基于共同视域的自监督立体匹配算法

王玉锋[1,2**]，王宏伟[3]，吴晨[1]，刘宇[2]，袁昱纬[4]，全吉成[1,2*]

[1]海军航空大学航空作战勤务学院，山东 烟台 264001；

[2]空军航空大学航空作战勤务学院，吉林 长春 130022；

[3]空军航空大学飞行研究所，吉林 长春 130022；

[4]中国人民解放军 91977 部队，北京 102200

摘要 提出了一种基于共同视域的自监督立体匹配算法，该算法根据视差的左右一致性来确定双目图像的共同可视区域，从而抑制被遮挡区域产生的噪声，为网络模型的学习提供了更加准确的反馈信号。研究结果表明：在没有任何标签数据的前提下，所提算法的预测误差降低了 11%～42%，且与有监督立体匹配算法的性能相当。

关键词 机器视觉；立体匹配；自监督学习；双目视觉

中图分类号 TP183 **文献标识码** A **doi**：10.3788/AOS201939.0215004

Self-Supervised Stereo Matching Algorithm Based on Common View

Wang Yufeng[1,2**], Wang Hongwei[3], Wu Chen[1], Liu Yu[2], Yuan Yuwei[4], Quan Jicheng[1,2*]

[1] *College of Operation Service on Aviation, University of Naval Aviation, Yantai, Shandong* 264001, *China*;

[2] *College of Operation Service on Aviation, Aviation University of Air Force, Changchun, Jilin* 130022, *China*;

[3] *Flight Institute, Aviation University of Air Force, Changchun, Jilin* 130022, *China*;

[4] *The* 91977 *Troops of the PLA, Beijing* 102200, *China*

Abstract A self-supervised stereo matching algorithm is proposed based on common view. In this algorithm, the common visible region of the binocular images is determined according to the left-right consistency of disparity and thus the noise generated in the occluded region is suppressed, which provides more accurate feedback signals for the network model learning. The research results show that the prediction error of the proposed algorithm can be reduced by 11%-42% without any label data, and the performance of the proposed algorithm is comparable to that of the supervised stereo matching algorithm.

Key words machine vision; stereo matching; self-supervised learning; binocular vision

OCIS codes 150.6910; 150.5670; 150.1135

1 引 言

立体匹配就是从已校正的双目图像中获取密集的视差图，是计算机视觉中的一个核心问题[1]，广泛应用于三维重构[2]、无人驾驶[3]及机器人导航[4]等多种领域。根据 Scharstein 等[5]的研究，典型的立体匹配方法包含 4 个部分：匹配损失计算、匹配损失聚合、视差计算和视差微调。传统的立体匹配方法需要设计较好的像素描述子，如 SIFT（Scale-Invariant Feature Transform）[6]、HOG（Histogram of Oriented Gradient）[7]等，使用像素描述子之间的差异来计算匹配损失，根据局部数据和平滑性约束进行全局优化[8]。

卷积神经网络（CNN）可以有效地理解语义信息，在目标分类[9]、目标检测[10]和语义分割[11]等任务中取得了优异的性能，在立体匹配算法中也广受关注[12-14]。Žbontar 等[12-13]将其应用于计算两个 9×9 区块之间的相似度，并通过一些传统的后处理步骤进行优化，取得了良好的效果。随后，这种方法不断地被改进，通过改进表达相似度函数的神经网络结构来提高算法性能[15-16]，为进一步消除可靠性低的匹配，建立自适应的平滑性约束、预测置信度等方法也取得了一定的成效[17-18]。端到端的学习在全局优化上往往能获得更好的性能，Mayer 等[19]提出一种“编码-解码”网络结构，并创建了一个大型的合成数据集来进行视差的端到端学习。以此

收稿日期：2018-07-20；**修回日期**：2018-09-27；**录用日期**：2018-10-10

* **E-mail**：jicheng_quan@126.com；** **E-mail**：wangyf_1991@163.com

视差预测网络为基础，Pang 等[20]通过级联另一个网络进行视差微调来提高精度。为更好地共享特征，Liang 等[21]提出一种融合立体匹配所有步骤的网络结构，从多尺度共享特征学习先验和后验的不变特征，并将 CNN 与贝叶斯推理相结合进行视差微调。从语义的上下文信息入手，Kendall 等[22]使用 3D 卷积自编码器进行匹配损失计算，再以可差分的函数进行亚像素的视差预测。在此基础上，Yu 等[23]引入一个明确的损失聚合子网络进行优化损失计算，使用双流网络分别进行损失聚合的提名和选择。Chang 等[24]采用金字塔池化来提高对全局特征的抽象能力，并用堆叠的三维卷积来处理语义的上下文信息。

目前，基于学习的视差预测方法大多使用真实视差作为监督信号，训练过程需要大范围高质量的视差数据，而这样的数据采集是一个极具挑战性的任务。对于双目图像，深度预测和视差预测十分相似，若已知相机内方位元素，可以用明确的公式表达它们之间的关系。近年来，在缺乏标签数据的情况下，可以使用双目图像的重构误差来实现自监督学习，这种方法利用视角合成技术，根据预测的深度图和右视角图像生成左视角图像，从而可以使用图像的重构误差来学习网络参数，在单目图像的深度预测领域已经有一定的研究成果[25-27]。Xie 等[25]采用一种可差分的方式从右视角图像重构左视角图像，从而实现自监督的深度预测模型。在此基础上，Garg 等[26]进一步引入平滑性约束损失来提高深度预测模型的稳定性。Godard 等[27]采用空间变换网络(STN)[28]来实现可差分的图像重构，并对左右视角图像进行同步预测，同时引入左右一致性约束损失来提高预测精度。这种自监督的深度预测模型只需要双目图像形成自反馈，不需要任何标签数据，很容易从自然场景扩展到其他应用场景，如医学领域的手术机器人[29]、内窥镜检查[30]等。

为方便扩展模型的应用场景，主要研究自监督立体匹配算法，这种方法只利用传感器的原始输出数据即可学习到合理的视差预测模型。与所提算法最相似的是文献[31]的算法，但该算法没有处理被遮挡的区域，而这些区域无法形成有效的反馈，反而是模型学习的噪声，降低了算法的整体性能。因此，所提算法通过检测被遮挡区域，只在共同视域内计算图像的重构损失以降低模型学习的噪声，从而提高自监督立体匹配算法的性能。

2　自监督立体匹配算法

有监督立体匹配算法可以利用有标签的数据驱动模型学习，合理的网络结构和大量的有标签数据是决定模型性能的关键，而对于自监督立体匹配算法，如何构建驱动模型学习的损失函数是问题的关键。自监督立体匹配算法基于左右视角图像的一致性，使用重构误差来构建模型训练的反馈信号。视角合成技术是从其他视角图像重构当前视角图像，为自监督立体匹配算法的实现提供了可能，利用观察图像和重构图像之间的差异来驱动模型的学习。空间变换网络(STN)[28]能够实现可差分的图像采样，是确保模型端到端学习的重要环节。

如图 1 所示，视差预测模型以双目图像为输入，输出左图像的密集视差图，STN 表达的重构函数以右图像和左视差图为输入，输出重构的左图像。根据左图的观察值和重构值即可构建损失函数并驱动视差预测模型的学习，这种回环反馈同样适用于右图像，具体流程如图 1 所示。

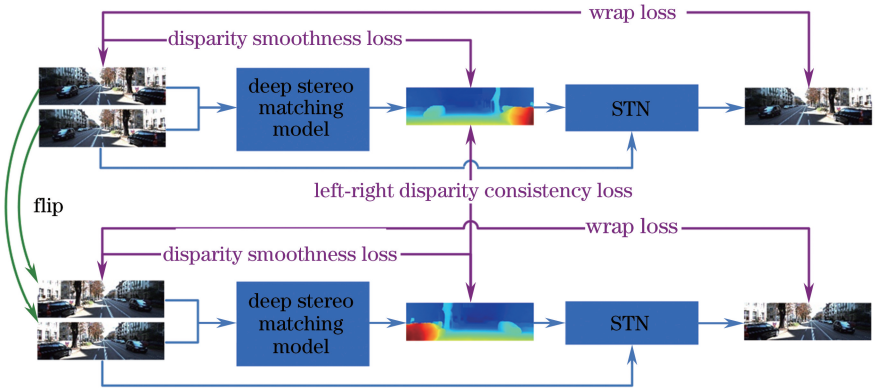

图 1　自监督立体匹配算法示意图

Fig. 1　Schematic of self-supervised stereo matching algorithm

设 $\boldsymbol{I}_{\mathrm{L}}$，$\boldsymbol{I}_{\mathrm{R}}$ 分别为左、右视角图像矩阵，则模型预测的左图像视差矩阵 $\boldsymbol{D}_{\mathrm{L}}$ 可表示为

$$\boldsymbol{D}_{\mathrm{L}}=F_{\mathrm{dsm}}(\boldsymbol{\theta};\boldsymbol{I}_{\mathrm{L}},\boldsymbol{I}_{\mathrm{R}}),\tag{1}$$

式中：F_{dsm} 为视差预测模型函数；$\boldsymbol{\theta}$ 为 F_{dsm} 的模型参数。同理可得右图像视差矩阵为

$$\boldsymbol{D}_{\mathrm{R}}=F_{\mathrm{filp}}\{F_{\mathrm{dsm}}[\boldsymbol{\theta};F_{\mathrm{filp}}(\boldsymbol{I}_{\mathrm{R}}),F_{\mathrm{filp}}(\boldsymbol{I}_{\mathrm{L}})]\},\tag{2}$$

式中 F_{filp} 为对图像进行左右翻转的函数。

使用 STN 对 $\boldsymbol{I}_{\mathrm{L}}$，$\boldsymbol{I}_{\mathrm{R}}$，$\boldsymbol{D}_{\mathrm{L}}$，$\boldsymbol{D}_{\mathrm{R}}$ 进行重构，得到

$$\begin{cases}\boldsymbol{I}_{\mathrm{L}}^{\mathrm{wrap}}=F_{\mathrm{STN}}[\boldsymbol{I}_{\mathrm{R}},F_{\mathrm{grid}}(\boldsymbol{D}_{\mathrm{L}})]\\ \boldsymbol{I}_{\mathrm{R}}^{\mathrm{wrap}}=F_{\mathrm{STN}}[\boldsymbol{I}_{\mathrm{L}},F_{\mathrm{grid}}(\boldsymbol{D}_{\mathrm{R}})]\\ \boldsymbol{D}_{\mathrm{L}}^{\mathrm{wrap}}=F_{\mathrm{STN}}[\boldsymbol{D}_{\mathrm{R}},F_{\mathrm{grid}}(\boldsymbol{D}_{\mathrm{L}})]\\ \boldsymbol{D}_{\mathrm{R}}^{\mathrm{wrap}}=F_{\mathrm{STN}}[\boldsymbol{D}_{\mathrm{L}},F_{\mathrm{grid}}(\boldsymbol{D}_{\mathrm{R}})]\end{cases},\tag{3}$$

式中：F_{STN} 为 STN 图像重构函数；F_{grid} 表示根据视差生成重采样网格的函数；上标 wrap 表示使用 STN 得到的对应矩阵变量的重构。

在没有任何遮挡的情况下，$\boldsymbol{I}_{\mathrm{L}}^{\mathrm{wrap}}$，$\boldsymbol{I}_{\mathrm{R}}^{\mathrm{wrap}}$ 和 $\boldsymbol{I}_{\mathrm{L}}$，$\boldsymbol{I}_{\mathrm{R}}$ 应当相同，以它们之间的差异、平滑性约束和左右一致性可构建损失函数：

$$C=F_{\mathrm{C}}(\boldsymbol{I}_{\mathrm{L}}^{\mathrm{wrap}},\boldsymbol{I}_{\mathrm{L}},\boldsymbol{I}_{\mathrm{R}}^{\mathrm{wrap}},\boldsymbol{I}_{\mathrm{R}},\boldsymbol{D}_{\mathrm{L}},\boldsymbol{D}_{\mathrm{L}}^{\mathrm{wrap}},\boldsymbol{D}_{\mathrm{R}},\boldsymbol{D}_{\mathrm{R}}^{\mathrm{wrap}}),\tag{4}$$

式中 F_{C} 表示损失函数。

3　损失函数

对于自监督立体匹配算法，如何构建驱动模型学习的损失函数是问题的关键，定义损失函数为

$$C=w_{\mathrm{ap}}(C_{\mathrm{ap}}^{\mathrm{L}}+C_{\mathrm{ap}}^{\mathrm{R}})+w_{\mathrm{ds}}(C_{\mathrm{ds}}^{\mathrm{L}}+C_{\mathrm{ds}}^{\mathrm{R}})+w_{\mathrm{lr}}(C_{\mathrm{lr}}^{\mathrm{L}}+C_{\mathrm{lr}}^{\mathrm{R}}),\tag{5}$$

式中：w 表示权重，用于权衡不同类型损失的重要性；下标 ap、ds 和 lr 分别为图像重构损失、平滑性约束损失和左右一致性约束损失。

3.1　图像重构损失

结合图像的绝对误差和结构相似度，定义基于图像颜色的重构损失函数为

$$C_{\mathrm{ap}}=F_{\mathrm{avg}}\left[\alpha\frac{1-F_{\mathrm{SSIM}}(\boldsymbol{I}^{\mathrm{wrap}},\boldsymbol{I})}{2}+(1-\alpha)(|\boldsymbol{I}^{\mathrm{wrap}}-\boldsymbol{I}|+|\nabla\boldsymbol{I}^{\mathrm{wrap}}-\nabla\boldsymbol{I}|)\right],\tag{6}$$

式中：F_{avg} 为对矩阵中所有元素求均值的函数；F_{SSIM}[32] 为图像的结构相似度函数；|·| 表示对矩阵中各个元素求绝对值；∇ 表示对图像在空间位置进行一阶差分；α 用来权衡两种图像重构损失的作用，$\alpha=0.85$。

由于被遮挡的范围不但不能形成有效的反馈，而且会产生模型学习的噪声，因此这些区域应当被排除，设双目图像的共同视域由彩色图像的像素点在共同视域内的概率矩阵 $\boldsymbol{M}_{\mathrm{ap}}$ 决定，则结合共同视域的基于图像颜色的损失函数应为

$$C_{\mathrm{ap}}^{\mathrm{M}}=F_{\mathrm{avg}}(\tilde{\boldsymbol{C}}_{\mathrm{ap}}\cdot\boldsymbol{M}_{\mathrm{ap}}),\tag{7}$$

式中 $\tilde{\boldsymbol{C}}_{\mathrm{ap}}$ 表示(6)式中求均值前的矩阵，上标 M 表示只在共同视域内计算相应变量。

3.2　平滑性约束损失

在无纹理或弱纹理的区域，多种视差的图像重构误差都很小，通常这些区域内部的视差变化很小，根据视差的这种平滑性约束构造相应的损失函数，可使驱动模型学习更加平滑的视差预测，一定程度上克服纹理一致情况下的视差多值性。

相对双目相机的焦平面，相机视域内任意局部平面的深度值随像素位置线性变化，深度值对像素位置的二阶差分为 0，在深度值突变的位置则为较大的模值。假设视差预测准确，根据这种特性可以构建自适应的平滑性约束。

像素点的视差和深度值满足

$$d(p)\times Z(p)=B\times f,\tag{8}$$

式中：p 为图像中一点的坐标；$d(p)$ 为 p 点处的视差值；$Z(p)$ 为 p 点处的深度值；B 为双目相机的基线长度；f 为相机焦距。

对于图像上任意一点 p 都满足 $d(p)>0$，$Z(p)>0$，若 p 点附近为平面，则

$$\frac{Z(p+1)}{Z(p)}+\frac{Z(p-1)}{Z(p)}-2=\frac{d(p)}{d(p+1)}+\frac{d(p)}{d(p-1)}-2=0,\tag{9}$$

式中：$p+1$ 表示点坐标在二维平面内沿某个方向变化 1 个单位，$p-1$ 表示点坐标沿与 $p+1$ 的相反方向变化 1 个单位。

定义基于平面约束的平滑性约束损失为

$$C_{\mathrm{ds}}^{\mathrm{p}}(p)=\left|\frac{d(p)}{d(p+1)}+\frac{d(p)}{d(p-1)}-2\right|=0,\tag{10}$$

式中 $C_{\mathrm{ds}}^{\mathrm{p}}(p)$ 为 p 点处基于平面约束的平滑性约束损失值。

通常，在视差变化较大的边缘，图像纹理的变化也比较明显。因此，定义自适应的平滑性约束损失

值为

$$C_{\mathrm{ds}}(p)=C_{\mathrm{ds}}^{\mathrm{p}}(p)\cdot\exp\left[-\beta\frac{|\nabla\boldsymbol{I}(p)|}{F_{\mathrm{avg}}(|\nabla\boldsymbol{I}|)}\right],\tag{11}$$

式中：$C_{\mathrm{ds}}(p)$为 p 点处自适应的平滑性约束损失值；$\nabla\boldsymbol{I}(p)$为$\nabla\boldsymbol{I}$ 在 p 点处的 RGB 均值；β 为平滑抑制系数，表示图像纹理边缘对视差平滑性的抑制强度，β 值越大则抑制强度越大，实验中取 $\beta=2$。

整幅图像的平滑性约束损失值为

$$C_{\mathrm{ds}}=F_{\mathrm{avg}}(\tilde{\boldsymbol{C}}_{\mathrm{ds}}),\tag{12}$$

(12)式中矩阵 $\tilde{\boldsymbol{C}}_{\mathrm{ds}}$ 中 p 点处的元素值可通过(11)式计算得到。

3.3　左右一致性损失

左右一致性检查常用于立体匹配算法的后处理步骤，从而得到更加合理的视差预测，这种一致性关系也能用于改善自监督立体匹配方法的性能，通过惩罚不一致性来引导模型得到更加合理的预测。常采用 L1 惩罚构造损失函数，对于左图像视差矩阵 $\boldsymbol{D}_{\mathrm{L}}$，$p$ 点处左右一致性损失可定义为

$$C_{\mathrm{lr}}^{\mathrm{L}}(p)=|d_{\mathrm{L}}(p)-d_{\mathrm{R}}[p+d_{\mathrm{L}}(p)]|,\tag{13}$$

式中：$d(p)$表示相应的视差图 $\boldsymbol{D}$ 在 p 点处的值；$p+d_{\mathrm{L}}(p)$表示 p 点处在图像宽度方向向右平移 $d_{\mathrm{L}}(p)$个像素之后的坐标。$C_{\mathrm{lr}}^{\mathrm{L}}$可表示为

$$C_{\mathrm{lr}}^{\mathrm{L}}=F_{\mathrm{avg}}(|\boldsymbol{D}_{\mathrm{L}}-\boldsymbol{D}_{\mathrm{L}}^{\mathrm{wrap}}|)。\tag{14}$$

与图像重构误差相似，只有在双目图像的共同视域内才能满足这种左右一致性关系，结合共同视域的左右一致性损失为

$$C_{\mathrm{lr}}^{\mathrm{M}}=F_{\mathrm{avg}}(|\boldsymbol{D}-\boldsymbol{D}^{\mathrm{wrap}}|\cdot\boldsymbol{M}_{\mathrm{lr}}),\tag{15}$$

式中 $\boldsymbol{M}_{\mathrm{lr}}$表示预测的视差点在共同视域内的概率矩阵。

3.4　共同视域的确定

自监督立体匹配算法通过左右视角的相关性构建损失函数，但对于被遮挡的区域并没有这种相关性信息，在训练过程中若不能排除被遮挡区域，这些错误反馈将对模型学习产生不利影响。

双目图像的左右视角图像都有各自视角特有的内容，如左视角图像的最左侧和右视角图像的最右侧，由于视场范围限制只能在一个视角内可见，因此无法形成相关性信息。对于双目图像的共同可视范围，若视差预测准确，左右图像预测的视差具有一致性，被遮挡的区域则不满足这种一致性，根据这种一致性约束可以对被遮挡区域进行检测。

假设模型预测的视差 $\boldsymbol{D}$ 准确，相应的重构图为 $\boldsymbol{D}^{\mathrm{wrap}}$，将无穷远区域的视差设为 $\delta>0$，若不在共同视角的整体范围，重构图像相应位置必然为 0，且其他位置均不为 0，因此，共同视角的整体范围为

$$\boldsymbol{M}_{\mathrm{lr}}^{\mathrm{A}}=(\boldsymbol{D}^{\mathrm{wrap}}>0),\tag{16}$$

式中：$\boldsymbol{M}_{\mathrm{lr}}^{\mathrm{A}}$表示共同视域的整体范围在相应视图的区域模板，矩阵中元素用 1 和 0 分别表示相应的视差点在共同视域内和不在共同视域内。

根据视差的左右一致性，共同视角内一点 p 是否被周围物体遮挡可以概率的形式表示为

$$m_{\mathrm{O}}(p)=\begin{cases}0.98 & \Delta d(p)\geqslant 5\\ 0.245\times[\Delta d(p)-1] & 1\leqslant\Delta d(p)<5\\ 0 & \Delta d(p)<1\end{cases},\tag{17}$$

式中：$m_{\mathrm{O}}(p)$表示点 p 被周围物体遮挡的概率，下标 o 是被周围物体遮挡的模板标记，$\Delta d(p)$是矩阵 $|\boldsymbol{D}^{\mathrm{wrap}}-\boldsymbol{D}|$在点 p 处的元素值。

对于图像的左右一致性损失，共同视域在相应视图的区域以概率的形式表示为

$$\boldsymbol{M}_{\mathrm{lr}}=\boldsymbol{1}-F_{\mathrm{clip}}\{[(\boldsymbol{1}-\boldsymbol{M}_{\mathrm{lr}}^{\mathrm{A}})+\boldsymbol{M}_{\mathrm{O}}],0,1\},\tag{18}$$

式中矩阵 $\boldsymbol{M}_{\mathrm{O}}$ 中 p 点处的元素值根据(17)式计算，$F_{\mathrm{clip}}(\cdot,0,1)$表示将矩阵中所有元素值限制到范围[0,1]。

为使视差预测模型在左边缘能够得到有效的反馈信号，在重构左图像时，使用原始的右图像，而不是已经裁剪的右图像。这样，对图像进行重构时在左边缘位置也能得到有效反馈，从而提高视差预测模型在左图像最左侧的预测效果。对于图像的重构损失，共同视域在相应视图的区域为

$$\boldsymbol{M}_{\mathrm{ap}}=\boldsymbol{1}-F_{\mathrm{clip}}\{[(\boldsymbol{1}-\boldsymbol{M}_{\mathrm{ap}}^{\mathrm{A}})+\boldsymbol{M}_{\mathrm{o}}],0,1\},\tag{19}$$

式中 $\boldsymbol{M}_{\mathrm{ap}}^{\mathrm{A}}$ 中的元素 $M_{\mathrm{ap}}^{\mathrm{A}}(p)=F_{\mathrm{or-RGB}}(\boldsymbol{I}^{\mathrm{wrap}}>0)$，$F_{\mathrm{or-RGB}}$表示 $\boldsymbol{I}^{\mathrm{wrap}}$的 RGB 值任意一个满足判断条件输出即为 1，否则为 0。

4　实　　验

为验证算法的有效性，使用 Scene Flow 数据集[19] 和 KITTI 数据集[33-34] 对算法进行评价。Scene Flow 数据集是仿真环境中生成的数据集，包含 35454 对训练图像和 4370 对测试图像，为训练一个没有过拟合的模型，数据量是足够大的。KITTI 数据集是在不同天气条件下对真实场景记录的数据集，包含 KITTI2012 和 KITTI2015 两个子集，前者包含 194 个训练图像对和 195 个测试图像对，后者包含 200 个训练图像对和 200 个测试图像对。

在实际应用中，需要综合考虑算法精度、运行效率和资源消耗。KITTI 数据集为各种算法的性能

提供了统一的对比，但运行各种算法使用的硬件、软件平台各不相同，并不能对算法的运行效率和资源消耗进行直接对比。在 Pytorch 上实现 5 种视差预测模型，源代码在 Github 上托管[35]，在 Quadro K6000 显卡上对它们的运行效率和 GPU 占用显存进行对比，结果如表 1 和图 2 所示。两组数据均为测试阶段的数据，训练阶段通常会消耗更多资源，其中，表 1 是对 100 对不同测试数据的平均运行时间，图 2 是将图片高度设为 256 pixel 的情况下改变图片宽度时 GPU 显存占用情况。

表 1　不同视差预测模型的运行时间

Table 1　Running time of different disparity prediction models

Image size /pixel×pixel	Running time /s				
	DispNet[19]	DispNetCorr[19]	GC-Net[22]	iResNet[21]	PSMnet[24]
375×1242	0.074	0.100	6.848	0.375	3.462
480×960	0.074	0.097	6.527	0.339	3.397

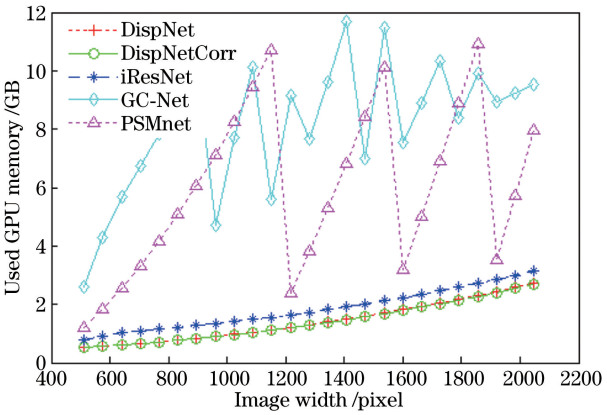

图 2　不同视差预测模型 GPU 显存占用情况

Fig. 2　GPU memory consumption of different disparity prediction models

综合考虑视差预测模型的精度、速度和资源消耗，实验选用 DispNetCorr[19] 作为视差预测模型。训练或微调过程中，均使用 Adam[36] 方法进行优化，批大小取 4，图片分辨率取 256 pixel × 512 pixel。数据增强包括空间变换和色彩变换，空间变换包括裁剪和缩放，色彩变换包括颜色、对比度和亮度变换。

在自监督立体匹配方法中，分别使用文献[27]和文献[31]的损失函数为基准与所提方法进行对比，其中损失函数中三种损失的权重初始值 $w_{ap}=1$，$w_{ds}=0.001$，$w_{lr}=0.001$，训练过程中动态调整 w_{ds} 和 w_{lr}，并使用相同的调整函数：

$$w=0.001+0.5\times\max(0,S_{IM}-0.75),\quad(20)$$

式中：$S_{IM}=F_{avg}[F_{SSIM}(\boldsymbol{I}^{wrap},\boldsymbol{I})]$，表示观察图像与重构图像的结构相似度的平均值；max 表示求两个变量的最大值。

使用的主要评价指标包括 EPE（end-point error）、D1 和 RPE（reconstruction pixel error）。EPE 表示预测视差与实际测量值之间的误差；D1 表示每组图像对中第一帧的评价区域错误像素百分比，其中 EPE 小于 3 pixel 或 EPE 小于实际测量值 5%的认为是正确像素，否则为错误像素。RPE 表示重构图像的像素值与观测值之间的误差。

4.1　自改善能力分析

在没有视差数据作为监督信号的情况下，自监督立体匹配方法以图像重构误差为模型训练提供反馈信号，从而改善模型的性能。为分析这种自我改善能力，以 KITTI2015 训练集对随机初始化的模型进行训练，学习率取 1×10^{-4}，共训练 2000 周期，每个周期处理完整个数据集需要 50 次迭代。训练完成后使用这两个数据集进行性能评价的结果如表 2 所示，训练过程中以 KITTI2012 训练集进行性能评价的结果如图 3 所示，其中对训练损失进行了归一化处理，-M 表示对应算法只在共同视域内计算损失时的结果。

表 2　随机初始化模型训练完成后对不同数据集进行评价的结果对比

Table 2　Comparison of evaluation results on different datasets after training with randomly initializing model

Method	KITTI2012 training set			KITTI2015 training set		
	D1 /%	EPE /pixel	RPE /pixel	D1 /%	EPE /pixel	RPE /pixel
Method in Ref. [19]	14.78	2.37	13.91	4.99	0.91	13.40
Method in Ref. [27]	14.53	2.90	11.75	11.39	1.95	9.92
Method in Ref. [27]-M	10.40	2.71	11.38	6.60	1.46	9.46
Method in Ref. [31]	12.98	2.69	11.13	10.80	1.89	9.48
Method in Ref. [31]-M	10.27	2.73	11.24	7.55	1.60	9.35
Proposed method	10.13	2.72	11.37	6.78	1.64	9.64

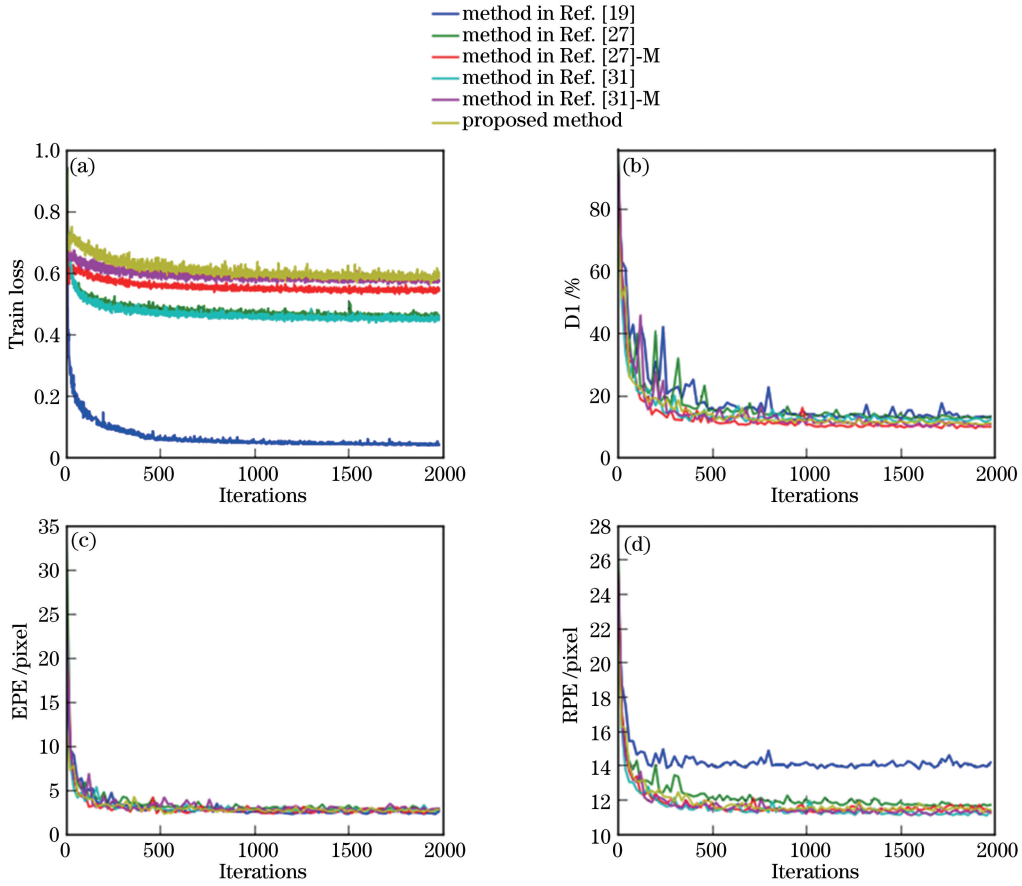

图 3　随机初始化模型训练过程中对 KITTI2012 训练集进行评价的结果对比。(a)训练损失;(b) D1;(c) EPE;(d) RPE

Fig. 3　Comparison of evaluation results on KITTI2012 training set with randomly initializing model during training process. (a) Training loss; (b) D1; (c) EPE; (d) RPE

从表 2 和图 3 可以看出:当训练数据集较小时,有监督学习立体匹配方法在训练集上获得了最佳性能,但并不能很好地适应新的数据集;与基准算法相比,所提算法可以明显改善自监督学习立体匹配方法的效果,在训练集上的 D1 评价指标降低了30%～42%,在验证集上的 D1 评价指标降低了21%～30%。

从图 3 也可以看出,无论采用哪种损失函数,最终模型训练都能收敛,并得到合理的视差图,说明了基于学习的立体匹配方法的有效性,同时也说明了自监督立体匹配方法具有驱动模型性能改善的能力。

4.2　微调能力分析

以预训练的模型作为初始值进行微调,能够分析不同损失函数对模型训练的精细调节能力。为使预训练的模型具有较高的准确度,使用所提方法在 Scene Flow 数据集上进行预训练,再将预训练的模型参数作为初始值,在 KITTI2015 训练集上进行微调,微调完成后使用不同数据集进行性能评价的结果如表 3 所示,训练过程中以 KITTI2012 训练集进行性能验证评价的结果如图4所示,各个参数含义

表 3　预训练模型微调后对不同数据集进行评价的结果对比

Table 3　Comparison of evaluation results on different datasets with pre-training model after fine tuning

Method	KITTI2012 training set			KITTI2015 training set		
	D1 /%	EPE /pixel	RPE /pixel	D1 /%	EPE /pixel	RPE /pixel
Pre-training	8.07	1.53	12.13	8.06	2.27	11.78
Method in Ref. [19]	5.71	1.31	12.74	2.15	0.69	12.75
Method in Ref. [27]	9.26	2.25	11.29	7.85	1.73	9.71
Method in Ref. [27]-M	7.29	1.70	10.84	5.91	1.31	9.46
Method in Ref. [31]	8.21	1.96	10.74	7.28	1.62	9.27
Method in Ref. [31]-M	7.30	1.59	10.57	5.95	1.33	9.21
Proposed method	6.86	1.61	10.64	5.96	1.35	9.31

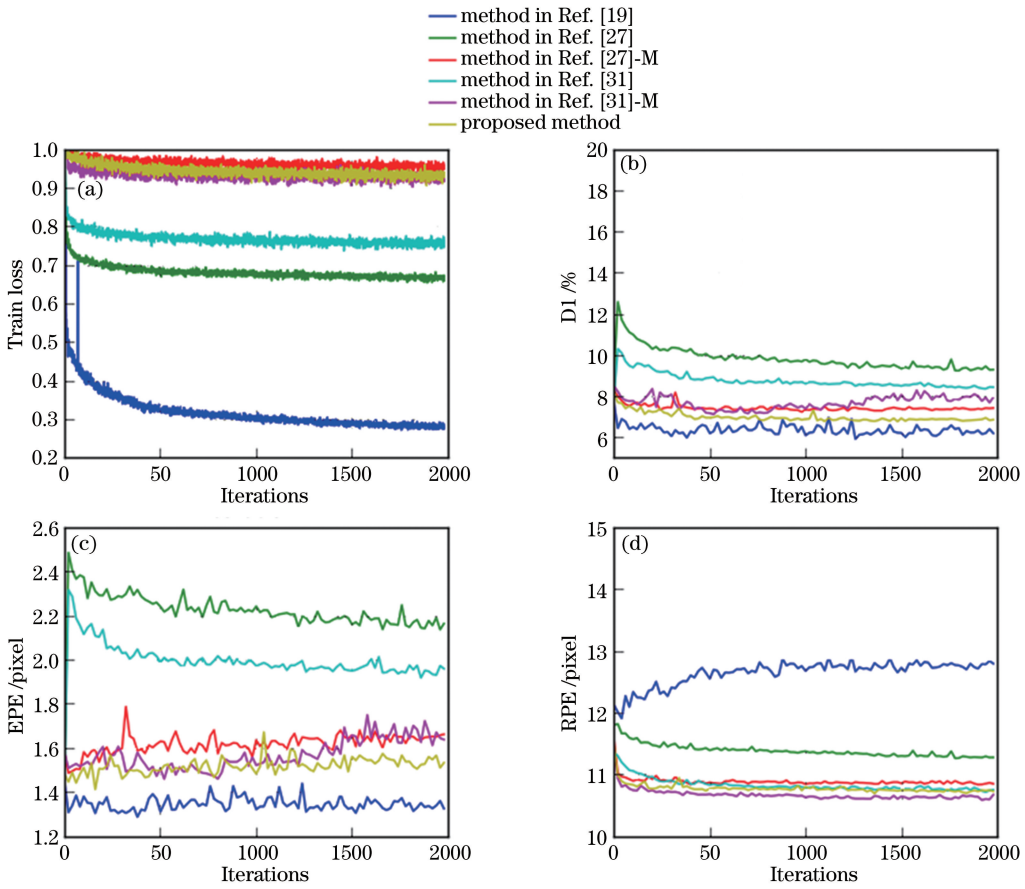

图 4　预训练模型微调过程中对 KITTI2012 训练集进行评价的结果对比。(a)训练损失;(b) D1;(c) EPE;(d) RPE

Fig. 4　Comparison of evaluation results on KITTI2012 training set with pre-training model during fine tuning process. (a) Training loss; (b) D1; (c) EPE; (d) RPE

与上一节相同。其中,在 Scene Flow 数据集上预训练时,初始学习率取 1×10^{-4},从第 40 个周期开始,每 20 个周期学习率减半,共训练 120 个周期。在 KITTI 数据集上微调时,初始学习率取 2×10^{-5},训练 400 个周期,调整学习率为 1×10^{-5} 再训练 1600 个周期。

从表 3 和图 4 中也可以看出,在视差预测的精度上,所提方法能够明显提高自监督学习方法的性能,在训练集上的 D1 评价指标降低了 18%～25%,在验证集上的 D1 评价指标降低了 11%～26%。由于自监督学习方法本质上无法对全部区域产生有效的反馈监督信号,与有监督学习方法的性能相比仍然有一定差距,但它不依赖标签数据,可以大大减小标签数据获取的成本和工作量。

同时也可以看出,对于有监督的立体匹配方法,重构图像的像素误差明显大于自监督的方法,这说明在 KITTI 数据集上视差预测越准确,重构误差不一定越小,产生这种现象的主要原因有两方面。首先,KITTI 数据集并没有提供全部像素的观测视差,在没有监督信号的区域,视差的预测误差可能会更大。其次,在产生遮挡的区域,使用不准确的视差得到的重构图像的像素误差可能会更小。由于自监督立体匹配方法将重构图像的像素误差作为反馈信号,因此第二种原因是限制其性能的主要因素。

对视差预测结果进行定性的对比分析,可以直观地分析模型的优点和缺陷,有助于进一步改进模型、提高算法性能。这里选择 KITTI2015 训练集中的两组图像进行对比分析,结果如图 5 和图 6 所示。图 5(a)、(c)为左视角图像,图 5(b)、(d)为视差的预测值。图 6(a)、(c)为不同方法预测视差的误差图,图 6(b)、(d)为不同方法的预测值。可以看出,汽车玻璃所在的区域,具有较强的反光,使得纹理很弱,自监督立体匹配方法的视差预测效果较差;细长的杆状物体所在的区域,纹理较明显的区域能够预测较准确的视差,但会使周围弱纹理的区域的误差增大,纹理不明显的区域将被误认为是背景。产生这种现象的主要原因是:自监督立体匹配方法的主要反馈信号就是图像的纹理,当纹理信号弱甚至是不

图 5　KITTI2015 训练集中的两组图像。(a)第一组图像的左视角图像;(b)第一组图像的视差预测值;(c)第二组图像的左视角图像;(d)第二组图像的视差预测值

Fig. 5　Two sets of images selected from KITTI2015 training set. (a) Left view image in first set. (b) predicted disparity in first set. (c) left view image in second set. (d) predicted disparity in second set

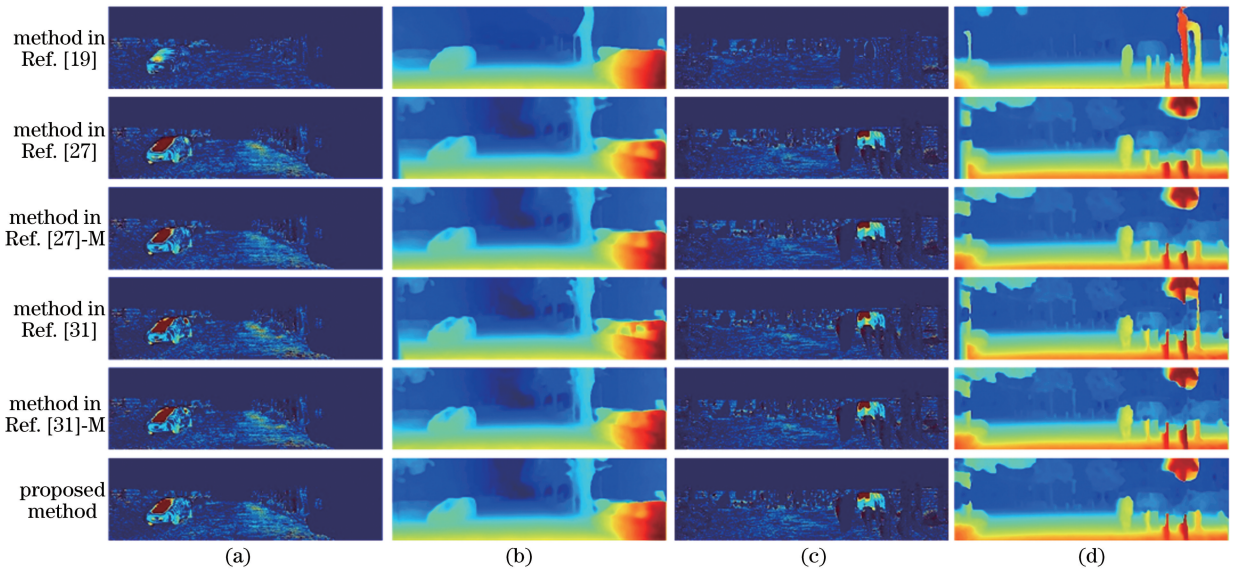

图 6　预训练模型微调后的实际应用效果对比。(a)第一组图像的误差图;(b)第一组图像的视差预测值;(c)第二组图像的误差图;(d)第二组图像的视差预测值

Fig. 6　Comparison of practical application effects with pre-training model after fine tuning. (a) Error map in first set; (b) predicted disparity in first set; (c) error map in second set; (d) predicted disparity in second set

可用时就无法学习到准确的视差,需要其他有效的约束条件来改善最终结果。

5　结　　论

提出了一种基于共同视域的自监督立体匹配算法,根据视差的左右一致性特点来判断遮挡区域,在损失函数中去除遮挡区域,减少了遮挡区域对模型训练产生的噪声。实验结果表明所提方法不需要任何标签数据就能达到接近有监督立体匹配方法的性能。但对于纹理弱的区域、具有较强反光和细长杆状物的区域,该算法无法得到有效的训练反馈信息,仍需要进一步的研究。

参　考　文　献

[1] Liu C, Yuen J, Torralba A. SIFT flow: dense correspondence across scenes and its applications[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(5): 978-994.

[2] Liu Y F, Cai Z J. Binocular stereo vision three-dimensional reconstruction algorithm based on ICP and SFM[J]. Laser and Optoelectronics Progress, 2018, 55(9): 091503.
刘一凡, 蔡振江. 基于 ICP 与 SFM 的双目立体视觉三维重构算法[J]. 激光与光电子学进展, 2018, 55(9): 091503.

[3] Sivaraman S, Trivedi M M. A review of recent developments in vision-based vehicle detection[C]. IEEE Intelligent Vehicles Symposium (IV), 2013: 310-315.

[4] Schmid K, Tomic T, Ruess F, *et al*. Stereo vision based indoor/outdoor navigation for flying robots[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems, 2013: 3955-3962.

[5] Scharstein D, Szeliski R, Zabih R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms[J]. International Journal of Computer Vision, 2002, 47(1): 7-42.

[6] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[7] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]. IEEE Computer Society

Conference on Computer Vision and Pattern Recognition (CVPR), 2005: 886-893.

[8] Hirschmuller H. Stereo processing by semiglobal matching and mutual information [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(2): 328-341.

[9] Liu D W, Han L, Han X Y. High spatial resolution remote sensing image classification based on deep learning [J]. Acta Optica Sinica, 2016, 36(4): 0428001.
刘大伟, 韩玲, 韩晓勇. 基于深度学习的高分辨率遥感影像分类研究[J]. 光学学报, 2016, 36(4): 0428001.

[10] Hou Y Q Y, Quan J C, Wei Y M. Valid aircraft detection system for remote sensing images based on cognitive models[J]. Acta Optica Sinica, 2018, 38(1): 0111005.
侯宇青阳, 全吉成, 魏勇明. 基于认知模型的遥感图像有效飞机检测系统[J]. 光学学报, 2018, 38(1): 0111005.

[11] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651.

[12] Žbontar J, LeCun Y. Computing the stereo matching cost with a convolutional neural network[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 1592-1599.

[13] Žbontar J, LeCun Y. Stereo matching by training a convolutional neural network to compare image patches[J]. Journal of Machine Learning Research, 2016, 17(1): 2287-2318.

[14] Xiao J S, Tian H, Zou W T, *et al*. Stereo matching based on convolutional neural network [J]. Acta Optica Sinica, 2018, 38(8): 0815017.
肖进胜, 田红, 邹文涛, 等. 基于深度卷积神经网络的双目立体视觉匹配算法[J]. 光学学报, 2018, 38(8): 0815017.

[15] Chen Z Y, Sun X, Wang L, *et al*. A deep visual correspondence embedding model for stereo matching costs [C]. IEEE International Conference on Computer Vision, 2015: 972-980.

[16] Park H, Lee K M. Look wider to match image patches with convolutional neural networks[J]. IEEE Signal Processing Letters, 2017, 24(12): 1788-1792.

[17] Güney F, Geiger A. Displets: resolving stereo ambiguities using object knowledge [C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 4165-4175.

[18] Seki A, Pollefeys M. Patch based confidence prediction for dense disparity map [C]. British Machine Vision Conference (BMVC), 2016: 23.

[19] Mayer N, Ilg E, Häusser P, *et al*. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 4040-4048.

[20] Pang J H, Sun W X, Ren J S, *et al*. Cascade residual learning: a two-stage convolutional neural network for stereo matching[C]. IEEE International Conference on Computer Vision Workshops, 2017: 887-895.

[21] Liang Z F, Feng Y L, Guo Y L, *et al*. Learning deep correspondence through prior and posterior feature constancy[EB/OL]. (2017-12-04)[2018-06-05]. http://cn.arxiv.org/abs/1712.01039.

[22] Kendall A, Martirosyan H, Dasgupta S, *et al*. End-to-end learning of geometry and context for deep stereo regression[C]. IEEE International Conference on Computer Vision, 2017: 66-75.

[23] Yu L D, Wang Y C, Wu Y W, *et al*. Deep stereo matching with explicit cost aggregation sub-architecture[EB/OL]. (2018-01-12)[2018-06-05]. http://cn.arxiv.org/abs/1801.04065.

[24] Chang J R, Chen Y S. Pyramid stereo matching network [EB/OL]. (2018-03-23) [2018-06-05]. http://cn.arxiv.org/abs/1803.08669.

[25] Xie J Y, Girshick R, Farhadi A. Deep 3D: fully automatic 2D-to-3D video conversion with deep convolutional neural networks [C]. European Conference on Computer Vision(ECCV), 2016: 842-857.

[26] Garg R, Vijay K B G, Carneiro G, *et al*. Unsupervised CNN for single view depth estimation: Geometry to the rescue[C]. European Conference on Computer Vision (ECCV), 2016: 740-756.

[27] Godard C, Aodha O M, Brostow G J. Unsupervised monocular depth estimation with left-right consistency [C]. IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2017: 6602-6611.

[28] Jaderberg M, Simonyan K, Zisserman A, *et al*. Spatial transformer networks[EB/OL]. (2015-06-05) [2018-06-05]. http://cn.arxiv.org/abs/1506.02025.

[29] Ye M L, Johns E, Handa A, *et al*. Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery[EB/OL]. (2017-05-17) [2018-06-05]. http://cn.arxiv.org/abs/1705.08260.

[30] LiuX, Sinha A, Unberath M, *et al*. Self-supervised

learning for dense depth estimation in monocular endoscopy[C]. European Conference on Computer Vision (ECCV), 2018, 11041: 128-138.

[31] Zhong Y R, Dai Y C, Li H D. Self-supervised learning for stereo matching with self-improving ability [EB/OL]. (2017-09-04) [2018-06-05]. http://cn.arxiv.org/abs/1709.00930.

[32] 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.

[33] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]. IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2012: 3354-3361.

[34] Menze M, Geiger A. Object scene flow for autonomous vehicles [C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 3061-3070.

[35] wyf2017: Pytorch implementation of the several deep stereo matching network [EB/OL]. https://github.com/wyf2017/DSMnet.

[36] Kingma D P, Ba J. Adam: a method for stochastic optimization[EB/OL]. (2014-12-22)[2018-06-05]. http://cn.arxiv.org/abs/1412.6980.