DepthDark：适用于低光环境的鲁棒单目深度估计

Longjian Zeng
245060073@hdu.edu.cn
Hangzhou Dianzi University
Hangzhou, China

Zunjie Zhu*
zunjiezhu@hdu.edu.cn
Hangzhou Dianzi University
Hangzhou, China

Rongfeng Lu
rongfeng-lu@hdu.edu.cn
Hangzhou Dianzi University
Hangzhou, China

Ming Lu
lu199192@gmail.com
Intel Labs China
Beijing, China

Bolun Zheng
blzheng@hdu.edu.cn
Hangzhou Dianzi University
Hangzhou, China

Chenggang Yan
cgyan@hdu.edu.cn
Hangzhou Dianzi University
Hangzhou, China

Anke Xue
akxue@hdu.edu.cn
Hangzhou Dianzi University
Hangzhou, China

Abstract

近年来，用于单目深度估计的基础模型受到越来越多的关注。目前的方法主要针对典型的白昼条件，但它们在低光照环境中的有效性显著下降。缺乏专门为低光照场景设计的稳健的单目深度估计基础模型。这主要源于缺乏大规模、高质量的低光照条件配对深度数据集以及有效的参数高效微调（PEFT）策略。为了解决这些挑战，我们提出了 DepthDark，这是一个用于低光照单目深度估计的稳健基础模型。我们首先引入了一个耀斑模拟模块和一个噪声模拟模块，以在夜间条件下准确模拟成像过程，从而生成高质量的低光照条件配对深度数据集。此外，我们提出了一种有效的低光照 PEFT 策略，该策略利用光照引导和多尺度特征融合以增强模型在低光照环境中的能力。我们的方法在具有挑战性的 nuScenes-Night 和 RobotCar-Night 数据集上实现了最先进的深度估计性能，验证了其在有限训练数据和计算资源条件下的有效性。

CCS Concepts

• **Computing methodologies** → **Computer vision**; **Image processing**; **Virtual reality**.

Keywords

Low-Light Monocular Depth Estimation, Parameter-Efficient Fine-Tuning, Foundation Model

ACM Reference Format:

Longjian Zeng, Zunjie Zhu, Rongfeng Lu, Ming Lu, Bolun Zheng, Chenggang Yan, and Anke Xue. 2025. DepthDark：适用于低光环境的鲁棒单目深度估计. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3746027.3754871

*Corresponding Author.

MM '25, Dublin, Ireland.

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754871

近年来，随着卷积神经网络 [12] 和视觉转换器 [7] 的快速发展，计算机视觉和自然语言处理领域正在经历一场重大变革。这场变革也大大促进了单目深度估计的发展，单目深度估计从单张图像中预测深度信息，使其在自动驾驶、增强现实和机器人等领域的应用成为可能。然而，大多数现有方法在光照充足的条件下表现良好，但在低光照条件下因信息大量丢失和噪声显著放大而表现较差。

尽管低光单目深度估计是一项极具挑战性的任务，但它在各种夜间应用中发挥着关键作用。然而，大多数现有关于低光单目深度估计的研究，如 RNW [29] 、ADDS [19] 、MonoVit [42] 、MonoFormer [1] 和 TDDC [32] ，主要集中于夜间自动驾驶，这限制了它们在其他应用中的适用性。因此，开发一个能够从单个夜间图像估计深度信息的基础模型已成为低光单目深度估计中的一个关键目标。虽然 Depth Anything [33] 、Depth Anything V2 [34] 和 Marigold [16] 等基础模型在典型的白天条件下表现出色，但由于训练数据覆盖范围的限制，它们在复杂的低光场景中的有效性仍然受到限制。这些模型的成功通常取决于大规模的高质量训练数据集。然而，构建一个包含数千万个低光条件下的深度标签的数据集是一项极具挑战性的任务。此外，训练一个基础模型需要大量的计算资源，这加剧了研究和实际应用的挑战。因此，基于参数高效微调（PEFT）开发一个用于低光单目深度估计的基础模型是紧迫且必要的。

在本研究中，我们通过引入一种创新的方法 DepthDark，全面解决了低光环境下单目深度估计中的挑战，以克服上述困难。具体而言，我们提出了一个耀斑模拟模块和一个噪声模拟模块，以从白天图像合成高度逼真的夜间图像。耀斑模拟模块技术有效地解决了由于光源分布不均导致的光度差异，而噪声模拟模块采用物理解耦框架来准确模拟低光场景中显著的噪声分布变化。因此，这些创新有效地克服了收集大规模高质量配对夜间深度数据的困难。

我们还提出了一种高效的参数高效微调 (PEFT) 策略，以适应预训练的基础模型用于低光深度估计。在此策略中，我们引入了光照引导与多尺度特征融合，这显著提高了模型在低光环境下的性能。借助光照引导，模型可以更好地适应光照条件的变化，从而能够专注于在低光场景中可能被遮挡的相关特征。此外，多尺度特征融合使得模型能够跨不同尺度整合信息，提升深度感知与准确性。这种结合使得 DepthDark 能够有效应对低光条件下的挑战，从而实现稳健的深度估计性能。此外，我

们的低光 PEFT 策略仅需在一个消费者级别的 GPU 上进行数小时训练，并利用少量合成的 RGB-D 训练数据即可实现基础模型的高效适应。这种高效性显著降低了研究人员和开发人员在低光单目深度估计领域的进入门槛，使其更加易于使用。通过利用这些进步，DepthDark 不仅在具有挑战性的场景中表现出色，还使低光环境中的深度估计能力更加普及，鼓励在该领域的进一步探索和创新。我们工作的主要贡献总结如下：

- 我们提出了一种通过真实地模拟光源和噪声特性来合成低光照图像的新方法。该方法解决了在低光场景中收集大规模配对深度数据的挑战。
- 我们设计了一种高效的参数高效微调（PEFT）策略，通过引入光照指导和多尺度特征融合，显著提高了模型在低光环境下的鲁棒性。
- DepthDark 在 nuScenes-Night 和 RobotCar-Night 数据集上使用单个消费级 GPU、短训练时间和极少量训练数据实现了最先进的性能。

1 相关工作

早期对单目深度估计 [13, 15, 18] 及其他方法的研究主要依赖于手工制作的特征和传统的计算机视觉技术，这些方法在处理包含遮挡和缺乏纹理的复杂场景时显得力不从心。然而，深度学习的兴起革新了单目深度估计，以 Eigen [9] 为代表的先驱者提出了多尺度网络，并证明深度结果可以转化为以单一传感器记录的数据集的度量深度。随后的研究主要集中在通过引入额外的先验知识 [17, 21, 26, 35] 和优化目标函数 [31, 37] 来提高精度。

1.1 低光单目深度估计

尽管这些方法提供了令人满意的深度估计结果，但由于昼夜分布的显著差异，它们在夜间场景中常常失败。为了解决低光深度估计的挑战，提出了多种采用领域适应或光度损失的方法。例如，ADFA [27] 引入了一种新的编码器，可以模拟白天的深度估计模型，通过领域适应从夜间图像生成白天特征。ITDFA [41] 使用迁移的图像对在特征和输出空间中约束训练过程。ADDS [19] 使用独立编码器将白天和夜间图像分解为不变和可变特征图，并利用不变信息估计深度。RNW [29] 通过基于先验的输出空间领域适应方法来规范夜间光度异常值。然而，上述方法的性能受限于低光图像中的亮度和光照不一致性带来的挑战。WSGD [28] 是第一个直接训练其提出的夜间图像分割框架的单阶段方法。TDDC [32] 通过应用物理先验来补偿关键的昼夜差异，提出了一种自监督的低光单目深度估计方法。

尽管现有的低光单目深度估计方法在夜间自动驾驶场景中显示出潜力，但它们的性能通常局限于特定的低光环境，这使得泛化到其他夜间应用具有挑战性。因此，训练一个用于低光单目深度估计的基础模型变得至关重要。这样的模型可以提高深度估计在不同低光场景中的泛化能力，为各种实际应用提供更稳定和准确的深度信息。

1.2 基础模型

视觉基础模型（VFMs）是经过大规模数据集训练的大型神经网络。VFMs 的显著扩展极大地推动了视觉理解的发展，使得在广泛的下游任务中进行高效的微调成为可能，仅需很少的努力。提示调优技术 [2, 20, 36, 39, 43] 显示，通过设计合适的提示，VFMs 可以有效地适应特定场景。特征适应方法 [2, 8, 11, 14, 44, 45] 进一步增强了 VFMs 在不同任务中的适用性。VPD [44] 展示了从预训练的文本到图像模型中提取特征进行领域特定深度估计的潜力。同时，I-LoRA [8] 展示了预训练图像生成器的多模态能力。Depth Anything [33] 率先使用教师模型训练来构建大规模数据集，成为单目深度估计的第一个基础模型。Marigold [16] 使用了稳定扩散 [25] 来生成具有完整深度信息的合成图像，进一步提高了合成图像标签的质量。Depth Anything V2 [34] 用合成图像替换了有标签的真实世界图像，并增强了教师模型的能力，开发出了一种先进的深度估计模型。然而，Depth Anything V2 在低光条件下表现不佳，表明该模型在具有挑战性的低光环境中的泛化能力有限。

为了解决上述挑战，我们提出了一种用于低光深度估计的高效基础模型，DepthDark。该方法通过引入眩光模拟模块和噪声模拟模块，构建了一个高质量的合成训练集，该训练集包括 74k 对昼夜深度数据。此外，我们还引入了一种高效的参数高效微调策略。

2 方法

单目深度估计基础模型，如 Depth Anything [33] 和 Depth Anything V2 [34]，利用数据引擎来收集和自动标注大规模未标记数据。然而，这些方法缺乏对夜间场景的有效建模，限制了它们在低光条件下的下游任务性能。为了克服这一局限，我们旨在开发一种专门为低光单目深度估计设计的强大基础模型。我们首次引入了低光数据集生成（LLDG），其集成了眩光模拟模块和噪声模拟模块技术，精确模拟了低光条件下的成像过程，从而生成高质量的低光场景成对深度数据集，详细内容见第 2.1 节。此外，为了进一步增强在低光场景中的鲁棒性和泛化能力，我们提出了一种高效的低光参数高效微调（LLPEFT）策略，专为低光场景量身定制。该策略利用光照引导和多尺度特征融合，显著增强模型在低光条件下的适应性和鲁棒性，详细内容见第 2.2 节。DepthDark 的整体框架如图 0.1 所示。

2.1 低光照数据集生成

为了应对昼夜之间图像分布的显著差异，我们提出了低光数据集生成（LLDG）。该框架通过融合两个关键组件，从白天图像中合成真实的夜间图像分布：耀斑模拟模块（FSM）和噪声模拟模块（NSM）。这些组件利用物理先验知识来模拟夜间场景中特有的噪声和光度特性，从而能够创建高质量的低光场景配对训练数据集。

光源的不均匀分布和夜间场景中显著的光学像差对深度估计模型提出了挑战。为了解决这个问题，我们引入了眩光模拟模块 (FSM) 作为低光照数据集生成的核心组件，它可以准确地模拟夜间图像中由于眩光、炫目和亮度峰值引起的光度不一致，从而生成更真实的合成数据集。这有助于弥合现有模型忽视光学缺陷所留下的差距，增强深度估计模型在低光环境下的鲁棒性和泛化能力。

在计算机图形学领域，一些方法使用 2D 傅里叶变换来近似这种光学现象。然而，这些方法通常会产生不可控的输出且计算成本很高。因此，我们不是直接模拟光源图像，而是基于 Flare7K 数据集构建了一个光源库。Flare7K 是目前唯一包含 7,000 个光源样本的大规模数据集。可以从该数据集中随机采样一个光源，并调整其大小以匹配输入图像的尺寸，以确保兼容性。另外，我们应用了简单的图像增强操作，如调整大小和裁剪，以多样化光源模式，增强后的光源记作 L_S 。

采样的光源在 3D 场景中随机定位，以模拟真实的深度关系。应用深度约束以限制光源的有效范围，确保远处的光源不

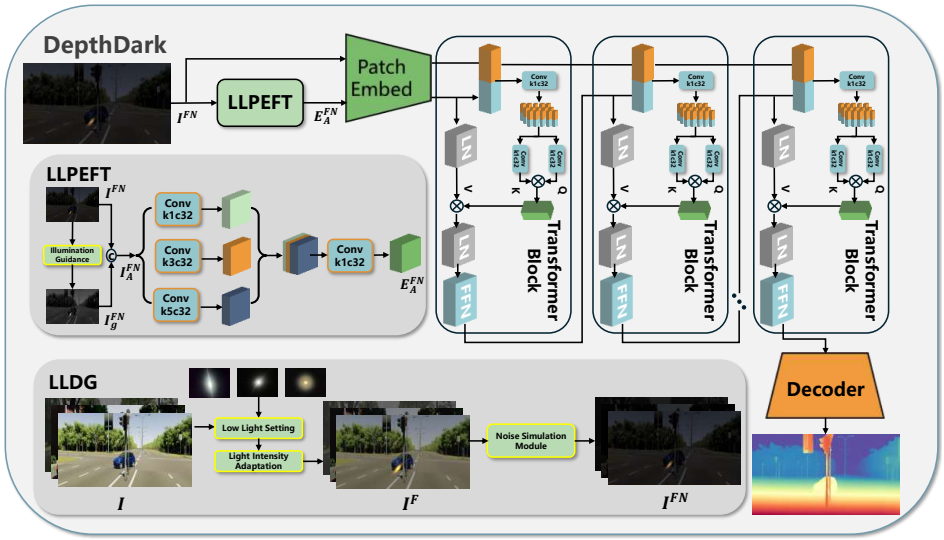

Figure 0.1: 我们 DepthDark 训练框架的概述。该框架由 LLDG 和 LLPEFT 组成。Patch Embed 模块的灵感来自视觉变换器 [7]，用于高效特征提取。

会产生过多的炫光伪影。光源的 3D 位置由

$$P_i = z_i \cdot K_I^{-1} \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix}, \tag{1}$$

给出，其中 z_i 表示光源的深度，其上限为 20 米，K_I 是内在相机矩阵，而 (u_i, v_i) 是光源的 2D 坐标。

由于白天的图像通常表现出足够的照明和高整体亮度，直接应用光源成像 L_S 往往难以产生显著的亮度峰值。

为了解决这个问题，我们对输入图像应用一个简单的随机变暗操作，其中亮度缩放因子 s_b 从均匀分布中采样。为了丰富光源的多样性，同时避免过于刺眼的效果，峰值亮度强度 F 定义为光源数量 N_F 与缩放因子 s_F 的乘积，其中 s_F 和 F 均从对数均匀分布中采样：

$$\begin{aligned} s_F &\sim \exp\left(U(\log s_F^{min}, \log s_F^{max})\right), \\ F &\sim \exp\left(U(\log F^{min}, \log F^{max})\right). \end{aligned} \tag{2}$$

，其中 s_F 是缩放因子，F^{min}, F^{max} 表示强度范围。光源的总数量 N_F 计算为：

$$N_F = \max\left(\left\lfloor \frac{F}{s_F} + 0.5 \right\rfloor, 1\right). \tag{3}$$

。此外，亮度比例 s_b 被采样为 $s_b \sim U(0.4, 1)$，伽马校正因子被采样为 $g_F \sim U(1.8, 2.2)$。使用 Phong 照明模型 [22, 32]，当前阶段的最终眩光图像 I^F 通过以下公式计算：

$$I^F = (s_b \cdot I)^{g_F} + \sum_{i=1}^{N_F} (ss(L_S, s_F, P_i))^{g_F}, \tag{4}$$

，其中 I 是白天图像，$ss(.)$ 表示通过缩放率 s_F 和二维坐标 P_i 对采样的 L_S 进行缩放和移动。

最终，FSM 通过结合现有的合成眩光图像、应用随机变暗操作以及采用自适应光源强度策略来生成更逼真的夜间眩光图像。该方法有效缓解了夜间图像中的光度不一致问题，并显著提升了训练数据集的多样性和质量。

2.1.1 **噪声模拟模块**. 在上一节中，我们推导了 I^F 的计算过程。然而，低光图像由于相机传感器中的低光子计数和高系统增益，往往会受到放大噪声的影响。受到噪声形成物理先验的启发，我们引入了噪声模拟模块（NSM）技术，以模拟夜间场景中的噪声分布。基于拍摄-读取噪声模型 [10, 30, 32]，我们将耀斑图像作为输入，在物理上解耦的方式下，在低光条件下精确模拟噪声。这种方法确保了合成的图像与实际低光场景中捕获的图像非常相似。因此，最终的低光图像 I^{FN} 可以通过以下公式表示：

$$I^{FN} = I^F + N, \tag{5}$$

，其中 I^F 表示真实场景的耀斑图像，N 指代所有物理噪声的总和。根据用于基础图像去噪的物理噪声模型 [40]，总体噪声 N 可以分解为四个组成部分：光子拍摄噪声 N_p、读取噪声 N_{read}、行噪声 N_r 和量化噪声 N_q：

$$N = KN_p + N_{read} + N_r + N_q. \tag{6}$$

其中，K 表示总系统增益，其值由相机的 ISO 设置决定，N_p 代表光子散粒噪声，它取决于入射光强度并遵循泊松分布。N_{read} 是一个统一术语，涵盖了多种噪声源，包括暗电流噪声、热噪声和源跟随器噪声。N_r 通常表现为图像中的水平或垂直线条，被建模为应用于每个列或行的固定偏移，从均值为零的高斯分布中采样，方差为 λ_{row}。N_q 是由传感器的有限位深引入的，被建模为像素值的均匀取样，其方差由 λ_{quant} 参数化。以上所有参数均遵循 ELD [30] 中概述的设计原则，并且我们在图 2.1 中给出了 I^{FN} 的一些配对视觉示例。

2.2 低光参数高效微调

虽然我们通过 LLDG 合成了大规模的白天和弱光场景的深度配对数据，但我们观察到，直接微调基础模型在训练中表现出不稳定的损失收敛，而完整的模型优化需要大量的计算资源。为了解决这些挑战，我们引入了一种低光参数高效微调策略，以优化基础模型用于弱光深度估计。LLPEFT 策略结合了光照引导和多尺度特征融合，以指导和增强模型在具有挑战性的弱光条件下的性能。

Figure 2.1: 不同行景的成对视觉示例，I 表示正常光照条件下的图像，I^F 表示在正常光照图像上添加 FSM 的可视化结果，I^{FN} 表示同时添加 FSM 和 NSM 到正常光照图像的可视化结果。

2.2.1 **光照指导**. 为了应对上述问题，我们分析主要因素 [5, 38] 来自两个方面。首先，在黑暗场景中使用高 ISO 和长时间曝光设置不可避免地引入噪声伪影。其次，图像中不均匀的亮度放大了噪声伪影，导致曝光不足/过度和颜色失真。为了缓解在低光照条件下训练基础模型所带来的挑战，我们的方法在训练流程中注入了光照引导，以指导模型解决低光照图像中的噪声伪影和不均匀光度分布的问题。具体而言，我们在 LLPEFT 策略中引入了光照引导项 I_g^{FN} ，其定义如下：

$$I_g^{FN} = mean_c(LLDG(I)), \tag{7}$$

操作 $mean_c$ 计算每个像素在通道维度上的平均值。如图 0.1 所示，低光图像 I^{FN} 是通过使用 LLDG 从白天图像 I 合成的，而 I_g^{FN} 是通过光照引导从 I^{FN} 生成的。这个过程通过使用简单而有效的方法将低光图像转换为灰度表示，来有效地简化了光照引导程序。这种转换有助于减少噪声、增强亮度分布并简化信息结构。虽然这种方法导致颜色信息的丢失，但对最终结果影响不大，因为颜色信息已经在低光图像 I^{FN} 中得到了保留。

最终，通过引入这种照明引导，模型可以专注于学习稳健的特征表示，从而减轻长时间曝光和高 ISO 设置导致的噪声放大、颜色失真和伪影。

2.2.2 **多尺度特征融合**. 为了有效地将输入的低光图像 I^{FN} 及其光照引导图像 I_g^{FN} 从图像空间转换到特征空间，我们提出了一种增强的特征融合方法，该方法结合了光照引导。具体而言，我们首先沿通道维度连接低光图像和光照引导图像，形成一个新的增强低光图像，记为低光辅助图像 I_A^{FN} 。随后，我们设计并采用了一种多尺度特征融合方法。该方法捕捉多尺度的上下文信息并动态调整特征权重，确保提取的特征能够全面代表嵌入在低光图像及其光照引导中的多层次信息。详细过程如下：

首先，将低光图像 $I^{FN} \in \mathbb{R}^{H\times W\times C_1}$ 和光照引导 $I_g^{FN} \in \mathbb{R}^{H\times W\times C_2}$ 按照通道维度进行拼接，以生成低光辅助图像 I_A^{FN} ，其定义为：

$$I_A^{FN} = Concat(I^{FN}, I_g^{FN}), \tag{8}$$

这个图像将低光图像的视觉信息与照明引导的统计信息结合在一起，为后续特征提取提供了更全面的输入。

随后，对低光辅助图像 I_A^{FN} 进行多尺度特征融合处理。该方法由三个平行的卷积层组成，卷积核大小分别为 1×1 、3×3 和 5×5 ，每个卷积层具有 C_3 个通道。相应的输出特征表示为：

$$E_1 = Conv_{1\times1}(I_A^{FN}), \tag{9}$$

$$E_2 = Conv_{3\times3}(I_A^{FN}), \tag{10}$$

$$E_3 = Conv_{5\times5}(I_A^{FN}), \tag{11}$$

其中，$E_1, E_2, E_3 \in \mathbb{R}^{H\times W\times C_3}$ 代表使用不同感受野提取的多尺度特征。

为了动态地整合多尺度特征，我们引入了 Softmax 函数来处理来自不同感受野的特征。具体来说，我们为每个尺度特征计算注意力权重 α_i ，如下所示：

$$\alpha_i = Softmax(W_i \cdot E_i + b_i), \quad i \in \{1, 2, 3\}, \tag{12}$$

其中 W_i 和 b_i 是可学习的参数，α_i 用于测量第 i 个尺度特征的重要性。

随后，融合特征表示为：

$$E_{fused} = \sum_{i=1}^{3} \alpha_i \cdot E_i, \tag{13}$$

其中 $E_{fused} \in \mathbb{R}^{H\times W\times C_3}$ 。最后，应用一个 1×1 卷积层来降低通道维度，生成最终的特征表示 E_A^{FN} ：

$$E_A^{FN} = Conv_{1\times1}(E_{fused}), \tag{14}$$

提取的低光特征图 $E_A^{FN} \in \mathbb{R}^{H\times W\times C_1}$ 整合了来自低光输入图像及其照明引导图像的多级信息，形成了一个鲁棒的表示，为后续模块提供了可靠的特征支持。此外，提取的特征图和低光图像 I_{FN} 都通过卷积层处理，然后输入到补丁嵌入模块，接着通过视觉变换器 [7] 生成最终的深度图。最终，多尺度特征融合将低光图像 I_{FN} 及其照明指导 I_{FN}^g 从图像空间转换到特征空间，提取出以低光特征图 E_A^{FN} 形式出现的全面多级信息，这显著增强了深度估计模型在低光条件下的鲁棒性。

3 实验

3.1 实现细节

为了在低光条件下训练单目深度估计的基础模型，我们使用 PyTorch 框架实现了 DepthDark。在训练过程中，我们采用了 Depth Anything V2 的训练设置，并在 ZoeDepth [3] 管道中使用了 DPT 解码器 [23] ，该管道建立在 DINOv2 编码器之上。训练数据集中的所有图像都被调整为 518×518 的分辨率进行训练，最终我们的模型被微调用于下游的单目深度估计任务。此外，还采用了数据增强技术，包括随机裁剪和水平翻转，以增强训练数据集，从而确保在单个 Nvidia RTX 3090 GPU 上高效使用内存资源。

3.2 数据集

3.2.1 **训练数据集**. 由于我们观察到 KITTI 训练数据集中的真实数据在某些情况下是不可靠的，并且存在数据量有限和场景泛化能力弱的问题，所以它并不理想用于基础模型的训练。因此，我们选择了 Hypersim [24] 室内数据集和 Virtual KITTI [4] 合成户外数据集作为我们的训练数据。由于这些数据集仅包含白昼图像，我们采用了第 2.1 节中描述的技术来合成一个高质量、大规模的低光深度对齐数据集，以训练我们提出的 DepthDark 模型。

Type	Method	Train on	Train Res.	Max depth	ABS rel ↓	Sq rel ↓	RMSE ↓	RMSE log ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Test on nuScenes-Night											
DT	MonoViT[42]	N $d\&n$	640 × 320	60	1.726	93.031	30.321	2.183	0.143	0.291	0.437
	WSGD[28]	N $d\&n$	640 × 320	60	0.663	9.573	15.200	0.755	0.199	0.388	0.567
DA	ITDFA[41]	N $d\&n$	640 × 320	60	0.337	4.511	10.118	0.403	0.515	0.767	0.890
	RNW[29]	N $d\&n$	640 × 320	60	0.341	5.516	11.152	0.406	0.531	0.789	0.902
	ADDS[19]	N $d\&n$	640 × 320	60	0.299	4.790	10.372	0.371	0.620	0.814	0.907
G	ITDFA[41]	R $d\&n$	640 × 320	60	0.362	3.760	10.252	0.441	0.418	0.702	0.867
	RNW[29]	R $d\&n$	640 × 320	60	0.376	4.732	11.193	0.506	0.451	0.712	0.835
	ADDS[19]	R $d\&n$	640 × 320	60	0.322	4.401	10.584	0.397	0.527	0.786	0.892
	MonoFormer[1]	K	768 × 256	60	0.307	3.591	10.162	0.413	0.521	0.762	0.872
	TDDC[32]	K	768 × 256	60	0.259	3.147	8.547	0.344	0.641	0.850	0.928
	Depth Anything[33]	UD & LD	-	60	0.302	3.051	9.233	0.321	0.487	0.784	0.972
	Depth Anything V2[34]	RD & SD	-	60	0.272	2.551	8.576	0.304	0.518	0.843	0.971
	Ours DepthDark	H, VK	-	60	0.210	1.910	7.764	0.260	0.630	0.914	0.976
Test on RobotCar-Night											
DT	MonoViT[42]	R $d\&n$	640 × 320	40	0.513	13.558	9.867	0.479	0.588	0.846	0.918
	WSGD[28]	R $d\&n$	640 × 320	40	0.202	1.835	5.985	0.231	0.737	0.934	0.977
DA	ITDFA[41]	R $d\&n$	640 × 320	40	0.266	3.010	8.293	0.287	0.567	0.888	0.962
	ADDS[19]	R $d\&n$	640 × 320	40	0.209	2.179	6.808	0.254	0.704	0.918	0.965
	RNW[29]	R $d\&n$	640 × 320	40	0.197	1.789	5.896	0.234	0.742	0.930	0.972
G	ITDFA[41]	N $d\&n$	640 × 320	40	0.302	3.692	8.642	0.327	0.548	0.852	0.938
	ADDS[19]	N $d\&n$	640 × 320	40	0.265	3.651	8.700	0.309	0.640	0.870	0.945
	RNW[29]	N $d\&n$	640 × 320	40	0.237	2.958	8.187	0.298	0.683	0.885	0.948
	MonoFormer[1]	K	768 × 256	40	0.289	2.893	7.468	0.302	0.543	0.873	0.964
	TDDC[32]	K	768 × 256	40	0.210	1.515	5.386	0.238	0.676	0.936	0.980
	Depth Anything[33]	UD & LD	-	40	0.302	3.331	6.622	0.314	0.635	0.822	0.918
	Depth Anything V2[34]	RD & SD	-	40	0.235	2.474	6.239	0.268	0.697	0.868	0.946
	Ours DepthDark	H, VK	-	40	0.157	1.063	4.284	0.202	0.760	0.941	0.985

Table 3.1: 最好结果用粗体表示，而次佳结果是 underline 。H、VK、K、N 和 R 分别表示 Hypersim、Virtual KITTI、KITTI、nuScenes-Night 和 RobotCar-Night 数据集。UD 和 LD 指的是 Depth Anything 在 6200 万无标签数据集和 150 万有标签数据集上训练。RD 和 SD 表示 Depth Anything V2 在 6200 万真实数据集和 50 万合成数据集上训练。d 和 n 指的是 RNW [29] 和 ADDS [19] 所提出的昼夜训练分割。最大深度是指真实深度值的上限。

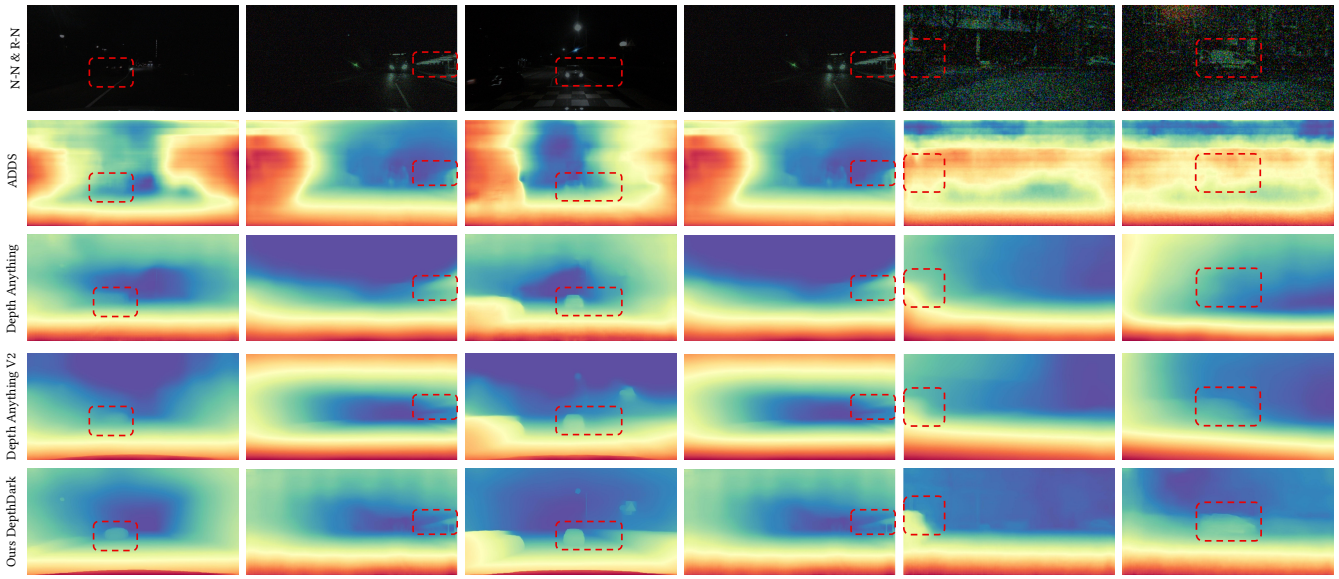

Figure 3.1: 不同单目深度估计方法在 nuScenes-Night 数据集上的定性比较结果。为简洁起见，我们展示了四种性能最具竞争力的先进方法的视觉比较，其中红色虚线框清晰地标示出了我们的方法表现出显著优势的区域。

3.2.2 **评估协议**. 为了评估我们的方法，我们选择了两个具有挑战性的基准数据集：nuScenes-Night 和 RobotCar-Night，据我们所知，它们是低光单目深度估计任务中最广泛使用的基准。nuScenes-Night 的评估设置遵循 RNW [29]，而 RobotCar-Night 的评估设置遵循 ADDS [19]。此外，在应用这些深度估计模型之前进行低光图像增强不仅增加了计算成本，还降低了模型的泛化能力。因此，我们主要将我们的实验结果与五种专为低光单目深度估计设计的方法进行比较。具体来说，直接训

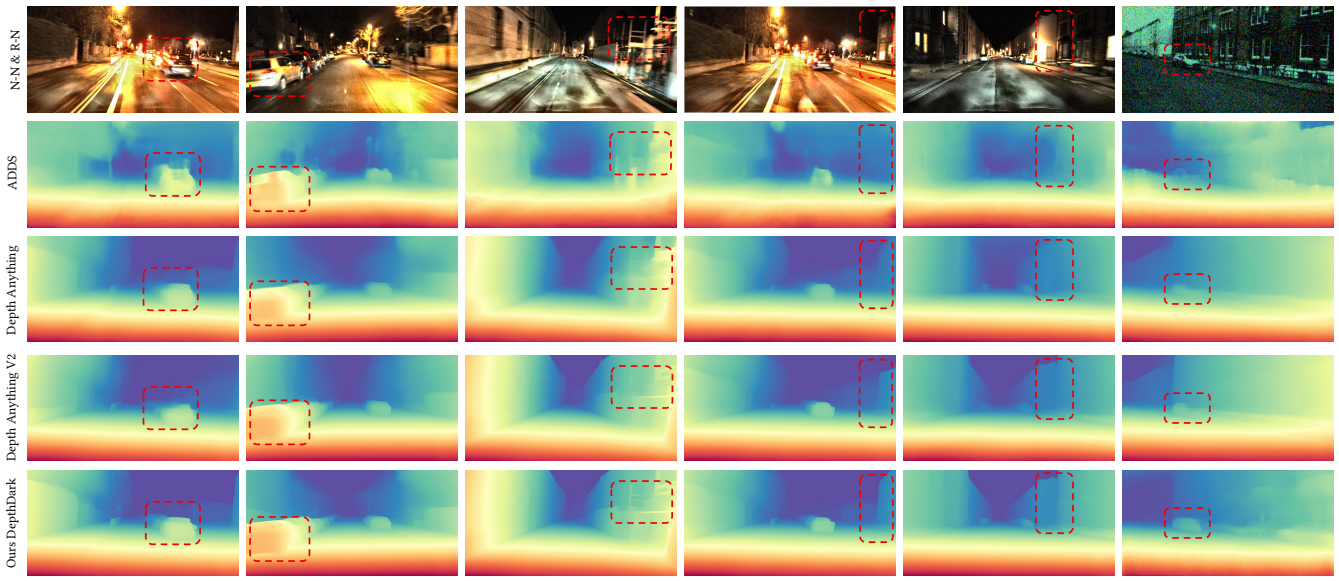

Figure 3.2: 不同单目深度估计算法在 RobotCar-Night 数据集上的定性比较结果。我们系统地评估了在不同噪声水平的场景下的性能，其中红色虚线框清楚地标出了我们的方法表现出显著性能优势的区域。

练（DT）方法包括 MonoViT [42] 和 WSGD [28]，而领域适应（DA）方法包括 RNW、ADDS 和 ITDFA [41]。评估设置基于 TDDC [32] 的协议。由于 TDDC 的官方代码未公开可用，我们使用他们论文中报告的实验结果。此外，我们尝试重现 TDDC 的方法，并为每种 DA 方法提供一般（G）版本的结果，以便进一步比较。我们的最终实验表明，G 版本的结果与 TDDC 中报告的结果大体一致。

为了进行更全面的比较，我们进一步将我们的实验结果与包括 Depth Anything、Depth Anything V2 和 TDDC 在内的最新方法进行比较。在 nuScenes-Night 和 RobotCar-Night 数据集上的评估表明，我们的方法在低光场景中表现出优越的鲁棒性。

3.3 与其他方法的比较。

3.3.1 **定量结果**。. 我们对 DepthDark 进行了定量评估，并与各种最先进的方法进行了比较。如表 3.1 所示，整体性能表明在 nuScenes-Night 数据集上的结果通常比 RobotCar-Night 更差。这主要是由于 RobotCar-Night 中 ISP 的强校正效果，这使得图像平面相比 nuScenes-Night 更加干净。

如表 3.1 所示，虽然我们的训练数据不包含自动驾驶领域的样本，但我们的方法在 nuScenes-Night 数据集上达到了最先进的性能（仅有 δ_1 指标排名第二），并在 RobotCar-Night 的所有指标上取得了全面领先的结果。这种显著的优势源于我们创新的微调策略，同时在 Depth Anything V2 上保留了在大规模合成数据集上预训练的核心参数，我们特别针对低光照条件优化了模型。

从详细结果来看，我们的方法在每一个评估指标上都优于其他所有方法，即使在完全未知的数据集上进行测试时也是如此，这对于低光条件下的深度估计尤为重要。此外，我们展示了其他方法的通用版本（G）的实验结果，这些方法在未知夜景场景中的准确性显著下降，表明其通用能力有限。相比之下，我们的方法成功克服了这一限制，展示了出色的鲁棒性和通用能力。

Method	ABS rel ↓	Sq rel ↓	RMSE ↓	RMSE log ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Test on nuScenes-Night							
Depth Anything V2	0.272	2.551	8.576	0.304	0.518	0.843	0.971
+ Only LLDG	0.264	2.956	9.209	0.371	0.581	0.813	0.914
+ Only LLPEFT	0.255	3.050	7.686	0.302	0.619	0.836	0.934
DepthDark	0.210	1.910	7.764	0.260	0.630	0.914	0.976
Test on RobotCar-Night							
Depth Anything V2	0.235	2.474	6.239	0.268	0.697	0.868	0.946
+ Only LLDG	0.183	1.679	6.769	0.251	0.702	0.893	0.972
+ Only LLPEFT	0.177	1.584	6.727	0.245	0.700	0.895	0.973
DepthDark	0.157	1.063	4.284	0.202	0.760	0.941	0.985

Table 3.2: 所提议模块的效能：在 nuScenes-Night 和 RobotCar-Night 上的消融研究表明，每个模块都能提升 DepthDark 的性能，结合使用可以显著提高稳健性。

为了全面评估我们的训练框架的有效性和效率，并确保与领域自适应方法进行公平比较，我们提供了各种单目深度估计方法在 RobotCar-Day 和 nuScenes-Day 数据集上的定量结果，以及它们在补充材料中的广义版本。

3.3.2 **定性结果**。. 为了进一步验证 DepthDark 的有效性，图 3.1 和 3.2 展示了在 nuScenes-Night 和 RobotCar-Night 低光数据集上的定性结果，对比了我们的方法与其他单目深度估计方法。

如 Figs. 3.1 和 3.2 的第一行所示，低光照条件由于光照不均和噪声伪影表现出显著的图像质量变化。这些导致亮斑和噪声斑块，严重影响物体轮廓和整体图像保真度，从而对单目深度估计构成挑战。然而，这种现象在实际的夜间环境中是常见的，并且对于实际应用至关重要。因此，我们选择了各种低光场景进行全面的测试和分析。为了确保对 DepthDark 的客观评估，我们仔细选择了对比方法。由于 TDDC 的官方代码不可用，我们选择了 ADDS，因为它是最先进和最具代表性的低光深度估计方法之一。

如图 3.1 和 3.2 所示，ADDS 生成的深度图与真实值显著偏离。尽管 Depth Anything 和 Depth Anything V2 能大致估计物体轮廓和深度信息，它们的预测仍呈现出明显的差异，特别是在车辆和建筑物的轮廓上。这是因为 ADDS 专为具有强人工照

Method	Parameter	ABS rel ↓	Sq rel ↓	RMSE ↓	RMSE log ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Test on nuScenes-Night								
Depth Anything V2	97.470M	0.264	2.956	9.209	0.371	0.581	0.813	0.914
+ AMFG	99.229M	0.243	2.351	5.416	0.267	0.674	0.862	0.945
+ LoRA	99.830M	0.233	2.112	5.220	0.259	0.685	0.865	0.947
+ LLPEFT	97.479M	0.210	1.910	7.764	0.260	0.630	0.914	0.976
Test on RobotCar-Night								
Depth Anything V2	97.470M	0.183	1.679	6.769	0.251	0.702	0.893	0.972
+ AMFG	99.229M	0.168	1.481	6.470	0.234	0.717	0.908	0.976
+ LoRA	99.830M	0.173	1.260	4.527	0.217	0.738	0.919	0.981
+ LLPEFT	97.479M	0.157	1.063	4.284	0.202	0.760	0.941	0.985

Table 3.3: 我们在 nuScenes-Night 和 RobotCar-Night 数据集上对 Depth Anything V2 的 PEFT 方法进行了全面的消融研究，并对每个微调变体的参数数量进行了定量分析。

明的夜间场景设计，而 Depth Anything 和 Depth Anything V2 则是在大规模白天图像数据集上训练的。

相比之下，我们的方法在各种低光条件下表现出色，生成更准确和更清晰的深度图。即使在存在强噪声和显著光度失真的情况下，我们的方法依然具有稳健性，强调了在低光场景中高效微调 Depth Anything V2 的必要性。

4 消融研究

4.1 每个组件的消融实验

如 Tab. 3.2 所示，本研究进行了系统的消融实验，以评估 LLDG 和 LLPEFT 模块引入的性能提升。实验结果表明，这两个模块在 nuScenes-Night 和 RobotCar-Night 数据集上的低光条件深度估计性能显著增强，特别是在 RobotCar-Night 上有显著的改进。具体而言，LLDG 生成的大规模配对低光数据集解决了低光基础模型训练数据收集的难题，使得此类模型的训练成为可能。LLPEFT 模块创新性地集成了光照引导和多尺度特征融合，显著增强了基础模型在极端低光条件下的鲁棒性。

4.2 参数高效微调的消融实验

为了全面评估 LLPEFT 的有效性，我们与具有代表性的 PEFT 方法进行了系统的比较。如表 3.3 所示，使用由 LLDG 生成的数据集，我们比较了各种 PEFT 方法对深度调整 V2 的效果，包括 RobustSam 用于低光 SAM 微调的 AMFG 方法 [6] 和经典的低秩适配（LoRA）模块 [14] 。在 RobotCar-Night 和 nuScenes-Night 数据集上的实验结果表明，LLPEFT 在性能上可以与其他先进方法媲美，同时引入的参数开销可以忽略不计。这一优势主要源于 LLPEFT 的新颖照明引导和多尺度特征融合机制，两者共同优化了光照感知和特征提取，能够有效解决低光条件下的噪声干扰和光照不均分布等关键挑战。

5 结论

在本文中，我们介绍了 DepthDark，这是一种在低光单目深度估计中的稳健基础模型。我们首先提出了眩光模拟模块和噪声模拟模块技术，以便准确模拟夜间条件下的成像过程，从而为低光环境生成高质量的配对深度数据集。此外，我们提出了一种有效的 PEFT 策略，该策略结合了照明引导和多尺度特征融合，增强了模型在低光条件下的稳健性和适应性。通过微调 Depth Anything V2，我们的方法 DepthDark 在 nuScenes-Night 和 RobotCar-Night 数据集上实现了最先进的性能，在低光场景的定性结果中超过了现有方法。最终，我们提出的 LLDG 和 LLPEFT 模块为微调大规模基础模型在低光单目深度估计中提供了一种新的可行方法。

6 致谢

This work was supported by the National Key Research and Development Program of China (U22A2047), the Key R & D Program of Zhejiang under Grant No. (2025C03001, 2023C01044), the Fundamental Research Funds for the Provincial Universities of Zhejiang (GK259909299001-023), the National Nature Science Foundation of China (62301198).

References

[1] Jinwoo Bae, Sungho Moon, and Sunghoon Im. 2023. Deep digging into the generalization of self-supervised monocular depth estimation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 187–196.

[2] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. 2022. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274* (2022).

[3] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288* (2023).

[4] Yohann Cabon, Naila Murray, and Martin Humenberger. 2020. Virtual kitti 2. *arXiv preprint arXiv:2001.10773* (2020).

[5] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. 2023. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12504–12513.

[6] Wei-Ting Chen, Yu-Jiet Vong, Sy-Yen Kuo, Sizhou Ma, and Jian Wang. 2024. RobustSAM: Segment Anything Robustly on Degraded Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4081–4091.

[7] Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[8] Xiaodan Du, Nicholas Kolkin, Greg Shakhnarovich, and Anand Bhattad. 2023. Generative models: What do they know? do they know things? let's find out! *arXiv preprint arXiv:2311.17137* (2023).

[9] David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems* 27 (2014).

[10] Hansen Feng, Lizhi Wang, Yuzhi Wang, and Hua Huang. 2022. Learnability enhancement for low-light raw denoising: Where paired real data meets noise modeling. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1436–1444.

[11] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision* 132, 2 (2024), 581–595.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[13] Derek Hoiem, Alexei A Efros, and Martial Hebert. 2007. Recovering surface layout from an image. *International Journal of Computer Vision* 75 (2007), 151–172.

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[15] Tao Huang, Songjiang Li, Xu Jia, Huchuan Lu, and Jianzhuang Liu. 2022. Neighbor2Neighbor: A self-supervised framework for deep image denoising. *IEEE Transactions on Image Processing* 31 (2022), 4023–4038.

[16] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. 2024. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9492–9502.

[17] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. 2015. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1119–1127.

[18] Ce Liu, Jenny Yuen, Antonio Torralba, Josef Sivic, and William T Freeman. 2008. Sift flow: Dense correspondence across different scenes. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part III 10*. Springer, 28–42.

[19] Lina Liu, Xibin Song, Mengmeng Wang, Yong Liu, and Liangjun Zhang. 2021. Self-supervised monocular depth estimation for all day images using domain separation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 12737–12746.

[20] Rongfeng Lu, Zunjie Zhu, Sheng Fu, Shenrong Chen, Tingyu Wang, Chenggang Yan, and Feng Xu. 2023. Self-supervised camera relocalization with hierarchical fern encoding. *IEEE Transactions on Instrumentation and Measurement* 73 (2023),

1–12.
[21] YiFan Lu, Ning Xie, and Heng Tao Shen. 2020. DMCR-GAN: adversarial denoising for monte carlo renderings with residual attention networks and hierarchical features modulation of auxiliary buffers. In *SIGGRAPH Asia 2020 Technical Communications*. 1–4.
[22] Bui Tuong Phong. 1998. Illumination for computer generated pictures. In *Seminal graphics: pioneering efforts that shaped the field*. 95–101.
[23] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. 2021. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*. 12179–12188.
[24] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. 2021. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10912–10922.
[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
[26] Shuwei Shao, Zhongcai Pei, Weihai Chen, Xingming Wu, and Zhengguo Li. 2023. Nddepth: Normal-distance assisted monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7931–7940.
[27] Madhu Vankadari, Sourav Garg, Anima Majumder, Swagat Kumar, and Ardhendu Behera. 2020. Unsupervised monocular depth estimation for night-time images using adversarial domain feature adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*. Springer, 443–459.
[28] Madhu Vankadari, Stuart Golodetz, Sourav Garg, Sangyun Shin, Andrew Markham, and Niki Trigoni. 2023. When the sun goes down: Repairing photometric losses for all-day depth estimation. In *Conference on Robot Learning*. PMLR, 1992–2003.
[29] Kun Wang, Zhenyu Zhang, Zhiqiang Yan, Xiang Li, Baobei Xu, Jun Li, and Jian Yang. 2021. Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark. In *Proceedings of the IEEE/CVF international conference on computer vision*. 16055–16064.
[30] Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. 2020. A physics-based noise formation model for extreme low-light raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2758–2767.
[31] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. 2020. Structure-guided ranking loss for single image depth prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 611–620.
[32] Haolin Yang, Chaoqiang Zhao, Lu Sheng, and Yang Tang. 2024. Self-Supervised Monocular Depth Estimation in the Dark: Towards Data Distribution Compensation. *arXiv preprint arXiv:2404.13854* (2024).
[33] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10371–10381.
[34] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth Anything V2. *arXiv preprint arXiv:2406.09414* (2024).
[35] Xiaodong Yang, Zhuang Ma, Zhiyu Ji, and Zhe Ren. 2023. Gedepth: Ground embedding for monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12719–12727.
[36] Hantao Yao, Rui Zhang, and Changsheng Xu. 2023. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6757–6767.
[37] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. 2019. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5684–5693.
[38] Jiaqi Yu, Yongwei Nie, Chengjiang Long, Wenjun Xu, Qing Zhang, and Guiqing Li. 2021. Monte Carlo denoising via auxiliary feature guided self-attention. *ACM Trans. Graph.* 40, 6 (2021), 273–1.
[39] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. 2023. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15211–15222.
[40] Yi Zhang, Hongwei Qin, Xiaogang Wang, and Hongsheng Li. 2021. Rethinking noise synthesis and modeling in raw denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4593–4601.
[41] Chaoqiang Zhao, Yang Tang, and Qiyu Sun. 2022. Unsupervised monocular depth estimation in highly complex environments. *IEEE Transactions on Emerging Topics in Computational Intelligence* 6, 5 (2022), 1237–1246.
[42] Chaoqiang Zhao, Youmin Zhang, Matteo Poggi, Fabio Tosi, Xianda Guo, Zheng Zhu, Guan Huang, Yang Tang, and Stefano Mattoccia. 2022. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *2022 international conference on 3D vision (3DV)*. IEEE, 668–678.
[43] Hengrun Zhao, Bolun Zheng, Shanxin Yuan, Hua Zhang, Chenggang Yan, Liang Li, and Gregory Slabaugh. 2022. CBREN: Convolutional Neural Networks for Constant Bit Rate Video Quality Enhancement. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 7 (2022), 4138–4149.
[44] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. 2023. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5729–5739.
[45] B Zheng, S Yuan, C Yan, X Tian, J Zhang, Y Sun, L Liu, A Leonardis, and G Slabaugh. 2022. Learning Frequency Domain Priors for Image Demoireing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2022), 7705–7717.