文章编号：1003-0530(2020)09-1450-07

# 基于自注意力机制的多阶段无监督单目深度估计网络

刘香凝[1,2] 赵 洋[2,3] 王荣刚[1,2]

（1. 北京大学深圳研究生院信息工程学院，广东深圳 518055；2. 鹏城实验室，广东深圳 518055；
3. 合肥工业大学计算机与信息学院，安徽合肥 230009）

**摘 要**：单幅图像的深度估计是场景几何理解过程中的一个重要步骤，但由于尺度模糊，也被计算机视觉领域普遍认为是一个典型的不适定问题。近年来，尽管监督学习方法在单目深度估计中取得了基本令人满意的效果，但需要对数据集进行大量真实深度值的标记，这是一项成本较高的工作。此外，由于物体的运动、遮挡、光照等常见问题，单目深度估计的表现并不尽如人意，尤其是在物体边缘和弱纹理区域。为了解决这些问题，本文提出了一种基于自注意力的多阶段无监督单目深度估计网络。该方法具有以下特点：1）多阶段网络结构对训练过程中的深度估计具有较强的约束和监督作用；2）通过掩模加权重构损失和左右视差一致性损失对网络进行优化；3）采用自注意力机制捕捉更多上下文信息，进而提升预测结果。实验结果表明，该方法在 KITTI 数据集上的深度估计效果达到甚至超过了已有方法。

**关键词**：无监督学习；单目深度估计；多阶段网络；自注意力

**中图分类号**：TP391 **文献标识码**：A **DOI**：10.16798/j.issn.1003-0530.2020.09.009

**引用格式**：刘香凝，赵洋，王荣刚. 基于自注意力机制的多阶段无监督单目深度估计网络[J]. 信号处理，2020，36(9)：1450-1456. DOI：10.16798/j.issn.1003-0530.2020.09.009.

**Reference format**：Liu Xiangning，Zhao Yang，Wang Ronggang. Self-attention Based Multi-stage Network for Unsupervised Monocular Depth Estimation[J]. Journal of Signal Processing，2020，36(9)：1450-1456. DOI：10.16798/j.issn.1003-0530.2020.09.009.

# Self-attention Based Multi-stage Network for Unsupervised Monocular Depth Estimation

Liu Xiangning[1,2] Zhao Yang[2,3] Wang Ronggang[1,2]

（1. School of Electronic and Computer Engineering，Shenzhen Graduate School，Peking University，Shenzhen，Guangdong 518055，China；2. Peng Cheng Laboratory，Shenzhen，Guangdong 518055，China；3. The School of Computer and Information，Hefei University of Technology，Hefei，Anhui 230009，China）

**Abstract**：Monocular depth estimation is an important but ill-posed procedure in the process of scene geometry understanding. Though recent supervised learning methods have achieved promising results for monocular depth estimation，they require vast amounts of ground truth depth data which is a costly task. Besides，previous works suffer from well-known problems such as moving objects，occlusions and lighting，which result in unsatisfactory performance，particularly in object edges and low-texture regions. To tackle these problems，we propose a self-attention based multi-stage network for unsupervised monocular depth estimation. Our method incorporates the following features：1）multi-stage network provides stronger constraint and supervision for depth estimation during training；2）the network is optimized with mask weighted reconstruction loss and left-right disparity consistency loss；3）self-attention module is adopted to capture more context information. Experimental results on the KITTI dataset show that the method can obtain state-of-the-art performance，which means the proposed method can effectively improve the performance of monocular depth estimation.

**Key words**：unsupervised learning；monocular depth estimation；multi-stage network；self-attention

收稿日期：2020-07-04；修回日期：2020-07-27

基金项目：国家自然科学基金资助项目(61672063)；深圳市研究项目(JCYJ20180503182128089，201806080921419290)

## 1　引言

从图像中准确地估计深度是许多场景理解任务的基本问题，如场景分割、场景对象检索及视觉跟踪等。深度估计可以分为单目深度估计、双目深度估计以及多目深度估计。单目深度估计就是从单张图像中恢复出场景的深度信息，而双目深度估计和多目深度估计就是从两张或多张图片中求取深度。双目深度估计和多目深度估计可以通过输入图像的视角变化得到视差信息，从而得到深度。对人类来说，单目深度估计和双目深度估计一样并不困难，因为人类可以利用透视、相对于已知物体的比例以及视觉经验等线索获取深度信息。然而，在计算机视觉领域，由于一幅二维图像可能对应无限个三维场景，单目深度估计是一项比双目深度估计和多目深度估计更加困难的任务。

单目深度估计的研究由来已久，早期的工作通过几何约束和手工特征从图像中推断深度信息。Saxena 等[1]提出了 Make3D 模型，该模型将输入图像分割成超像素，然后通过马尔可夫随机场推断每个超像素三维表面的三维位置和方向。

近年来，基于深度学习的方法在单目深度估计方面取得了令人瞩目的成绩。Eigen 等[2-3]首先提出了一种多尺度深度神经网络（DNN），该方法利用额外的网络对原始网络的预测结果进行细化，它证明了基于深度神经网络的深度估计方法可以有效地提取局部和全局信息。基于该方法提出的改进方案包括：深度卷积神经场模型[4-5]；将深度估计视为像素级别的分类任务而非回归问题[6]；采用鲁棒性更强的损失函数如 Huber 损失[7]。Kendall 等[8]通过构造匹配代价卷来分析几何特征并采用 3D 卷积合并上下文信息。Chang 等[9]提出使用金字塔池化模块来捕捉更多的全局信息。虽然这些基于有监督学习的方法取得了很好的效果，但是需要对数据集进行大量的真实深度值标记，对于深度估计任务而言，深度值的标记工作是比较困难的。另外，由于有监督学习方法比较依赖真实的深度值数据，所以模型在新场景中的泛化能力很差。因此，近年来出现了一些用于单目深度估计的无监督框架。基于无监督学习的深度估计方法不需要对数据集进行真实深度值的标记，且对新场景的泛化能力更好。Garg 等[10]提出的网络从左图像中预测出一张视差图，然后利用该视差图和原始右图重建出合成左图，通过最小化原始左图和合成左图的光度误差训练网络。Godard 等[11]提出了左右视差一致性损失，从而保持单目深度估计的几何一致性。Pilzer 等[12-13]首先提出将循环网络结构应用于深度估计。

本文提出了一种新的单目深度估计算法——基于自注意力机制的多阶段网络。该方法是完全无监督的，即训练过程中无需标记出真实深度值。该方法将双目图像对中的左图作为网络输入进行训练和测试。Godard 等[11]提出的左右视差一致性损失函数有效地提升了深度图预测效果，因此网络也输出两张视差图用于计算左右视差一致性损失函数来加强两张预测视差图之间的一致性。利用两张预测视差图对双目图像对进行采样可以得到重建的左右两张图，用于计算左右两张图的重建损失函数以及对应的平滑度损失函数。值得注意的是，卷积网络的输入仅为单独的左图，而只有训练过程中需要使用右侧图像。整个网络结构由两个子网络构成，形成循环。正向子网络预测了两张视差图，重建的右图输入到反向子网络中预测两张新的视差图，重建出新的左右图像对，通过最小化新的重建图像对和原始图像的误差训练反向子网络。此外，该模型采用了自注意力机制，通过计算图像中任意两个位置之间的相互作用来捕获更多图像信息。网络利用左右视差一致性损失、平滑度损失以及掩模加权重建损失进行优化。综上所述，本文提出了以下几点创新：1）多阶段网络结构为单目深度估计提供了更强的约束和监督；2）自注意力特征提取器有利于提取更多的上下文信息；3）本文所提出的方法在 KITTI 数据集上获得了令人满意的预测效果。

## 2　主要方法

本文方法的详细架构如图 1 所示。立体图像对中的左图 $I_l$ 输入到正向子网络中预测两个视差图 $\{d_r, d_l\}$。值得注意的是，同时估计两个视差图是为了计算左右视差一致性损失作为网络的约束。因此，尽管训练过程中需要使用立体图像对来计算左右两个视角的损失函数，测试时仅需要输入单张图像就可以估计出深度。原始图像对 $\{I_l, I_r\}$ 基于预测视差图 $\{d_r, d_l\}$ 采样后重建出新的图像对 $\{\hat{I}_r, \hat{I}_l\}$，重建图像对被用于计算左右两个视角的重建损失等损失函数。重建右图 $\hat{I}_r$ 输入到反向子网络中预测出新的视差图 $\{d'_r, d'_l\}$。正向子网络得到的重建图像对 $\{\hat{I}_r, \hat{I}_l\}$ 基

于$\{d'_r,d'_l\}$采样后得到了新的重建图像对$\{\hat{I}'_r,\hat{I}'_l\}$。反向子网络通过最小化新的重建图像对$\{\hat{I}'_r,\hat{I}'_l\}$和原始输入图像对$\{I_l,I_r\}$的误差而得到优化。

如图 1 所示,整个网络经过了多阶段的训练。首先对正向子网络进行单独训练,直到正向子网络收敛,然后固定正向子网络的参数,再训练反向子网络。最后,联合训练整个网络直到收敛。测试时只使用正向子网络来估计单个图像的视差图。

下文分别介绍网络结构的细节,自注意力模块以及损失函数。

### 2.1 网络结构

如前所述,本文提出的网络结构包括正向子网络和反向子网络。正向子网络 $G_f$从输入左图 $I_l$中预测两个视差图$\{d_r,d_l\}$:

$$\{d_r,d_l\}=G_f(I_l) \tag{1}$$

其中 $d_r$和 $d_l$分别表示从左到右和从右到左的视差。原始输入图像对$\{I_l,I_r\}$基于预测视差$\{d_r,d_l\}$采样($f_w$)后合成新的图像对$\{\hat{I}_r,\hat{I}_l\}$:

$$\hat{I}_r=f_w(d_r,I_l) \tag{2}$$

$$\hat{I}_l=f_w(d_l,I_r) \tag{3}$$

合成图像对中的右图 $\hat{I}_r$输入到反向子网络 $G_b$中预测新的视差图$\{d'_r,d'_l\}$:

$$\{d'_l,d'_r\}=G_b(\hat{I}_r) \tag{4}$$

接下来用新的视差图$\{d'_l,d'_r\}$和合成图像对$\{\hat{I}_r,\hat{I}_l\}$重建出新的图像对$\{\hat{I}'_l,\hat{I}'_r\}$:

$$\hat{I}'_l=f_w(d'_l,\hat{I}_r) \tag{5}$$

$$\hat{I}'_r=f_w(d'_r,\hat{I}_l) \tag{6}$$

新的合成图像对$\{\hat{I}'_l,\hat{I}'_r\}$和原始图像对$\{I_l,I_r\}$之间的差异被用来优化反向子网络。

测试时不需要反向子网络。单张图像输入到正向子网络中可以直接预测出视差图。两个子网络均采用编码器-解码器结构。如图 1 所示,本文采用 Densenet169[14] 作为编码器结构,包括一个卷积层、一个池化层以及四个下采样模块,每个下采样模块由多个 DenseBlock(DB)和一个卷积层组成。解码器网络结构由一系列上采样层组成,每个上采样层后面连接一个卷积层。在上采样过程中,网络会预测出四个不同尺度的视差图。此外,编码器和解码器对应尺度之间的跳跃连接可以一定程度上保留图像的高分辨率细节。

在训练过程中,正向子网络从原始输入图像中学习视差,而反向子网络从合成图像中学习视差。

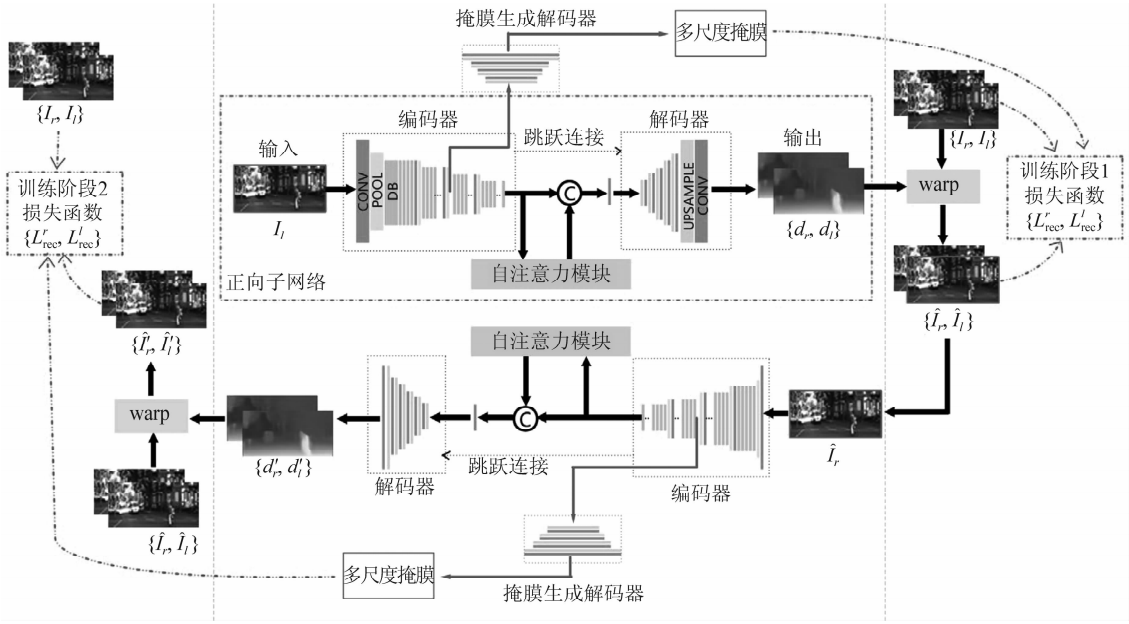


图 1 本文方法的详细框架(参见 2)。Warp 是用来合成左右视图图像的重建函数(参见 2.1)。自注意力模块是 2.2 中介绍的自注意力特征提取器。掩膜生成解码器用于产生多尺度掩模,并计算掩模加权重建损失,详见 2.3

Fig. 1 Framework overview of proposed method (see Sec 2). Warp is the warping operation to synthesize left and right view images (see Sec 2.1). Self-attention denotes the self-attention feature extractor introduced in Sec 2.2. Mask decoder is used to produce multi-scale masks and compute mask weighted reconstruction loss detailed in Sec 2.3

训练反向子网络时，为了最小化重新合成的图像与原始输入图像之间的差异，反向子网络的输入应当是准确的图像对。因此，反向子网络对正向子网络的深度估计结果有很强的约束。与文献[12]中提出的用于双目深度估计的网络结构不同，本文方法用于单目深度估计，即两个子网络的输入都是单幅图像。

## 2.2　自注意力特征提取器

近年来，深度学习与注意力机制相结合的研究工作在很多领域得到了广泛应用[15]。在计算机视觉领域，注意力机制可以在提取特征的过程中获取更多的全局信息。本文方法中两个子网络都采用了自注意力特征提取器。自注意力机制[16-17]的作用是通过保证相似像素具有相似的深度特征来提高弱纹理区域的效果。自注意力模块被连接到编码器结构的最后一层，生成自注意力特征图，并将其与之前的特征图连接起来。

本方法所采用的自注意力模块的细节如图 2 所示。首先，将给定的特征图 $x \in \mathbb{R}^{C\times N}$ 输入两个 1×1 卷积层中，得到两个新的特征图 $\{\boldsymbol{Q},\boldsymbol{K}\} \in \mathbb{R}^{C\times N}$，其中 $C$ 表示图像通道数，$N=H\times W$ 表示特征图的大小。然后在矩阵 $\boldsymbol{Q}$ 与 $\boldsymbol{K}^{\mathrm{T}}$ 之间进行矩阵乘法运算，再对上述结果的每一行进行 Softmax 激活，得到注意力图 $\boldsymbol{S}$：

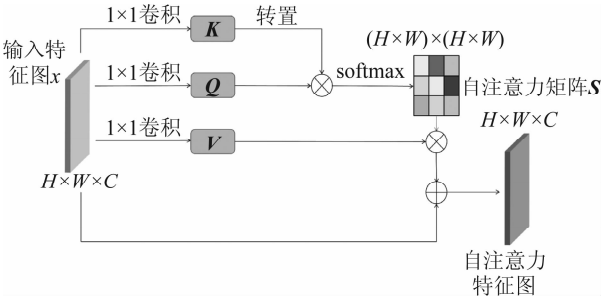


图 2　自注意力模块实现细节

Fig. 2　Details of the self-attention module

$$\beta_{j,i}=\frac{\exp(s_{i,j})}{\sum_{i=1}^{N}\exp(s_{i,j})},\ s_{i,j}=(\boldsymbol{K}_j)^{\mathrm{T}}\cdot\boldsymbol{Q}_i \tag{7}$$

$\beta_{j,i}$ 代表位置 $i$ 对位置 $j$ 的影响。其中，相关性更强的两个位置的特征表示的相似性更高。

与此同时，将特征图 $x$ 输入另一个 1×1 卷积层，生成新的特征图 $\boldsymbol{V}\in\mathbb{R}^{C\times N}$，然后对 $\boldsymbol{S}$ 与 $\boldsymbol{V}$ 做矩阵乘法运算。最后得到的自注意力模块的输出如下：

$$y_i=\alpha\sum_{i=1}^{N}(\beta_{j,i}\boldsymbol{V}_i)+x_j \tag{8}$$

尺度参数 $\alpha$ 初始化为 0，在训练过程中逐渐分配更多的权重。如公式(8)所示，输出的特征图是自注意力特征图和之前的特征图的叠加。

自注意力模块通过计算每个位置与其他所有位置之间的相似度，并将具有相似特征或外观的像素聚合在一起，保证了相似像素具有相似的深度特征，从而提高了弱纹理区域的预测效果。值得注意的是，自注意力模块被连接到编码器的最后一层，因此它在增强了特征表示的基础上并没有大幅提升计算复杂度。

## 2.3　损失函数

本文的方法不使用任何真实深度值数据，因此主要由重建损失进行监督。本文的单目深度估计框架所作的假设是：场景中没有运动的物体；左右视图之间没有遮挡关系。然而，在真实场景中，这些假设并不成立，因此训练过程可能会受到一些负面的影响。受到文献[18]的启发，本文采用一个额外的网络来估计掩模，作为重建损失的权重。如图 1 所示，掩模估计网络与每个子网络共享部分编码器权重，其解码器部分包含与主网络结构的编码器类似的四个上采样层和卷积层。掩膜生成网络预测出四个不同尺度的像素级别的掩膜 $\hat{\boldsymbol{M}}$ 作为重建损失函数的权重来提高网络的鲁棒性。

掩模加权重构损失采用了 L1 和 SSIM 组合的形式。对于正向子网络，该重建损失最小化 $\{\hat{I}_r,\hat{I}_l\}$ 和 $\{I_l,I_r\}$ 之间的差异，而对于反向子网络，该重建损失最小化 $\{\hat{I}'_l,\hat{I}'_r\}$ 和 $\{I_l,I_r\}$ 之间的差异。网络估计了四种不同尺度下的视差和掩模，四种尺度下的损失之和为最终损失，掩膜加权重建损失的计算方法如下：

$$L^l_{\mathrm{rec}}=\sum_{n=0}^{4}\left(\lambda\frac{1-\mathrm{SSIM}(I_l,\hat{I}_l)}{2}+(1-\lambda)\hat{M}\|I_l-\hat{I}_l\|\right)_n \tag{9}$$

设置 $\lambda$ 为 0.85。

此外，该方法引入平滑度损失，平滑度损失通过参考输入图像的梯度使视差图更加平滑，平滑度损失的计算方法如下：

$$L^l_{\mathrm{smooth}}=\sum_{n=0}^{4}\left(|\partial_x d_l|\cdot e^{-\|\partial_x I_l\|}+|\partial_y d_l|\cdot e^{-\|\partial_y I_l\|}\right)_n \tag{10}$$

本文还采用了左右视差一致性损失[11]，该损失加强了两张视差图之间的一致性，有利于无监督网络的收敛：

$$L^l_{lr}=\sum_{n=0}^{4}\left(\frac{1}{N}\sum_p|d_l(p)-d_r(p+d_l(p))|\right)_n \tag{11}$$

其中 $p$ 表示视差图中的像素位置。公式(10)和公式(11)中,$d_l$ 和 $d_r$ 用于训练正向子网络,而对于反向子网络,这些损失函数中应使用 $d_l'$ 和 $d_r'$。

完整的训练损失函数包括掩模加权重建损失、平滑度损失和左右视差一致性损失的组合。所有的损失函数均包含左视角和右视角两项:

$$L=\alpha_{\text{rec}}(L_{\text{rec}}^{l}+L_{\text{rec}}^{r})+\alpha_{\text{smooth}}(L_{\text{smooth}}^{l}+L_{\text{smooth}}^{r})+\alpha_{lr}(L_{lr}^{l}+L_{lr}^{r}) \tag{12}$$

## 3 实验过程及结果

KITTI 数据集[19]是一个用于深度估计的大规模公开数据集,本文算法主要在该数据集上进行评估。另外,Cityscapes 数据集[20]用于泛化能力的评估。

### 3.1 实验设置

#### 3.1.1 数据集

本文算法在 KITTI 数据集上进行训练和测试,使用 Eigen 等[3]提出的数据集分割方式。测试集包括 697 对分辨率为 1242×375 的图像对,共覆盖 29 个场景。其余 23488 对图像用于训练和交叉验证,其中 22600 对用于训练,剩余的用于验证。由于 KITTI 原始数据集的真实视差图非常稀疏,本文另外在包含 200 组图像对的 KITTI 2015 的训练集上进行了测试。

为了评估该单目深度估计模型的泛化能力,本文将 KITTI 数据集上训练的模型直接对 Cityscapes 数据集进行测试。Cityscapes 测试集包括 1525 对分辨率较高的图像对。

#### 3.1.2 训练细节

本文提出的网络结构使用 Pytorch 框架实现。训练时将输入图像大小调整为 512×256。网络由 Adam[21]优化器进行优化,优化器参数设置为 $\beta_1=0.9$, $\beta_2=0.999$。初始学习率设置为 0.0001,mini-batch 尺寸设置为 8。损失函数各个部分的权重设置如下:$\alpha_{\text{rec}}=1$, $\alpha_{\text{smooth}}=0.1$, $\alpha_{lr}=1$。

整个网络的训练过程分为多个阶段。首先,对正向子网络训练 25 次迭代,然后固定正向子网络的参数,再对反向子网络训练 20 次迭代。最后,对整个网络训练 10 次迭代,得到最终的模型。

#### 3.1.3 评价指标

基于先前工作[3,11],本文采用如下客观评价指标来评估深度估计模型:

平均相对误差 (Abs Rel): $\frac{1}{T}\sum_{i=1}^{T}\frac{\|d_i^*-d_i\|}{d_i}$

相对均方误差 (Sq Rel): $\frac{1}{T}\sum_{i=1}^{T}\frac{\|d_i^*-d_i\|^2}{d_i}$

均方根误差 (RMSE): $\sqrt{\frac{1}{T}\sum_{i=1}^{T}(d_i^*-d_i)^2}$

均方根对数误差 (RMSE log):

$$\sqrt{\frac{1}{T}\sum_{i=1}^{T}(\log d_i^*-\log d_i)^2}$$

准确率——满足如下条件的 $d_i^*$ 的百分比:

$$\delta=\max\left(\frac{d_i^*}{d_i},\frac{d_i}{d_i^*}\right)<\text{thr},\text{thr}=1.25,1.25^2,1.25^3$$

其中,$d_i^*$ 和 $d_i$ 分别为像素 $i$ 的预测深度和真实深度,$T$ 为图像的像素总数。

### 3.2 实验结果与分析

表 1 所示为本文方法与几个先进的基于有监督学习(Eigen et al.[3], Liu et al.[5], Gan et al.[22])、基于半监督学习(Kuznietsov et al.[23])以及基于无监督学习(Zhou et al.[18], Garg et al.[10], Yin et al.[24], Pilzer et al.[12], Godard et al.[11])的方法的对比结果。结果表明,本文算法的深度估计效果明显优于其他无监督学习方法的效果,同时也达到甚至超过了有监督学习方法的效果。为了证明该算法各部分的有效性,表 1 提供了在 KITTI 数据集上的消融实验结果。仅对正向子网络进行训练的客观指标证明了多阶段网络结构的有效性。另外,通过对比有无自注意力机制的实验结果,验证了自注意力特征提取器的有效性。

图 3 展示了 KITTI 数据集上的深度估计效果图。与其他方法相比,该模型在具有挑战性区域的预测效果更好,如细小物体、运动的车辆以及遮挡区域等。本文算法得到的深度图整体上看起来更加平滑,同时细节的预测效果也更好,特别是前景边缘部分的深度图预测得更加锐利清晰。

为了评估本文方法的泛化能力,该模型直接被用于测试 Cityscapes 数据集,并将结果同其他基于无监督学习的方法进行对比,如图 4 所示。该模型只在 KITTI 数据集上进行训练,而没有用到 Cityscapes 数据集。尽管两个数据集在摄像机参数和场景特征上存在一定差异,该模型仍然可以产生视觉上令人信服的深度图像。

表 1　KITTI 数据集上的测试结果。K 代表 KITTI[19]，CS 代表 Cityscapes[20]。深度的最大值是 80 米

Tab. 1　Results on KITTI dataset. K represents KITTI[19] and CS is Cityscapes[20]. Depth predictions are capped at 80 meters

| 方法 | 有/无监督 | 数据集 | Abs Rel | Sq Rel | RMSE | RMSE log | $\delta<1.25$ | $\delta<1.25^2$ | $\delta<1.25^3$ |
|---|---|---|---|---|---|---|---|---|---|
| | | | 越小越好 | | | | 越大越好 | | |
| Eigen et al.[3] | Y | K | 0.203 | 1.548 | 6.307 | 0.282 | 0.702 | 0.890 | 0.958 |
| Liu et al.[5] | Y | K | 0.202 | 1.614 | 6.523 | 0.275 | 0.678 | 0.895 | 0.965 |
| Kuznietsov et al.[23] | Semi | K | 0.113 | 0.741 | 4.621 | 0.189 | 0.862 | 0.960 | 0.986 |
| Gan et al.[22] | Y | K | **0.098** | **0.666** | **3.933** | **0.173** | 0.890 | **0.964** | **0.985** |
| Zhou et al.[18] | N | K | 0.208 | 1.768 | 6.856 | 0.283 | 0.678 | 0.885 | 0.957 |
| Zhou et al.[18] | N | CS+K | 0.198 | 1.836 | 6.565 | 0.275 | 0.718 | 0.901 | 0.960 |
| Garg et al.[10] | N | K | 0.169 | 1.080 | 5.104 | 0.273 | 0.740 | 0.904 | 0.962 |
| Yin et al.[24] | N | K | 0.155 | 1.296 | 5.857 | 0.233 | 0.793 | 0.931 | 0.973 |
| Pilzer et al.[12] | N | K | 0.152 | 1.388 | 6.016 | 0.247 | 0.789 | 0.918 | 0.965 |
| Godard et al.[11] | N | K | 0.148 | 1.344 | 5.927 | 0.247 | 0.803 | 0.922 | 0.964 |
| Godard et al.[11] | N | CS+K | 0.124 | 1.076 | 5.311 | 0.219 | 0.847 | 0.942 | 0.973 |
| 本文算法(正向子网络) | N | K | 0.119 | 1.147 | 5.008 | 0.213 | 0.861 | 0.947 | 0.975 |
| 本文算法(不包括 SA) | N | K | 0.105 | 0.852 | 4.441 | 0.183 | **0.897** | 0.962 | 0.982 |
| 本文算法 | N | K | 0.099 | 0.736 | 4.435 | 0.180 | **0.897** | 0.962 | 0.982 |

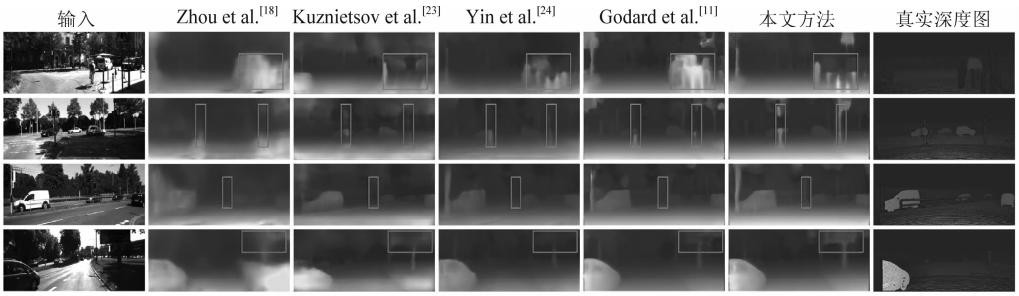


图 3　本文方法在 KITTI 2015 上的深度预测结果示例

Fig. 3　Results on the KITTI 2015 training set containing 200 images

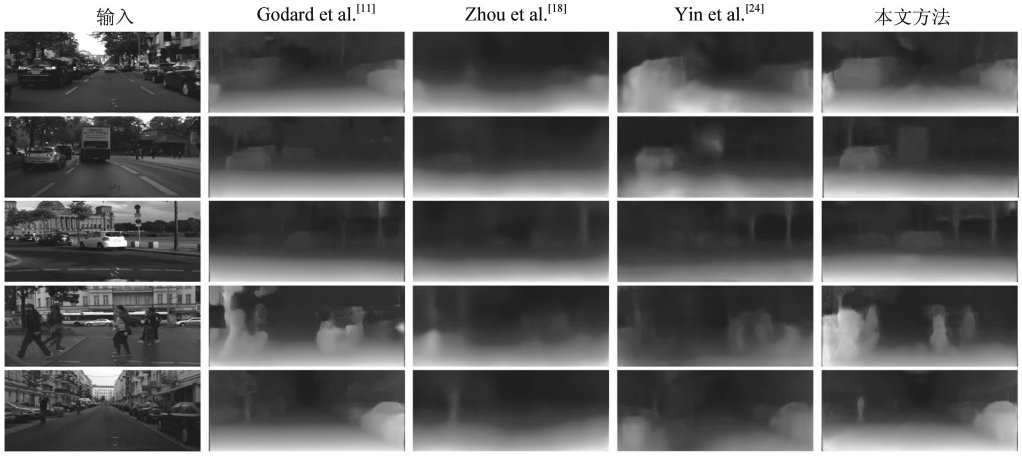


图 4　本文方法在 Cityscapes 上的深度预测结果示例。该模型仅在 KITTI 数据集上进行训练，直接在 Cityscapes 上测试

Fig. 4　Our example depth predictions on the Cityscapes. Our model is trained on KITTI only, and directly tested on Cityscapes

## 4　结论

本文提出了一种基于自注意力机制的多阶段单目深度估计网络。网络结构由两个子网络构成，形成一个循环，为深度估计提供更强的约束和监督。此外，本文采用了掩模加权重建损失和左右一致性损失等损失函数，提高了深度估计的准确性。该方法在 KITTI 数据集上获得了较为满意的深度估计结果。但是，该方法对弱纹理区域预测结果的改进并没有达到理想的效果，天空等远景部分仍存在一定程度的伪影，这也是今后工作需要重点解决的问题。另外，今后的工作将会由图像深度估计扩展

到视频深度估计领域进行更加深入和系统的研究。

## 参考文献


[1] Saxena A, Sun M, Ng A Y. Make3d: Learning 3-d scene structure from a single still image[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(5): 824-840.

[2] Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture[C]//ICCV, 2015: 2650-2658.

[3] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[C]//NIPS, 2014: 2366-2374.

[4] Liu Fayao, Shen Chunhua, Lin Guosheng. Deep convolutional neural fields for depth estimation from a single image[C]//CVPR, 2015: 5162-5170.

[5] Liu Fayao, Shen Chunhua, Lin Guosheng, 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.

[6] Cao Yuanzhouhan, Wu Zifeng, Shen Chunhua. Estimating depth from monocular images as classification using deep fully convolutional residual networks[J]. arXiv preprint arXiv: 1605.02305, 2016.

[7] Laina I, Rupprecht C, Belagiannis V, et al. Deeper depth prediction with fully convolutional residual networks[C]//3DV, 2016: 239-248.

[8] Kendall A, Martirosyan H, Dasgupta S, et al. End-to-end learning of geometry and context for deep stereo regression[C]//ICCV, 2017: 66-75.

[9] Chang Jiaren, Chen Yongsheng. Pyramid stereo matching network[C]//CVPR, 2018: 5410-5418.

[10] Garg R, Kumar BG V, Carneiro G, et al. Unsupervised CNN for single view depth estimation: Geometry to the rescue[C]//ECCV, 2016: 740-756.

[11] Godard C, Mac Aodha O, Brostow G J. Unsupervised monocular depth estimation with left-right consistency[C]//CVPR, 2017: 270-279.

[12] Pilzer A, Xu Dan, Puscas M, et al. Unsupervised adversarial depth estimation using cycled generative networks[J]. arXiv preprint arXiv: 1807.10915, 2018.

[13] Pilzer A, Lathuiliere S, Sebe N, et al. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation[C]//CVPR, 2019: 9768-9777.

[14] Huang Gao, Liu Zhuang, Weinberger K Q, et al. Densely connected convolutional networks[C]//CVPR, 2017: 4700-4708.

[15] 肖易明, 张海剑, 孙洪, 等. 引入注意力机制的视频声源定位[J]. 信号处理, 2019, 35(12): 1969-1978.
Xiao Yiming, Zhang Haijian, Sun Hong, et al. Video Sound Source Localization with Attention Mechanism[J]. Journal of Signal Processing, 2019, 35(12): 1969-1978. (in Chinese)

[16] Wang Xiaolong, Girshick R, Gupta A, et al. Non-local neural networks[C]//CVPR, 2018: 7794-7803.

[17] Zhang Han, Goodfellow I, Metaxas D, et al. Self-attention generative adversarial networks[J]. arXiv preprint arXiv: 1805.08318, 2018.

[18] Zhou Tinghui, Brown M, Snavely N, et al. Unsupervised learning of depth and ego-motion from video[C]//CVPR, 2017: 1851-1858.

[19] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//CVPR, 2012: 3354-3361.

[20] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding[C]//CVPR, 2016: 3213-3223.

[21] Kingma D, Ba Adam J. A method for stochastic optimization[J]. arXiv preprint arXiv: 1412.6980, 2014.

[22] Gan Yukang, Xu Xiangyu, Sun Wenxiu, et al. Monocular depth estimation with affinity, vertical pooling, and label enhancement[C]//ECCV, 2018: 224-239.

[23] Kuznietsov Y, Stuckler J, Leibe B. Semi-Supervised Deep Learning for Monocular Depth Map Prediction[C]//CVPR, 2017: 6647-6655.

[24] Yin Zhichao, Shi Jianping. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose[C]//CVPR, 2018: 1983-1992.


## 作者简介


**刘香凝** 女, 1996年生, 辽宁人。北京大学信息工程学院硕士研究生, 主要研究方向为计算机视觉。
E-mail: liuxiangning@pku.edu.cn

**赵 洋** 男, 1987年生, 安徽人。合肥工业大学计算机与信息学院副研究员。2013年获中国科学技术大学博士学位。研究方向为视频与图像处理。
E-mail: zhaoyang@pkusz.edu.cn

**王荣刚(通信作者)** 男, 1976年生, 黑龙江人。北京大学信息工程学院教授、博士生导师。2006年获中科院计算技术研究所计算机应用专业博士学位。研究方向为多媒体信息处理技术。
E-mail: rgwang@pkusz.edu.cn