先进驾驶辅助系统中基于单目视觉的场景深度估计方法

丁萌[1,2*]，姜欣言[1]

[1]南京航空航天大学民航学院，江苏 南京 211106；

[2]中国民用航空局飞机健康监测与智能维护重点实验室，江苏 南京 211106

摘要 针对先进驾驶辅助系统对车辆前视景深信息的需求，在无监督学习框架下提出了一种基于单目视觉的场景深度估计方法。为了降低不同尺寸的前视目标对景深估计结果的影响，采用金字塔结构对输入图像进行预处理；在训练过程中，将深度估计问题转化为图像重建问题，利用双目图像设计了新的损失函数代替真实深度标签，解决了真实场景景深数据难以获取的问题；将中间多尺度的视差图与原输入图像的尺寸统一，改善了深度图中的空洞现象，提升了景深估计精度。在 KITTI 和 Make3D 数据集上的定量与定性对比结果表明，本方法可以获得准确度较高的绝对景深数据，且具有良好的泛化能力。在真实道路场景下的实验结果表明，本方法可以利用单张车载前视图像得到对应的像素级景深信息。

关键词 深度估计；卷积神经网络；无监督学习；多尺度统一

中图分类号 TP301.6 **文献标志码** A **doi**：10.3788/AOS202040.1715001

Scene Depth Estimation Based on Monocular Vision in Advanced Driving Assistance System

Ding Meng[1,2*], Jiang Xinyan[1]

[1] *College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu* 211106, *China*;

[2] *Key Laboratory of Aircraft Health Monitoring and Intelligent Maintenance, Civil Aviation Administration of China, Nanjing, Jiangsu* 211106, *China*

Abstract Aiming at that the requirements advanced driving assistance system for vehicle forward looking depth of field information, this paper proposes a scene depth estimation method based on monocular vision under the framework of unsupervised learning. In order to reduce the influence of forward looking targets with diverse sizes on the depth estimation results, the proposed method uses a pyramid structure to preprocess the input image. In the training process, the depth estimation problem is transformed into an image reconstruction problem, and a new loss function is designed using binocular images instead of the true depth label, which solves the problem that the depth data of the real scene is difficult to obtain. The size of disparity map and original input image is unified, which improves the hole phenomenon in depth map and improves the accuracy of scene depth estimation. The quantitative and qualitative comparison results on the KITTI and Make3D datasets show that the proposed method can obtain high accuracy absolute depth of field data and has good generalization ability, Experimental results in real road scenes show that the proposed method can obtain pixel level depth of field information from a single vehicle forward looking image.

Key words depth estimation; convolutional neural network; unsupervised learning; multiscale unification

OCIS codes 150.0155; 110.6880; 330.1400

1 引 言

前视场景深度(景深)信息在先进车辆驾驶辅助系统(ADAS)中具有重要的作用，通过前视景深信息可以精确感知车辆的运行环境，获得道路交通环境中车辆、行人等交通参与者以及路灯建筑等障碍

收稿日期：2020-05-06；**修回日期**：2020-05-21；**录用日期**：2020-05-29

基金项目：国家自然科学基金(61673211)、国家自然科学基金联合基金(U1633105)、中央高校基本科研业务费专项(NS2020049)

* **E-mail**：nuaa_dm@nuaa.edu.cn

物与车辆本体的距离等，从而实现 ADAS 的避障和行人保护功能[1-3]。目前，车辆前视景深信息主要通过激光雷达(lidar)、雷达(radar)以及超声传感器获取，但这类传感器的成本较高，且得到的多是点阵、线阵或稀疏的面阵深度信息[4]。由于 CCD 摄像机能够获取丰富的色彩、纹理等信息，且价格相对低廉，被广泛应用于 ADAS 中[5]。

传统基于图像的景深估计方法多是利用拍摄环境假设的几何约束和手工特征[6-7]，如运动恢复结构(SFM)方法，但这类方法受特征提取与匹配误差的影响，且只能获得较为稀疏的局部景深数据[6]。随着计算机运算能力的提高，以卷积神经网络(CNN)为代表的深度学习方法在计算机视觉领域得到了广泛的应用[8-12]。基于深度学习的景深估计方法中，根据训练过程是否输入真实景深数据作为标签，可以分为监督(supervised)和无监督(unsupervised)方法；根据得到的景深数据可以分为绝对景深估计和相对景深估计方法。

Eigen 等[14]提出的 Coarse-Fine 方法，将两种尺度的 CNN 看作一个整体结构，将粗尺度 CNN 估计场景的全局深度与图像输入到精尺度 CNN 中，对局部细节特征进行优化。在此基础上，Eigen 等[14]提出了新的多尺度网络架构，将深度估计、表面法线估计和语义分割统一在神经网络中，优化了模型的性能。Liu 等[15]提出了深度卷积神经场，将条件随机场与 CNN 相结合；Li 等[16]将 CNN 与条件随机场、超像素相结合，提出用 CNN 回归超像素的深度，用条件随机场进行后处理，再通过超像素尺度优化深度结果。但监督方法需要输入图像的真实景深数据作为训练标签，而场景的真实景深数据难以获取，且真实深度图是稀疏的，无法与输入图像完全拟合。

无监督的单张图像景深估计方法不需要真实景深数据作为标签。其中一种无监督方法为自监督方法，使用单目视频的时序信息作为监督信息。Zhou 等[17]利用单目视频中前后帧之间的时空线索作为监督约束，完成了无监督的深度估计任务，但该方法只能得到相对深度结果，无法满足 ADAS 的要求。另一种无监督方法的思路是将立体图像对的空间约束关系作为监督信息，由于立体图像对的训练过程中，两个相机的相对位姿是已知的，因此，并不需要额外训练位姿估计网络。这类方法在训练过程中利用同步的左右两张图像，逐像素地预测图像对之间的视差，并在测试阶段对单张图像进行深度估计。Xie 等[18]提出的 Deep3D 网络解决了从单张图像到立体图像对的合成问题。在此基础上，Garg 等[19]提出了基于 CNN 的无监督单目视觉深度估计方法，该方法在编码阶段输入左视图，通过 CNN 生成深度图，根据双目图像中视差与深度的关系得到对应的视差图，在解码过程中利用视差图与原输入的右视图重建左视图。但该方法为了使损失函数可以反向传播，用泰勒级数展开的方法计算梯度，增加了网络的复杂度。Godard 等[20]在文献[19]的基础上，提出将左右视图的一致性检查加入训练过程中，并用双线性采样方法确保整个过程的连续性，降低了网络的复杂度。但该方法经过编码器产生的特征图中，尺度较小的低纹理区域比较模糊，导致后续深度图中出现了虚假纹理和空洞现象，难以满足 ADAS 的要求。

为了得到车载前视图像中的景深信息，本文利用车载 CCD 摄像机提供的前视图像信息，提出了一种基于单张图像的前视景深估计方法，为 ADAS 提供与前视图像相匹配的像素级景深信息。将 RGB(Red、Green、Blue)图像转换为 RGB-D(depth)图像，得到的景深信息可快速进行目标检测、识别并与跟踪算法相结合，提高了 ADAS 的环境感知能力。在此基础上，提出了一种利用双目视觉合成原理的无监督景深估计方法，并采取尺度统一化方法减少目标表面深度图的空洞现象。在训练过程中同时预测左右视图，并在损失函数中引入左右一致性损失，加强对视差图一致性的约束。

2　本文算法

2.1　双目图像重建方法的基本原理

在双目相机成像过程中，基于双目图像左右视图重建方法的原理如图 1 所示。用左右两个相机拍摄同一个物体，得到两张不同的图像[21]，其中，f 为两相机的焦距，b 为两相机之间的距离，即基线

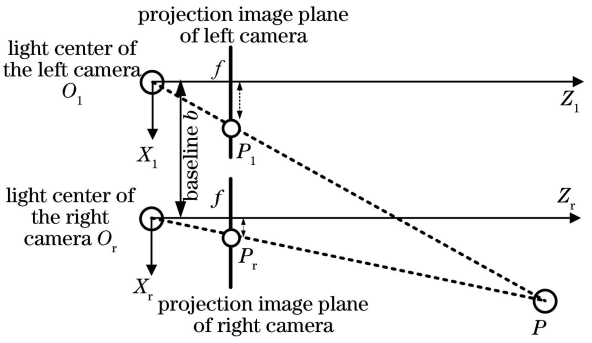

图 1　基于双目重建方法的原理

Fig. 1　Principle based on binocular reconstruction method

(baseline)长度。物点 P 在左相机成像平面上的像点为 P_l，在坐标系 O_l-$X_\mathrm{l}Z_\mathrm{l}$ 下的坐标为(x_l, f)；在右相机成像平面上的像点为 P_r，在坐标系 O_r-$X_\mathrm{r}Z_\mathrm{r}$ 下的坐标为(x_r, f)。

根据相似三角形定理得到

$$\frac{x_\mathrm{p}^\mathrm{l}}{x_\mathrm{l}}=\frac{z_\mathrm{p}^\mathrm{l}}{f},\frac{x_\mathrm{p}^\mathrm{r}}{x_\mathrm{r}}=\frac{z_\mathrm{p}^\mathrm{r}}{f},\tag{1}$$

式中，$(x_\mathrm{p}^\mathrm{l},z_\mathrm{p}^\mathrm{l})$，$(x_\mathrm{p}^\mathrm{r},z_\mathrm{p}^\mathrm{r})$ 分别为物点 P 在左、右相机坐标系下的坐标。令 $x_\mathrm{p}^\mathrm{l}-x_\mathrm{p}^\mathrm{r}=b$，$z_\mathrm{p}^\mathrm{r}=z_\mathrm{p}^\mathrm{l}$，可将(1)式改写为

$$\frac{x_\mathrm{p}^\mathrm{l}}{x_\mathrm{l}}=\frac{z_\mathrm{p}^\mathrm{l}}{f},\frac{x_\mathrm{p}^\mathrm{l}-b}{x_\mathrm{r}}=\frac{z_\mathrm{p}^\mathrm{l}}{f},\tag{2}$$

$$z_\mathrm{p}^\mathrm{l}=\frac{fb}{x_\mathrm{l}-x_\mathrm{r}}。\tag{3}$$

令 $d=x_\mathrm{l}-x_\mathrm{r}$ 为视差，表示物点 P 在左相机和右相机中成像的偏离值，即左视图中的像素点需要平移 d 才能得到右视图中的对应像素，即

$$z_\mathrm{p}^\mathrm{l}=\frac{fb}{d}。\tag{4}$$

可以发现，已知两相机之间的基线距离 b，相机焦距 f 以及两相机中物点的视差 d，就能恢复出物点 P 在左相机坐标系下的像素深度 z_p^l。基于此，可将景深估计问题转化为求解双目图像视差的问题。

2.2　基于左右视图的无监督景深估计

基于左右视图的无监督景深估计网络训练过程中，分别从左右两个相机中获得两张图像 $\boldsymbol{I}^\mathrm{l}$ 和 $\boldsymbol{I}^\mathrm{r}$，将左视图 $\boldsymbol{I}^\mathrm{l}$ 输入景深估计网络后，逐像素地预测从左视图转换为右视图对应的视差图 $\boldsymbol{d}^\mathrm{r}$。为了得到从右视图到左视图的视差图 $\boldsymbol{d}^\mathrm{l}$，需对左视图进行逆向采样，得到左右反向的视图，作为预定右视图，再将预定右视图送入神经网络预测右视图到左视图的视差图 $\boldsymbol{d}^\mathrm{l}$[19-20]。由于左视图逆向采样得到的预定右视图与真实右视图的差距较大，导致网络前期的重建误差较大，模型收敛缓慢且容易陷入局部极小值。因此，实验在训练过程中将左视图与右视图同时输入网络中，用神经网络同时对输入的左视图和右视图进行逐像素预测，得到 $\boldsymbol{d}^\mathrm{r}$ 和 $\boldsymbol{d}^\mathrm{l}$。根据原左视图 $\boldsymbol{I}^\mathrm{l}$ 与左视角到右视角的右视差 $\boldsymbol{d}^\mathrm{r}$ 重建右视图 $\tilde{\boldsymbol{I}}^\mathrm{r}$，根据右视图 $\boldsymbol{I}^\mathrm{r}$ 和右视角到左视角的左视差 $\boldsymbol{d}^\mathrm{l}$ 重建左视图 $\tilde{\boldsymbol{I}}^\mathrm{l}$，最后将重建的左视图和右视图分别与原始左右视图进行对比，具体流程如图 2 所示。

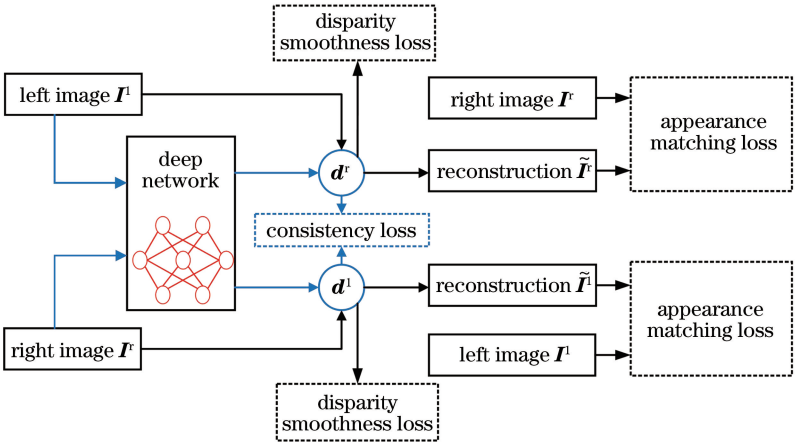

图 2　基于左右视图的景深估计及损失函数

Fig. 2　Depth of field estimation and loss function based on left and right views

2.3　损失函数

从图 2 中可以发现，本方法的损失函数由外观匹配损失 C_ap，左右一致性损失 C_lr 以及视差平滑损失 C_ds 组成。

1) 外观匹配损失 C_ap：外观匹配损失由衡量图像重建质量的结构相似度(SSIM)[22]和对异常点不敏感的 L1 损失函数[23]组成。SSIM 是评价重建图像质量的重要指标，相比均方误差(MSE)，不仅可以计算两图像之间对应像素点的灰度差值，还可以衡量图像的结构相似性，可表示为

$$X_\mathrm{SSIM}(\boldsymbol{x},\boldsymbol{y})=\frac{2\mu_x\mu_y+C_1}{\mu_x^2+\mu_y^2+C_1}\times\frac{2\sigma_{xy}+C_2}{\sigma_x^2+\sigma_y^2+C_2},\tag{5}$$

式中，μ_x 和 μ_y 分别为图像 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的像素值均值，σ_x 和 σ_y 分别为图像 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的像素值标准差，σ_{xy} 为图像 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的像素值协方差，C_1 和 C_2 为避免分母为零增设的常数。SSIM 越大，表明重建图像与原图像之间的差距越小，当两图像完全相同时，SSIM 为 1。L1 损失函数为最小绝对值误差，与 L2 损失函数相比，鲁棒性更高，且对异常点不敏感。以重建的左视图 $\tilde{\boldsymbol{I}}^\mathrm{l}$ 与

输入的左视图 $\boldsymbol{I}^{\mathrm{l}}$ 为例，外观匹配损失 $C_{\mathrm{ap}}^{\mathrm{l}}$ 可表示为

$$C_{\mathrm{ap}}^{\mathrm{l}}=\frac{\alpha}{N}\sum_{i,j}\frac{1-X_{\mathrm{SSIM}}(\boldsymbol{I}_{ij}^{\mathrm{l}},\tilde{\boldsymbol{I}}_{ij}^{\mathrm{l}})}{2}+(1-\alpha)\left\|\boldsymbol{I}_{ij}^{\mathrm{l}}-\tilde{\boldsymbol{I}}_{ij}^{\mathrm{l}}\right\|_1,\tag{6}$$

式中，$\boldsymbol{I}_{ij}^{\mathrm{l}}$、$\tilde{\boldsymbol{I}}_{ij}^{\mathrm{l}}$ 为对应第 i 行、第 j 列的像素点，$\|\ \|_1$ 为 L1 范数，$\alpha=0.85$ 为权值系数，N 为图像像素点数目。

2）左右一致性损失 C_{lr}：本方法将双目图像的左右视图同时输入神经网络，可同时预测出两个视角的视差图。为了保证两个视差图与输入图像具有相同的双目图像转换关系，在损失函数中加入左右一致性损失，确保左右视差图的空间一致性，提高了计算场景深度的精度。以输入的左视图 $\boldsymbol{I}^{\mathrm{l}}$ 为例，左右一致性损失 $C_{\mathrm{lr}}^{\mathrm{l}}$ 可表示为

$$C_{\mathrm{lr}}^{\mathrm{l}}=\frac{1}{N}\sum_{i,j}\left|d_{ij}^{\mathrm{l}}-d_{ij+d_{ij}^{\mathrm{l}}}^{\mathrm{r}}\right|,\tag{7}$$

式中，d_{ij}^{l} 为左视差图第 i 行、第 j 列的像素值，$d_{ij+d_{ij}^{\mathrm{l}}}^{\mathrm{r}}$ 为右视差图第 $i+d_{ij}^{\mathrm{l}}$ 行，第 $j+d_{ij}^{\mathrm{l}}$ 列的像素值。

3）视差平滑损失 C_{ds}：C_{ds} 可以解决像素点的深度不适定问题，对视差图的梯度 $\partial\boldsymbol{d}$ 进行 L1 惩罚，使视差在局部具有平滑性。针对图像梯度中出现的深度不连续情况，用图像梯度 $\partial\boldsymbol{I}$ 的边缘感知项对 C_{ds} 进行加权，输入左视图的平滑损失 $C_{\mathrm{ds}}^{\mathrm{l}}$ 可表示为

$$C_{\mathrm{ds}}^{\mathrm{l}}=\frac{1}{N}\sum_{i,j}\left|\partial_x d_{ij}^{\mathrm{l}}\right|\exp(-\left\|\partial_x\boldsymbol{I}_{ij}^{\mathrm{l}}\right\|_1)+\left|\partial_y d_{ij}^{\mathrm{l}}\right|\exp(-\left\|\partial_y\boldsymbol{I}_{ij}^{\mathrm{l}}\right\|_1)。\tag{8}$$

式中，∂_x、∂_y 分别为 x 与 y 方向上梯度，将左右视图的每一项损失相加，得到的总损失函数为

$$C_z=C_{\mathrm{ap}}^{\mathrm{l}}+C_{\mathrm{ap}}^{\mathrm{r}}+C_{\mathrm{ds}}^{\mathrm{l}}+C_{\mathrm{ds}}^{\mathrm{r}}+C_{\mathrm{lr}}^{\mathrm{l}}+C_{\mathrm{lr}}^{\mathrm{r}}。\tag{9}$$

2.4　多尺度视差图

为了防止训练陷入局部极小值，将双目图像输入神经网络之前，先对输入图像进行金字塔结构处理，分别将输入图像下采样至原图像的 1，1/2，1/4，1/8 尺度，形成金字塔结构。然后将下采样得到的图像输入编解码网络拟合视差图，经编码网络提取的特征图尺度分别为原输入图像的 1/16，1/32，1/64，1/128。将编码阶段获得的四个尺度特征图输入解码网络中，并对输入特征进行逐层反卷积，使其恢复至原输入图像的 1，1/2，1/4，1/8 尺度的金字塔结构，得到四个尺度的视差图。在四个尺度上分别对原图进行重建，并计算四个尺度的损失。计算总损失时，将四个尺度上的损失相加，以减轻多尺度目标对景深估计的影响。由于低纹理区域的模糊性，处理较低尺度的视差图时，会使该区域的外观误差不稳定，从而导致生成的深度图中产生空洞和纹理错误现象，降低景深估计的精度。

在金字塔结构处理的基础上，提出了多尺度统一方法。将编码网络获得的四个尺度视差图统一上采样至输入图像的尺度，在原始高分辨率图像下进行图像重建以及损失计算，具体流程如图 3 所示。用四个输入的右视差图 $\boldsymbol{d}^{\mathrm{r}}$ 与输入的左视图 $\boldsymbol{I}^{\mathrm{l}}$ 重建

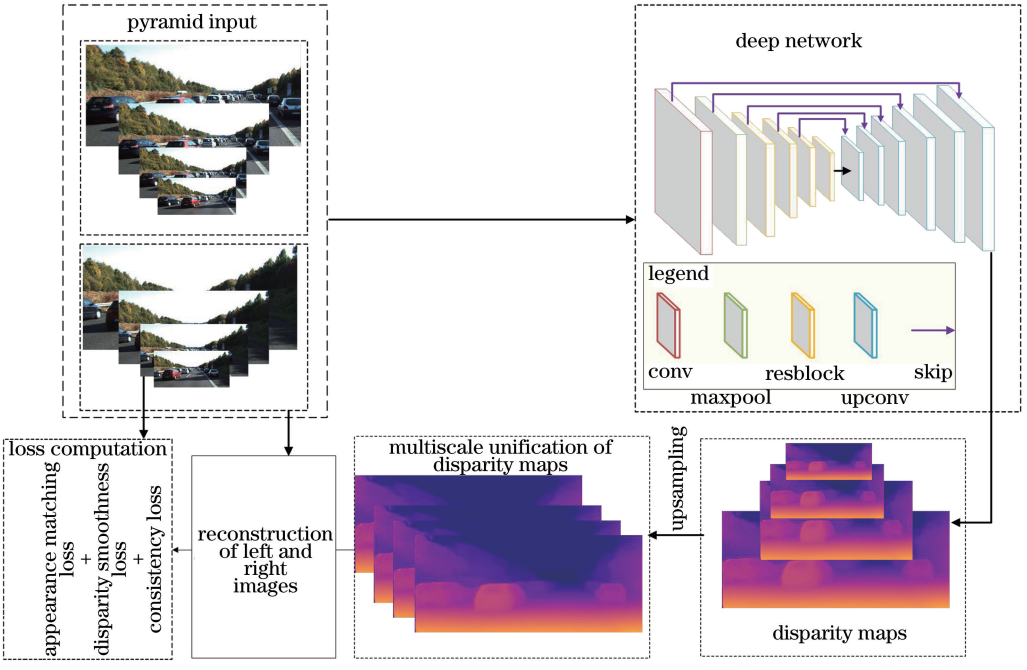

图 3　多尺度统一操作原理

Fig. 3　Principle of multi scale unified operation

出右视图 $\tilde{\boldsymbol{I}}^{\mathrm{r}}$，用四个输入的左视差图 $\boldsymbol{d}^{\mathrm{l}}$ 与输入的右视图 $\boldsymbol{I}^{\mathrm{r}}$ 重建出左视图 $\tilde{\boldsymbol{I}}^{\mathrm{l}}$。多尺度统一操作可以确保每一尺度的视差图都能准确地重建出与输入图像尺寸一致的目标图像，从而提高重建图像的准确率，改善低尺度视差图中出现的空洞现象。

2.5　本方法原理

本方法以编解码网络为主要框架，在编码阶段采用残差网络（ResNet-50）作为特征提取网络[24]，解码阶段的网络为反方向的 ResNet-50。在残差网络中，网络层数越多，CNN 越深，模型的复杂度越高，计算量越大，对硬件的要求也就越高。因此，在保证网络深度的前提下，采用计算量较小的 ResNet-50 作为编码阶段的主干网络，以兼顾网络的性能与实时性。

在训练过程中，将双目图像的左右视图同时输入景深估计网络，利用 CNN 强大的拟合能力同时生成左右两个视差图，结合视差图与输入的原始图像，根据双目图像立体匹配原理，重建出左右视图，并将重建的图像与原始图像进行对比，计算出误差。将误差反向传播到网络前端，调整网络参数，改进生成的视差图，反复进行该过程，直到重建误差降到最小。在测试过程中，只需将单张测试图像输入预训练景深估计网络，通过训练好的模型参数拟合输入图像对应的视差图，再结合双目相机的焦距和基线距离信息计算出输入图像中像素点在相机坐标系下的深度值 z，本网络的具体流程如图 4 所示。

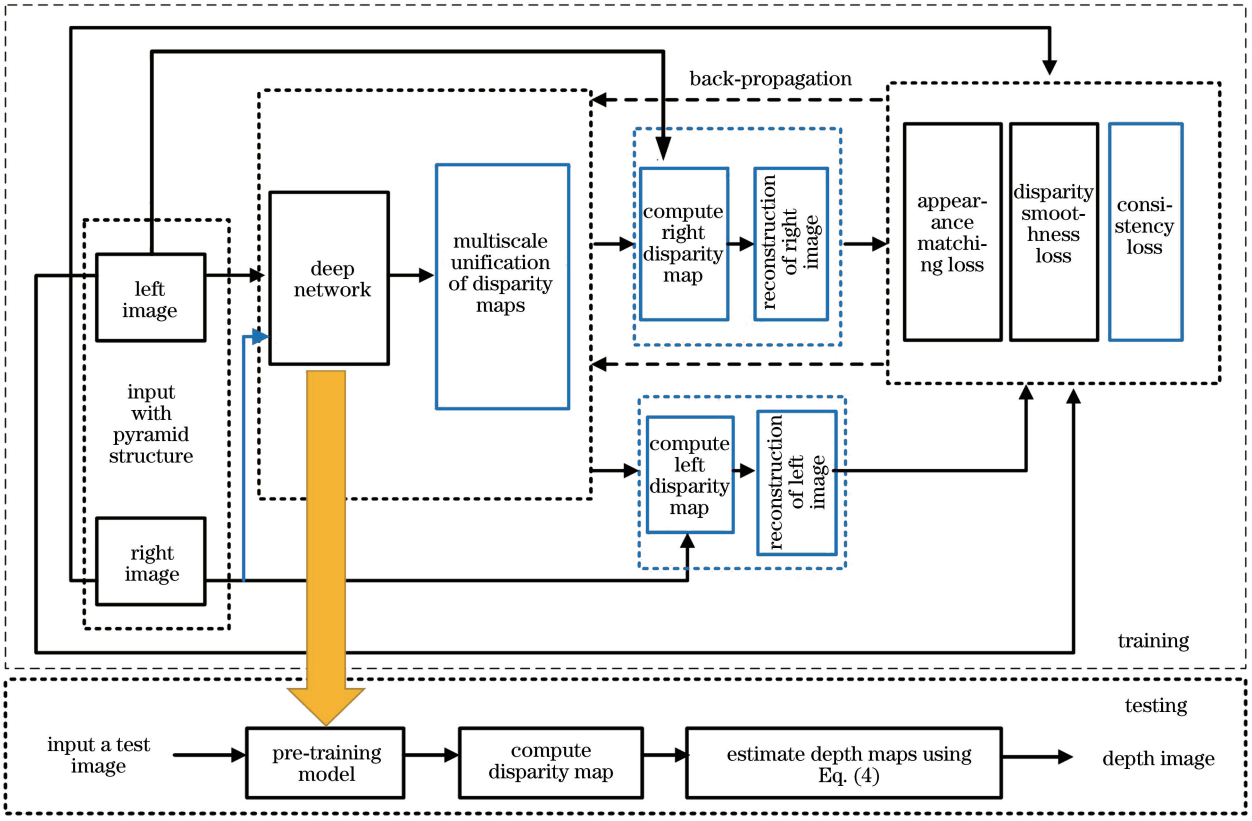

图 4　本网络的流程图

Fig. 4　Flow chart of our network

3　实验结果及分析

实验平台配置：显卡为 NVIDIA GeForce GTX 1080Ti，操作系统为 Ubuntu14.04，采用 TensorFlow1.4.0 框架搭建平台，在经典驾驶数据集 KITTI 上进行训练。将本方法与包括监督方法和无监督方法在内的其他景深估计方法进行定量与定性分析，定量分析以数据集中车载雷达获取的稀疏真实景深数据作为基准值（ground truth）。在立体测试数据集 KITTI 2015 上验证本方法的有效性，在 Make3D 数据集中的 150 张测试样本上定量描述本方法的泛化能力，同时在真实的驾驶道路场景图像集上进行测试。对比方法包括文献[14]中用真实数据作为标签的有监督景深估计方法，文献[17]中以单目视频序列作为输入的无监督景深估计方法，文献[20]中利用双目图像进行无监督景深估

计的方法。

3.1　实验指标

在训练过程中，用具有双目图像的 KITTI 数据集中 18000 张图像对网络进行训练。初始学习率 $l_r=0.0001$，在 40 个迭代周期后，每 10 个周期将学习率变为当前学习率的一半，共训练 70 个周期。批处理大小为 8，使用 Adam 优化器进行优化，其中，$\beta_1=0.9$，$\beta_2=0.999$。定量分析采用的评价指标如下[25]。

1）绝对值相对误差(AbsRel)可表示为

$$X_{\mathrm{AbsRel}}=\frac{1}{N}\sum_{i=1}^{N}\frac{|y_i-y_i^*|}{y_i^*},\tag{10}$$

式中，y_i 为预测得到的某像素点深度值，y_i^* 为该像素点的真实深度值。

2）均方根误差(RMSE)可表示为

$$X_{\mathrm{RMSE}}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\|y_i-y_i^*\|^2}。\tag{11}$$

3）阈值精度(ThrAcc)与正确识别率(CRR)。

$$\max\left(\frac{y_i}{y_i^*},\frac{y_i^*}{y_i}\right)=\delta。\tag{12}$$

根据(12)式计算 δ，统计图像中 $\delta<X_{\mathrm{thr}}$ 的像素点占总像素点的比例，即正确率 X_{CRR}，阈值精度 X_{thr} 一般取 1.25，1.25^2，1.25^3。根据不同的 X_{thr} 可得到不同的 X_{CRR}。上述评价指标中，X_{bsRel}，X_{RMSE} 越小，表明景深估计结果的准确度越高；X_{CRR} 越大，表明景深估计的结果越好。

3.2　KITTI 数据集上的实验结果与分析

首先，在 KITTI 数据集上进行定量对比实验，结果如图 5 和表 1 所示，图 5(a)、图 5(b)分别为四种方法的 AbsRel 和 RMSE，表 1 为 X_{thr} 分别取 1.25，1.25^2，1.25^3 时的 X_{CRR}。可以发现，尽管没有使用真实景深数据作为监督信号，相比文献[14]中的有监督方法，本方法的景深估计结果更好。由于文献[17]中的方法只利用单一视角的视频序列完成景深估计，只能得到测试图像的相对场景深度图。为了进行定量对比，用测试图像的相对景深估计结果与其对应的真实景深数据的中位数比值作为尺度因子，将估计出的相对场景深度恢复到绝对场景深度上。结果表明，本方法不仅可以直接估计出绝对场景深度，且在各项评价指标上均优于文献[17]的方法。此外，在评价指标 AbsRel 和三个不同阈值的 X_{CRR} 上，本方法也优于文献[20]的方法。这表明基于双目重建的框架下，本方法的景深估计精度较好。

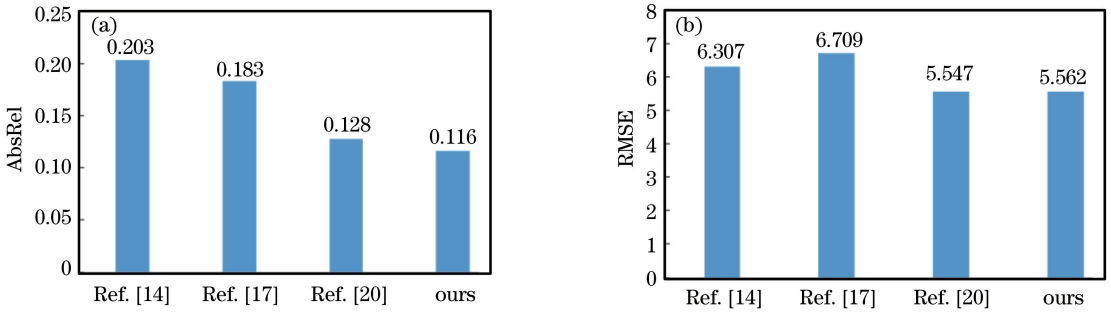

图 5　四种方法的评价指标(KITTI 数据集)。(a) AbsRel；(b) RMSE

Fig. 5　Evaluation indicators of the four methods (KITTI dataset). (a) AbsRel; (b) RMSE

表 1　不同的阈值得到的 X_{CRR}(KITTI 数据集)

Table 1　X_{CRR} obtained by different thresholds (KITTI dataset)　　unit: %

X_{thr}	Ref. [14]	Ref. [17]	Ref. [20]	Ours
1.25	70.2	73.4	81.5	**82.6**
1.25^2	89.0	90.2	92.2	**92.8**
1.25^3	95.8	95.9	96.8	**97.4**

为了直观评估景深估计的结果，给出了四种方法对 3 组图像的景深估计结果，如图 6 所示。可以发现，使用稀疏的真实数据插值得到的稠密深度图出现深度图不连续、空洞等现象，且物体边缘不够平滑。从图 6(d)～图 6(f)中的矩形框区域发现，本方法可以有效改善深度图空洞现象，且能平滑地预测图像中车辆等物体的边缘。

3.3　Make3D 数据集上的实验结果与分析

为了验证本方法的泛化性能，将 KITTI 数据集训练得到的模型，应用到户外场景 Make3D 数据集上，用 150 个测试样本得到四种方法的定量对比结果，如图 7 和表 2 所示。图 7 为不同方法得到的 AbsRel 和 RMSE，表 2 为 X_{thr} 取 1.25，1.25^2，1.25^3 时的 X_{CRR}。

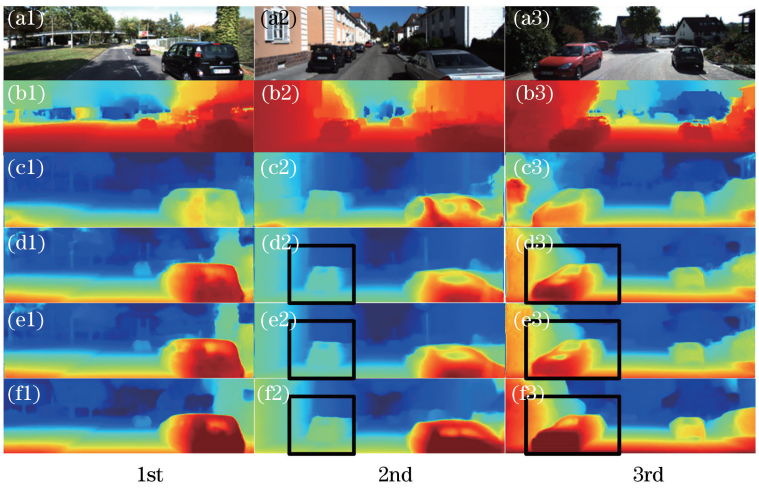

图 6　KITTI 数据集上的景深估计结果。(a)原始图像;(b)真实值;(c)文献[14]的方法;(d)文献[17]的方法;(e)文献[20]的方法;(f)本方法

Fig. 6　Depth estimation results on the KITTI dataset. (a) Original image; (b) true value; (c) method of Ref. [14]; (d) method of Ref. [17]; (e) method of Ref. [20]; (f) our method

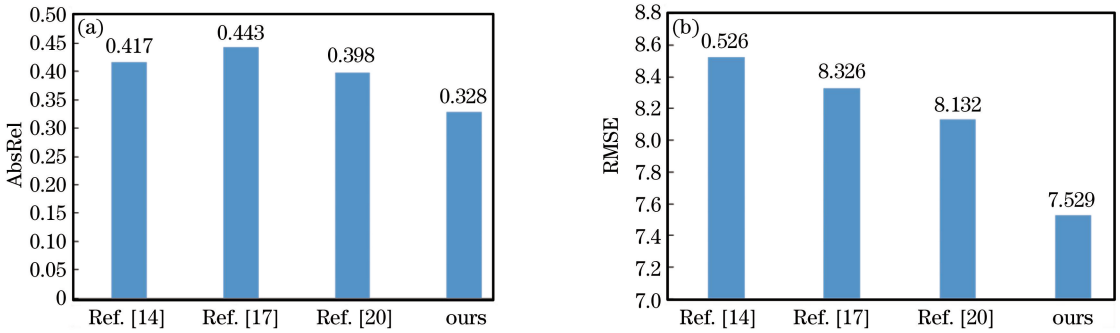

图 7　四种方法的评价指标(Make3D 数据集)。(a) AbsRel;(b) RMSE

Fig. 7　Evaluation indicators of the four methods (Make3D dataset). (a) AbsRel; (b) RMSE

表 2　不同的阈值得到的 X_{CRR}(Make3D 数据集)

Table 2　X_{CRR} obtained by different thresholds (Make3D dataset)　　unit: %

X_{thr}	Ref. [14]	Ref. [17]	Ref. [20]	Ours
1.25	69.2	66.2	72.1	**75.2**
1.25^2	89.9	88.5	90.7	**91.2**
1.25^3	94.8	93.2	95.1	**96.2**

与 KITTI 数据集不同,Make3D 数据集多为户外风景以及建筑物。对比图 5 和图 7、表 1 和表 2 可以发现,在 Make3D 数据集中,本方法的 AbsRel 由 KITTI 数据集中的 0.116 上升到 0.328。原因是 Make3D 与 KITTI 数据集中图像的场景分布差异较大,但本方法在 Make3D 数据集上的性能依然优于其他三种景深估计方法。

3.4　真实道路场景实验结果

为了进一步验证本方法的实际性能,在真实拍摄的室外道路场景图像上进行测试,得到的结果如图 8 所示。可以发现,在未有相似视角的道路驾驶场景图像加入训练的情况下,本方法训练得到的景深估计网络依然能够反映测试场景中各物体的深度,辨识出近距离目标,且基本能完整地恢复驾驶场

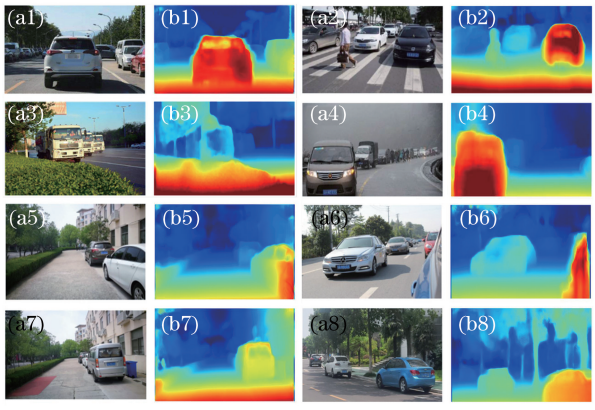

图 8　室外道路景深估计结果。(a)原始图像;(b)景深估计结果

Fig. 8　Depth estimation results of outdoor roads. (a) Original image; (b) estimated result of depth of field

景中的三维环境信息，可以满足 ADAS 中近距离障碍物距离测量的要求。

4 结 论

针对真实景深数据难以获取的情况，提出了一种无监督的场景深度估计方法。本方法在训练过程中使用双目图像进行训练，在测试过程中以双目视差图为基础，根据双目图像的三角测量原理，利用已知的相机焦距和基线对输入的单目图像进行深度估计。在重建过程中增加了平滑误差，以减轻深度不适定问题；同时通过多尺度统一方法在输入分辨率上计算损失，使各个尺度的优化目标一致，减轻了深度图空洞的问题。相比其他基于深度学习的深度估计方法，本方法在 KITTI 数据集上的估计精度更好；在 Make3D 数据集上的定量分析结果表明，本方法具有良好的泛化能力；在真实道路场景下的实验结果表明，本方法能满足 ADAS 的要求。

参 考 文 献

[1] Ranft B, Stiller C. The role of machine vision for intelligent vehicles[J]. IEEE Transactions on Intelligent Vehicles, 2016, 1(1): 8-19.

[2] Bengler K, Dietmayer K, Farber B, et al. Three decades of driver assistance systems: review and future perspectives[J]. IEEE Intelligent Transportation Systems Magazine, 2014, 6(4): 6-22.

[3] Zaklouta F, Stanciulescu B. Real-time traffic-sign recognition using tree classifiers[J]. IEEE Transactions on Intelligent Transportation Systems, 2012, 13(4): 1507-1514.

[4] Arnold E, Al-Jarrah O Y, Dianati M, et al. A survey on 3D object detection methods for autonomous driving applications[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(10): 3782-3795.

[5] Ding M, Zhang X, Chen W H, et al. Thermal infrared pedestrian tracking via fusion of features in driving assistance system of intelligent vehicles[J]. Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering, 2019, 233(16): 6089-6103.

[6] Zheng T X, Huang S, Li Y F, et al. Key techniques for vision based 3D reconstruction: a review[J]. Acta Automatica Sinica, 2020, 46(4): 631-652.
郑太雄, 黄帅, 李永福, 等. 基于视觉的三维重建关键技术研究综述[J]. 自动化学报, 2020, 46(4): 631-652.

[7] Liu X M, Du M Z, Ma Z B, et al. Depth estimation method of light field image based on occlusion scene [J]. Acta Optica Sinica, 2020, 40(5): 0510002.
刘晓旻, 杜梦珠, 马治邦, 等. 基于遮挡场景的光场图像深度估计方法[J]. 光学学报, 2020, 40(5): 0510002.

[8] Wu X W, Sahoo D, Hoi S C H. Recent advances in deep learning for object detection[J]. Neurocomputing, 2020, 396: 39-64.

[9] Chang L, Deng X M, Zhou M Q, et al. Convolutional neural networks in image understanding[J]. Acta Automatica Sinica, 2016, 42(9): 1300-1312.
常亮, 邓小明, 周明全, 等. 图像理解中的卷积神经网络[J]. 自动化学报, 2016, 42(9): 1300-1312.

[10] Meng L, Yang X. A survey of object tracking algorithms [J]. Acta Automatica Sinica, 2019, 45(7): 1244-1260.
孟琭, 杨旭. 目标跟踪算法综述[J]. 自动化学报, 2019, 45(7): 1244-1260.

[11] Li P X, Wang D, Wang L J, et al. Deep visual tracking: review and experimental comparison [J]. Pattern Recognition, 2018, 76: 323-338.

[12] 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.

[13] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network [EB/OL]. (2014-6-09) [2020-04-25]. https://arxiv.org/abs/1406.2283.

[14] Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture [C] // 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 2650-2658.

[15] Liu F Y, Shen C H, Lin G S, et al. Learning depth from single monocular images using deep convolutional neural fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(10): 2024-2039.

[16] Li B, Shen C H, Dai Y C, et al. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs [C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 1119-1127.

[17] 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: 6612-6619.

[18] Xie J Y, Girshick R, Farhadi A. Deep3D: fully automatic 2D-to-3D video conversion with deep convolutional neural networks[M]//Leibe B, Matas J, Sebe N, et al. Computer Vision-ECCV 2016. Lecture Notes in Computer Science. Cham: Springer 2016, 9908: 842-857.

[19] Garg R, Vijay K B G, Carneiro G, et al. Unsupervised CNN for single view depth estimation: geometry to the rescue[M]//Leibe B, Matas J, Sebe N, et al. Computer Vision-ECCV 2016. Lecture Notes in Computer Science. Cham: Springer, 2016, 9912: 740-756.

[20] 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.

[21] Wang S, Xu X. 3D reconstruction based on horopter [J]. Acta Optica Sinica, 2017, 37(5): 0515004.
王珊, 徐晓. 基于双目单视面的三维重建[J]. 光学学报, 2017, 37(5): 0515004.

[22] 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: a Publication of the IEEE Signal Processing Society, 2004, 13(4): 600-612.

[23] Zhao H, Gallo O, Frosio I, et al. Loss functions for image restoration with neural networks[J]. IEEE Transactions on Computational Imaging, 2017, 3(1): 47-57.

[24] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[25] Jiang X Y, Ding M. Unsupervised monocular depth estimation with scale unification[C]// 2019 12th International Symposium on Computational Intelligence and Design (ISCID), December 14-15, 2019, Hangzhou, China. New York: IEEE, 2019: 284-287.