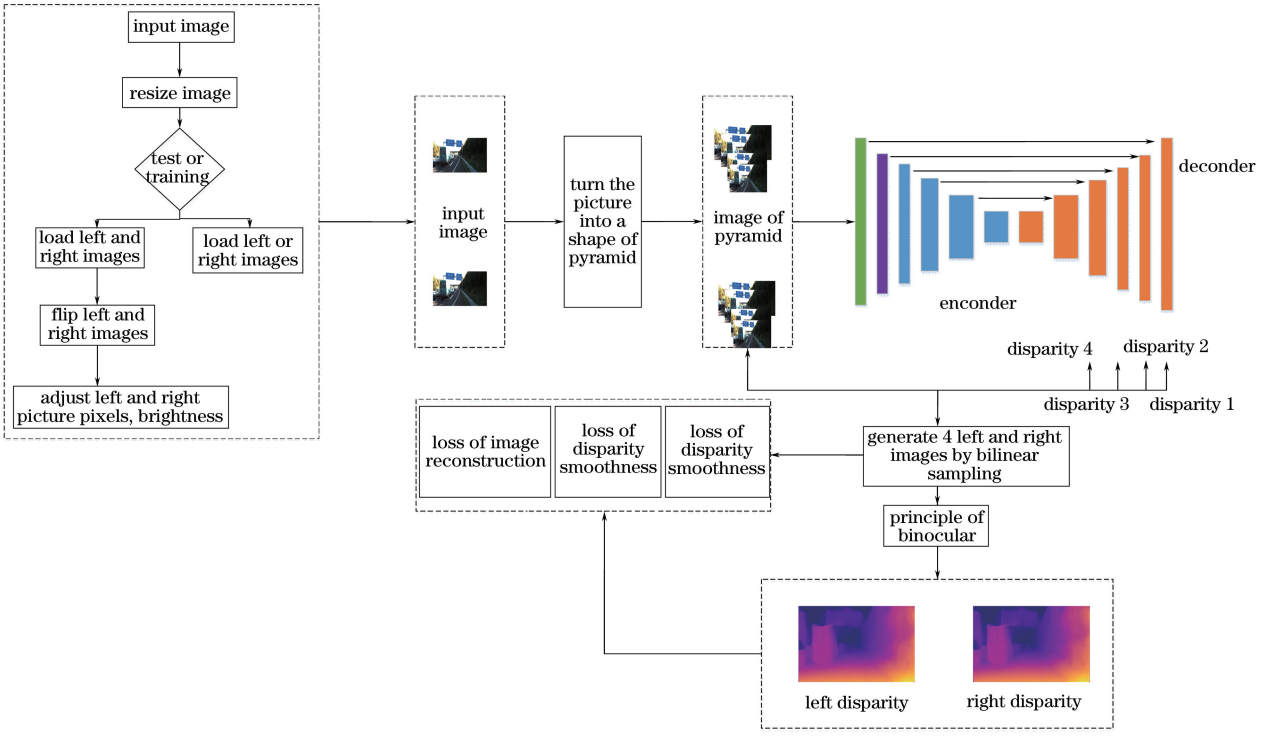

图 2 无监督单目深度估计结构图

Fig. 2 Structural diagram of unsupervised monocular depth estimation

1)先进行图像处理，即将输入的图像采用面积插值法处理成 256×512 的固定尺寸，然后对图像进行去燥、调整亮度、左右翻转等处理，最后将输入的图像通过金字塔处理成 256×512、128×256、64×128、32×64 的 4 种分辨率；2)将处理好的图像通过图 3 所示的自编码神经网络进行图像重构，该网络主要由两部分构成，即基于 ResNet-50 进行特征提取的编码神经网络和图像重构的解码反卷积神经网络，通过该网络生成 4 种不同分辨率的近似视差图；3)基于双线性插值法，根据生成的视差图和对应输入的左视图或右视图重构出与其相对应的右视图或左视图；4)依据双目成像的原理生成视差图，最终得出深度图。

3.1 图像重构网络

基于左视图或右视图重构出相应的右视图或左视图是本模型进行单目深度估计的核心。本文借鉴自编码神经网络进行图像重构的思想，同时考虑到网络层数过多会导致梯度消失和梯度爆炸的问题，提出一种基于 ResNet-50 进行特征提取、基于反向卷积神经网络进行图像重构的网络模型。本模型结构如图 3 所示，该模型主要由两部分构成：一是基于 ResNet-50 进行特征提取；二是基于反向卷积神经网络实现图像重构。

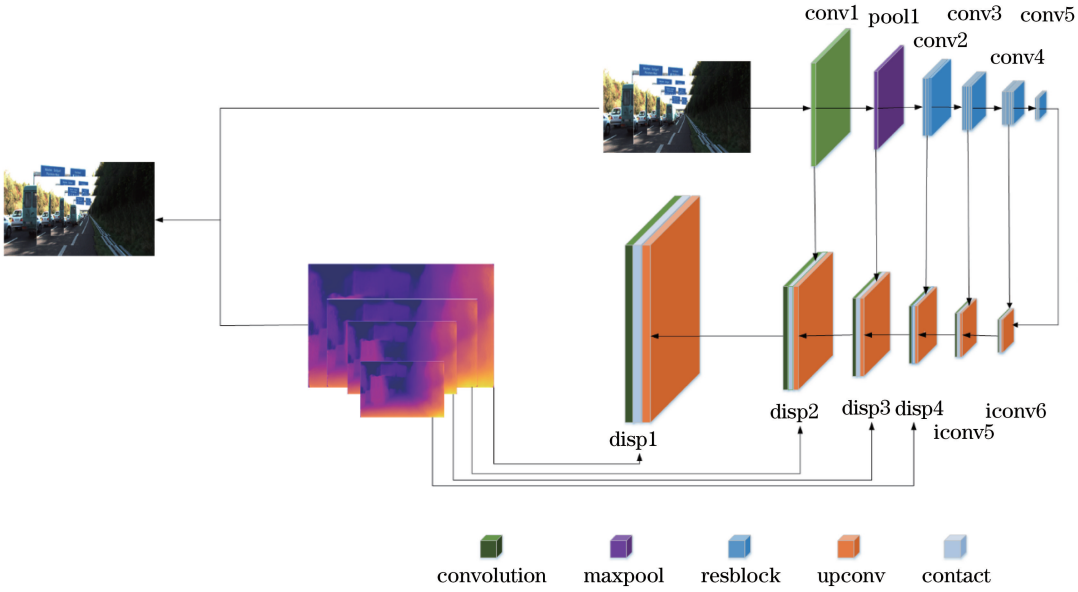

图 3 图像重构模型

Fig. 3 Model of image reconstruction

ResNet-50 由于引入残差块，能有效解决传统网络因层数过多而产生梯度消失和退化的问题。本模型采用基于三层恒等残差块的 ResNet-50 对输入的金字塔图像进行特征提取，三层恒等残差块的数学表达式为

$$a_{l+3}=R\{C\{C[C(a_l,n,f_1,s_1,R),n,f_2,s_2,R],4n,f_3,s_3\}+a_l\},\tag{5}$$

$$R(X)=\begin{cases}0, & X\leqslant 0\\ 1, & X\geqslant 0\end{cases},\tag{6}$$

式中：a_l 表示第 l 层的输出；a_{l+3} 表示第 $l+3$ 层的输出；R 表示 ReLU 激活函数，该激活函数能使学习周期大大缩短；C 表示卷积函数；n 表示输出层大小；f_i 表示滤波器的大小，$i=1,2,3$；s 表示步长，本模型中 $f_1=1,s_1=1,f_2=3,s_2=2,f_3=1,s_3=1$。基于 ResNet-50 进行粗特征提取的具体过程为：首先对输入图像 x 采用金字塔原理处理成 256×512、128×256、64×128、32×64 四种尺度；然后将输入图像进行 $C(x,64,7,2,R)$ 卷积操作输出原图像1/2尺寸的 Conv1，对 Conv1 进行最大池化（maxpool）操作输出原图像 1/4 尺寸的 Pool1，对 Pool1 进行 Resblock(Pool1,64,3)卷积操作输出原图像 1/8 尺寸的 Conv2 ，其中 64 代表输出卷积层数目，3 代表采用 3 个(5)式所示的残差块，对 Conv2 进行 Resblock(Conv2,128,4)卷积操作输出原图像 1/16 的 Conv3，对 Conv3 进行 Resblock(Conv3,256,6)卷积操作输出原图像 1/32 尺寸的 Conv4；最后对 Conv4 进行 Resblock(Conv4,512,3)卷积操作输出原图像 1/64 尺寸的 Conv5。经过以上基于 ResNet-50 进行特征提取后，输出特征图变为输入图像的 1/64。

仅通过 ResNet-50 进行粗特征提取无法实现精确的图像重构，本模型借鉴自编码神经图像重构的

思想，采用反向卷积神经网络进行图像重构（见图 3 中的 iconv6，iconv5，Disp4，Disp3，Disp2 和 Disp1）。反卷积操作的目的是通过卷积神经特征提取恢复其输入的特征尺寸，其原理为

$$a_{l+1}=C\{[\mathrm{Upsample}(a_l,S_{\mathrm{cale}})],n,f,s\},\quad(7)$$

式中：Upsample 表示上采样操作；S_{cale} 表示图像变化比例；s 表示卷积函数的步长，在本模型中 $s=1$，将(7)式的反卷积操作简写为 $a_{l+1}=\tilde{C}(a_l,n,f,s_{\mathrm{cale}})$。为确保反卷积神经网络中特征图尺寸与 ResNet-50 特征图尺寸相对应，提出设置 Skip 作用域，将 ResNet-50 编码过程中的部分特征图直接拼接到反卷积神经网络解码过程中，具体拼接过程如图 3 所示。受双目深度估计原理的启发，利用双目深度估计的逆原理在反卷积神经网络中进行图像重构，利用卷积神经网络生成近似视差图，然后利用视差图和其对应的左视图或者右视图，采用双线性插值法重构出右视图或左视图。视差图生成的数学表达式为

$$y_{\mathrm{Disp}}=\mathrm{Upsample}[\alpha C(a_l,n,f,s,S),S_{\mathrm{cale}}],\quad(8)$$

$$S=\frac{1}{1+\mathrm{e}^{-x}},\quad(9)$$

式中：y_{Disp} 表示近似视差图；α 表示视差比例系数；S 表示 Sigmoid 激活函数，在本模型中，视差生成函数 $\alpha=0.3$，$n=2$，$f=1$，$s=1$。反卷积神经网络进行图像重构的步骤为：1）将 ResNet-50 特征提取得到的 Conv5 利用 (7)式进行 $\tilde{C}$(Conv5，515，3，2)反卷积操作输出原图像 1/32 尺寸的 upconv6，然后对 upconv6 与 ResNet-50 特征提取中的 Conv5 进行第 4 维度的张量拼接得到 concat6，最后对拼接后的特征图采用 C(concat6，512，3，1，R)输出 iconv6；2）同理对 iconv6 进行 $\tilde{C}$(iconv6，256，3，2)反卷积操作、与 Conv4 进行张量拼接得到 concat5、采用 C(concat5，256，3，1，R)卷积操作输出原图像 1/16 尺寸的 iconv5；3）对 iconv5 进行 $\tilde{C}$(iconv5，128，3，2)反卷积操作、与 Conv3 进行张量拼接得到 concat4、采用 C(concat4，128，3，1，R)卷积操作输出原图像 1/8 尺寸的 iconv4。然后采用(8)式进行 Upsample[0.3C(iconv4，2，1，1，S)，2]操作输出视差图 4(Disp4)，重复视差图 4 的生成方法对 iconv4 进行操作输出原图像 1/4 尺寸 iconv3 和视差图 3(Disp3)，对 iconv3 进行操作输出原图像 1/2 尺寸 iconv2 和视差图 2(Disp2)，对 iconv2 进行操作输出原图像一样尺寸 iconv1 和视差图 1(Disp1)，在生成视差图 3、视差图 2、视差图 1 的过程中对应的卷积层输出层数依次为 64、32、16，其余参数与生成视差 4 一样。最后基于金字塔视差图（视差 1、视差 2、视差 3、视差 4）与其对应输入处理的左视图或右视图金字塔图像重构出右视图或左视图金字塔图像。

3.2 训练损失函数

为提高单目深度估计的精度，提出一种新的训练损失函数，该损失函数主要包括 4 种不同分辨率左右视图的重构损失函数、视差图平滑性损失函数、基于双目原理生成视差图一致性损失函数。函数表达式为

$$C_s=\alpha_{\mathrm{ap}}(C_{\mathrm{apl}}+C_{\mathrm{apr}})+\alpha_{\mathrm{ds}}(C_{\mathrm{dsl}}+C_{\mathrm{dsr}})+\alpha_{\mathrm{lr}}C_{\mathrm{lr}},\quad(10)$$

$$C=\sum_{s=1}^{4}C_s,\quad(11)$$

式中：C_{apl} 表示左视图重构损失；C_{apr} 表示右视图重构损失；C_{dsl} 表示左视图产生视差图的平滑性损失；C_{dsr} 表示右视图产生视差图的平滑性损失；C_{lr} 表示左视图重构右视图产生视差与右视图重构左视图产生视差的一致性损失；α_{ap}、α_{ds}、α_{lr} 分别表示图像重构损失函数、视差图平滑性损失函数、双目原理生成视差图一致性损失函数所占的比例。

视差图平滑性损失函数主要是降低通过卷积神经网络生成近似视差图的平滑性对图像重构的影响。为提高视差图的平滑性，提出对视差图的 x、y 进行梯度求导，该函数的数学表达式为

$$C_{\mathrm{dsl}}=\frac{1}{N}\sum_{i,j}\left|\partial_x\boldsymbol{d}_{ij1}\right|\mathrm{e}^{-\|\partial_x\boldsymbol{I}_{ij1}\|}+\left|\partial_y\boldsymbol{d}_{ij1}\right|\mathrm{e}^{-\|\partial_y\boldsymbol{I}_{ij1}\|},\quad(12)$$

式中：$\partial_x\boldsymbol{d}_{ij1}$ 表示视差图在 x 方向上的梯度；$\partial_y\boldsymbol{d}_{ij1}$ 表示视差图在 y 方向上的梯度；$\boldsymbol{I}_{ij1}$ 表示输入的金字塔图像；$\mathrm{e}^{-\|\partial_x\boldsymbol{I}_{ij1}\|}$ 表示 x 方向上的平滑度系数；$\mathrm{e}^{-\|\partial_y\boldsymbol{I}_{ij1}\|}$ 表示 y 方向上的平滑度系数；N 表示训练中使用图像对的数量。

在训练中本模型通过输入一张左视图或右视图，根据与其对应的视差图生成相对应的右视图或者左视图，借鉴文献[20-21]在图像重构中将结构相似性（SSIM）引入图像重构损失函数的影响，提出在训练中采用生成图像与原图像之差绝对值和 SSIM 联合作为图像重构损失函数，该函数的数学表达式为

$$C_{\alpha\mathrm{pl}}=\frac{1}{N}\sum_{i,j}\alpha\ \frac{1-S_{\mathrm{SSIM}}(\boldsymbol{I}_{ij\mathrm{l}},\tilde{\boldsymbol{I}}_{ij\mathrm{l}})}{2}+(1-\alpha)\parallel\boldsymbol{I}_{ijl}-\tilde{\boldsymbol{I}}_{ijl}\parallel,\tag{13}$$

式中：$\tilde{\boldsymbol{I}}_{ij\mathrm{l}}$表示重构的金字塔图像。

双目原理生成视差图一致性损失函数主要是确保左视图重构右视图产生的视差图与右视图重构左视图产生的视差图一致，本损失函数借鉴(13)式，将SSIM引入双目原理生成视差图一致性损失函数中，联合左右视差之差绝对值、左右SSIM作为双目原理生成视差图一致性损失函数，该损失函数学表达式为

$$C_{\mathrm{lr}}=\frac{1}{N}\sum_{i,j}\alpha\ \frac{1-S_{\mathrm{SSIM}}(\boldsymbol{d}_{ij\mathrm{l}},\boldsymbol{d}_{ij\mathrm{r}})}{2}+(1-\alpha)\left|\boldsymbol{d}_{ij\mathrm{l}}-\boldsymbol{d}_{ij\mathrm{r}}\right|\tag{14}$$

式中：$\boldsymbol{d}_{ij\mathrm{l}}$表示通过左视图重构右视图产生的视差图；$\boldsymbol{d}_{ij\mathrm{r}}$表示通过右视图重构左视图产生的视差图。

4 实验过程与结果分析

4.1 实验设置

本实验训练无监督单目深度估计模型的硬件和软件配置为：处理器选用Intel(R)Core(TM)i7-7700K CPU 4.20 GHz；显卡采用NVIDIA GeForce GTX 1080 8 GB；内存为16 GB；训练系统为Ubuntu16.04；深度学习框架为谷歌开源的Tensorflow。训练数据集采用KITTI，将该数据集划分为280000对训练集、1600对验证集。本文定义的最大训练步长为182000次，batchsize为8，epoch为50；训练中每100步保存一次训练损失、每10000次保存一次训练权重；训练初始学习率为0.0001，并在训练总步长的60%时将学习率衰减至原来的50%，在训练总步长的80%时将学习率衰减至原来的25%；在训练中损失函数的系数$\alpha_{\mathrm{ap}}=1$，$\alpha_{\mathrm{ds}}=0.85$，$\alpha_{\mathrm{lr}}=1$。

4.2 模型训练

本模型采用Adam Optimizer对网络权重不断优化，整个训练总共用了35.5 h，每张图片大约0.05 s，训练具体结果如图4所示。图4(a)为重构图像与原图一致性损失函数；图4(b)为重构图像与原图之差绝对值损失函数；图4(c)为总的图像重构损失函数，即为图4(a)和(b)损失函数之和；图4(d)为视差图平滑性损失函数；图4(e)为基于双目原理生成左右视差图一致性损失函数；图4(f)为本模型总的损失函数，即为图4(c)～(e)训练损失函数之和。观察图4可知，经过训练损失函数，整体有小幅度振荡并趋于稳定，且总损失经过训练降到0.4左

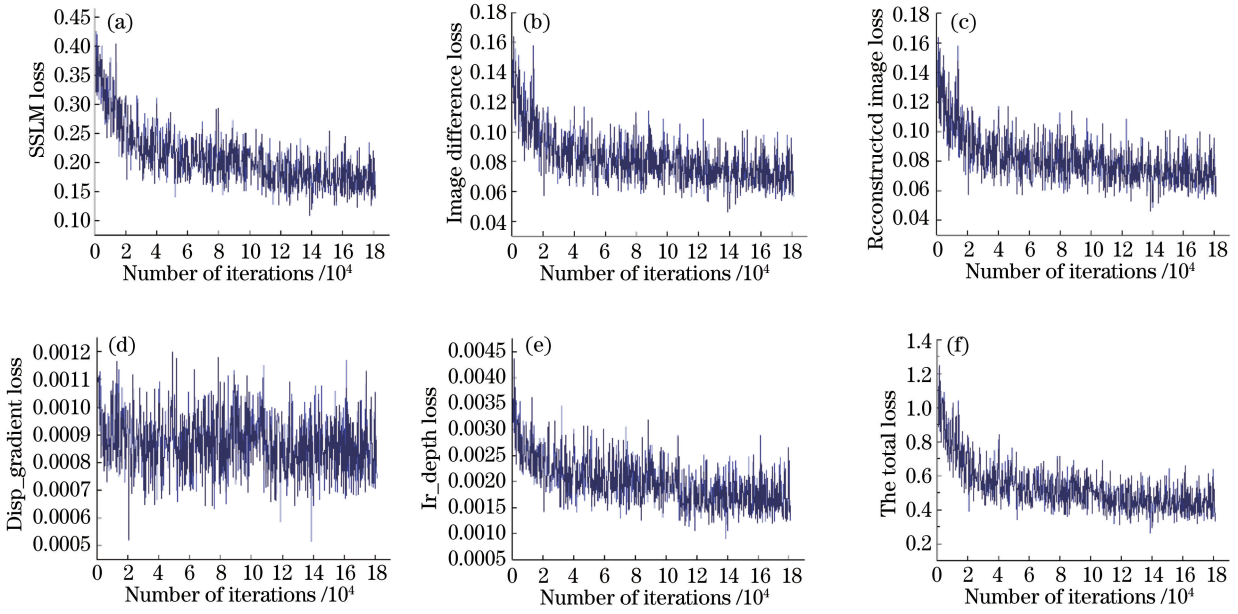

图4 训练过程中各部分损失函数。(a)重构图像与原图的结构相似性损失；(b)重构图像与原图之差绝对值损失；(c)总的图像重构损失；(d)视差图平滑性损失；(e)左右视差图一致性损失；(f)本模型的总损失

Fig. 4 Loss function of each part of training process. (a) Structural similarity loss of reconstructed image and original image; (b) absolute value loss of difference between reconstructed image and original image; (c) total image reconstruction loss; (d) loss of disparity smoothness; (e) loss of consistency in left and right disparity maps; (f) total loss of our model

右，表明本文基于无监督单目深度估计模型的训练效果比较理想。

4.3 无人机实验平台

针对目前市场上无人机因其技术封装性不利于二次开发的问题，选用开源的 Pixhawk 飞控以及相关无人机组件组装无人机，如图 5(a)所示。为实现数据的处理，选用搭载 Auvidea J120 载板的 NVIDIA Jeston TX2，NVIDIA Jeston TX2 是一款嵌入式领域的人工智能(AI)计算机，该计算机使用 6 核 Tegra 处理器和 256 核 Pascal 架构核心 GPU，具备极强的 AI 计算能力。NVIDIA Jeston TX2 与无人机的连接如图 5(b)所示。

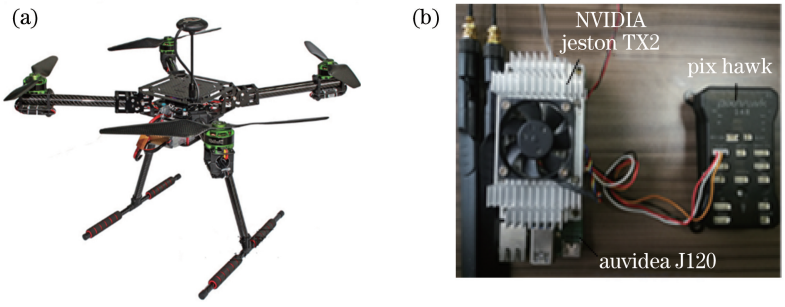

图 5 无人机实验平台。(a)无人机；(b) NVIDIA Jeston TX2 与 Pixhawk 连接

Fig. 5 Platform of drone experiment. (a) Drone; (b) connection of NVIDIA Jeston TX2 and Pixhawk

4.4 实验结果及分析

为验证本文基于无监督单目深度估计模型在无人机上的准确性和实时性，将训练好的模型移植到无人机上的 NVIDIA Jeston TX2 进行实验验证；为验证本模型的泛化能力，将本模型与比较经典的深度估计方法在 KITTI、Make3D 上进行比较，并且在真实室外环境下进行模型测试。以经常采用的评价指标作为参考，评价指标主要包括平均相对误差(REL)、方均根误差(RMSE)、对数空间平均误差、阈值 δ_{th}。前三者的表达式分别为

$$E_{\text{REL}}=\frac{1}{M}\sum_{i}\frac{|\boldsymbol{y}_i^{*}-\boldsymbol{y}_i|}{\boldsymbol{y}_i^{*}}, \tag{15}$$

$$E_{\text{RMSE}}=\sqrt{\frac{1}{M}\sum_{i}(\boldsymbol{y}_i^{*}-\boldsymbol{y}_i)^2}, \tag{16}$$

$$\lg E_{\text{RMSE}}=\sqrt{\frac{1}{M}\sum_{i}\|\lg \boldsymbol{y}_i^{*}-\lg \boldsymbol{y}_i\|^2}, \tag{17}$$

式中：$\boldsymbol{y}_i^{*}$ 表示真实深度值；$\boldsymbol{y}_i$ 表示预测的深度值；M 表示测试中的图像像素总和。阈值 δ_{th} 满足条件 $\max=\left(\frac{\boldsymbol{y}_i^{*}}{\boldsymbol{y}_i},\frac{\boldsymbol{y}_i}{\boldsymbol{y}_i^{*}}\right)=\delta<\delta_{\text{th}}$ 的像素数量所占像素的百分比。

如表 1 所示，本文采用 KITTI 数据集进行了 5 组对比实验，前 2 组分别为文献[12]、文献[15]提出的基于监督学习的单目深度估计，第 3 组和第 4 组分别为文献[19]、文献[20]提出的无监督学习的单目深度估计(输入图像未进行金字塔处理且训练损失函数比较单一)，第 5 组为本文模型采用 VGG-16 进行特征提取的无监督学习的单目深度估计。根据表中的数据可以得出以下结论：1)本文模型进行深度估计的准确率接近甚至优于基于监督学习的单目深度估计模型，这说明基于无监督进行深度估计具有较高的可行性；2)第 3 组、第 4 组和第 5 组数据表明，图像金字塔处理可以实现不同大小目标的检测，进而提高深度估计的准确率，联合视差图平滑性损失、图像重构损失、视差图一致性损失作为训练的总损失有助于提升模型的训练效果，进而提升深度估计的准确率；3)第 5 组和本文模型对比得出基于 ResNet-50 进行特征提取可以降低模型的训练损失，进而提高深度估计的准确率，综上可以得出，本模型提出采用 ResNet-50 进行特征提取、联合视差图平滑性损失、图像重构损失、视差图一致性损失作为训练的总损失函数有助于提升深度估计的准确率；4)本模型每一帧图片的检测时间少于其他模型，平均每帧图片检测时间为 0.048 s，即每秒检测 21 frame，本无人机平台选用的单目相机每秒拍摄 20 frame，因此本模型满足无人机自主飞行中实时深度估计的要求。

为测试本模型深度估计的检测效果，图 6 对比了表 1 中 3 组深度估计模型与本模型。图 6(a)为输入图片，图 6(b)为真实深度图，图 6(c)为文献[15]提出的基于监督学习的单目深度估计深度图，图 6(d)为文献[20]提出的基于无监督单目深度估计深度图，图6(e)是本模型采用VGG-16进行特

表 1 KITTI 数据集上实验结果对比

Table 1 Comparison of experimental results on KITTI dataset

Method	Supervised	Error (lower is better)			Accuracy (higher is better)			Time /s
		E_{REL}	E_{RMSE}	Log E_{RMSE}	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$	
Ref. [12]	Yes	0.203	6.307	0.282	0.702	0.890	0.958	0.051
Ref. [15]	Yes	0.202	6.523	0.275	0.678	0.895	0.965	0.045
Ref. [19]	No	0.208	6.856	0.283	0.678	0.885	0.957	0.062
Ref. [20]	No	0.159	5.789	0.234	0.796	0.923	0.963	0.057
Our (VGG-16)	No	0.148	5.496	0.226	0.812	0.912	0.960	0.056
Our (RseNet-50)	No	0.124	5.331	0.219	0.847	0.945	0.975	0.048

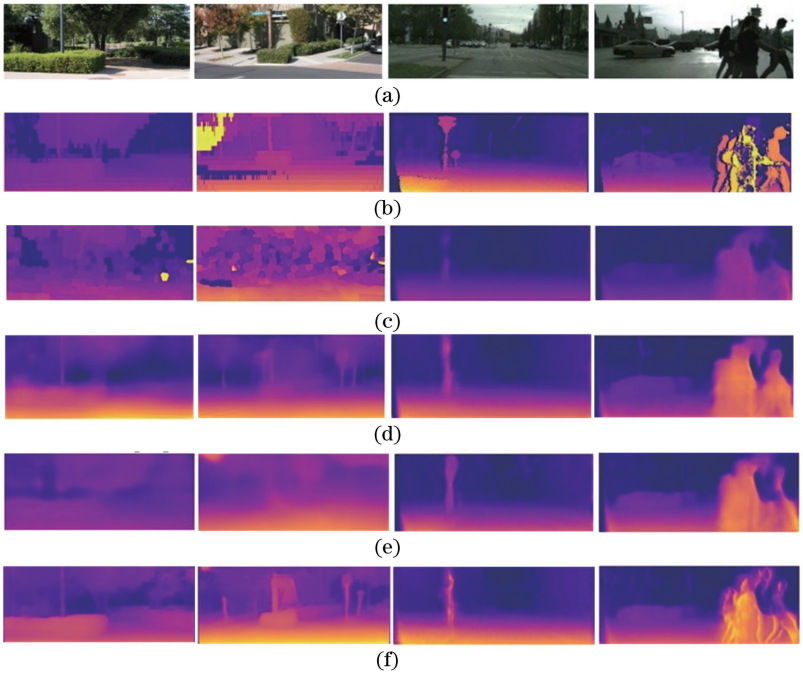

图 6 KITTI 数据集上预测深度图实例。(a)输入的图片;(b)真实深度图;(c)文献[15]预测的深度图;(d)文献[20]预测的深度;(e) 本模型基于 VGG-16 预测的深度;(f)本模型基于 ResNet-50 预测的深度

Fig. 6 Examples of depth map predicted on KITTI dataset. (a) Input image; (b) ground truth depth map; (c) depth map predicted by Ref. [15]; (d) depth map predicted in Ref. [20]; (e) depth map predicted by our model based on VGG-16; (f) depth map predicted by our model based on ResNet-50

征提取的无监督单目深度估计的深度图,图 6(f)是本模型进行深度估计的深度图。4 组实验与真实深度图比较可知:1)本模型预测深度图的清晰度优于其他 3 组实验,且本模型预测的深度图与真实深度图比较接近;2)本模型可以检测比较小的物体轮廓,比如图中的车、人、电线杆等;3)预测近距离深度的准确率优于远处,说明本模型深度估计类似于人眼估计深度。综上可得,本文基于无监督单目深度估计具有较高的准确率,满足无人机自主飞行中实时三维感知的要求。

为验证本模型的泛化能力,将本模型与其他深度估计方法在 Make3D 上进行实验验证,实验结果如表 2 所示。表 2 与表 1 对比得知:1)采用 KITTI 数据集训练得到的模型在 KITTI 上深度估计准确率优于在 Make3D 上的深度估计准确率;2)本模型单目深度估计的准确率优于其他 5 种单目深度估计方法,且准确率基本满足无人机自主飞行对深度估计准确率的要求;3)本模型每帧图片的检测时间少于其他 5 种方法,且每帧平均检测时间为 0.053 s,即平均每秒检测 19 frame,无人机相机每秒拍摄 20 frame,因此本模型针对 Make3D 数据集基本满足无人机实时深度估计的要求。

此外,为更好体现本模型的泛化能力,采用真实的室外场景进行实验验证,图 7 为本模型在真实室外环境进行深度估计的结果。由图 7 可知:1)本模型在距离较近时可以很好地预测出景物轮廓,当距

表 2 Make3D 数据集上实验结果对比

Table 2 Comparisonof experimental results on Make3D dataset

Method	Supervised	Error (lower is better)			Accuracy (higher is better)			Time/s
		E_{REL}	E_{RMSE}	Log E_{RMSE}	$\delta<1.25$	$\delta<1.252$	$\delta<1.252$	
Ref. [12]	Yes	0.417	8.526	0.403	0.692	0.899	0.948	0.068
Ref. [15]	Yes	0.462	9.972	0.456	0.656	0.887	0.945	0.048
Ref. [19]	No	0.443	8.326	0.398	0.662	0.885	0.932	0.074
Ref. [20]	No	0.387	7.895	0.354	0.704	0.899	0.946	0.054
Our (VGG16)	No	0.361	8.102	0.377	0.727	0.905	0.958	0.061
Our (RseNet-50)	No	0.328	7.529	0.348	0.751	0.924	0.962	0.053

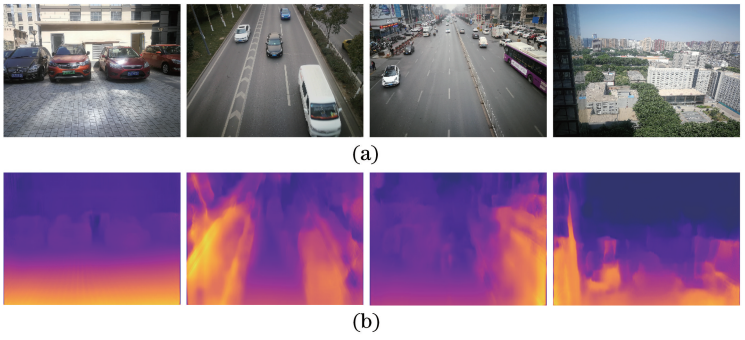

图 7 真实室外场景上预测深度图实例。(a)输入的图片;(b)真实深度图

Fig. 7 Examples of depth map predicted in real outdoor scenes. (a) Input images; (b) ground truth depth maps

离较远时景物轮廓不清晰,其深度估计精度类似于人眼深度估计的精度;2)与 KITTI 数据集上深度估计检测结果进行比较,得知本模型在真实室外场景中仍能取得较好的深度估计效果;3)每帧图片的深度估计时间为 0.058 s,即每秒检测 18 frame。综上得知,本模型对真实室外场景的深度估计仍能取得较好的估计效果,由此证明本模型具有较强的泛化能力。

5 结 论

针对目前单目深度估计需要大量深度图进行训练以及双目深度估计体积大、成本高的问题,为实现无人机自主飞行过程中的三维实时环境感知,提出一种基于无监督学习的单目深度估计模型。本模型借鉴双目深度估计和基于自编码神经网络进行图像重构的原理,将无监督单目深度估计问题转换为图像重构问题。为避免网络层数过多而导致梯度消失和退化问题,该模型基于 ResNet-50 进行特征提取;为实现对小目标的检测,将输入的图像进行金字塔化处理;为提高模型的训练效果,提出联合视差图平滑性损失、图像重构损失、视差图一致性损失作为训练的总损失,并且将结构相似性原理引入到图像损失函数、视差图一致性损失函数中。基于 KITTI 数据集、Make3D 数据集以及真实室外场景进行实验,实验验证与一些经典的基于监督学习的单目深度估计、基于无监督学习的单目深度估计方法相比,本文方法显示出了有效性和优越性,满足无人机自主飞行中实时深度估计准确性和实时性的要求。

参 考 文 献

[1] Li Y Y, Wang H M, Zhang Y F, et al. Structured deep learning based depth estimation from a monocular image[J]. Robot, 2017, 39(6): 812-819.
李耀宇, 王宏民, 张一帆, 等. 基于结构化深度学习的单目图像深度估计[J]. 机器人, 2017, 39(6): 812-819.

[2] Bao Z Q, Li A H, Cui Z G, et al. Research progress of deep learning in visual localization and three-dimensional structure recovery [J]. Laser & Optoelectronics Progress, 2018, 55(5): 050007.
鲍振强, 李艾华, 崔智高, 等. 深度学习在视觉定位与三维结构恢复中的研究进展[J]. 激光与光电子学进展, 2018, 55(5): 050007.

[3] Liao B, Li H W. Image depth estimation model based on atrous convolutional neural network [J]. Journal of Computer Applications, 2019, 39(1): 267-274.
廖斌, 李浩文. 基于多孔卷积神经网络的图像深度估计模型[J]. 计算机应用, 2019, 39(1): 267-274.

[4] Bi T T, Liu Y, Weng D D, et al. Survey on

supervised learning based depth estimation from a single image[J]. Journal of Computer-Aided Design & Computer Graphics, 2018, 30(8): 1383-1393.
毕天腾, 刘越, 翁冬冬, 等. 基于监督学习的单幅图像深度估计综述[J]. 计算机辅助设计与图形学学报, 2018, 30(8): 1383-1393.

[5] He T N, You J G, Chen D F. Depth estimation from single monocular images based on DenseNet[J]. Computer Measurement & Control, 2019, 27(2): 233-236.
何通能, 尤加庚, 陈德富. 基于 DenseNet 的单目图像深度估计[J]. 计算机测量与控制, 2019, 27(2): 233-236.

[6] Gu T T, Zhao H T, Sun S Y. Depth estimation of infrared image based on pyramid residual neural networks[J]. Infrared Technology, 2018, 40(5): 417-423.
顾婷婷, 赵海涛, 孙韶媛. 基于金字塔型残差神经网络的红外图像深度估计[J]. 红外技术, 2018, 40(5): 417-423.

[7] Yuan J Z, Zhou W J, Pan T, et al. Road scene depth estimation based on deep convolutional neural networks[J]. Laser & Optoelectronics Progress, 2019, 56(8): 081501.
袁建中, 周武杰, 潘婷, 等. 基于深度卷积神经网络的道路场景深度估计[J]. 激光与光电子学进展, 2019, 56(8): 081501.

[8] Snavely N, Seitz S M, Szeliski R. Skeletal graphs for efficient structure from motion[C]// 2008 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2008, Anchorage, AK, USA. New York: IEEE, 2008: 10139983.

[9] Zhang R, Tsai P S, Cryer J E, et al. Shape-from-shading: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999, 21(8): 690-706.

[10] Nayar S K, Nakagawa Y. Shape from focus[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1994, 16(8): 824-831.

[11] Favaro P, Soatto S. A geometric approach to shape from defocus[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(3): 406-417.

[12] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems, December 8-13, 2014, Montreal, Canada. USA: MIT Press, 2014: 2366-2374.

[13] Xu L, Zhao H T, Sun S Y. Monocular infrared image depth estimation based on deep convolutional neural networks[J]. Acta Optica Sinica, 2016, 36(7): 0715002.
许路, 赵海涛, 孙韶媛. 基于深层卷积神经网络的单目红外图像深度估计[J]. 光学学报, 2016, 36(7): 0715002.

[14] Cao Y, Wu Z F, Shen C H. Estimating depth from monocular images as classification using deep fully convolutional residual networks[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(11): 3174-3182.

[15] Liu F Y, Shen C H, Lin G S. Deep convolutional neural fields for depth estimation from a single image[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 5162-5170.

[16] Chen W F, Fu Z, Yang D W, et al. Single-image depth perception in the wild[C]// Advances in Neural Information Processing Systems 29 (NIPS 2016), December 5-10, 2016, Barcelona, Spain. Canada: NIPS, 2016: 730-738.

[17] Mayer N, Ilg E, Hausser P, et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 4040-4048.

[18] Godard C, Aodha O M, Brostow G J. Unsupervised monocular depth estimation with left-right consistency[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 6602-6611.

[19] Zhou T H, Brown M, Snavely N, et al. Unsupervised learning of depth and ego-motion from video[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 6612-6621.

[20] Wang C Y, Buenaposada J M, Zhu R, et al. Learning depth from monocular videos using direct methods[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York:

IEEE, 2018: 2022-2030.

[21] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks [C] // Advances in Neural Information Processing System 28 (NIPS 2015), December 7-12, 2015, Montreal, Quebec, Canada. Canada: NIPS, 2015: 2017-2025.