

# 自监督单目深度估计网络 MLDA-Net 复现

## 摘要

基于监督学习的深度估计方法依赖于高精度的真实深度信息，使用传感器获取深度信息成本高昂，限制了其发展。在这一背景下，自监督深度估计方法成为了一种经济高效的替代方法。但现有自监督深度估计方法存在特征提取不足的问题，这容易导致深度图模糊并且丢失深度细节。因此，为了克服这些限制，论文提出了一种名为 MLDA-Net 的新颖框架，获得具有更清晰的边界和更丰富的深度细节的深度图。具体来说，改论文使用了多级特征提取 (MLFE) 策略，它可以学习丰富的层次表示。然后，提出了一种结合全局注意和结构注意的双重注意策略，来强化所获得的全局和局部特征，从而得到具有更清晰边界的改进深度图。最后，提出了基于多级输出的重新加权损失策略，对自监督深度估计进行有效监督。本文对 MLDA-Net 框架进行了复现，使用 KITTI 数据集以单目加立体 (MS) 进行实验。结果表明，复现的 MLDA-Net 框架在自监督单目深度估计上得到了比原文略好的结果。同时，在原文基础上，对 MLDA-Net 框架改进并进行实验，最终实验结果有所提升。

**关键词：**深度估计；自监督；双重注意力；特征融合

## 1 引言

近年来，随着深度传感器（如 Lidar）的发展，深度信息变得更加容易获取，也在场景理解 [1, 2]、自动驾驶 [3, 4]、增强现实 [5, 6] 等领域得到了广泛的应用。尽管深度传感器技术不断进步，已经能够获取到高质量、高清的深度信息，但是目前的深度传感器仍然十分昂贵，因此要获取高精度的深度信息成本较高。因此，单目深度估计方法作为一种更经济高效的替代方法，成为了一个研究热点。

单目深度估计方法可以分为两类：有监督方法和自监督方法 [7, 8]。传统的有监督单目深度估计方法依赖于昂贵的深度传感器捕获的真实深度数据，使用这种高精度的深度信息作为监督信号，其成本高昂限制了它的广泛应用。而自监督的深度估计方法不需要真实的深度信息，它利用连续帧之间的几何关系来进行深度预测，减少了成本。但是，自监督的深度估计方法仍存在一些不足，比如特征提取不足，这会导致生成的深度图存在模糊现象，丢失深度细节。如图 1 所示，图中的深度图以颜色模式编码，从黄色到紫色的颜色表示从 0 到  $\infty$  的深度。图 1(a) 是输入的彩色图像，图 1 (b)、(c)、(d) 是三种不同方法估计深度得到的深度图，可以看到图 1 (b)、(c) 中的植物和门是模糊的，而图 1 (d) 中这些细节则比较清晰，这说明 MLDA-Net 可以估计更好的深度信息。

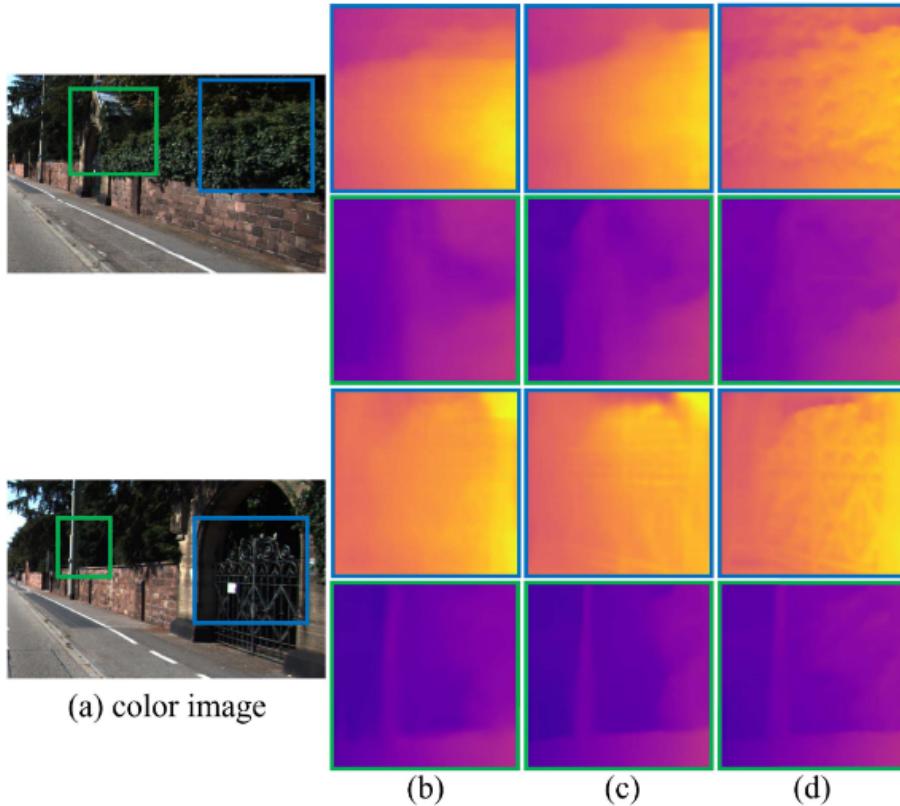


图 1. 不同方法的深度图比较。其中 (a) 是彩色图像, (b)、(c)、(d) 分别是 Monodepth2 [7]、Monodepth2 dh [8] 和 MLDA-Net 方法估计的深度。

为了解决这些问题，本文复现的论文作者注意到特征注意力机制 [9, 10] 可以强化模型在重要特征区域的表达，因此引入了特征关注机制来提高深度图的细节保留能力。基于这一思想，论文作者提出了一个新的网络架构——多层次双重注意力网络（MLDA-Net），该网络旨在通过多层次的特征融合和注意力机制来强化深度估计的精度，保留深度边界清晰度和小物体的深度细节。

本文旨在复现和分析 MLDA-Net，探索其在自监督单目深度估计中的优势，并验证其在公开数据集上的表现。通过对比实验，我们将展示 MLDA-Net 如何有效缓解当前方法中存在的深度模糊和细节丢失问题，并对其进行改进，提升其在自监督单目深度估计中的性能。

## 2 相关工作

### 2.1 单目深度估计

单目深度估计是指从单一图像或者图像序列中推断出场景的深度信息。自监督的单目深度估计方法通常通过连续图像帧之间的光度约束来进行训练，最初的方法 [11] 使用一种基于图像序列自监督学习的框架，通过同时预测深度图和相机的位姿来构建训练目标，这一方法利用相邻图像之间的几何关系和光度一致性，实现了可以在没有标注数据的情况下进行深度估计，后续被广泛应用。

在单目深度估计中，除了需要估计深度，还需要进行相机位姿估计。通常是通过估计连

续帧之间的相对位姿，使用光度一致性损失来训练深度网络。[11] 和 [12] 都使用了相机姿态估计来辅助深度估计。然而，在处理具有动态物体的场景时，传统深度估计方法表现较差，因此，一种将运动分解为刚性和非刚性的分量的方法 [13] 被提出，它利用深度和光流信息的组合来描述物体的运动，有效地处理了动态物体带来的问题。[14] 则利用预先计算的实例分割掩码作为先验信息，将运动信息和深度信息结合，帮助模型在复杂场景下分辨动态物体和静态物体，提高了深度估计的精度。[15] 方法则引入了语义信息，通过结合语义标签和深度估计，在动态物体和复杂场景中提供了更精确的深度预测，解决了由物体移动引起的深度估计不准确问题。

为了进一步提高单目深度估计的精度，许多约束和正则化方法被提出，例如，[7] 提出了最小投影损失方法，该方法通过处理图像遮挡和视角变化来提高深度估计在实际场景的鲁棒性。[16] 则提出利用表面法线作为额外的几何约束，并引入了边缘感知深度-法线方法来改善深度图的细节表现。[17] 提出了基于几何匹配的损失函数，通过强化时序一致性来提升深度估计的稳定性和精度。通过这些额外的几何约束和正则化方法，单目深度估计模型能够在复杂的场景中获得更为准确和鲁棒的深度预测。

## 2.2 立体图像深度估计

立体图像估计通过利用成对的立体图像推断场景的深度信息，与单目深度估计相比，立体图像提供了更多的几何信息，能够获得更高的估计，并且立体图像中已经包含了相机位姿，不需要再进行相机姿态估计。

为了提高立体图像的深度估计精度，[18] 通过从不同视角合成特征并执行立体匹配来推断深度信息。[19] 则使用左右图像深度一致性约束，通过最小化左右图像之间的视差差异来提升深度估计的结果。[20, 21] 将深度估计、相机运动估计、光流估计、运动分割和语义分割等任务结合进同一框架，通过任务间的一致性约束来优化深度估计结果。[8] 则提出了结合法线深度信息来改进深度预测性能的方法，有效提高了深度图的细节表现。[22] 则研究了如何通过法线估计模型来优化深度质量，利用预测的法线图来改善深度图的精度。

此外，[23] 通过循环一致性和蒸馏操作来推断深度信息，提出了基于深度循环一致性损失的框架。[24] 则提出了通过一系列二分类操作来估算深度信息，进一步增强了深度估计的精度。不确定性建模也可以提高立体图像深度估计性能。[25] 探讨了如何估算深度预测的不确定性，并分析这一不确定性对深度估计精度的影响，在推断深度时为模型提供更可靠的决策依据。

现有的许多传统方法依赖于单一的训练模式，通常仅通过使用单目或立体图像输入来训练深度网络，这导致模型往往无法有效捕捉到图像中的全局和局部信息，从而影响深度估计的精度。此外，现有的许多自监督深度估计方法都存在特征提取不充分问题。论文提出的端到端的自监督深度估计框架——MLDA-Net 可以同时处理单目、立体以及单目加立体的输入，这种多输入类型的设计能够充分利用数据中的丰富信息。同时，MLDA-Net 网络使用多级特征提取策略，可以有效缓解特征提取不足的问题。

### 3 本文方法

#### 3.1 本文方法概述

复现论文提出了 MLDA-Net 深度估计网络，该网络的框架图如图 2 所示。它将特征注意策略结合到一个新的框架中，是一种多级双重注意网络。该网络首先通过一种多级特征提取策略 (MLFE)，来获得有效的特征，再利用双重注意策略以全局和局部方式强化提取的特征。全局注意力以全局方式增强 MLFE 中获得的特征，结构注意力以局部方式增强全局注意力在边缘和结构部分获得的特征。最后，利用重新加权损失策略以自我监督的方式进一步细化深度信息。

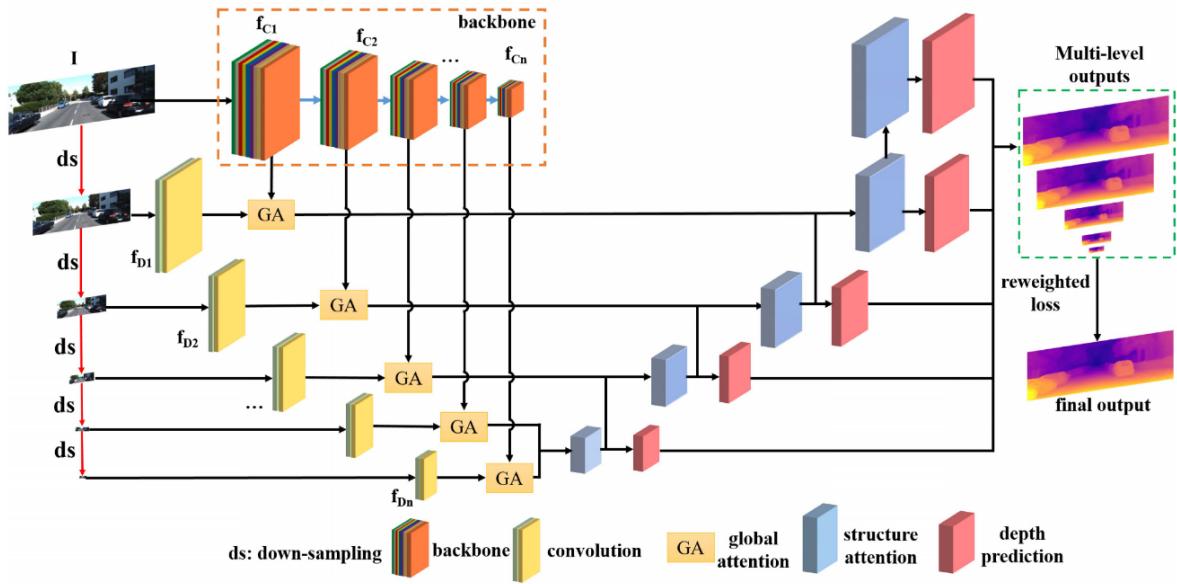


图 2. MLDA-Net 网络框架图，引用自 [26]

#### 3.2 多级特征提取块

该模块以彩色图像作为输入，并进行五次下采样。原始图像通过 backbone 网络（通常为 Resnet18 或者 Resnet50）提取特征，而下采样版本则通过卷积操作提取特征。 $f_C$  特征的提取如公式 1 所示， $f_D$  特征的提取如公式 2 所示：

$$\begin{aligned} f_{C_1} &= B_1(I) \dots \\ f_{C_n} &= B_n(f_{C_{n-1}}) \end{aligned} \quad (1)$$

其中， $f_C$  表示以原图像为输入从 backbone 网络中提取的特征， $B_i$  表示  $B$  中的第  $i$  个卷积运算。

$$\begin{aligned} f_{D_1} &= Conv(D \downarrow_1 (I)) \dots \\ f_{D_n} &= Conv(D \downarrow_n (I)) \end{aligned} \quad (2)$$

其中， $f_D$  表示从输入图像的下采样版本中提取的特征， $D \downarrow_i$  表示下采样操作，它将原始图像从  $(W, H)$  转换为  $(\frac{W}{2^i}, \frac{H}{2^i})$ ，其中  $i$  从 1 变化到  $n$ ， $Conv$  表示卷积操作。

### 3.3 双重注意力块

多级特征提取模块提取的特征  $f_C$  和  $f_D$  作为双重注意力块的输入。在该模块中，使用全局注意力和结构注意力来强化特征。其中，全局注意力对提取到的特征的每个通道都回归一个权重，从而强化重要性更高的特征通道。全局注意力的过程如图 3 (A) 所示，而结构注意力则通过对低分辨率的特征图和相应的高分辨率特征图进行引导过滤来恢复结构和边缘信息。结构注意力能够滤除噪声并处理上采样引起的边界模糊问题，其目的是强化每个特征图中的边缘区域，从而获得边界更清晰的更好的深度图，结构注意力的过程如图 3 (C) 所示。

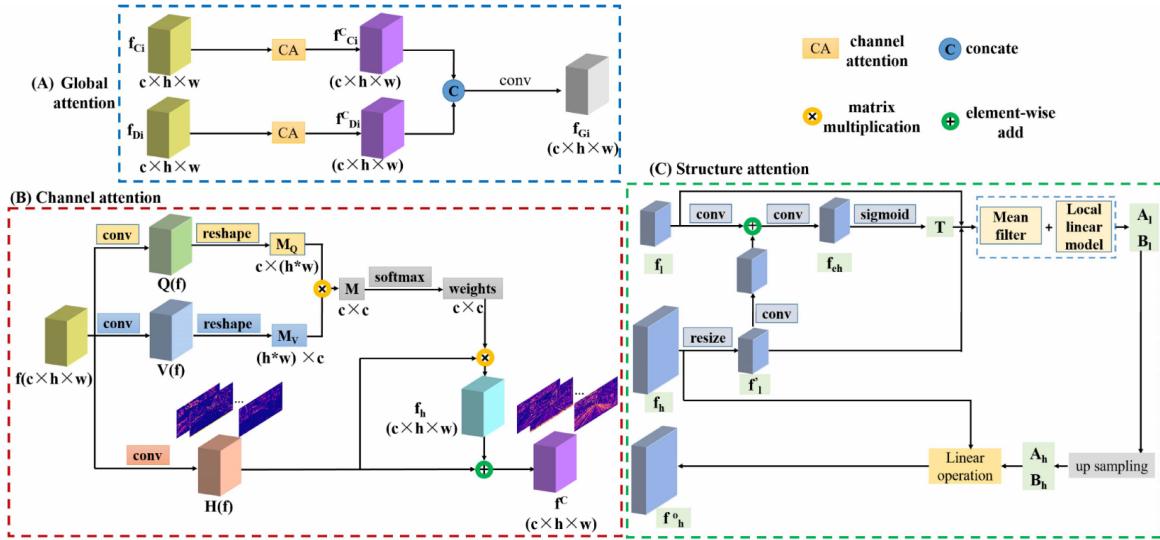


图 3. GA、CA 和 SA 的流程图，引用自 [26]

#### 3.3.1 全局注意力

全局注意力首先使用通道注意力以全局方式来强化特征。这一过程可以用式 3 来表示：

$$\begin{aligned} f_C^C &= CA(f_C) \\ f_D^C &= CA(f_D) \end{aligned} \quad (3)$$

其中， $CA$  表示通道注意力， $f_C^C$  和  $f_D^C$  分别表示  $f_C$  和  $f_D$  经通道注意力操作强化后的特征。接着将具有相同分辨率的强化特征融合，得到  $f_G = \{f_{G_1}, \dots, f_{G_n}\}$ 。这一过程可以用式 4 表示：

$$\begin{aligned} f_{G_1} &= Conv(Concat(f_{C_1}^C, f_{D_1}^C)) \dots \\ f_{G_n} &= Conv(Concat(f_{C_n}^C, f_{D_n}^C)) \end{aligned} \quad (4)$$

其中， $Concat$  表示连接操作， $Conv$  表示卷积操作， $f_G$  表示经过全局注意力操作加强后的特征。

#### 3.3.2 通道注意力

在全局特征加强操作中，主要用到了通道注意力。通道注意力的流程如图 3 (B) 所示。它的输入是大小为  $(c \times h \times w)$  的特征  $f$ ，分别通过卷积运算  $H(\cdot)$ 、 $Q(\cdot)$  和  $V(\cdot)$  得到三个分

量  $H(f)$ 、 $Q(f)$  和  $V(f)$ ，其大小为  $(c \times h \times w)$ 。接着利用重塑操作将  $Q$  和  $V$  转换为  $(c \times hw)$  和  $(hw \times c)$ ，这一过程如 5 所示：

$$\begin{aligned} M_Q &= R_S(Q(f)) \\ M_V &= R_S(V(f)) \end{aligned} \quad (5)$$

其中， $Q(\cdot)$  和  $V(\cdot)$  是用于学习通道注意力的权重参数。 $R_S$  表示重塑操作， $M_Q$  和  $M_V$  表示  $f$  分别经  $Q$  和  $V$  卷积再重塑后得到的分量。接着，将  $M_Q$  和  $M_V$  经点积运算  $M = M_Q \odot M_V$  得到大小为  $c \times c$  的  $M$ 。接着将  $M$  输入到 softmax 操作中得到通道注意力权重  $\theta$ ，然后  $H(f)$  和  $\theta$  经过点积运算得到增强分量  $f_h$ ，最后将增强分量  $f_h$  和  $H(f)$  逐元素相加得到通道注意力的输出  $f^C$ 。该过程可以表示为式 6：

$$\begin{aligned} \theta &= \text{softmax}(M) \\ f_h &= \theta \odot H(f) \\ f^C &= f_h + H(f) \end{aligned} \quad (6)$$

其中， $\theta$  表示通道注意力权重， $H(\cdot)$  表示卷积操作， $f^C$  表示强化特征。通道注意力通过利用特征图的所有位置信息来获得每个通道的有效注意力权重，以全局的方式增强相应的特征。

### 3.3.3 结构注意力

MLDA-Net 框架使用结构注意力来获取更清晰的深度边界。具体而言，结构注意力通过对不同分辨率的特征图进行过滤来恢复结构和边缘信息，以局部方式增强特征，从而获得边界清晰的深度图。如图 3(C) 所示，结构注意力使用大小为  $c \times h \times w$  的特征  $f_l$  和大小为  $c \times 2h \times 2w$  的  $f_h$  作为输入，首先将  $f_h$  下采样到大小为  $c \times h \times w$  的  $f'_l$ 。接着通过卷积和相加运算得到特征  $f_{eh}$ ，最后通过 sigmoid 运算计算出权重  $T$ 。该过程如式 7：

$$\begin{aligned} f'_l &= D \downarrow_i (f_h) \\ f_{eh} &= \text{Conv}(f'_l) + \text{Conv}(f_l) \\ T &= \text{sigmoid}(f_{eh}) \end{aligned} \quad (7)$$

其中， $D \downarrow_i$  是最近邻方式的下采样操作。然后，以  $T$ 、 $f'_l$  和  $f_l$  为输入，最小化重构误差，然后利用平均滤波器和局部线性模型获得  $f_l$  和  $f'_l$  之间的线性系数： $A_l$  和  $B_l$ 。接着使用上采样操作来获得  $f_h$  的线性系数  $A_h$  和  $B_h$ 。最后，利用线性运算来获得特征  $f_h^o$ ，该过程如式 8：

$$\begin{aligned} \min_{a_k, b_k} (E(a_k, b_k)) &:= \sum_{i \in w_k} (T_i^2 (a_k f_{l_i} + b_k - f'_{l_i})^2 + \lambda a_k^2) \\ \hat{f}_i &= \frac{1}{N_k} \sum_{k \in \Omega_i} a_k f'_{l_i} + \frac{1}{N_k} \sum_{k \in \Omega_i} b_k = A_l * f'_l + B_l \\ f_h^o &= A_h * f_h + B_h \end{aligned} \quad (8)$$

其中， $w_k$  是指结构注意过程中的一个方形窗口，该方形窗口中每个位置  $k$  的半径为  $r$ ， $\Omega_i$  是指位置  $i$  的所有窗口集合。对于  $f_l$  中的一个像素，可以通过线性变换  $\hat{f}_{ki} = a_k f_{l_i} + b_k, \forall i \in w_k$  求得  $\hat{f}_{ki}$ 。通过最小化所有像素  $\hat{f}_{ki}$  和  $f'_{l_i}$  之间的差异来得到  $a_k$  和  $b_k$ 。之后通过对每个与  $i$  位

置的所有窗口的系数进行平均。最后，对  $A_l$  和  $B_l$  进行上采样得到  $A_h$  和  $B_h$ ，再由线性计算得到最终的输出  $f_h^o$ 。在 MLDA-Net 网络中，首先对低分辨率的特征进行上采样，再将其与高分辨率的特征连接起来，然后应用全局注意力，这样做可以增强低分辨率特征的作用。

### 3.4 深度预测块

深度预测块的输入是结构注意力的输出特征  $f_{SA}$ ，不同特征输入得到不同尺寸的深度图。

$$\begin{aligned} d_1 &= DP(f_{SA_1}) \dots \\ d_n &= DP(f_{SA_n}) \end{aligned} \quad (9)$$

它可表示为式9，其中，DP 表示深度预测操作，包含卷积层和激活层，卷积核的大小为  $3 \times 3$ 。

### 3.5 重新加权损失

训练数据由图像对  $I$  和  $I'$  组成，通过预测的深度图和相机位姿将图像  $I$  重投影到  $I'$ ，以此来训练网络重建图像  $I''$ ，最后计算  $I'$  和  $I''$  之间的损失。损失使用光度一致性损失，其公式如式 10， 11：

$$l(I', \bar{I}) = \frac{\alpha}{2}(1 - SSIM(I', \bar{I})) + (1 - \alpha)||I', \bar{I}||_1 \quad (10)$$

$$\begin{aligned} SSIM(I_x, I_y) &= l(I_x, I_y)cs(I_x, I_y) \\ l(I_x, I_y) &= \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \\ cs(I_x, I_y) &= \frac{2\delta_{xy} + c_2}{\delta_x^2 + \delta_y^2 + c_2} \end{aligned} \quad (11)$$

其中， $\alpha$  根据经验设置， $SSIM$  是结构性相似度量函数，通过亮度、对比度和结构来评估两个图像之间的结构相似性。 $\mu_x$  是  $I_x$  的平均值， $\mu_y$  是  $I_y$  的平均值， $\delta_x^2$  是  $I_x$  的方差， $\delta_y^2$  是  $I_y$  的方差， $\delta_{xy}$  是  $I_x$  和  $I_y$  的协方差矩阵。 $c_1 = (k_1 L)^2$  和  $c_2 = (k_2 L)^2$  是稳定弱分母除法的两个变量， $L$  是像素值的范围， $k_1 = 0.01$ ， $k_2 = 0.03$ 。

之前由多级特征  $f_{SA}$  得到的深度图  $d$  具有不同的分辨率，重新加权损失模块对不同的深度图设置不同的权重，来得到最终的损失，其公式如式 12：

$$loss = \frac{1}{n} \sum_{i=1}^n \frac{1}{2^i} loss_i(I'_i, \bar{I}_i) \quad (12)$$

其中， $n$  是多级特征输出的数量， $loss_i$  是第  $i$  个深度图的损失。

## 4 复现细节

### 4.1 与已有开源代码对比

本文所选论文无官方开源代码，在生成深度提示信息时，参考了 depth-hints 代码<sup>1</sup>，在复现时参考了对所选论文进行复现的代码<sup>2</sup>，并在其基础上进行了以下几个方面的工作：

<sup>1</sup><https://github.com/nianticlabs/depth-hints>

<sup>2</sup><https://github.com/bitcjm/MLDA-Net-repo>

- 参考代码仅复现了图像尺寸在  $640 \times 192$  下 MLDA-Net 的结果，复现代码参照原论文，在其基础上增加了图像尺寸在  $1024 \times 320$  下 MLDA-Net 的结果。由于分辨率提高，适当降低了初始学习率。同时高分辨率下模型的训练需要更多的迭代来收敛，因此增大了步长，确保模型最终成功收敛。
- 参考代码仅复现了在 Resnet18 网络下 MLDA-Net 的结果，复现代码参照原论文，增加了 MLDA-Net 框架在 Resnet50 网络下的训练。
- 损失权重调整与优化：原有的损失函数采用固定权重的方式来组合多个不同类型的损失项，然而这种固定权重设置可能无法适应训练过程中不同阶段以及不同数据样本的复杂特性，导致模型训练效果受限。因此，在计算总损失时，使用自适应的权重调整损失组合的方法，根据实际情况为不同的损失项分配权重，从而提高模型性能。

## 4.2 数据集与评价指标

本文使用到的数据集为 KITTI [27]，该数据集由经过激光雷达测量配准后的校准视频组成，包含了自动驾驶场景中的多种典型场景和环境，如城市街道、高速公路和乡村道路等。深度信息的评估是在 LiDAR 点云上完成的。数据集被划分成 39810 个单目三元组用于训练，4424 个用于测试，697 个用于验证。

论文中的自监督深度估计有单目视觉、立体视觉以及单目加立体这三种训练模式。具体而言，单目训练模式需要同时预测相机姿态和深度，而立体模式的原图像中有已知的相机姿态，只需预测深度。本文采用了第 3 种单目 + 双目立体视觉的训练模式。此外，由于 [8] 已经证明深度提示信息有助于深度估计。因此，本文在实验时均使用了深度提示信息。

本文使用的评估指标与原论文一致。具体而言，使用了“Abs Rel”、“Sq Rel”、“RMSE”，“RMSE log”、“ $\delta < 1.25$ ”、“ $\delta < 1.25^2$ ”、“ $\delta < 1.25^3$ ”这七个评估指标来评估模型预测深度信息的性能。其中，“Abs Rel”用来衡量预测深度与真实深度之间的绝对值相对误差，“Sq Rel”用来衡量预测深度与真实深度之间的平方相对误差，“RMSE”用来衡量误差的整体水平，“RMSE log”则反映预测深度和真实深度对数值之间的均方根误差，最后三个指标分别表示预测深度与真实深度之比在  $1.25$ 、 $1.25^2$ 、 $1.25^3$  以内的像素比例。

## 4.3 创新点

- 支持高分辨率的图像输入。改进后的 MLDA-Net 支持在  $1024 \times 320$  分辨率下的训练和测试，提升模型在高分辨率场景中的性能。
- 引入 Resnet50 网络。增加了支持更深层次特征提取的 ResNet50 网络，对比分析不同网络架构对模型性能的影响。
- 损失权重调优。将固定权重组合的损失函数改为自适应权重优化策略。通过动态调整训练过程中各损失项的权重，使模型在不同训练阶段能够更好地适应多样化数据特性，显著提升训练效果和模型性能。
- 优化训练超参数设置。针对高分辨率实验和新增骨干网络，调整了学习率、权重衰减等训练参数，同时优化批量大小以平衡显存占用与训练效率，确保不同实验条件下模型性

能的稳定性。

## 5 实验结果分析

本项目基于 paddlepaddle 深度学习框架复现。实验环境为 Ubuntu22.04，使用的显卡是 Nvidia 4090GPU。在输入图像尺寸为  $640 \times 192$  时，模型训练批量数设置为 2，训练轮数为 10，学习率初始值设置为  $1e-4$ ，优化器使用的是 Adam，学习率衰退因子设置为 0.1，每训练两轮进行学习率衰退。在输入图像尺寸为  $1024 \times 320$  时，将初始学习率改为  $5e-5$ ，引入 Resnet50 后，根据训练过程对模型的训练参数进行修改，以获得最佳结果。具体的实验结果如表 1。

表 1. MLDA-Net 深度估实验结果

Method	Mode	W×H	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
MLDA-Net(ResNet18)	MS	$640 \times 192$	0.105	0.816	4.690	0.190	0.878	0.959	0.981
Retrain(ResNet18)	MS	$640 \times 192$	<b>0.098</b>	<b>0.691</b>	<b>4.299</b>	<b>0.178</b>	0.900	<b>0.965</b>	<b>0.984</b>
Ours(ResNet18)	MS	$640 \times 192$	<b>0.098</b>	0.735	4.401	<b>0.178</b>	<b>0.901</b>	<b>0.965</b>	<b>0.984</b>
MLDA-Net(ResNet18)	MS	$1024 \times 320$	0.099	0.724	4.415	0.183	0.887	0.963	<b>0.983</b>
Retrain(ResNet18)	MS	$1024 \times 320$	0.096	0.791	4.413	0.182	0.911	0.965	0.982
Ours(ResNet18)	MS	$1024 \times 320$	<b>0.088</b>	<b>0.682</b>	<b>4.184</b>	<b>0.176</b>	<b>0.919</b>	<b>0.967</b>	<b>0.983</b>
MLDA-Net(ResNet50)	MS	$1024 \times 320$	<b>0.097</b>	<b>0.658</b>	<b>4.278</b>	<b>0.178</b>	<b>0.889</b>	<b>0.965</b>	<b>0.984</b>
Retrain(ResNet50)	MS	$1024 \times 320$	0.115	0.760	4.711	0.197	0.872	0.957	0.981
Ours(ResNet50)	MS	$1024 \times 320$	0.111	0.833	4.682	0.194	0.882	0.959	0.982

表 1 中前四个指标越小，表示实验结果越好，后三个指标越大，表示实验结果越好。在使用 Resnet18 骨干网络，图像尺寸为  $640 \times 192$  时，重新训练 MLDA-Net 网络已经得到了比原论文更好的结果，经我们的方法改进后，实验结果在除 RMSE 的指标上均有所改善。图像尺寸为  $1024 \times 320$  时，重新训练的 MLDA-Net 网络与原论文结果相差无几。经我们的方法改进后，各评估指标的结果得到了极大的改善。前四个指标均有所降低：Abs Rel 降低了 0.008，Sq Rel 降低了 0.109，RMSE 降低了 0.229，RMSE log 降低了 0.006。后三个指标均有所升高： $\delta < 1.25$  升高了 0.008， $\delta < 1.25^2$  升高了 0.002， $\delta < 1.25^3$  升高了 0.001。但是在使用 Resnet50 骨干网络训练时，重新训练的 MLDA-Net 网络和经我们的方法改进后训练的网络，各评估指标经的结果都不如原论文中 Resnet50 的结果，猜测可能的原因是：超参数欠优化、网络结构改变造成初始化不当等等，后续会进行更细致的参数调优并检查网络结构改变后可能出现的问题。综上，我们的改进方法对 MLDA-Net 网络有着一定的改善。

实验结果可视化图如图 4 所示，图中从上到下三张图依次是：输入图像，生成的深度图、真实深度图。从图中可以看到，生成的深度图可以较好地反映输入图像的深度信息，特别是在简单场景中，生成的深度图与真实深度图高度吻合。同时，生成的深度图也能较好地捕捉到物体的边缘信息，但与真实深度图相比，仍然存在一定程度的模糊。

此外，在复杂场景中（如多个物体重叠或存在遮挡关系的区域），生成的深度图虽然能够识别不同深度层次，但对细节的区分能力仍有一定不足。例如，在遮挡区域，深度过渡可能过于平滑，导致预测结果无法完全反映真实深度信息。

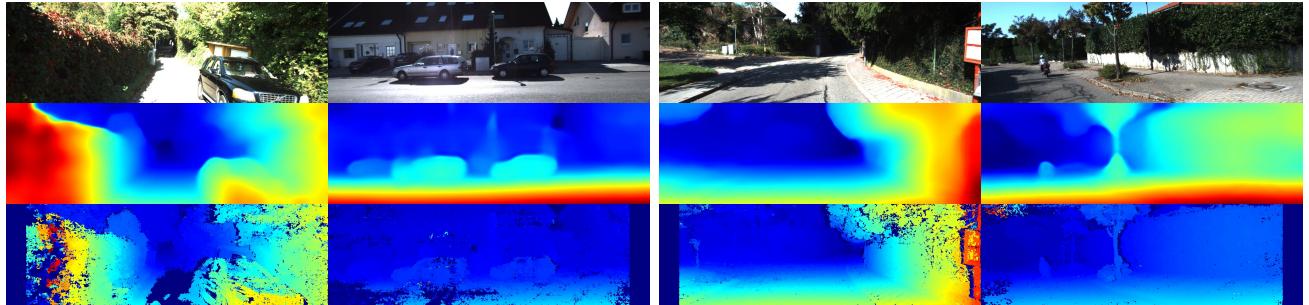


图 4. 实验结果图

## 6 总结与展望

MLDA-Net 框架利用 MLFE 模块，从多尺度层获取丰富的特征，为深度预测奠定了基础，接着使用双重注意力策略分别以全局和局部方式增强特征，最后使用重新加权损失策略对多级特征进行有效监督。MLDA-Net 有效克服了自监督深度估计中特征提取不足和深度图模糊问题。本文成功复现了 MLDA-Net 框架，并在此基础上进行了改进和优化。通过引入高分辨率输入 ( $1024 \times 320$ )、探索不同主干网络（如 ResNet50）以及优化损失函数权重的方式，在一定程度上提升了模型的性能表现。

## 参考文献

- [1] Haoyu Ren, Mostafa El-Khamy, and Jungwon Lee. Deep robust single image depth estimation neural network using scene understanding. In *CVPR Workshops*, volume 2, page 2, 2019.
- [2] Po-Yi Chen, Alexander H. Liu, Yen-Cheng Liu, and Yu-Chiang Frank Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [3] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q. Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [4] Feng Xue, Guirong Zhuo, Ziyuan Huang, Wufei Fu, Zhuoyue Wu, and Marcelo H. Ang. Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2330–2337, 2020.
- [5] Jinmeng Rao, Yanjun Qiao, Fu Ren, Junxing Wang, and Qingyun Du. A mobile outdoor augmented reality method combining deep learning object detection and spatial relationships for geovisualization. *Sensors*, 17(9), 2017.

- [6] Megha Kalia, Nassir Navab, and Tim Salcudean. A real-time interactive augmented reality depth estimation technique for surgical robotics. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8291–8297. IEEE, 2019.
- [7] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019.
- [8] Jamie Watson, Michael Firman, Gabriel J Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2162–2171, 2019.
- [9] Zhi-Song Liu, Li-Wen Wang, Chu-Tak Li, Wan-Chi Siu, and Yui-Lam Chan. Image super-resolution via attention based back projection networks. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3517–3525. IEEE, 2019.
- [10] Xibin Song, Yuchao Dai, Dingfu Zhou, Liu Liu, Wei Li, Hongdong Li, and Ruigang Yang. Channel attention based iterative residual learning for depth map super-resolution. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 5631–5640, 2020.
- [11] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017.
- [12] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. Sfm-net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804*, 2017.
- [13] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1983–1992, 2018.
- [14] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8001–8008, 2019.
- [15] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 582–600. Springer, 2020.
- [16] Zhenheng Yang, Peng Wang, Wei Xu, Liang Zhao, and Ramakant Nevatia. Unsupervised learning of geometry from videos with edge-aware depth-normal consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

- [17] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5667–5675, 2018.
- [18] Fabio Tosi, Filippo Aleotti, Matteo Poggi, and Stefano Mattoccia. Learning monocular depth estimation infusing traditional stereo knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9799–9809, 2019.
- [19] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017.
- [20] Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Learning monocular depth estimation with unsupervised trinocular assumptions. In *2018 International conference on 3d vision (3DV)*, pages 324–333. IEEE, 2018.
- [21] Po-Yi Chen, Alexander H Liu, Yen-Cheng Liu, and Yu-Chiang Frank Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 2624–2632, 2019.
- [22] Uday Kusupati, Shuo Cheng, Rui Chen, and Hao Su. Normal assisted stereo depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2189–2199, 2020.
- [23] Andrea Pilzer, Stephane Lathuiliere, Nicu Sebe, and Elisa Ricci. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9768–9777, 2019.
- [24] Abhishek Badki, Alejandro Troccoli, Kihwan Kim, Jan Kautz, Pradeep Sen, and Orazio Gallo. Bi3d: Stereo depth estimation via binary classifications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1600–1608, 2020.
- [25] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3227–3237, 2020.
- [26] Xibin Song, Wei Li, Dingfu Zhou, Yuchao Dai, Jin Fang, Hongdong Li, and Liangjun Zhang. Mlda-net: Multi-level dual attention-based network for self-supervised monocular depth estimation. *IEEE Transactions on Image Processing*, 30:4691–4705, 2021.
- [27] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.