北京大學

本科生毕业论文

题目：**无监督学习范式下域迁移与场景重建技术研究**

Domain Adaptation and Scene Reconstruction under the Unsupervised Learning Paradigm

姓　　名：罗润冬

学　　号：2000013063

院　　系：信息科学技术学院

本科专业：计算机科学与技术

指导老师：刘家瑛 副教授

二〇二四年五月

北京大学本科毕业论文导师评阅表

<table>
<tr><td>学生姓名</td><td>罗润冬</td><td rowspan="2">本科专业</td><td rowspan="2">计算机科学与技术</td><td rowspan="2">论文成绩
（等级制）</td><td rowspan="2"></td></tr>
<tr><td>学生学号</td><td>2000013063</td></tr>
<tr><td>导师姓名</td><td>刘家瑛</td><td>导师单位/
所在学院</td><td>王选计算机研究所</td><td>导师职称</td><td>副教授</td></tr>
<tr><td rowspan="2">论文题目</td><td>中文</td><td colspan="4">无监督学习范式下域迁移与场景重建技术研究</td></tr>
<tr><td>英文</td><td colspan="4">Domain Adaptation and Scene Reconstruction under the Unsupervised Learning Paradigm</td></tr>
<tr><td colspan="6">导师评语
（包含对论文的性质、难度、分量、综合训练等是否符合培养目标的目的等评价）

该论文详细研究了无监督学习范式下的昼夜域迁移与三维物体发现两个计算机视觉任务。针对昼夜域迁移这一问题，该论文提出了相似度最小最大化的解决方案。该方案求解了一个双重优化问题，采用了涉及图像层面和模型层面的两阶段训练方法，并在多个夜间视觉任务和数据集上取得了最优的结果；针对三维物体发现的问题，该论文提出了一种以物体为中心的、基于位置无关表征的“无监督物体中心神经辐射场”框架，在无监督学习范式下取得了鲁棒的三维物体发现与场景重建效果。论文写作规范、行文准确、内容详实、条理清晰，对应两篇英语论文投稿或发表在计算机领域内顶级会议。该毕业论文体现了该生独立科研和学术表达的能力，达到本科学士学位培养目标。

导师签名：

年　　月　　日</td></tr>
</table>

版权声明

摘　要

无监督学习是机器学习的主要范式之一，其特点是不依赖于标注数据进行模型训练，而是通过发现数据内在的结构和关系来学习表征。在众多机器学习任务中，无监督学习的方法和理论都发挥着重要作用，尤其是在数据标注成本高昂或标注困难的情况下。本文聚焦于无监督学习范式下的低光高层视觉任务和三维物体发现与场景重建任务，即研究如何通过无监督学习的手段提升模型在低光数据上的表现，以及如何从单一图像中提取出以物体为中心的三维场景表示。本文的主要创新点如下：

1. 针对提升深度神经网络在夜间下游任务性能的问题，本文提出了一个不需要访问目标域（即夜间域）数据的、涉及两个层面的“相似性最小–最大”框架。该框架在图像层面生成一个与白天域特征相似性最小的合成夜间域，以放大域间隙；在模型层面最大化两个域中图像的特征相似性，以实现昼夜域适应。结合这一框架，本文还提出了一个稳定的训练流程，包括在图像层面使用曝光度引导的亮度调整模块和在模型层面通过多任务对比学习来对齐来自昼夜域图像的表征。实验结果表明，这一方法能在未接触真实夜间图像的情况下实现昼夜适应，并在多个高级夜间视觉任务上显示出显著的性能提升。

2. 针对从单一图像中进行三维物体发现与场景重建的问题，本文提出了“无监督物体中心神经场”框架。已有的研究在简单合成图像上进行无监督三维物体重建方面虽有所进展，但当面对背景复杂且物体种类多样的真实世界场景时，其性能往往受到限制。这一局限主要来源于现有方法所采用的物体表示，即将物体的内在属性（如形状和外观）与外在、观察者中心的属性（如 3D 位置）绑定。本研究通过解耦物体的内在属性和外在位置属性，显著提高了模型在复杂场景下的泛化能力。这一技术不仅模型能够从稀疏的真实世界图像中无监督地学习到准确的、以物体为中心三维场景表征，还支持了三维物体分割和场景操纵等应

用。进一步，本方法展现了从真实图像中识别并重建物体的零样本泛化能力，进而为场景理解和三维建模等领域的研究提供了新的视角和可能性。

关键词: 无监督学习、域泛化、物体发现、场景重建

Domain Adaptation and Scene Reconstruction under the Unsupervised Learning Paradigm

Rundong Luo (Computer Science and Technology)

Supervised by Professor Jiaying Liu

ABSTRACT

Unsupervised learning is one of the main paradigms in machine learning, referring to model training that does not rely on labeled data. Instead, it learns representations by discovering the intrinsic structures and relationships within data. In many machine learning tasks, the methods and theories of unsupervised learning play a critical role, and its value becomes even more prominent when the cost of data labeling is high or labeling is difficult. This paper focuses on low-light high-level vision tasks and 3D object discovery and scene reconstruction tasks under the unsupervised learning paradigm, *i.e.,* how to improve model performance on low-light data through unsupervised learning, and how to extract object-centric 3D scene representations from single images. The major contributions of this work are as follows:

1. Regarding the problem of enhancing deep neural networks' performance on nighttime scenarios, this work proposes a "similarity min-max" framework that does not require access to target domain (i.e., nighttime domain) data, involving two levels. The framework generates a synthetic nighttime domain that shares the minimum similarity to daytime domain features at the image level; at the model level, it achieves day-night domain adaptation by maximizing the feature similarity of images from the two domains. Combined with this framework, this paper also proposes a stable training process, including an exposure-guided illumination adjustment module at the image level and feature alignment of images from day and night domains through

multi-task contrastive learning at the model level. Experimental results show that the proposed method can achieve day-night adaptation without accessing real nighttime images and significantly improve performance on multiple nighttime high-level vision tasks.

2. Regarding the problem of 3D object discovery and scene reconstruction from single images, this paper proposes an "unsupervised discovery of object-centric neural field" (uOCF) framework. Although existing research has made progress in unsupervised 3D object discovery on simple synthetic images, difficulties arise when attempting to generalize these approaches to complex real-world scenes with complex backgrounds and diverse object types. The limitation lies in their object representations: they represent each object in the viewer's frame, entangling intrinsic object attributes such as shape and appearance with extrinsic properties such as the object's 3D location. Unlike existing methods, this research focuses on learning the intrinsics of objects and models the extrinsics separately. By decoupling the object's intrinsic properties from its external attributes, this method significantly improves the model's generalizability in complex scenes, allowing models to learn object-centric scene representations with high fidelity from sparse real-world images and supports applications like scene manipulation. More importantly, this method demonstrates the zero-shot generalization ability to identify and reconstruct objects from real images, thereby offering new perspectives and possibilities for research in scene understanding and 3D modeling.

Key Words: Unsupervised Learning, Domain Adaptation, Object Discovery, Scene Reconstruction

目录

第一章　引言

无监督学习是机器学习的主要范式之一，其目标是使模型在没有标注数据的情况下学习到数据的内在结构和分布。这种学习方式不仅减少了对大量手动标注数据的需求，还能探索数据中未被明确定义的模式和关系。在实际应用中，无监督学习特别适合处理那些难以获取大量标注信息的复杂数据，如图像、视频和文本等。特别是在计算机视觉领域，通过无监督学习技术训练得到的数据表征被广泛应用于各下游任务，如图像分类、目标检测、语义分割等，并展现出了强大泛化能力。在无监督学习的范式下，本文重点关注了昼夜域适应和三维物体发现与场景重建两个计算机视觉任务。

本章首先介绍选题的背景与研究意义。接着围绕本文研究的两个任务，分别介绍现有的技术和这些技术存在的问题，并概述本文提出的解决方案。最后将阐述本文的结构安排。

1.1　课题背景与研究意义

区别于监督学习需要使用大量的有标注数据训练模型，无监督学习旨在从无标注数据中学习到数据的内在结构和分布。传统无监督学习常用于聚类、降维、特征学习等任务，主要方法包括 K-Means 聚类, T-SNE 降维等；进入深度学习时代，无监督学习被广泛应用于计算机视觉、自然语言处理等多个领域。主要方法包括自编码器 (Autoencoders)、生成对抗网络 (Generative adversarial networks, GANs)、对比学习 (Contrastive learning) 等等。

自编码器通过学习输入数据的压缩表示来重构输出，进而达到特征抽取和数据降维的目的。在自然语言处理领域，从早期的词袋模型到现在的词嵌入技术如 Word2Vec 和 GloVe，无监督学习帮助模型捕捉到词汇之间的复杂语义关系。近年来，Transformer 架构的出现，尤其是预训练模型如 BERT 和 GPT 系列，进一步推动了无监督学习在语言模型中的应用；而在计算机视觉领域，自编码器也构成了潜在扩散模型 (Latent diffusion models) 等应用的模型基础。此外，自编码器也常用于去噪，即学习从带有噪声的数据中恢复出

清晰的数据。生成对抗网络则常用于图像生成、图像翻译等应用场景。例如，在医学图像处理中，生成对抗网络能够生成足够用于训练其他机器学习模型的高质量合成医疗图像。

对比学习作为一种新兴的无监督学习技术，通过比较不同样本之间的相似性或差异性来学习特征表示。这种方法在计算机视觉领域尤为有效，例如用于图像识别和图像检索。通过对比学习，模型能够无需显式标注即可辨识出图像中的相似和不同之处，这使得它非常适合处理未标注的大规模数据集。

此外，无监督学习也在改进传统模型和算法中发挥了重要作用。例如，聚类算法不仅限于传统的 K-Means 等方法，还发展出了基于密度的 DBSCAN、基于图的 Spectral Clustering 等新方法。这些算法在处理非均匀数据集、发现复杂多样的数据模式方面显示出更高的灵活性和效率。

总之，作为机器学习的一个重要分支，无监督学习技术与应用领域在近年来随着深度学习技术发展已经取得了巨大进步，并在处理未标注数据、理解数据内在结构、提升模型泛化能力方面均展现出了巨大潜力。未来，随着算法的进一步发展和计算能力的提升，无监督学习将无疑在人工智能领域发挥更加重要的作用。

1.2 低光场景下的计算机视觉任务

深度神经网络在图像处理和计算机视觉领域已经取得了巨大的成功，但其性能往往受到输入数据质量的限制。光照条件的变化，特别是低光环境，不仅影响了图像的可视性，也严重影响了深度学习模型的识别和分析能力，并对计算机视觉应用的安全性构成了重大威胁。如图 1.1所示，在主要由正常光照图像构成的数据集上训练的模型，在部署到低光图像上时，通常会给出错误的识别结果。直观上，低光增强方法 [1–6] 可以通过将低光图像恢复至正常光照状态在输入端初步解决识别准确率较低的问题。然而，这些低光增强模型往往是为了优化人类视觉感知而设计，而没有充分考虑机器视觉系统的需求，因此并不一定有助于下游的高级视觉任务。

在现有的研究中，无监督学习范式下的域适应技术已在提升视觉模型在

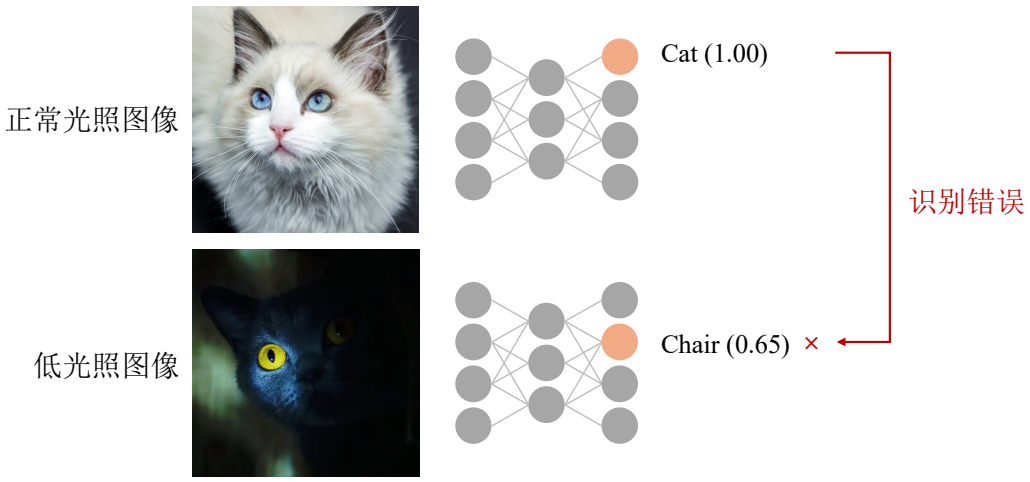


图 1.1 低光照条件导致的模型性能下降

夜间环境下的性能上取得显著进展。这些方法主要是通过图像转换 [7–9]、自监督学习 [10,11] 或多阶段算法 [12–14] 等技术手段来对齐夜间和白天数据集之间的分布统计数据，极大地提升了模型在夜间环境下的表现。这些域适应方法的一个核心假设是，目标域的数据是容易获取的。然而，在诸如深空探索和深海分析等特殊应用场景中，从任务特定的目标域中获取数据依然是一个巨大的挑战。

为减少对目标域数据的要求，**零样本**域适应已成为一个有前景的研究方向。其核心思想是在不访问目标域数据的情况下进行模型适应操作。当然，实现零样本域适应需要源域及目标域差异的先验知识，例如在本文所探讨的昼夜域适应问题中，零样本域适应要求已知目标域和源域的差异为光照条件的差异。表 1.1介绍了有监督学习、域适应、无监督域适应这三种学习范式间的差异。

在零样本域适应范式下，Lengyel 等人 [15] 提出了一种用于处理照明变化的颜色不变卷积。Cui 等人 [16] 设计了一个逆 ISP 流程，生成了带有伪标签的合成夜间图像。然而，图像级方法只是简单地将合成夜间视为伪标签数据，而忽略了模型；模型级方法关注调整模型架构，但忽略了光照变化对图像特征的影响。因此，这两种方法都不足以捕获那些能够弥合复杂的昼夜域间隙的、对照明鲁棒的表示。鉴于此，本文构建了一个涵盖两个层面的 “相似性最小–最大” 框架。

表 1.1 不同模型训练范式间可用训练数据的差异

模型训练范式	源域数据与标签	目标域数据	目标域标签
监督学习	✓	✓	✓
域适应	✓	✓	
零样本域适应	✓		

1.3 三维场景理解与重建

构建分解的、以物体为中心的三维场景表示是人类视觉的基本能力，也是计算机视觉和机器学习领域长期关注的话题。如图 1.2所示，三维物体发现与场景重建这一任务要求从二维图像中发现其中的物体并重建其三维结构，因而不仅对模型的视觉感知能力提出了较高的要求，还需要模型具备较强的生成能力。为充分研究模型在三维场景理解与物体重建任务上的潜力，近期的一些工作 [17–20] 开始探索从图像中无监督学习三维分解场景表示的可能性，并在简单的合成场景中取得了一定的成果。

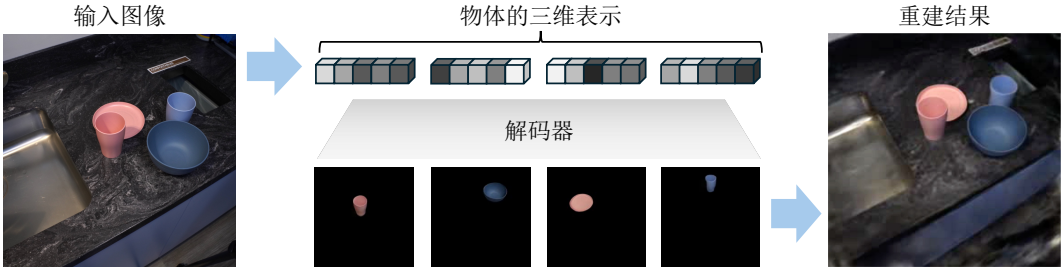


图 1.2 三维物体发现与场景重建任务

然而，这些方法在推广到复杂的真实世界场景时遇到了困难。这些困难主要源于物体的表征方式：这些方法将每个物体表示在观察者的视角中，将物体的内在属性（如形状和外观）与外在属性（如物体的 3D 位置）绑定在一起。这种范式意味着，物体位置的轻微改变或相机位置的微小调整都会显著改变物体的表征。直觉上，物体的内在属性应该与位置无关保持一致，但在现有的 3D 以物体为中心的学习模型中，这种不变性被忽视了。正如卷积网络 [21] 所示，考虑这种不变性对于泛化至关重要 [22,23]。而已有方法对于物

体内在和外在属性的绑定大大阻碍了模型的泛化能力。基于此，本文提出了一种“以物体为中心”的三维物体发现与场景重建方法。与现有方法不同，本方法专注于学习物体的内在属性，并将外在属性单独建模。

1.4　本文的结构安排

本文研究了无监督学习范式下的昼夜域适应和三维物体发现与场景重建两个任务，结构安排如下：

第一章为引言，首先概述了无监督学习的重要性和在机器学习中的应用背景，并明确了本文研究的主要内容及选题意义。

第二章综述了在无监督学习范式下，特别是在低光高层视觉和三维场景理解与重建方面的国内外研究现状和进展。这一章详细回顾了相关领域的重要研究成果和方法，分析了现有技术的优势与局限，并指出了当前研究中存在的问题和挑战。

第三章和第四章分别详细介绍了本文的两项主要研究工作。第三章聚焦于通过域适应技术提升视觉模型在夜间环境下的性能，详细描述了相似性最小-最大框架的设计思路、实现方法及实验结果，并给出了本方法和其他方法的比较与分析。第四章则针对三维物体发现与场景重建任务，提出了无监督发现物体中心神经场框架，并详细介绍了这一框架的技术细节，同时通过充分的实验展现了其优越性。

第五章总结全文，并对未来的研究方向进行展望。

第二章　国内外研究现状与工作基础

无监督学习是机器学习的核心范式之一，其目的是使模型在没有标记数据的情况下识别数据的内在结构和分布。该方法减少了对大量人工标注数据的依赖，并能发现数据中未明确定义的模式和联系。无监督学习尤其适用于处理难以获得大量标注信息的复杂数据集，例如图像、视频和文本等。在计算机视觉领域，利用无监督学习技术获得的数据表示广泛应用于多种下游任务，例如图像分类、目标检测、语义分割等，显示出了卓越的泛化能力。本章首先简要介绍无监督学习的基本内涵和主要技术；随后总结了低光高层视觉，包括基于低光增强的方法、基于域适应的方法、和基于零样本域适应的方法；最后梳理了三维物体发现与场景重建的算法，包括传统无监督物体发现方法，以物体为中心的场景重建方法，及生成式神经辐射场。

2.1　无监督学习

无监督学习是机器学习中的一大分支，它不依赖于标注数据进行学习，而是试图自动发现数据中的结构，可以用于聚类、降维、特征学习等多种任务。传统方法包含 K-Means，层次聚类，主成分分析、T-SNE 等。其中，K-means 是最基础、最广泛使用的无监督聚类算法之一。其核心思想是将数据集分成 K 个聚类，每个聚类由其内部点到聚类中心的平均距离最小化确定；PCA 则是一种常用的数据降维技术，被广泛应用于数据预处理、数据可视化以及发现数据的内在结构。该方法通过线性变换将数据转换到新的坐标系统中，使数据在各方向上的投影中方差最大者成为新的第一个坐标，第二大方差成为第二个坐标，依此类推。

近年来，深度学习的崛起促进了无监督学习，特别是其自监督学习分支的进一步发展。在计算机视觉领域，早期基于深度学习的自监督学习方法是将图像分割成多个块，然后将它们打乱，要求模型重新排列这些块到正确的位置。这种基于解决拼图任务的方法 [24] 能够使模型学习图像中的内在表示，从而更好地理解图像内容。基于对比学习的方法 [25–28] 通过拉近同一类样

本间表征的距离、拉远不同类样本间表征的距离进行模型训练。这种方法已被广泛应用于无监督特征学习，尤其是在图像和自然语言处理领域。近年来，基于掩码的学习方法 [29,30] 等通过随机遮挡输入数据的一部分，再要求模型预测被遮挡的部分以学习数据的表征。在 NLP 领域，基于 Transformer 的大模型如 GPT 和 BERT [31–33] 等充分展现了了无监督学习的强大能力。这些模型通过在大规模文本数据上进行无监督预训练以学习高质量的语言表征，从而在多种下游任务中取得了卓越的性能；此外，多模态模型如 CLIP 通过通过在大量的图像-文本对上进行无监督训练，将图像和语言数据的表征嵌入同一个隐空间，使模型能够理解图像内容与自然语言之间的关系。伴随着人工智能领域的高速发展，无监督学习将在进一步发挥其重要作用。

2.2 低光场景下的计算机视觉任务

给定在正常光照数据集上训练并服务于某一特定高层视觉任务的模型，其在夜间或低光环境数据集上的测试表现往往较差。直观上，低光增强技术可以通过将低光图像恢复至正常光照状态在输入端初步解决识别准确率较低的问题。然而，这些低光增强模型往往是为了优化人类视觉感知而设计，而没有充分考虑机器视觉系统的需求，因此并不一定有助于下游的高级视觉任务。

然而，从零开始构筑一个夜间数据集并在此基础上训练新的模型费时费力，相对地，正常光照下数据集资源非常丰富。因此，如何在不使用夜间数据标签的情况下将模型从正常光照迁移至低光照，即昼夜域适应，受到了学界的大量关注。昼夜域适应的主要思想是在仅使用来自目标夜间域的无标注数据的情况下，对模型中的参数进行微调。更进一步，在已知目标域为夜间低光域的情况下，零样本域自适应研究了考虑了一个更为困难的场景，即不适用来自夜间域的、和目标任务相关的任何数据，以减少对任务相关数据的需求。本节将对这三类方法分别展开介绍。

2.2.1 低光照增强

低光照增强旨在改善在光线不足的条件下拍摄的图像的人眼视觉体验。传统方法采用非学习技术，包括如直方图均衡化 [34]，伽玛校正以及基于 Retinex 理论 [35] 的算法等。以直方图均衡化为例，直方图用于表示数字图像中像素的每一灰度级别的出现频次，即图像像素的概率分布函数，其描述了图像灰度级别的分布情况。利用直方图可以观察图像的灰度范围、每个灰度级别的频度、整幅图像的平均明暗程度和对比度等图像灰度分布特性。具体而言，给定数字图像 I，其像素值取值范围为 $\{0,1,2,\cdots,i_{max}-1\}$（通常取 i_{max} 为 256），则图像 I 对于像素值 n 的直方图值 p_n 定义为：

$$p_n = \frac{\sum_{i\in I}\mathbf{1}_{i=n}}{N},\ n=0,1,\cdots,i_{max}-1. \tag{2.1}$$

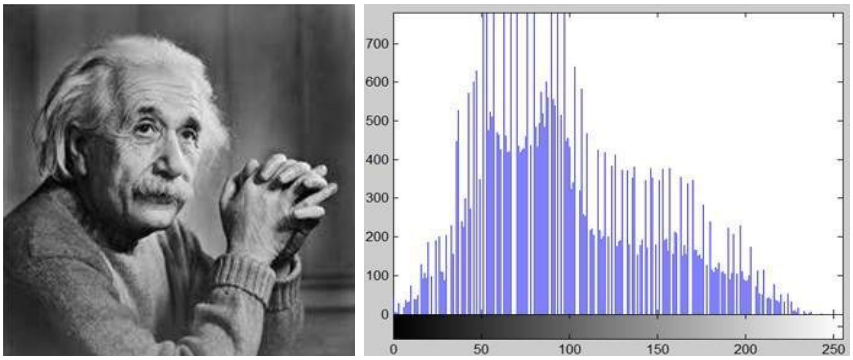
其中 N 为图像的像素总数。形式上，直方图通常以柱状图来表示。

直方图均衡化算法通过对图片的直方图进行全局调整来调整图片亮度。在低光照图像中大部分像素集中在低灰度区域，直方图均衡化对图像进行非线性拉伸，重新分配图像像素的灰度值，使一定灰度范围内像素的数量大致相等。具体而言，给定数字图像 I 及其直方图，直方图均衡化算法过程为：

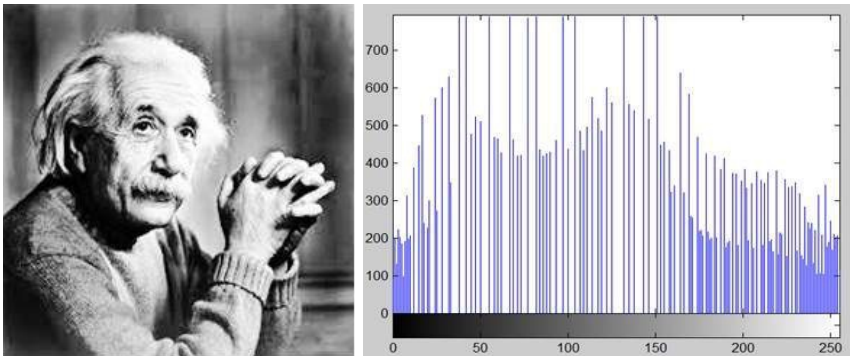
$$\hat{I}(x) = \text{floor}\left\{(i_{max}-1)\sum_{n=0}^{I(x)} p_n\right\}, \tag{2.2}$$

其中 floor$(\cdot)$ 表示向下取整，x 为像素所在位置而 $I(x),\hat{I}(x)$ 分别为直方图均衡化前后像素 x 的值。如图 2.1所示，直方图均衡化能有效地对图像进行全局调整，改善画面亮度和对比度，然而对于局部调整能力不佳。为克服这一弊端，研究者们引入了自适应的局部调整机制。Pizer 等人提出自适应直方图均衡化 [36]，即对于图像 I 以及任意给定的位置 x，统计其邻域窗口 $\delta(x)$ 的直方图 $p_n(x)$ 来进行像素值的转换，其中邻域窗口大小可以自定义，如 2×2□4×4 等等。然而，这一算法的时间复杂度较高。为解决这一问题，Pizer 等人提出在统计 $p_n(x)$ 后对窗口 $\delta(x)$ 中心的 $\delta(x)$ 个像素均进行转换，以此降低时间开

销，如统计 32×32 窗口内的直方图，而后将中心 16×16 个像素按照该直方图进行转换。事实上，自适应直方图均衡化在取全图为窗口，调整窗口内所有像素时退化为全局的直方图均衡化算法。



(a) 直方图均衡化前的图像及其直方图



(b) 直方图均衡化后的图像及其直方图

图 2.1 直方图均衡化结果示例

随着深度学习技术的发展，最新的研究大多基于深度学习。[37] 等研究采用了成对的训练数据模拟 Retinex 分解过程；EnlightenGAN [2] 采用了对抗学习范式来消除对成对数据的需求；DRBN [38] 设计了一个半监督框架，结合了监督和无监督方法的优点；RUAS [39] 展开了 Retinex 启示模型的优化过程，并使用神经架构搜索找到更好的网络架构；Zero-DCE [40] 为增强引入了二次曲线，其参数可以在没有正常光照图像进行参考的情况下学习。

而除了上述基于卷神经网络的方法，得益于注意力机制，Transformer 结构在图像增强及复原领域的各类任务上展现出更为优越的性能。Dudhane 等人 [41] 提出使用 Transformer 结构对于多张低光照图像的输入进行融合及增

强。其框架结合了邻域及全局特征，并设计了多尺度分层对齐模块来对多张图像进行预处理，并且提出了无参考特征增强模块来渐进地融合特征。在完成特征对齐和特征增强后，该方法还提出使用循环采样机制来加强帧间的信息融合，最终在多帧超分辨率、去噪以及低光照增强领域均获得了性能增益；扩散模型基于多步小尺度的加噪去噪过程完成分布间的转换，研究者通常使用扩散模型完成复杂分布如自然图像分布与简单分布如标准高斯分布之间的转换，从而构建从高斯噪声到自然图像的生成模型。得益于扩散模型展现的强大的生成能力，其产生的图像往往符合自然图像分布且具有较高质量。Kawar 等人 [42] 首次提出将扩散模型的强大生成能力应用于图像复原领域；Nguyen 等人 [43] 则将扩散模型应用于低光照场景下的文字图像增强，借由扩散模型的预测能力和高频信号生成能力还原文字的细节，该模型有效的提升了图像的质量以及文字可识别度，挖掘了扩散模型在病态逆问题求解方面的潜力。此外，包括频率分解 [44]，特征金字塔 [5,6]，和流模型 [45] 等技术也在近期的论文中被采用。除了上述针对低光 RGB 图像设计的提亮方法，也有部分工作研究了针对低光 RAW 图像 [46] 和视频 [47,48] 等数据形式的提亮。

2.2.2 昼夜域适应

除使用低光照增强模型进行预处理外，昼夜域适应，即使用无标注的夜间数据集对模型进行微调，也是一个被广泛探讨的解决方案。针对 RAW 图像上的物体检测任务，YOLO-in-the-Dark [13] 最早提出使用粘合层 (glue layer) 和对抗式生成模型组合不同域的预训练模型。如图 2.2所示，该框架首先在两域 A 和 B 上分别预训练模型，要求域 A 上训练的模型从输入 RAW 图像预测对应的 RGB 图像，而域 B 上训练的模型在 RGB 图像上进行物体检测，根据两个模型的特征边界提取一部分模型，然后通过粘合层进行接合。训练过程使用了基于生成模型的知识蒸馏，如图 2.3所示，通过将特征表示作为输入来训练粘合层，不需要借助其他的数据集。生成模型输出域 A 的特征，G1e-G1d 是教师模型，而 G2e 是学生模型。

同样针对低光目标检测任务，MAET [16] 框架将相机拍摄技术与机器学习融合，通过将 RAW 图像到 RGB 图像的变换与物体检测预测共同编码来学

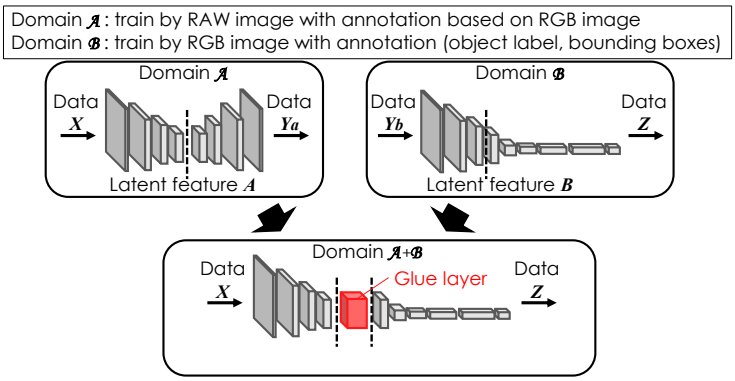


图 2.2 YOLO-in-the-Dark [13] 域迁移总框架图

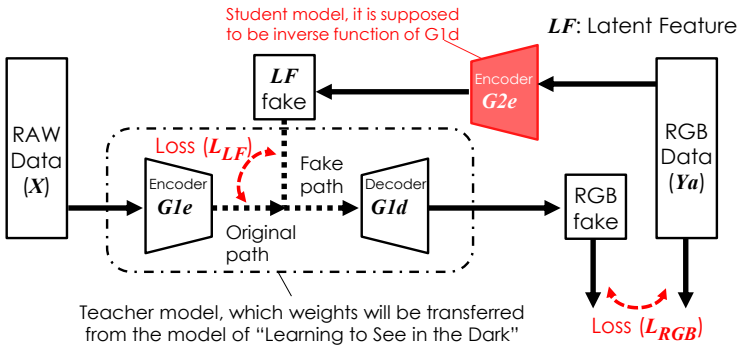


图 2.3 YOLO-in-the-Dark [13] 中基于知识蒸馏的粘合层训练方法

习特征，使模型能够捕获光照变换中的隐式特征。具体而言，MAET 框架假设图像信号处理过程 (image signal processing, ISP) 中的参数分布已知，依此从正常光照的 RGB 图像出发，通过反向 ISP 过程和 RAW 域的线性亮度降低操作获取低光照 RGB 图像。模型训练时，MAET 采用了多任务训练范式，即在优化物体检测任务的同时，要求模型解耦图像降质中使用的参数，以学习光照相关的信息学习视觉结构。为了避免两个任务特征的过度纠缠，MAET 采用了正交切线正则化约束特征，最大化不同任务输出的切线之间的正交性。MAET 框架可以广泛地应用于各类目标检测模型上，并取得显著的性能提升。

针对夜间街景识别，DANNet [49] 采用对抗学习来调整模型，并借助了暗光、微光和正常光照的不同亮度图像，如图 2.4所示。DANNet 的任务为将街景分割模型从仅包含白天数据的数据集 Cityscapes 迁移到同时包含白天、暗光和夜间数据的 Dark Zurich 数据集，且 Dark Zurich 数据集还包含了无标签

的利用 GPS 记录信息大致对齐的白天-暗光和白天-夜间图像对。研究者们提出首先将模型从 Cityscapes 迁移到 Dark Zurich 中的白天数据，然后将在 Dark Zurich 白天数据上的预测值作为 Dark Zurich 夜间数据的伪标签监督信息，用于训练域迁移模型。此外，DANNet 采用了权重共享的语义分割网络结构，还使用了提亮网络、对抗学习、和权重重新赋值策略。提亮网络让不同光照的图像趋于一致，让语义分割网络对光照不要过于敏感。判别器通过输出空间的对抗学习，判断语义预测结果是来自源域还是目标域。小物体的标注也更少，为了弥补这种偏差，使用了权重重新赋值策略。

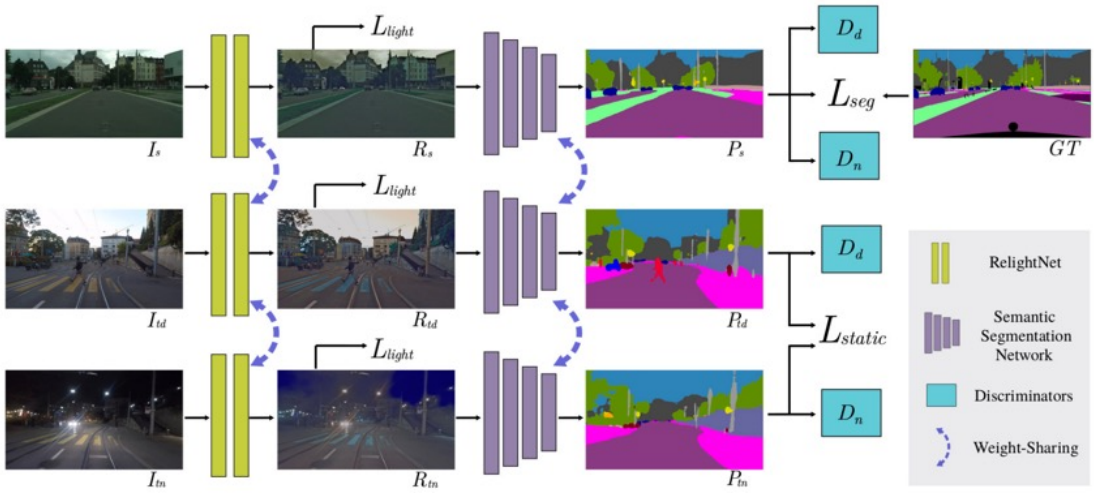


图 2.4 DANNet [49] 算法的模型框架

针对通用任务，Wang 等人提出了一种针对高层次视觉的可学习照明增强模型 SACC [11]。这一模型受到真实相机响应函数的启发，假设照明增强功能应为凹曲线，并通过离散积分来实现这一凹性。此外，为了适应机器视觉的照明需求，研究者们设计了一种非对称的跨域自监督训练策略，无需特定任务的标注数据。这一模型架构和训练设计的相互促进，形成了一个强大的无监督正常光至低光适应框架，如图 [11] 所示。

此外，一些研究工作也关注了低光照图像搜索 [50] 夜间深度估计 [51]，低光照图像匹配 [52] 等任务。

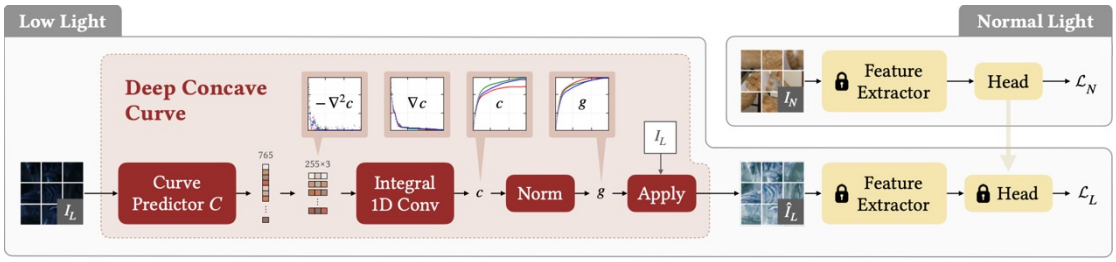


图 2.5 SACC [11] 的可学习照明增强曲线与自监督训练框架

2.2.3 零样本昼夜域适应

在传统域适应方法之外，零样本昼夜域适应方法考虑了一个更严格的条件，即无法获取真实的夜间图像。针对这一更严格的假设，Lengyal 等人 [15] 基于 Kubelka-Munk 理论的图像信息模型，推导出一种用来提取光照不变表征的边缘检测器。其描述物体反射出的光线 E 的幅度：

$$E(\lambda, x) = e(\lambda, x)((1 - \rho_f(x))^2 R_\infty(\lambda, x) + \rho_f(x)), \tag{2.3}$$

其中 x 是图像平面上的空间位置，λ 是光的波长，$e(\lambda, x)$ 是光源的辐照度，R_∞ 是材料反射性，ρ_f 是 Fresnel 反射系数。对上式进行简化，可以得到不同的形式。最终，将提取光照不变因子的过程嵌入神经网络，便能得到光照不变卷积层 (color invariant convolution, CIConv)：

$$\text{CIConv(x,y)} = \frac{(\log(CI^2(x, y, \sigma = 2^s) + \epsilon) - \mu_S}{\sigma_S}, \tag{2.4}$$

其中 CI 是光照不变因子，μ_S 和 σ_S 是在 $\log(CI^2 + \epsilon)$ 上的样本均值和方差，ϵ 是维持数值计算稳定性的小量。

此外，前文所述 MAET 框架 [16] 在仅适用合成夜间数据微调模型时，也可以视为零样本方法。此外，域泛化方法 [53–59] 也适用于这一设定。

尽管有了这些进展，低光增强方法注重提升图像的视觉效果，但忽略了下游的夜间视觉任务；传统域适应方法需要任务特定的夜间数据集，给数据收集带来了额外的负担，并限制了它们对多个任务的泛化能力；已有的零样本适应方法的性能则较为一般。与已有方法不同，本方法从从图像层面和模

型层面综合建模了昼夜域迁移的问题，并基于此提出了一个有效的相似度最小-最大框架。

2.3　三维场景理解与重建

2.3.1　无监督物体发现

在深度学习兴起之前，用于物体发现 (object discovery) 的传统方法主要旨在从图像中定位相似的物体 [60, 61]，其中物体被定义为视觉词 (visual words) 或图像块的聚类 (cluster of patches) [62, 63]。这种聚类概念后来被纳入深度学习技术中，以改善分组结果 [64, 65]。深度概率推理模型的引入推动了该领域向分解场景表示学习发展 [66]。这些方法将视觉场景分解为几个组成部分，其中物体通常被建模为可以解码成图像块 [67–70]、场景混合物 [71–78] 或层 [79] 的潜在表征（latent code)。尽管它们在场景分解中非常有效，但它们没有模拟物体的 3D 性质。

为了模拟场景和物体的 3D 性质，一些最近的工作尝试从单一场景的多视角图像 [80] 或大型数据集中学习 3D 感知表示，以实现泛化 [81–83]，而最新的研究强调单图像推理的以物体为中心的分解场景表示 [17–19]。值得注意的是，Yu 等人 [18] 提出了从单一图像中发现物体并重建场景的方法 uORF。后续文献改进了效率 [19] 和分割 [20]。然而，它们的物体表示受到外部纠缠的影响，并且在复杂场景中的泛化能力有限。与之相反，我们的方法明确地将物体外观与它们的外部属性分离。

2.3.2　以物体为中心的场景重建

构建分解的、以物体为中心的三维场景表示是人类视觉的基本能力，也是计算机视觉和机器学习领域长期关注的话题。近年来，一系列工作 [84–87] 初步探索了逐物体分解视觉场景并估计这些物体的语义/几何属性这一任务。其中的代表性工作，如 AutoRF [88]，成功地从标注图像中重建特定物体（如汽车)。其他方法将视觉场景分解为背景和由神经场表示的单个物体物体 [89, 90]。然而，这些工作都需要有标注数据进行训练，与本文所关注的无监督学

习范式不同。另一系列最新的工作探究了如何分割由神经辐射场或高斯点染表示的 3D 场景，例如 [91–93]。与之相反，我们的工作仅需要单张图像进行推理，而这些研究集中于多视角重建。

2.3.3 生成式神经场

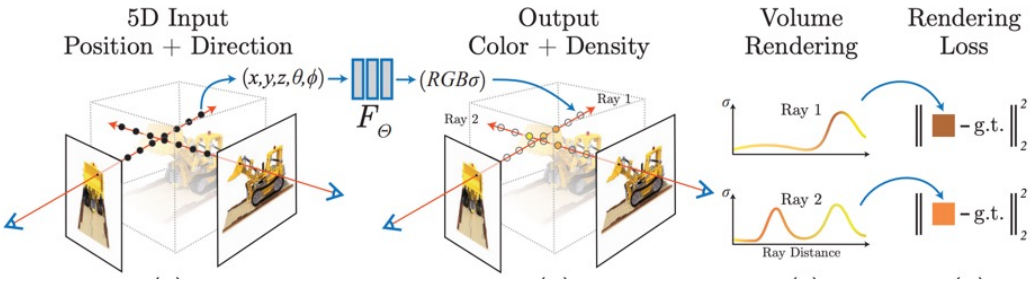


图 2.6 神经辐射场（NeRF）技术框架 [94]

神经场（Neural Fields）深刻地改变 3D 场景建模这一领域的研究重心。早期神经场方面的工作 [95, 96] 在三维场景表征上取得了初步令人满意的结果。神经辐射场（NeRF） [94] 和三维高斯点染（3DGS） [97] 两项开创性技术的提出引发了关于神经场的大量研究，并进一步促进了 3D 场景建模领域的高速发展。如图 2.6所示，NeRF 利用神经网络来建模和渲染复杂场景的辐射场（radiance fields），该神经网络以空间中任意点的坐标和观察视角作为输入，输出该点的颜色和密度，进而通过体渲染（volume rendering）技术从任何视角渲染出新的图片。NeRF 需要多视角图像作为训练数据，并通过最小化同一视角下渲染出的图像和真实图像之间的差异进行模型训练。本文所提出的方法采用 NeRF 作为物体的三维表征，并支持从单张图像中进行推理。

2.4 本章小结

本章以无监督学习为切入点，介绍了低光高层视觉及三维物体发现与场景重建两个研究领域的代表性工作进行了简要介绍，并分析了这些方法的特点和问题，为后文所作的改进工作进行铺垫。

第三章　基于相似性最小最大化的零样本昼夜域适应方法

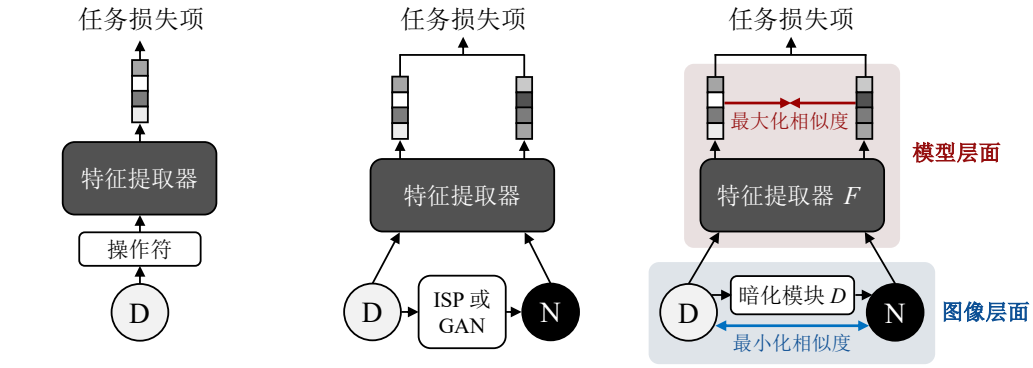
本章将探究如何通过无监督昼夜域适应的方法提升模型在夜间视觉任务上的性能。与传统昼夜域适应算法不同，本章重点关注零样本昼夜域适应这一情形，即在域适应的过程中不访问任何目标域数据。昼夜域适应的主要挑战在于学习对照明鲁棒的表示，以便泛化至昼夜两种模态。为在零样本约束下实现这一目标，[15] 引入一种用于处理照明变化的颜色不变卷积。[16] 则设计了一个逆 ISP 流程以生成带有伪标签的合成夜间图像。然而，图像层面方法只是简单地将合成夜间视为伪标签数据，并忽略了模型级特征提取；模型层面方法则关注调整模型架构，而忽略了图像的夜间特征。这两者都不足以捕获那些能够弥合复杂的昼夜域间隙的、对光照条件鲁棒的表示。

基于这一视角，本方法构建了一个涉及两个层面的相似性最小–最大框架。在图像层面上，本方法生成一个合成的夜间域，该域与白天域的特征相似性最小，以放大域间隙。在模型层面上，本方法通过最大化两个域中图像的特征相似性，以实现昼夜域适应。

3.1　问题定义与建模

如图 3.1所示，现有昼夜域适应方法通常分为基于运算符和基于图像暗化两类，但它们均存在一定缺陷。基于运算符的方法 [15] 依赖于**模型层面**手动设计的运算符来处理光照变化，但这些运算符无法适应真实的复杂场景。基于图像暗化的方法通过在**图像层面**应用 ISP [16] 或 GAN [7, 8, 12, 98] 等技术，将带标签的白天数据转换为夜晚数据。然而，前者依赖于传感器并且不能在设备和数据集之间泛化，而后者需要来自任务特定夜间域的数据，因此无法适用于本方法的零样本设定中。

本质上，现有方法的最关键问题是忽略了**像素**和**特征**之间的相互影响。本工作首次对这个问题进行了系统性的研究，并提出了一个充分利用双方信息的相似性最小最大框架。具体来说，**在像素（图像）层面**，本方法通过白天到夜晚的转换最小化原始图像和暗化图像之间的特征相似性。而**在特征（模**



(a) 基于运算符的方法　(b) 基于图像暗化的方法　(c) 本章节提出的方法

图 3.1 本章节提出的网络架构示意图

型）层面，本方法通过表示对齐来最大化特征相似性。这种联合优化带来了对光照变化更鲁棒的表示。

如图 3.2所示，本方法构建了包含 D 和 F 的统一框架。具体而言，令下游模型的特征提取器为 $F(\cdot)$。对光照具有鲁棒性要求白天图像 I 及其夜晚版本 $D(I)$ 的提取特征相似，其中 $D(\cdot)$ 代表一个暗化过程。为解决现有基于暗化的方法中忽略 D 对于 F 的影响的问题，本方法对 D 引入额外的约束，即要求 D 最小化白天特征 $F(I)$ 和夜晚特征 $F(D(I))$ 之间的相似性。

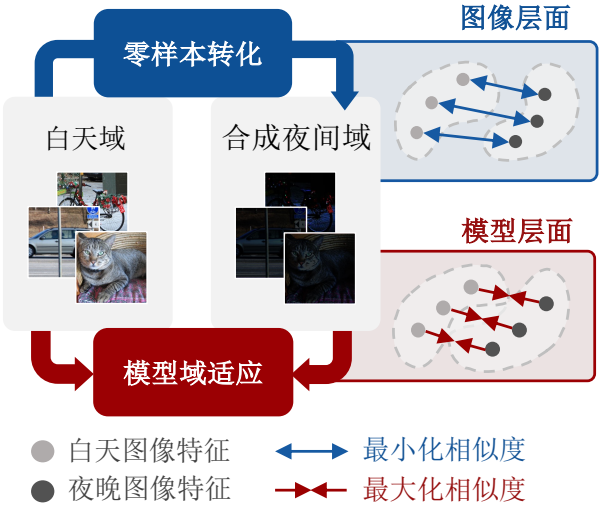


图 3.2 本章节提出的网络架构示意图

本方法可以将 D 和 F 集成为一个最小-最大优化问题：

$$\max_{\theta_F} \min_{\theta_D} \quad \mathrm{Sim}(F(I), F(D(I))), \tag{3.1}$$

其中 θ_D 和 θ_F 分别表示 D 和 F 中的参数，$\mathrm{Sim}(\cdot,\cdot)$ 测量特征之间的相似性。

然而，在公式 (3.1) 中存在平凡解，例如 D 生成完全黑色的图像，以及 F 为所有输入提取相同的特征。为解决这一问题，本方法分别为 D 和 F 添加了正则项：

$$\max_{\theta_F} \min_{\theta_D} \ \mathrm{Sim}(F(I), F(D(I))) + \mathcal{R}_D(\theta_D) - \mathcal{R}_F(\theta_F), \tag{3.2}$$

其中 $\mathcal{R}_D$ 和 $\mathcal{R}_F$ 旨在防止模型崩溃。

如何合适地设计 $\mathcal{R}_D$ 和 $\mathcal{R}_F$ 是解决公式 (3.2) 的关键。以下将介绍本方法如何设计 $\mathcal{R}_D$ 和 $\mathcal{R}_F$，以及构建整个学习框架。

3.2　基于相似性最小最大化的零样本昼夜域适应框架

3.2.1　图像层面：相似性最小化

本节描述暗化模块 D 的设计。D 应当满足三个属性：

- **稳定性.** 公式 (3.2) 中应使用适当的 $\mathcal{R}_D$ 以防止模型训练崩溃。
- **泛化性.** D 应代表一个泛化的暗化过程，这样下游模型可以从 $D(I)$ 中学习对处理未见过的夜间场景有用的知识。
- **灵活性.** 本方法希望对暗化的程度有灵活的控制，以生成对优化 F 有利的多样输入。

本方法设计了一个曝光度引导的逐像素映射算法来满足上述属性。与依赖实际夜间图像的广泛使用的图像到图像的暗化方法 [7,12,98] 不同，逐像素映射使用一个具有可学习参数的预选函数来调整图像。

本方法经验性地发现，通过对映射函数设置适当的约束，本方法可以自然地避免在相似性最小–最大优化中获得平凡的解（稳定性），并保证 D 遵循

一个典型的暗化过程（泛化性）。最后，本方法为了更好的灵活性添加了一个曝光度引导机制。详细设计将如下所示。

暗化过程. 本方法首先定义一个通用的色调映射函数。给定图像 $I \in [0,1]^{C\cdot H\cdot W}$，本方法使用一个非线性映射 $f : [0,1] \to [0,1]$ 和一个逐像素的调整映射 $\mathcal{A} \in [0,1]^{C\cdot H\cdot W}$ 来处理图像：

$$D^0(I) = f(I, \mathcal{A}), \tag{3.3}$$

通常，f 应当是单调递增的以保持对比度，并满足对于所有的 α 都有 $f(1,\alpha) = 1$ 以避免信息丢失（例如，伽马校正）。然而，对于暗化，后一个约束 $f(1,\alpha) = 1$ 不再适用。因此，本方法采用一个额外的映射单调递增函数 $g : [0,1] \to [0,1]$ 作为辅助，该函数由 $\mathcal{B} \in [0,1]^{C\cdot H\cdot W}$ 参数化。整体的暗化过程可以表述为：

$$D(I) = g^{-1}(f(g(I, \mathcal{B}), \mathcal{A}), \mathcal{B}), \tag{3.4}$$

其中 $\mathcal{A}$ 和 $\mathcal{B}$ 都由一个以 I 为输入的映射估计器决定。

为了保证 D 代表一个暗化过程（即，$D(I) < I$），f 还应满足凸性。具体来说，本方法采用 f 为迭代二次曲线 [40]：$f(x) = h^{(8)}(x)$，$h(x,\alpha) = \alpha x^2 + (1-\alpha)x$，并且采用 g 为除法操作：$g(x,\beta) = x/\beta$。

相似性最小化. 模块 D 的训练目标包含两部分：相似性最小化和正则项。对于前者，本方法直接减小特征之间的距离：

$$\mathcal{L}_D^{sim} = \frac{\langle F(I), F(D(I))\rangle}{||F(I)||_2 \cdot ||F(D(I))||_2}, \tag{3.5}$$

其中 $\langle\cdot,\cdot\rangle$ 是两个向量之间的内积。

正则化项由四个损失组成。除了修正色彩偏差的色彩一致性损失 $\mathcal{L}_{col}$ [40] 外，本方法提出了三个额外的正则损失：

首先，采用条件曝光度控制来对齐曝光度和合成的图像：

$$\mathcal{L}_{c-exp} = \sum_{1\le i\le H, 1\le j\le W} |\hat{D}_{i,j}(I,E) - E_{i,j}|, \tag{3.6}$$

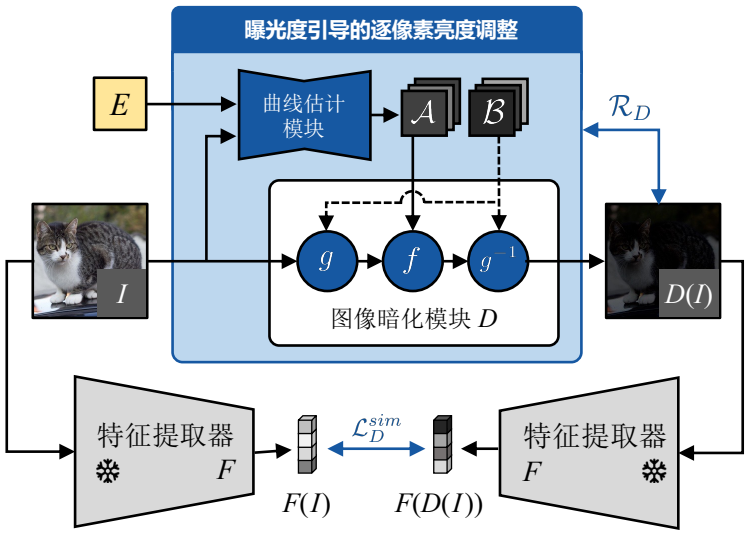


图 3.3 图像层面的相似性最小化模块

其中 $\hat{D}(I,E)$ 是 $D(I,E)$ 的通道间平均值。在训练过程中，E 的所有元素在 $[0,0.5]$ 之间随机抽取。

此外，本方法在 $\mathcal{A}$ 上添加约束。直觉上，$\mathcal{A}$ 表示亮度降低的程度。图像的亮度通常在一个场景中缓慢变化，但在图像中实体的边缘则变化较快。遵循这一属性，本方法添加了平滑总变分损失：

$$\begin{aligned} \mathcal{L}_{ltv}(\mathcal{A}) &= \sum_{c\in\{R,G,B\}} (h(|\nabla_x \mathcal{A}^c|)^2 + h(|\nabla_y \mathcal{A}^c|)^2), \\ h(x) &= \max(\alpha - |x-\alpha|, 0), \end{aligned} \tag{3.7}$$

其中 ∇_x, ∇_y 分别是沿着水平轴和垂直轴的梯度运算，α 是一个超参数。与原始的总变分损失相比，其中 h 是恒等函数，平滑后的版本允许网络为相邻像素预测更大的差值，这在物体的边界上是常见的。

最后，本方法加入 $\mathcal{L}_{flex}(\mathcal{B}) = 1 - \mathcal{B}$ 以避免模型仅通过 g 拟合曝光条件。综合以上设计，本方法设计了如图 3.3所示的网络结构，并通过优化如下损失函数进行模型训练：

$$\begin{aligned}\mathcal{L}_D &= \lambda_D^{sim}\mathcal{L}_D^{sim} + \mathcal{R}_D, \\ \mathcal{R}_D &= \lambda_{c-exp}\mathcal{L}_{c-exp} + \lambda_{col}\mathcal{L}_{col} + \lambda_{ltv}\mathcal{L}_{ltv} + \lambda_{flex}\mathcal{L}_{flex}.\end{aligned} \tag{3.8}$$

3.2.2 模型层面：相似性最大化

对比学习 [25,26] 是自监督学习范式之一，其训练思路是在降低正样本对的特征距离的同时增大负样本对的特征距离。然而，分类中的同类图像或分割中的相邻场景将形成伪负样本对，从而损害模型的性能。为了解决这一问题，BYOL [27] 提出了一个不采用负样本对的变体，它仅在正图像对 $\{v, v^+\}$ 之间对特征进行对齐：

$$\mathcal{L}_{\text{BYOL}}(v, v^+) = 2 - \frac{2 \cdot \langle z(q(F(v))), q'(F'(v^+))\rangle}{||z(q(F(v)))||_2 \cdot ||q'(F'(v^+))||_2}, \tag{3.9}$$

其中 q, q' 是映射模块，而 z 是预测模块。它们都是具有单个隐藏层的多层感知机。F' 和 q' 与 F 和 q 共享相同的架构和权重初始化，但不接收梯度，并通过指数移动平均（EMA）进行更新。

相似性最大化. 本方法采用 BYOL 中提出的方案，通过对比学习最大化白天域与合成夜晚域之间的特征相似性。给定白天图像 I 和曝光图 E，本方法将训练目标公式化如下：

$$\mathcal{L}_F^{sim} = \mathcal{L}_{\text{BYOL}}(I, D(I, E)) + \mathcal{L}_{\text{BYOL}}(D(I, E), I). \tag{3.10}$$

注意，特征相似性的度量在公式 (3.5) 和公式 (3.10) 之间是不同的。直接应用公式 (3.5) 训练 F 会引起特征退化并带来次优的结果。相对地，非对称的模块设计和停止梯度的策略能够防止特征提取器 F 崩溃，即作为公式 (3.2) 中的正则化项 $\mathcal{R}_F$ 与任务损失协同优化模型。

此外，与公式 (3.6) 中的 E 不同，本方法使用一个复合曝光图 E' 代替。首先，E' 的元素均匀采样于 $[0, 0.2]$ 之间，以模拟夜间照明。该范围在所有下游任务中一致，进而不引入与任务相关的先验。此外，本方法向 E 添加像素

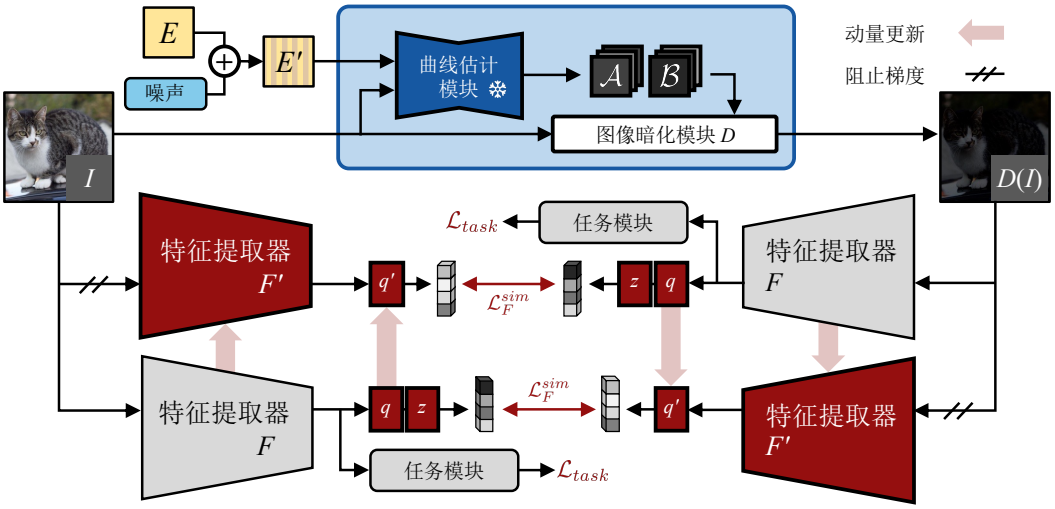


图 3.4 模型层面的相似性最大化模块

级噪声 z_1 和块级噪声 z_2 以模拟曝光差异。总的来说，E' 可以表示为：

$$E' = \mathcal{U}(0, 0.2) + z_1 + z_2. \tag{3.11}$$

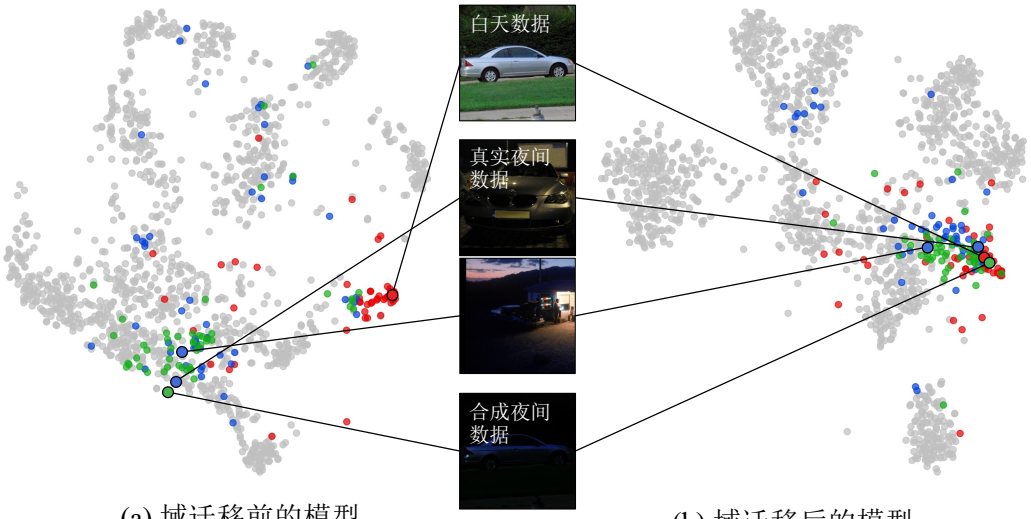
此外，本方法在白天和合成夜间域上添加了任务特定的监督信号 $\mathcal{L}_{task}$。总结而言，本节提出的网络结构如图 3.4所示，并以如下损失函数作为优化目标：

$$\mathcal{L}_F = \lambda_F^{sim}\mathcal{L}_F^{sim} + \lambda_{task}\mathcal{L}_{task}. \tag{3.12}$$

3.2.3　模型训练流程

前文介绍了图像层面的相似性最小化（第 3.2.1节）和模型层面的相似性最大化（第 3.2.2节）。本节讨论训练的总体流程。

采用 GAN [99,100] 中交替训练的技术的一种直观的方法。然而，平衡 D 和 F 增加了参数调整的难度，并使优化过程不稳定。本方法采用了一个简单但有效的两步策略来解决该问题：先训练模型 D 并保持模型 F 的权重固定，然后训练模型 F 并保持模型 D 的权重固定。具体而言，与交替策略相比，分步训练策略使得模型在夜间图像分类上的性能从 63.84% 提高到了 65.87%。



(a) 域迁移前的模型　　(b) 域迁移后的模型

图 3.5 白天数据（红色）、真实夜晚数据（蓝色）、合成夜晚数据（绿色）特征的降维可视化结果

在所有下游任务中，特征提取器和任务模块由白天预训练的模型初始化。本方法首先固定特征提取器并训练暗化模块（图像级转换）。然后，本方法保持暗化模块固定，并共同训练特征提取器和任务模块（模型级适应）。

3.2.4 暗化模块的经验性证明

在不接触实际夜间图像的情况下模拟夜间条件是本方法框架的关键。特别地，夜间条件除了照明变化外还带来语义变化，例如，图 3.5中第二张实际夜间图像的黑暗环境与人工灯光。然而，由于本方法的先验知识仅限于“低照度”，精确的模拟极其困难。幸运的是，与针对人类视觉体验的典型昼夜图像合成过程 [101] 不同，本方法只关心特征空间中暗化图像的分布。撇开视觉质量，本方法发现本方法合成的夜间域的特征分布与真实夜间域的特征分布相似，如图 3.5(a) 所示。这个观察表明，本方法的暗化过程可以从模型层面表征夜间域。

3.3　实验结果与分析

3.3.1　实验设置

本方法广泛应用于各种夜间视觉任务。下文选取了四个代表性任务进行评估：图像分类、语义分割、视觉位置识别和视频动作识别。只有白天的数据可用于训练和验证，而夜间数据只在评估期间使用。本文比较的方法包含三类：低光照增强、域泛化、零样本昼夜域适应。这些方法在训练时都不需要目标域的数据。本方法方法的结果是三次独立试验的平均值。

3.3.2　夜间图像分类

本方法首先考虑最基础的视觉任务之一：图像分类。CODaN [15] 是一个包含有 10000 张白天图像的训练集和一个分别有 2500 张白天和夜间图像的测试集的 10 类数据集。本方法在白天测试集上验证模型，并在夜间测试集上评估它们。实验采用的主干网络是 ResNet-18 [102]。

基准测试结果显示在表 3.1中。增强方法从人类视觉的角度恢复输入的低光图像，同时保持模型不变，从而获得有限的性能增益。域泛化方法被设计用于一般任务，而在未见过的夜间环境中表现不佳。MAET [16] 依赖于带有传感器特定参数的降级转换，从而在泛化能力上表现不佳。CIConv [15] 采用可学习的颜色不变边缘探测器，这些在真实场景中的复杂光照变化下并不鲁棒。相比之下，本方法大幅超越了此前最先进的方法（60.32% v.s. 65.87%），表明本方法的统一框架可以获得对光照变化更鲁棒的特征。

消融实验. 为了进一步验证本方法设计的框架，本节在 CODaN [15] 夜间测试集上进行了实验实验，并报告了实验的 Top-1 分类准确度。实验结果如表 3.2 所示。首先研究如何设计暗化模块 D。本实验考虑采用启发式的图像调整方法替换本方法的暗化模块，例如亮度调整（`PIL`[1] 中的 `Brightness`）和伽马校正（$D(I) = I^{\gamma}$）。如表 3.2所示，两种方法取得的实验结果均不如本方法设计的暗化模块。

[1] https://pillow.readthedocs.io/en/stable/reference/ImageEnhance.html

表 3.1 CODaN [15] 夜间测试集上的 Top-1 分类准确度

方法类别	方法	Top-1 (%)
基线方法	ResNet-18 [102]	53.32
低光照增强	EnlightenGAN [2]	56.68
	LEDNet [5]	57.40
	Zero-DCE++ [3]	57.96
	RUAS [39]	58.36
	SCI [4]	58.68
	URetinexNet [1]	58.72
域泛化	MixStyle [55]	53.12
	IRM [54]	54.52
	AdaBN [53]	54.25
零样本域适应	MAET [16]	56.48
	CIConv [15]	60.32
	本方法	**65.87**

随后，本实验探究了曲线 f 的其他可能形式。如表 3.2所示，伽马曲线（$f(x,\alpha)=x^{\frac{1}{\alpha}},\alpha\in(0,1]$）和倒数曲线（$f(x,\alpha)=\frac{(1-\alpha)\cdot x}{1-\alpha\cdot x},\alpha\in[0,1)$）均带来了比本方法中采用的迭代二次曲线略差实验结果。

表 3.2 暗化模块 D 和相似性损失的消融实验

实验类别	方法	Top-1 (%)
基线方法	ResNet-18	53.32
模块 D (启发式设计)	亮度调整	57.96
	伽马校正	63.96
模块 D (可学习设计)	双曲线	62.60
	指数函数曲线	64.16
相似性损失	w/o $\mathcal{L}_D^{sim}$ and $\mathcal{L}_F^{sim}$	64.08
	w/o $\mathcal{L}_D^{sim}$	64.56
	w/o $\mathcal{L}_F^{sim}$	64.88
本方法	-	**65.87**

3.3.3 夜间语义分割

本节探讨一个更具挑战性的夜间视觉任务：语义分割。基线方法为以 ResNet-101 为骨干的 RefineNet [103]。白天的训练数据集为 Cityscapes [104]，其包含 2975 张用于训练的图像和 500 张用于验证的图像，所有图像都配有

表 3.3 Nighttime Driving 和 Dark-Zurich 数据集上的语义分割结果，以 mIoU 形式报告

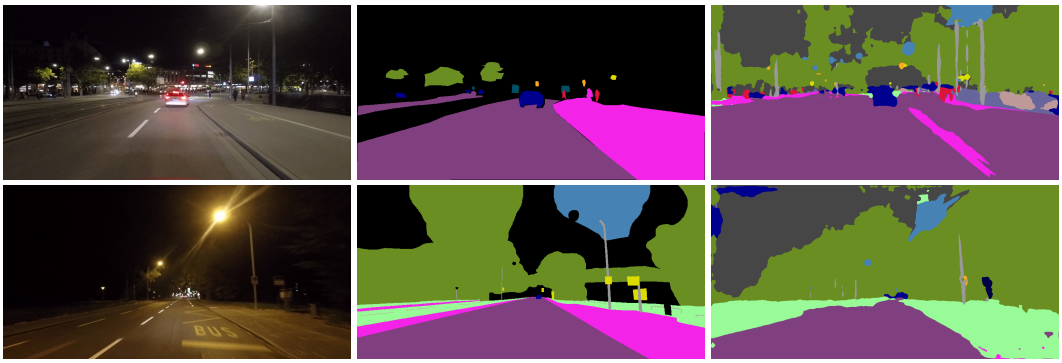
方法类别	方法	Nighttime Driving	Dark-Zurich
基线方法	RefineNet [103]	34.3	30.6
低光照增强	EnlightenGAN [2]	25.2	24.9
	Zero-DCE++ [3]	32.7	28.3
	RUAS [39]	25.1	23.4
	SCI [4]	28.6	25.7
	URetinexNet [1]	28.1	24.0
	LEDNet [5]	27.6	26.6
域泛化	AdaBN [53]	37.2	31.1
	RobustNet [56]	33.0	34.5
	SAN-SAW [57]	28.1	16.0
零样本昼夜域适应	MAET [16]	28.1	26.4
	CIConv [15]	41.2	34.5
	本方法	**44.9**	**40.2**

密集注释。夜间测试数据集是 Nighttime Driving [105] 和 Dark-Zurich [12]。这两个数据集包含 50 张粗略标注和 151 张密集标注的夜间街景图像。

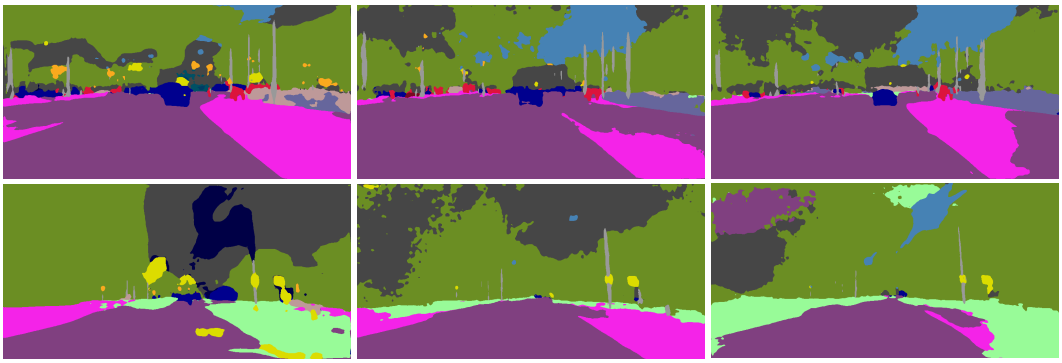
测试结果如表 3.3所示。低光照增强方法比基线的结果更差，因为它们在带有复杂光源的街景上表现不佳。域泛化方法难以处理巨大的昼夜域差距，从而导致了不令人满意的结果。注意 RobustNet [56] 采用了 DeepLab-v3 [106] 架构，该架构优于本方法实现中采用的 RefineNet [103]。

在零样本域适应方法中，MAET [16] 向图像中注入过多的噪声，导致严重的性能下降。CIConv 获得了更好的结果，但提升有限。相比之下，本方法将 Nighttime Driving 的 mIoU 分数提高到 44.9%，Dark-Zurich 提高到 40.2%。

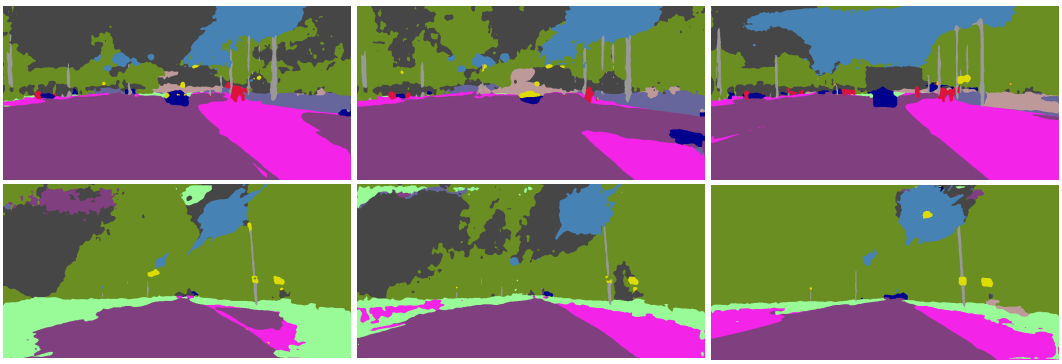
图 3.6展示了两个夜间数据集上的定性分割结果。低光照增强方法在夜间街道场景上表现不佳。本方法更好地提取了被黑暗掩盖的信息，因此生成了更精确的语义图。



(a) 输入图像　(b) 基准分割　(c) RefineNet



(d) MAET　(e) CIConv　(f) Zero-DCE++



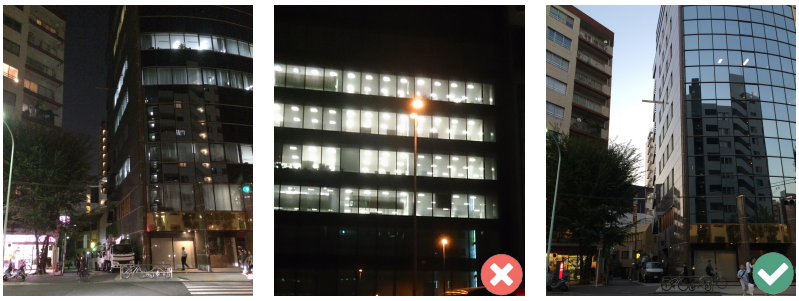
(g) SCI　(h) URetinexNet　(i) 本方法

图 3.6 夜间语义分割可视化实验结果

3.3.4　夜间视觉位置识别

本节探讨视觉位置识别任务，其目标是从图像池中检索出展示与查询图像相同场景的图像。与分类和分割不同，位置识别方法在推断过程中不是端到端的。本方法基于 GeM [107]（采用 ResNet-101 作为骨干网络）进行改进。在 GeM 中，网络接收一组图像 $\{p, q, n_1, \cdots, n_k\}$ 作为输入，其中查询 q 只匹配 p。网络在一个对比损失上进行训练，类似于本方法框架中的模型级阶段。因此，方法实现中保持了图像级阶段不变，并修改了模型级阶段：具体而言，先和其他任务一样训练暗化模块 D。随后，将 $D(p)$ 视为 p 的一个额外匹配，即，一个输入元组包含两个正样本（而不是一个）和 k 个负样本。模型采用 Retrieval-SfM 数据集 [107] 进行训练，并在包含多种照明条件和观看方向的城市视图的 Tokyo 24/7 数据集上对其进行评估 [108]。

性能以平均精度（mAP）的形式报告在表 3.4中。比较方法的结果借用自 [50] 和 [15]。本方法优于所有零样本方法，并能与传统的域适应方法相媲美。此外，图 3.7展示了一项可视化结果：基线方法被图像的低光外观迷惑并输出了错误的识别结果，而本方法则给出了正确预测。



(a) 查询图像　(b) 基线方法　(c) 本方法

图 3.7 视觉位置识别任务可视化结果。基线方法为 GeM [107]

3.3.5　低光视频动作识别

在图像任务外，本方法也适用于视频任务，如视频动作识别。训练数据由来自 HMDB51 [110]、UCF101 [111]、Kinetics-600 [112] 和 Moments in Time [113] 的约 2600 个正常光照视频片段组成。本方法在 ARID 数据集的官

表 3.4 Tokyo 24/7 数据集 [108] 上的视觉位置识别结果

方法类别	方法	mAP (%)
昼夜域适应 (训练时可用夜晚图像)	U-Net jointly [50]	86.5
	EdgeMAC + CLAHE [50]	90.5
	EdgeMAC + U-Net jointly [50]	90.0
零样本昼夜域适应	EdgeMAC [109]	75.9
	U-Net jointly [50]	79.8
	GeM [107]	85.0
	CIConv-GeM [15]	88.3
	本方法	**90.4**

方测试集上进行评估 [114]。动作识别器是基于 3D-ResNet [115] 的 I3D [116]。

表 3.5 ARID 数据集 [114] 上的视频动作识别结果

方法类别	方法	Top-1 (%)
基线方法	I3D [116]	47.02
低光视频增强	StableLLVE [117]	45.08
	SMOID [48]	47.27
	SGZ [118]	46.42
域泛化、零样本昼夜域适应	AdaBN [53]	46.17
	本方法	**51.52**

为从图像扩展至视频应用场景，需要对方法进行如下优化。在训练暗化模块时，模型以视频片段中的帧作为输入。对于每帧，模型对公式 (3.4) 中的 $\mathcal{A}$ 和 $\mathcal{B}$ 进行独立估计。计算损失函数时，$\mathcal{L}_D^{sim}$ 在视频片段之间计算，而其余损失在视频帧之间计算。在生成低光视频时，各视频帧被分别输入到曲线估计器中，同时共享相同的曝光图 E'。

如表 3.5所示，低光视频增强方法 StableLLVE [117]、SMOID [48] 和 SGZ [118] 带来的性能增益有限。与此同时，本方法将模型的性能提升了 4.38%，展现了其在视频上的优越性。图 3.8展示了可视化结果，相比于其他低光视频增强方法，本方法给出了正确的动作预测。

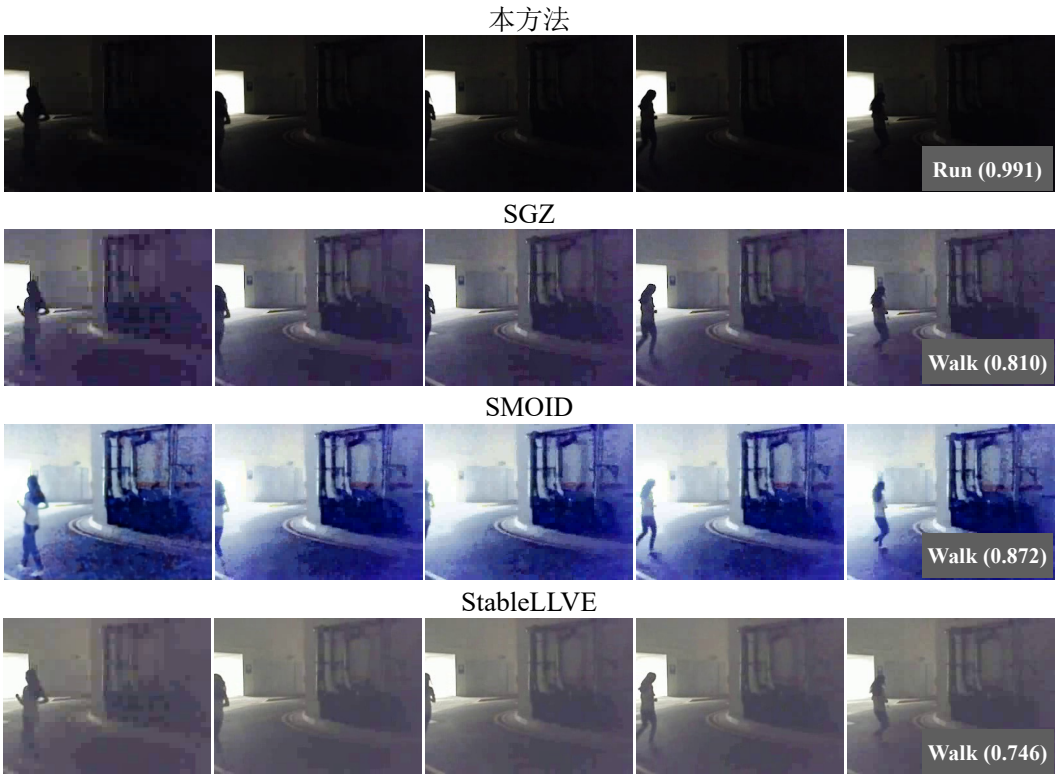


图 3.8 低光视频动作识别的可视化结果

3.4　本章小结

本章节提出了一种基于相似度最小最大化的零样本昼夜域适应算法。具体地，本算法主要包含两部分：（1）在图像层面上合成夜间域，该域与白天域的特征相似性最小，以放大域间隙；（2）在模型层面上最大化两个域中图像的特征相似性，以实现昼夜域适应。在各夜间下游任务上的大量实验证明了本方法的有效性。

第四章　以物体为中心的三维物体发现与场景重建方法

本章将探究如何通过无监督学习构建分解的、以物体为中心的三维场景表示。传统方法将每个物体表示在观察者的视角中，将物体的内在属性（如形状和外观）与外在属性（如物体的 3D 位置）紧密地绑定在一起。因此，物体位置的轻微改变或相机位置的微小调整都会显著改变物体的潜在表示。直觉上，物体的内在属性应该与位置无关保持一致，但在现有的 3D 以物体为中心的学习模型中，这种不变性被忽视了。正如卷积网络 [21] 所示，考虑这种不变性对于泛化至关重要 [22,23]。而已有方法对于物体内在和外在属性的混合大大阻碍了模型的泛化能力。如图 4.1所示，虽然两张输入图像所示物体仅有位置上的不同，但模型最终给出的渲染结果却不仅仅包含物体位置的差异。这说明了，若将物体自身的内在属性（内参）与位置等外在属性（外参）耦合在物体的表征中，出现在不同位置的同一物体会被模型提取出不同的表征，进而影响模型的泛化性。

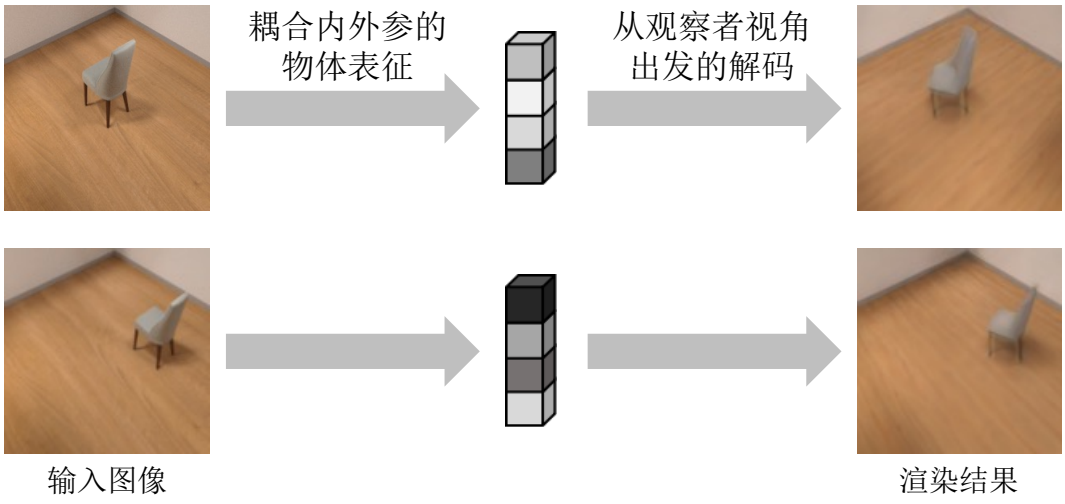


图 4.1 已有方法的物体表征及其问题

基于以上观察，本文提出了一种学习从单一图像推断 3D 以物体为中心的场景表示的方法，称为无监督物体中心辐射场（Unsupervised Discovery of Object-Centric Neural Fields, uOCF）。与现有方法不同，本问题提出的方法专注于学习物体的内在属性，并将外在属性单独建模，也即位置无关表征。如

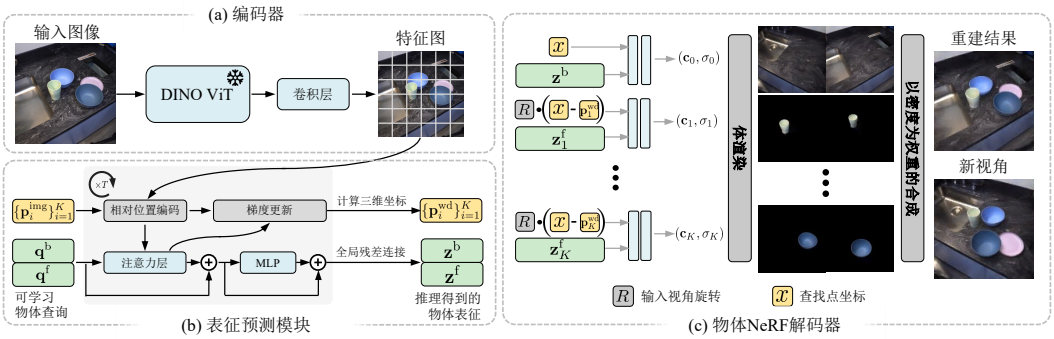


图 4.2 本章节提出的模型框架

图 4.9所示，本方法能够从单张输入图像准确地发现其中的物体，并支持从任意视角重建场景或移动场景中的物体。

4.1 问题定义与建模

给定单个输入图像，本方法的目标是预测以场景中物体为中心的神经辐射场，即每个被发现的物体都被表示在以其自身为中心的坐标而非世界或观察者坐标，并获取这些物体的 3D 位置。如图 4.2所示，本方法提出的框架包括一个编码器，一个表征推理模块和一个解码器。

编码器. 编码器从输入图像 $\mathbf{I}$ 中提取一个特征图 $\mathbf{f} \in \mathbb{R}^{N \cdot C}$，其中 $N = H \cdot W$ 是特征图的空间大小，C 代表通道数。本框架采用的编码器为一个冻结的 DINOv2-ViT [119]，后接两个卷积层。

表征预测模块. 表征预测模块从特征图中推断出底层 3D 场景中物体的表征及其位置。本方法假定场景由一个背景环境和不超过 K 个前景物体组成。因此，输出包括一个背景表征 $\mathbf{z}^{\mathrm{b}} \in \mathbb{R}^{1\times D}$ 和一组前景物体表征 $\mathbf{z}^{\mathrm{f}} = [\mathbf{z}_1^{\mathrm{f}T}\ \mathbf{z}_2^{\mathrm{f}T}\ \cdots\ \mathbf{z}_K^{\mathrm{f}T}]^T \in \mathbb{R}^{K\times D}$ 及其相应位置 $\{\mathbf{p}_i^{\mathrm{wd}}\}_{i=1}^K$，其中 $\mathbf{p}_i^{\mathrm{wd}} \in \mathbb{R}^3$ 表示世界坐标中的位置。注意，当场景中的物体少于 K 个时，某些物体表征可能不代表任何物体。

解码器. 本框架的解码器采用了带条件的 NeRF：$g(\mathbf{x}|\mathbf{z})$，它以 3D 位置 $\mathbf{x}$ 和表征 $\mathbf{z}$ 作为输入，生成用于渲染的辐射颜色和密度。本框架采用两个 MLP，g^{b} 和 g^{f}，分别用于背景环境和前景物体。

4.2　从图像中预测物体表征及位置

获取物体表征. 表征预测模块的目标是将一组可学习的物体查询（$\mathbf{q}^{\mathrm{f}} = [\mathbf{q}_1^{\mathrm{f}T}\ \mathbf{q}_2^{\mathrm{f}T}\ \cdots\ \mathbf{q}_K^{\mathrm{f}T}]^T \in \mathbb{R}^{K\times D}$）与每个前景物体的视觉特征绑定，以及另一个查询绑定到背景特征（$\mathbf{q}^{\mathrm{b}} \in \mathbb{R}^{1\times D}$）。这一绑定操作通过交叉注意力机制和可学习的线性函数 $\mathcal{K}^{\mathrm{b}}, \mathcal{K}^{\mathrm{f}}, \mathcal{Q}^{\mathrm{b}}, \mathcal{Q}^{\mathrm{f}}, \mathcal{V}^{\mathrm{b}}, \mathcal{V}^{\mathrm{f}}$ 建模：

$$\begin{aligned}\mathbf{A}_{i,j} =& \frac{\exp(\mathbf{M}_{i,j})}{\sum_k \exp(\mathbf{M}_{i,k})},\ \text{其中}\\ \mathbf{M} =& \frac{1}{\sqrt{D^s}} \begin{bmatrix} \mathcal{Q}^{\mathrm{b}}(\mathbf{q}^{\mathrm{b}}) \cdot \mathcal{K}^{\mathrm{b}}(\mathbf{f})^T \\ \mathcal{Q}^{\mathrm{f}}(\mathbf{q}^{\mathrm{f}}) \cdot \mathcal{K}^{\mathrm{f}}(\mathbf{f})^T \end{bmatrix}^T \in \mathbb{R}^{N\times(K+1)}.\end{aligned} \tag{4.1}$$

随后通过输入的注意力加权平均计算物体查询的更新信号：

$$\begin{aligned}\mathbf{u}^{\mathrm{b}} &= (\mathbf{W}_{(:,1)})^T \cdot \mathcal{V}^{\mathrm{b}}(\mathbf{f}) \in \mathbb{R}^{1\times D},\\ \mathbf{u}^{\mathrm{f}} &= (\mathbf{W}_{(:,2:)})^T \cdot \mathcal{V}^{\mathrm{f}}(\mathbf{f}) \in \mathbb{R}^{K\times D},\end{aligned} \tag{4.2}$$

其中 $\mathbf{W}_{i,j} = \frac{\mathbf{A}_{i,j}}{\sum_l \mathbf{A}_{l,j}}$ 是在空间维度上标准化后的注意力图。然后通过以下方式更新查询：

$$\begin{aligned}&\mathbf{q}^{\mathrm{b}} \leftarrow \mathbf{q}^{\mathrm{b}} + \mathbf{u}^{\mathrm{b}},\ \mathbf{q}^{\mathrm{f}} \leftarrow \mathbf{q}^{\mathrm{f}} + \mathbf{u}^{\mathrm{f}},\\ &\mathbf{q}^{\mathrm{b}} \leftarrow \mathbf{q}^{\mathrm{b}} + t^{\mathrm{b}}(\mathbf{q}^{\mathrm{b}}),\ \mathbf{q}^{\mathrm{f}} \leftarrow \mathbf{q}^{\mathrm{f}} + t^{\mathrm{f}}(\mathbf{q}^{\mathrm{f}}),\end{aligned} \tag{4.3}$$

其中 t^{b} 和 t^{f} 是 MLP。本方法对上述过程进行共计 T 次迭代，随后将更新后的物体查询与输入特征图 $\mathbf{f}$ 的相应注意力加权平均（即全局残差）拼接，最终得到背景表征 $\mathbf{z}^{\mathrm{b}}$ 和前景物体的表征 $\{\mathbf{z}_i^{\mathrm{f}}\}_{i=1}^K$。

本方法中的表征推理模块与已有文献中的 Slot-Attention 机制有关 [76]，但在几个关键方面有所不同：(1) 物体查询是完全可学习的，而不是从可学习的高斯分布中采样，以增强训练稳定性；(2) 本方法联合提取物体位置及其潜在表示，并添加物体特定的位置编码以利用提取的位置信息；(3) 本方法用

Transformer 架构替代了 GRU [120]，以平滑梯度流。

预测物体在图像中的位置. 为了预测物体的表征及其位置，本方法为每个物体查询分配图像坐标上的归一化位置 $\mathbf{p}_i^{\text{img}} \in [-1,1]^2$，初始设置为零，然后通过动量 m 与归一化 2D 网格 $\mathbf{E}^{\text{abs}} \in [-1,1]^{N\times 2}$ 上的注意力加权平均迭代进行迭代更新：

$$\mathbf{p}_i^{\text{img}} \leftarrow (\mathbf{W}_{(:,i+1)})^T \cdot \mathbf{E}^{\text{abs}} \cdot (1-m) + \mathbf{p}_i^{\text{img}} \cdot m. \tag{4.4}$$

为利用预测的物体位置，本方法采用如下形式的相对位置编码 [77]：

$$\mathbf{E}_i^{\text{pos}} := \text{concat}([\mathbf{E}^{\text{abs}} - \mathbf{p}_i^{\text{img}}, \mathbf{p}_i^{\text{img}} - \mathbf{E}^{\text{abs}}]) \in \mathbb{R}^{N\times 4} \tag{4.5}$$

其中 concat 是沿张量最后一个维度的拼接。于是，公式 (4.1) 中的 M 可以改写为：

$$M = \frac{1}{\sqrt{D^s}} \begin{bmatrix} \mathcal{Q}^{\text{b}}(\mathbf{q}^{\text{b}}) \cdot \mathcal{K}^{\text{b}}(\mathbf{f} + h_1(\mathbf{E}^{\text{abs}}))^T \\ \mathcal{Q}^{\text{f}}(\mathbf{q}_1^{\text{f}}) \cdot \mathcal{K}^{\text{f}}(\mathbf{f} + h_1(\mathbf{E}_1^{\text{pos}}))^T \\ \cdots \\ \mathcal{Q}^{\text{f}}(\mathbf{q}_K^{\text{f}}) \cdot \mathcal{K}^{\text{f}}(\mathbf{f} + h_1(\mathbf{E}_K^{\text{pos}}))^T \end{bmatrix}^T, \tag{4.6}$$

其中 $h_1 : \mathbb{R}^4 \to \mathbb{R}^D$ 是一个线性函数。

总体而言，表征推理模块通过物体查询及其位置的迭代更新，实现了这些查询与场景中的物体之间的绑定。为解决同一物体被重复识别的问题，若在最后一次迭代开始时存在一对余弦相似度较高且位置接近的物体查询，其中之一会被设置为无效。最后，为了处理可能的遮挡问题，预测得到的位置被添加上了一个小的偏差项，即 $\mathbf{p}_i^{\text{img}} \leftarrow \mathbf{p}_i^{\text{img}} + \tanh(h_2((W_{(:,i+1)})^T)) \cdot \alpha$，其中比例超参数 $\alpha = 0.2$，$h_2 : \mathbb{R}^N \to \mathbb{R}^2$ 是一个线性函数。

从图像二维坐标推断空间三维坐标. 在预测得到物体在图像上的二维坐标 $\mathbf{p}_i^{\text{img}}$ 后，需要将其投影到三维世界坐标中以获取 $\mathbf{p}_i^{\text{wd}}$。本方法考虑两种情况：若物体位于已知平面（例如，使用平面检测 [121]），则直接将射线延伸至已知平面以获取物体坐标（将相机中心与 $\mathbf{p}_i^{\text{img}}$ 之间的射线延伸以与地面平

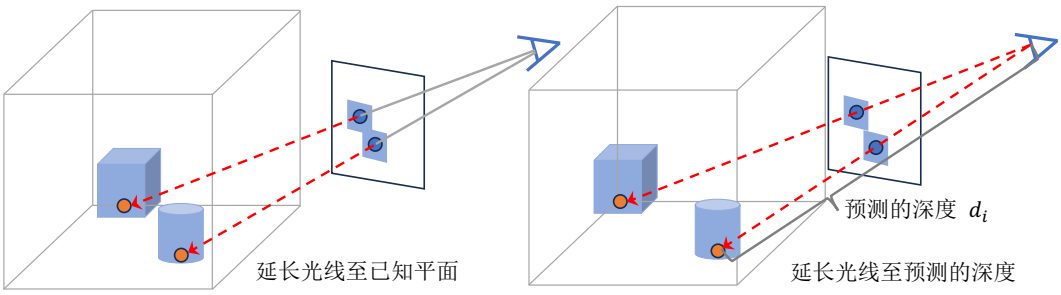


图 4.3 uOCF-P（左）和 uOCF-N（右）中推断物体三维坐标的不同方式

面相交)；若没有这样的平面，则将射线延长至深度 $d \cdot s_i$，其中 d 是相机与场景中心之间的距离，$\{s_i\}_{i=1}^{K}$ 是通过使用相机参数和物体潜在作为输入的线性层预测的缩放项。上述两种方法分别称为 uOCF-P 和 uOCF-N，其中 uOCF 默认表示 uOCF-P。图 4.3展示了它们的具体流程。

组合神经渲染. 预测得到的物体位置允许本方法将物体放置在以它们自身为中心的坐标而不是观察者或世界坐标中，从而获得以物体为中心的神经场。技术上，对于世界坐标中的每个点 $\mathbf{x} \in \mathbb{R}^3$，先通过 $\mathbf{x}_i = R \cdot (\mathbf{x} - \mathbf{p}_i^{\text{wd}})$ 将其转换到第 i 个物体的本地坐标中，其中 R 表示输入相机旋转矩阵。再计算 $\mathbf{x}$ 在前景辐射场中的颜色和密度 $(\mathbf{c}_i, \sigma_i) = g^{\text{f}}(\mathbf{x}_i|\mathbf{z}_i^{\text{f}})$ 和背景辐射场中为 $(\mathbf{c}_0, \sigma_0) = g^{\text{b}}(\mathbf{x}|\mathbf{z}^{\text{b}})$，最后将这些值通过密度加权平均得到场景的密度和颜色 $(\bar{\mathbf{c}}, \bar{\sigma})$:

$$\bar{\sigma} = \sum_{i\geq 0} \omega_i \sigma_i,\ \bar{\mathbf{c}} = \sum_{i\geq 0} \omega_i \mathbf{c}_i,\ \text{其中}\ \omega_i = \frac{\sigma_i}{\sum_{j\geq 0} \sigma_j}, \tag{4.7}$$

并通过体渲染计算图像中像素的颜色。本框架的全部流程都可微，因此可以同时对模型的所有参数同时进行反向传播与梯度更新。

4.3　模型训练

4.3.1　物体先验学习

在复杂的组合场景中进行无监督三维物体发现的难点在于其内在的歧义性。例如，当一个物体被视觉上相似的物体部分遮挡时，推理模型需要找到物体边界，以将被遮挡的物体与遮挡物分开，并为被遮挡的物体生成合理的

补全。

解决这种挑战的一种方式是学习物体先验。然而，现有方法在学习通用的物体先验方面存在困难，因为它们的物体表示对空间配置敏感：相机姿态或物体位置的轻微变化都可能导致物体表征的剧烈变化。因此，这样学到的物体先验在遇到未见过的空间配置时泛化性较差。与现有方法不同，通过本方法提取得到的物体表征因具有平移不变性而具有较强的泛化性，因此能够支持学习物体先验。学习物体先验的主要思想是从简单场景（例如，仅包含单一物体的合成场景）中学习可泛化的物体先验（如空间连续性），然后利用学习得到的先验从可能具有非常不同的场景几何和空间布局的更复杂场景中学习。图 4.4中展示了物体先验学习的流程。

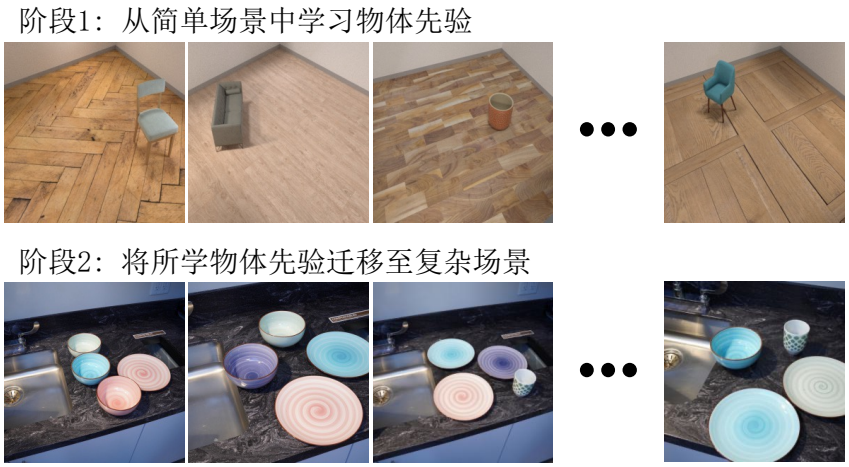


图 4.4 本方法采用的两阶段训练流程

4.3.2 物体中心采样

为了提高重建质量并充分利用物体中心辐射场的性质，本方法提出了物体中心采样的技术，以将采样点集中到物体位置附近。如图 4.5所示，在训练模型一定周期使其能够正确区分前景物体并预测它们的位置后，距离预测位置较远的采样点会被直接丢弃。采用这种技术可以在增加四倍采样点数量的同时维持相近的计算开销，从而显著提高模型的鲁棒性和场景的视觉效果。

在上述两个训练阶段中，训练数据集均包含多个场景，而每个场景均含标定后的稀疏多视角图像。在每个训练步骤中，模型以单张图像作为输入，推

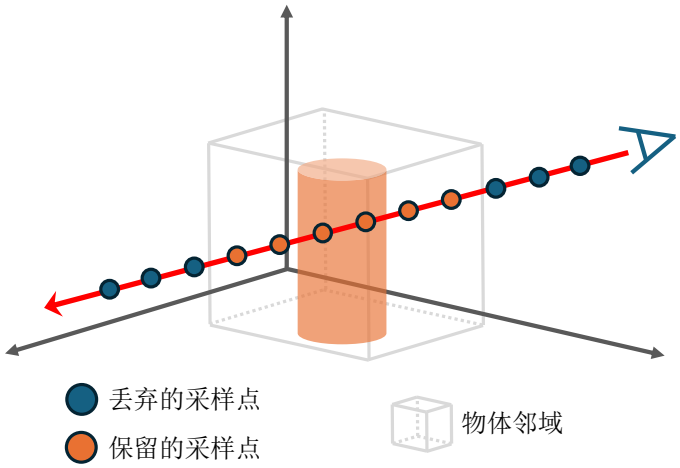


图 4.5 物体中心采样：距离物体较远的采样点被直接丢弃，以节省计算开销

断物体的表征和位置，从输入和参考视角渲染图像，并将它们与真实图像进行比较以计算损失。模型监督信号包含均方误差损失 ℓ_{recon} 和重建图像与真实图像之间的感知损失 ℓ_{perc} [122] 组成。此外，本方法还利用预训练单目深度估计器预测深度图，并引入深度排序损失 [123] 和背景遮挡正则化损失 [124] 来减少少样本 NeRF 中常见的漂浮遮挡现象。

因此，总体损失函数如下所述：

$$\mathcal{L} = \ell_{\text{recon}} + \lambda_{\text{perc}}\ell_{\text{perc}} + \lambda_{\text{depth}}\ell_{\text{depth}} + \lambda_{\text{occ}}\ell_{\text{occ}}. \tag{4.8}$$

4.4　实验结果及分析

4.4.1　实验设置

下文在无监督物体分割、新视角合成和 3D 场景操作等方面评估了本章节提出的方法。

数据集. 实验部分构建了两个真实世界数据集以评估提出的方法，以及一个合成数据集用于学习物体先验，图 4.6展示了训练数据集中的部分样本。

Room-Texture 为合成数据集，其中的物体来自 ABO [125] 数据集中“扶手椅”类别的 324 个模型。每个场景包含一个物体，而背景是从一系列地板纹理中随机选择的。为促进与现有方法的比较，该数据集另有一个包含多物

(a) Room-Texture (b) Kitchen-Matte (c) Kitchen-Shiny

图 4.6 实验采用的数据集

体场景的版本，每个场景包含 2-4 个物体。单物体版本包含 1296 个场景，而多物体版本包含训练用的 5000 个场景和测试用的 100 个场景。每个场景包含四张图像，均朝向场景的中心。

Kitchen-Matte 数据集的前景物体为单色哑光餐具，并包含两种背景环境：简单的桌面和复杂的厨房背景。共有 735 个场景用于训练，102 个用于测试。每个场景包含 3-4 个随机位置的物体，并包含 3 个（对于桌面场景）或 2 个（对于厨房背景）视角的图像。

Kitchen-Shiny 数据集的前景物体为具有纹理光泽的餐具。与 Kitchen-Matte 类似，背景环境有简单的桌面背景和复杂的厨房背景两种。该数据集共有 324 个场景用于训练，56 个用于测试。

物体先验学习. 本方法首先在只有一个椅子的合成房间场景 Room-Texture 上训练模型。这种合成数据易于生成和扩展。值得注意的是，尽管存在域间隔，uOCF 可以学习类别不可知的物体先验，并泛化到不同的物体类别。例如，从合成椅子数据上学到的物体先验有效地泛化到真实厨房场景中的餐具。

定性指标. 有关新视角合成的实验报告了 PSNR、SSIM 和 LPIPS 指标；场景分割的实验则采用了 ARI 指数的三个变体：常规 ARI（在所有输入图像像素上计算）、前景 ARI（FG-ARI，在前景输入图像像素上计算）和新视角 ARI（NV-ARI，在新视角像素上计算）。若未额外说明，所有指标都在 128×128 分辨率的图像上计算。

基线方法. 基线方法包括 uORF [18]、BO-QSA [20] 和 COLF [19]。在测试时，模型使用具有已知相机内参的单个图像作为输入，并从目标姿态输出

重建/分割结果。为了公平比较，实验中增加了基线方法的表征维数和训练迭代次数，以保证相近的计算资源开销。默认情况下，所有方法的物体查询数设置为 $K = 4$。

4.4.2 无监督三维物体分割

本节在无监督三维物体分割任务上对 uOCF 进行测试。记第 i 物体查询通过体渲染得到的密度图为 $\mathbf{d}^i$，则对于图像中任意像素 p，其分割标签 $s_p = \arg\max_{i=0}^{K} \mathbf{d}_p^i$。

实验结果见表 4.1和图 4.7。在合成的 Room-Texture 场景中，本方法在所有指标上均超过了现有方法。值得注意的是，没有一个基线方法能在真实世界的 Kitchen-Shiny 场景中产生合理的分割结果。具体来说，uORF 将所有物体绑定到背景上，导致空的物体分割；BO-QSA 未能区分不同的物体实例；COLF 在新视角上产生无意义的结果，因为光场不保证多视角的一致性。相比之下，得益于物体中心建模，uOCF 更好地区分了前景物体，并一致地产生了令人满意的场景分割结果，展示了其在物体中心表示学习中的有效性。此外，uOCF 能够处理物体相互遮挡的场景。

表 4.1 Room-Texture 数据集上三维场景分割与新视角生成的定性结果

方法	场景分割			新视角生成		
	ARI↑	FG-ARI↑	NV-ARI↑	PSNR↑	SSIM↑	LPIPS↓
uORF [18]	0.670	0.093	0.578	24.23	0.711	0.254
BO-QSA [20]	0.697	0.354	0.604	25.26	0.739	0.215
COLF [19]	0.235	0.532	0.011	22.98	0.670	0.504
uOCF-N (本方法)	**0.791**	**0.584**	**0.722**	28.81	0.796	0.138
uOCF-P (本方法)	0.785	0.563	0.704	**28.85**	**0.798**	**0.136**

4.4.3 新视角合成

本节在新视角合成任务上对本方法进行测试。每个测试场景使用单视角图像作为输入，并使用其余图像作为参考。如表 4.2和图 4.8所示，本方法在所有指标上显著超过基线方法。值得注意的是，区别于已有方法难以区分前

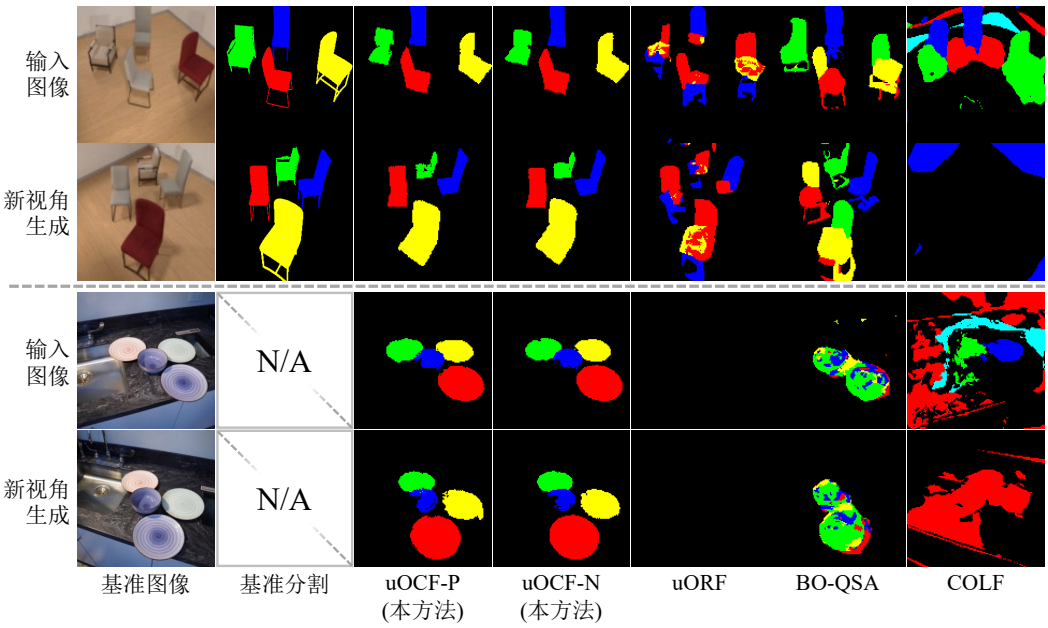


图 4.7 在 Room-Texture 和 Kitchen-Shiny 数据集上的三维场景分割定性实验结果

景物体和背景环境，本方法能够准确地发现三维场景中的物体并生成高保真的场景重建和新视角合成结果。

表 4.2 Kitchen-Shiny 和 Kitchen-Matte 数据集上的新视角合成定量实验结果

方法	Kitchen-Shiny			Kitchen-Matte		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
uORF [18]	19.23	0.602	0.336	26.07	0.808	0.092
BO-QSA [20]	19.78	0.639	0.318	27.36	0.832	0.067
COLF [19]	18.30	0.561	0.397	20.68	0.643	0.236
uOCF-N (本方法)	27.87	0.842	0.055	28.25	0.841	0.055
uOCF-P (本方法)	**28.58**	**0.862**	**0.049**	**29.40**	**0.867**	**0.043**

4.4.4 三维场景操纵

本节评估本方法在场景操作方面的能力。由于本方法支持以下场景编辑功能：1）物体平移，通过修改发送到解码器的物体位置实现；2）物体移除，通过在组合渲染时排除某些物体实现。下文先在 Room-Texture 数据集上对本方法进行定量评估，随后在真实世界的 Kitchen-Shiny 数据集上进行定性评估。

定量实验随机选择场景内的一个物体并移动其位置或将其移除它。在测

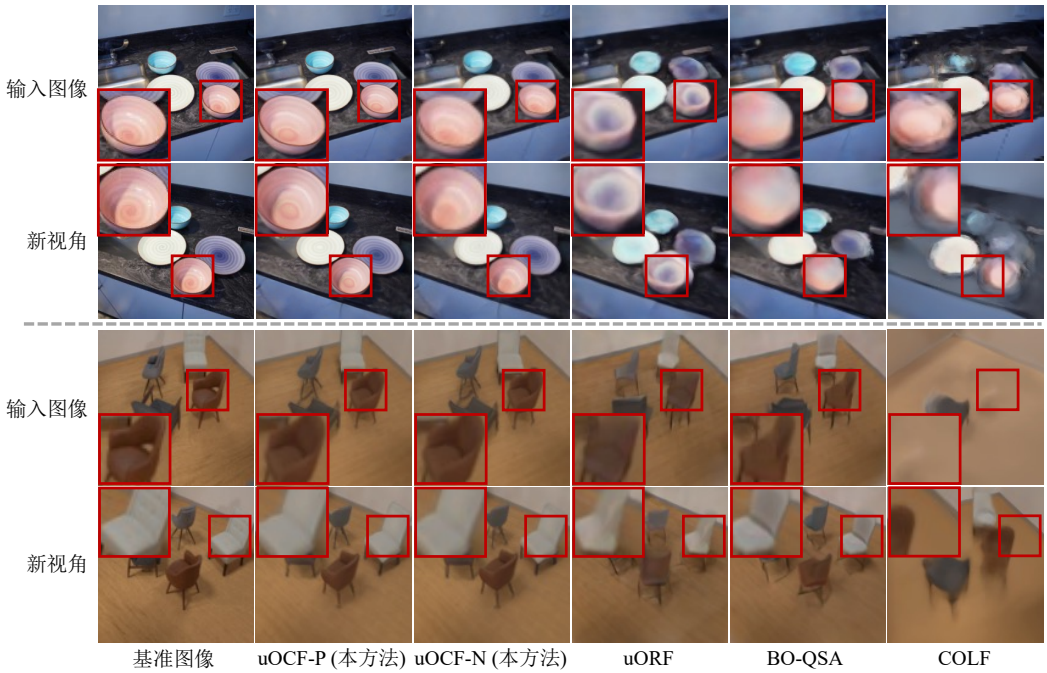


图 4.8 在 Room-Texture 和 Kitchen-Shiny 数据集上的三维场景分割定性实验结果

试时，各方法将渲染出的分割结果与基准分割中的被操作物体计算 IoU，并将分数最高者确定为待操作的物体。值得注意的是，现有方法的应用限于物体的相对位置调整，而本方法由于物体位置和表示的解耦，可以允许绝对坐标移动。如表 4.3所示，本方法在所有指标上均优于基准线，证明了其在三维场景操纵方面的有效性。此外，如图 4.9所示，本方法能够从单张图像中提取物体表示，重建这些物体，并允许我们操纵整个三维场景。

图 4.10展示了 Kitchen-Shiny 数据集上单图像三维场景操纵的定性实验结果。如图所示，uORF 将所有物体合并到背景中，因此不支持场景操纵，即操作结果与原始重建相同；BO-QSA 未能正确区分前景物体，也导致操作结果错误。相比之下，本方法给出了更合理且视觉效果更好的结果。

4.4.5　泛化能力分析

实验最后将探究本方法的零样本和少样本泛化能力。零样本泛化性测试实验要求将 Kitchen-Matte 上训练的模型通过测试期优化泛化到 Kitchen-Shiny 中的未见场景，或将 Room-Texture 上训练的泛化到室内扫描数据集

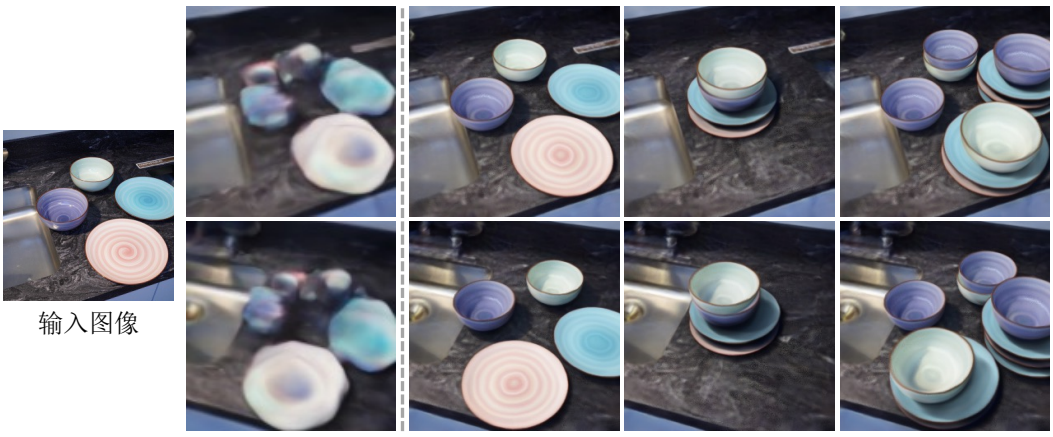


图 4.9 本方法在场景操纵任务的效果

表 4.3 Room-Texture 数据集上三维场景操纵定量实验结果

方法	PSNR↑	SSIM↑	LPIPS↓
物体平移			
uORF	23.65	0.654	0.284
BO-QSA	25.21	0.700	0.226
uOCF-N (本方法)	**27.70**	0.773	**0.156**
uOCF-P (本方法)	27.66	**0.774**	**0.156**
物体移除			
uORF	23.81	0.664	0.282
BO-QSA	24.58	0.698	0.247
uOCF-N (本方法)	**29.20**	**0.803**	**0.132**
uOCF-P (本方法)	28.99	0.802	0.136

HM3D [126] 中的未见场景。实验结果如 4.11所示，其中对于每一组实验，第一、二、三行分别为场景重建结果，背景重建结果和物体重建结果。实验结果表明，已有方法难以泛化到未见物体或未能将前景物体与背景分离，而本方法只需要快速的测试期优化就能从一个数据集泛化到新数据集中更复杂的物体，因此具备出色的泛化能力。

下述实验则在 Kitchen-Shiny 数据集上考察本方法的少样本泛化能力。与前文所述实验利用整个数据集进行训练不同，本实验考虑了一个更具挑战性的情景，即仅使用较少数量的场景对模型进行训练，并在模型未见过的新场景上进行测试。如图 4.12所示，随着训练场景数量的减少，模型给出的场景

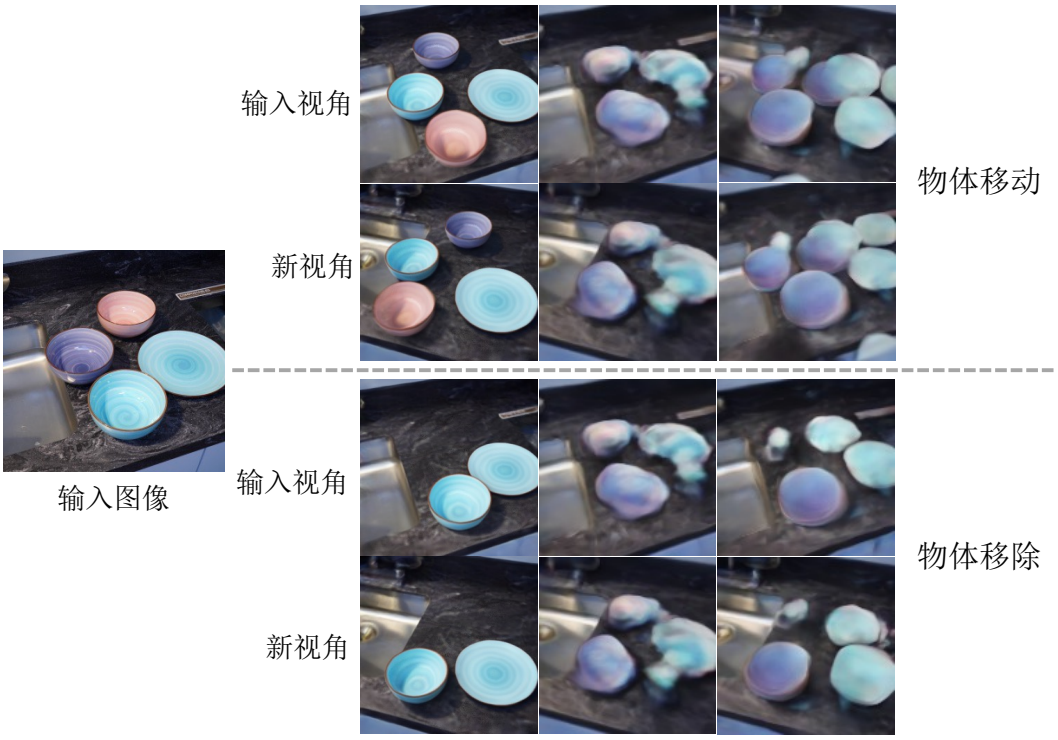
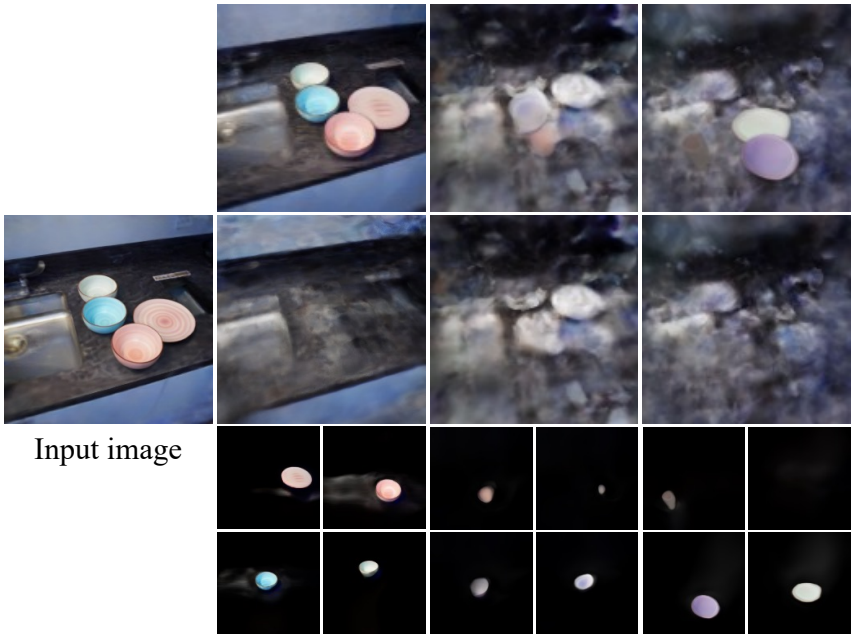
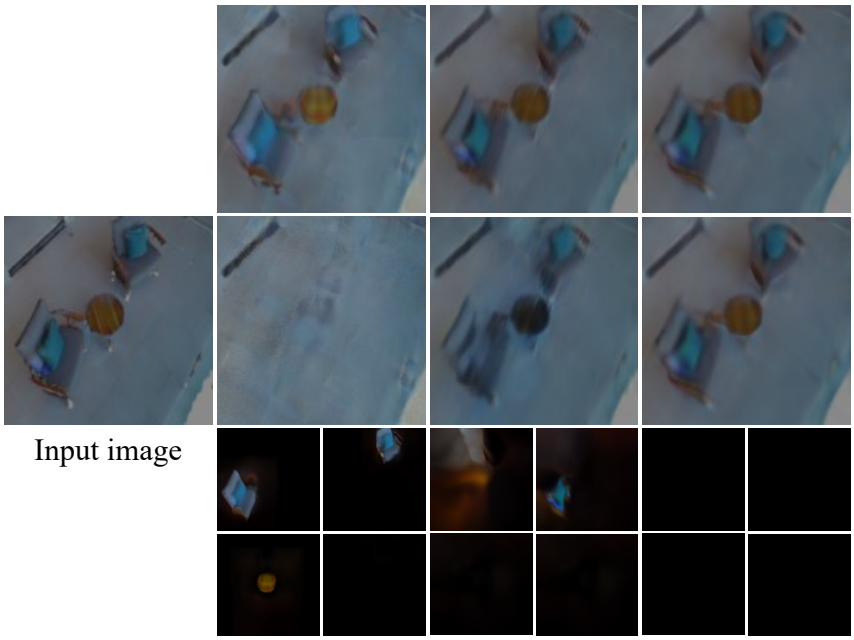


图 4.10 在 Kitchen-Shiny 数据集上的三维场景操纵定性实验结果

重建结果的质量也随之下降。特别地，在极其有限的训练集上，模型难以将物体与背景区分开。这一趋势展现了了数据集大小对于无监督物体发现任务的重要性。



(a) Kitchen-Matte → Kitchen-Shiny



(b) Room-Texture → HM3D

图 4.11 零样本泛化定性实验结果

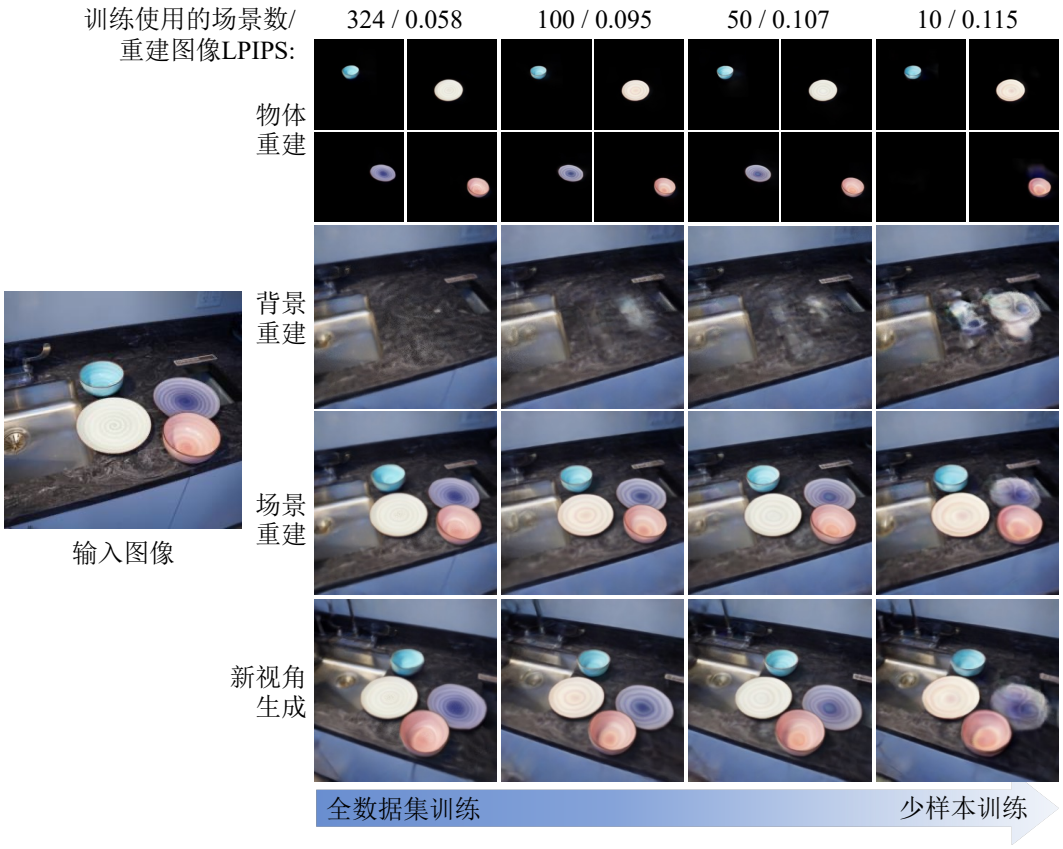


图 4.12 Kitchen-Shiny 数据集上的少样本泛化实验结果

4.5　本章小结

本章节研究了三维场景中物体发现与场景重建的算法，重点关注了位置无关表征对这一问题的重要性，并提出了无监督物体中心神经辐射场（uOCF）这一框架。为了评估这一方法，本研究收集了三个具有挑战性的新数据集，每个数据集的场景中都包含设置在复杂背景之前的多个类别的物体。实验结果表明，以物体为中心的设计和物体先验学习能够显著提高无监督三维物体发现模型的泛化能力。

第五章　总结与展望

本文以机器学习中的无监督学习作为切入点，重点关注了昼夜域适应和三维物体发现与场景重建着两个计算机视觉领域的热点问题。针对昼夜域适应任务，本文提出了一种基于“相似性最小-最大化”的零样本昼夜域适应方法，通过图像层面与模型层面的同时优化，显著增强了在正常光照数据集上训练的模型在低光环境下的表现；针对三维物体发现与场景重建任务，本文提出了一种以物体为中心的物体发现与场景重建算法，称为无监督物体中心辐射场，其主要创新点是通过引入平移不变性显著增强了模型的泛化能力。

本章首先对以上两个工作进行总结，随后对未来工作进行展望。

5.1　本文工作总结

1. 针对昼夜域适应问题，本文提出了一个基于“相似性最小-最大化”的解决方案，且不需要使用任何夜间数据。该方案综合了图像层面的亮度调整和模型层面的多任务对比学习，减少了对目标域数据的依赖，并且在多个夜间视觉任务上展现出显著的性能提升。

2. 针对三维物体发现与场景重建物体，本文提出了一种以物体为中心的物体发现方法，称为无监督物体中心辐射场。该方法将物体表示在以其自身为中心而非以观察者为中心的神经辐射场中，并通过引入平移不变性增强了模型的泛化能力。为评估此方法，本研究收集了三个具有挑战性的数据集。在数据集上的优异结果表明，以物体为中心的设计和物体先验学习能够显著提高无监督三维物体发现模型的泛化能力。

5.2　未来研究展望

基于无监督学习的范式，本文围绕昼夜域适应与三维物体发现与场景重建两个问题做出了一系列创新性设计，并取得了显著的性能提升。在此基础之上，还有很多潜在的方法值得研究，以进一步提升模型的效果，例如：

复杂光照条件变化中的域泛化. 如前文所述，“相似性最小-最大框架”目前只能用于处理光照变化，而不能处理真实世界中更为复杂的光照条件变化，例如多种光源、动态光源等。因此，未来的工作可以进一步优化模型框架设计及训练技术以增强模型对于更复杂光照条件的适应性。

复杂真实世界场景中的三维物体发现. 如实验部分所示，本研究提出的三维物体发现算法虽然相比已有的方法取得了显著的泛化性提升，但仍然无法扩展到真实世界中包含多种类别、且背景极为复杂的场景中。在本研究的基础上，未来的方法可以探索在大型但物体数据集（如 ObjVerse [127], MVImgNet [128] 等）上学习泛化性更强的物体先验，以实现复杂场景中的三维物体发现。

参考文献

[1] W. Wu, J. Weng, P. Zhang, X. Wang, W. Yang, and J. Jiang, "Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[2] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "Enlightengan: Deep light enhancement without paired supervision," *IEEE Transactions Image Processing*, vol. 30, pp. 2340–2349, 2021.

[3] C. Li, C. Guo, and C. C. Loy, "Learning to enhance low-light image via zero-reference deep curve estimation," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 44, no. 8, pp. 4225–4238, 2021.

[4] L. Ma, T. Ma, R. Liu, X. Fan, and Z. Luo, "Toward fast, flexible, and robust low-light image enhancement," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[5] S. Zhou, C. Li, and C. Change Loy, "Lednet: Joint low-light enhancement and deblurring in the dark," in *Proceedings of European Conference on Computer Vision*, 2022.

[6] Y. Zhao, Y. Xu, Q. Yan, D. Yang, X. Wang, and L.-M. Po, "D2HNet: Joint denoising and deblurring with hierarchical network for robust night image restoration," in *Proceedings of European Conference on Computer Vision*, 2022.

[7] V. F. Arruda, T. M. Paixão, R. F. Berriel, A. F. De Souza, C. Badue, N. Sebe, and T. Oliveira-Santos, "Cross-domain car detection using unsupervised image-to-image translation: From day to night," in *Proceedings of International Joint Conference on Neural Networks*, 2019.

[8] E. Romera, L. M. Bergasa, K. Yang, J. M. Alvarez, and R. Barea, "Bridging the day and night domain gap for semantic segmentation," *IEEE Intelligent Vehicles Symposium*, pp. 1312–1318, 2019.

[9] H. Gao, J. Guo, G. Wang, and Q. Zhang, "Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[10] W. Wang, X. Wang, W. Yang, and J. Liu, "Unsupervised face detection in the dark," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 45, no. 1, pp. 1250–1266, 2022.

[11] W. Wang, Z. Xu, H. Huang, and J. Liu, "Self-aligned concave curve: Illumination enhancement for unsupervised adaptation," in *Proceedings of ACM International Conference on Multimedia*, 2022.

[12] C. Sakaridis, D. Dai, and L. V. Gool, "Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2019.

[13] Y. Sasagawa and H. Nagahara, "Yolo in the dark-domain adaptation method for merging multiple models," in *Proceedings of European Conference on Computer Vision*, 2020.

[14] X. Deng, P. Wang, X. Lian, and S. Newsam, "Nightlab: A dual-level architecture with hardness detection for segmentation at night," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[15] A. Lengyel, S. Garg, M. Milford, and J. C. van Gemert, "Zero-shot day-night domain adaptation with a physics prior," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2021.

[16] Z. Cui, G.-J. Qi, L. Gu, S. You, Z. Zhang, and T. Harada, "Multitask aet with orthogonal tangent regularity for dark object detection," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2021.

[17] K. Stelzner, K. Kersting, and A. R. Kosiorek, "Decomposing 3d scenes into objects via unsupervised volume segmentation," *arXiv:2104.01148*, 2021.

[18] H.-X. Yu, L. J. Guibas, and J. Wu, "Unsupervised discovery of object radiance fields," in *Proceedings of International Conference on Learning Representations*, 2022.

[19] C. Smith, H.-X. Yu, S. Zakharov, F. Durand, J. B. Tenenbaum, J. Wu, and V. Sitzmann, "Unsupervised discovery and composition of object light fields," *Transactions on Machine Learning Research*, 2023.

[20] B. Jia, Y. Liu, and S. Huang, "Improving object-centric learning with query optimization," in *Proceedings of International Conference on Learning Representations*, 2023.

[21] R. Zhang, "Making convolutional networks shift-invariant again," in *Proceedings of International Conference for Machine Learning*, 2019.

[22] P. Chattopadhyay, Y. Balaji, and J. Hoffman, "Learning to balance specificity and invariance for in and out of domain generalization," in *Proceedings of European Conference on Computer Vision*, 2020.

[23] W. Deng, S. Gould, and L. Zheng, “On the strong correlation between model invariance and generalization,” in *Proceedings of Advances in Neural Information Processing Systems*, 2022.

[24] M. Noroozi and P. Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles,” in *Proceedings of European Conference on Computer Vision*, 2016.

[25] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proceedings of International Conference for Machine Learning*, 2020.

[26] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[27] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, “Bootstrap your own latent: A new approach to self-supervised learning,” in *Proceedings of Advances in Neural Information Processing Systems*, 2020.

[28] X. Chen and K. He, “Exploring simple siamese representation learning,” in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[29] R. Wang, D. Chen, Z. Wu, Y. Chen, X. Dai, M. Liu, Y.-G. Jiang, L. Zhou, and L. Yuan, “Bevt: Bert pretraining of video transformers,” in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[30] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[31] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, “Improving language understanding by generative pre-training,” 2018.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[33] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, “Learning transferable visual models from natural language supervision,” in *Proceedings of International Conference for Machine Learning*, 2021.

[34] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Computer Vision, Graphics, and Image Processing*, vol. 39, no. 3, pp. 355–368, 1987.

[35] Z.-u. Rahman, D. J. Jobson, and G. A. Woodell, "Retinex processing for automatic image enhancement," *Journal of Electronic Imaging*, 2004.

[36] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Computer Vision, Graphics, and Image Processing*, 1987.

[37] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," in *Proceedings of British Machine Vision Conference*, 2017.

[38] W. Yang, S. Wang, Y. Fang, Y. Wang, and J. Liu, "From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[39] R. Liu, L. Ma, J. Zhang, X. Fan, and Z. Luo, "Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[40] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[41] A. Dudhane, S. W. Zamir, S. Khan, F. S. Khan, and M.-H. Yang, "Burstormer: Burst image restoration and enhancement transformer," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[42] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," in *Proceedings of Advances in Neural Information Processing Systems*, 2022.

[43] C. M. Nguyen, E. R. Chan, A. W. Bergman, and G. Wetzstein, "Diffusion in the dark: A diffusion model for low-light text recognition," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

[44] Y. Jin, W. Yang, and R. T. Tan, "Unsupervised night image enhancement: When layer decomposition meets light-effects suppression," in *Proceedings of European Conference on Computer Vision*, 2022.

[45] Y. Wang, R. Wan, W. Yang, H. Li, L.-P. Chau, and A. Kot, "Low-light image enhancement with normalizing flow," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2022.

[46] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[47] C. Chen, Q. Chen, M. N. Do, and V. Koltun, "Seeing motion in the dark," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2019.

[48] H. Jiang and Y. Zheng, "Learning to see moving objects in the dark," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2019.

[49] X. Wu, Z. Wu, H. Guo, L. Ju, and S. Wang, "Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[50] T. Jenicek and O. Chum, "No fear of the dark: Image retrieval under varying illumination conditions," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2019.

[51] B. Mahasseni and S. Todorovic, "Regularizing long short term memory with 3d human-skeleton sequences for action recognition," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[52] W. Song, M. Suganuma, X. Liu, N. Shimobayashi, D. Maruta, and T. Okatani, "Matching in the dark: A dataset for matching image pairs of low-light scenes," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2021.

[53] Y. Li, N. Wang, J. Shi, J. Liu, and X. Hou, "Revisiting batch normalization for practical domain adaptation," in *Proceedings of International Conference on Learning Representations Workshops*, 2017.

[54] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv:1907.02893*, 2019.

[55] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," in *Proceedings of International Conference on Learning Representations*, 2021.

[56] S. Choi, S. Jung, H. Yun, J. T. Kim, S. Kim, and J. Choo, "Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[57] D. Peng, Y. Lei, M. Hayat, Y. Guo, and W. Li, "Semantic-aware domain generalized segmentation," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[58] J. Huang, D. Guan, A. Xiao, and S. Lu, "Fsdr: Frequency space domain randomization for domain generalization," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[59] N. Kim, T. Son, C. Lan, W. Zeng, and S. Kwak, "Wedge: web-image assisted domain generalization for semantic segmentation," *arXiv:2109.14196*, 2021.

[60] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2005.

[61] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman, "Using multiple segmentations to discover objects and their extent in image collections," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2006.

[62] K. Grauman and T. Darrell, "Unsupervised learning of categories from sets of partially matching image features," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2006.

[63] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2010.

[64] B. Li, Z. Sun, Q. Li, Y. Wu, and A. Hu, "Group-wise deep object co-segmentation with co-attention recurrent neural network," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[65] H. V. Vo, P. Pérez, and J. Ponce, "Toward unsupervised, multi-object discovery in large-scale image collections," in *Proceedings of European Conference on Computer Vision*, 2020.

[66] S. Eslami, N. Heess, T. Weber, Y. Tassa, D. Szepesvari, K. Kavukcuoglu, and G. E. Hinton, "Attend, infer, repeat: Fast scene understanding with generative models," in *Proceedings of Advances in Neural Information Processing Systems*, 2016.

[67] A. R. Kosiorek, H. Kim, I. Posner, and Y. W. Teh, "Sequential attend, infer, repeat: Generative modelling of moving objects," in *Proceedings of Advances in Neural Information Processing Systems*, 2018.

[68] E. Crawford and J. Pineau, "Spatially invariant unsupervised object detection with convolutional neural networks," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2019.

[69] J. Jiang, S. Janghorbani, G. de Melo, and S. Ahn, "Scalor: Generative world models with scalable object representations," in *Proceedings of International Conference on Learning Representations*, 2020.

[70] Z. Lin, Y.-F. Wu, S. V. Peri, W. Sun, G. Singh, F. Deng, J. Jiang, and S. Ahn, "Space: Unsupervised object-oriented scene representation via spatial attention and decomposition," in *Proceedings of International Conference on Learning Representations*, 2020.

[71] K. Greff, A. Rasmus, M. Berglund, T. H. Hao, J. Schmidhuber, and H. Valpola, "Tagger: Deep unsupervised perceptual grouping," in *Proceedings of Advances in Neural Information Processing Systems*, 2016.

[72] K. Greff, S. Van Steenkiste, and J. Schmidhuber, "Neural expectation maximization," in *Proceedings of Advances in Neural Information Processing Systems*, 2017.

[73] K. Greff, R. L. Kaufman, R. Kabra, N. Watters, C. Burgess, D. Zoran, L. Matthey, M. Botvinick, and A. Lerchner, "Multi-object representation learning with iterative variational inference," in *Proceedings of International Conference for Machine Learning*, 2019.

[74] C. P. Burgess, L. Matthey, N. Watters, R. Kabra, I. Higgins, M. Botvinick, and A. Lerchner, "Monet: Unsupervised scene decomposition and representation," *arXiv:1901.11390*, 2019.

[75] M. Engelcke, A. R. Kosiorek, O. P. Jones, and I. Posner, "Genesis: Generative scene inference and sampling with object-centric latent representations," *arXiv:1907.13052*, 2019.

[76] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf, "Object-centric learning with slot attention," in *Proceedings of Advances in Neural Information Processing Systems*, 2020.

[77] O. Biza, S. van Steenkiste, M. S. Sajjadi, G. F. Elsayed, A. Mahendran, and T. Kipf, "Invariant slot attention: Object discovery with slot-centric reference frames," in *Proceedings of International Conference for Machine Learning*, 2023.

[78] A. Didolkar, A. Goyal, and Y. Bengio, "Cycle consistency driven object discovery," *arXiv:2306.02204*, 2023.

[79] T. Monnier, E. Vincent, J. Ponce, and M. Aubry, "Unsupervised layered image decomposition into object prototypes," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2021.

[80] S. Liang, Y. Liu, S. Wu, Y.-W. Tai, and C.-K. Tang, "Onerf: Unsupervised 3d object segmentation from multiple views," *arXiv:2211.12038*, 2022.

[81] S. A. Eslami, D. J. Rezende, F. Besse, F. Viola, A. S. Morcos, M. Garnelo, A. Ruderman, A. A. Rusu, I. Danihelka, K. Gregor, *et al.*, "Neural scene representation and rendering," *Science*, vol. 360, no. 6394, pp. 1204–1210, 2018.

[82] C. Chen, F. Deng, and S. Ahn, "Learning to infer 3d object models from images," *arXiv:2006.06130*, 2020.

[83] M. S. Sajjadi, D. Duckworth, A. Mahendran, S. van Steenkiste, F. Pavetic, M. Lucic, L. J. Guibas, K. Greff, and T. Kipf, "Object scene representation transformer," in *Proceedings of Advances in Neural Information Processing Systems*, 2022.

[84] J. Wu, J. B. Tenenbaum, and P. Kohli, "Neural scene de-rendering," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[85] S. Yao, T. M. H. Hsu, J.-Y. Zhu, J. Wu, A. Torralba, W. T. Freeman, and J. B. Tenenbaum, "3d-aware scene manipulation via inverse graphics," in *Proceedings of Advances in Neural Information Processing Systems*, 2018.

[86] A. Kundu, Y. Li, and J. M. Rehg, "3d-rcnn: Instance-level 3d object reconstruction via render-and-compare," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[87] J. Ost, F. Mannan, N. Thuerey, J. Knodt, and F. Heide, "Neural scene graphs for dynamic scenes," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[88] N. Müller, A. Simonelli, L. Porzi, S. R. Bulo, M. Nießner, and P. Kontschieder, "Autorf: Learning 3d object radiance fields from single view observations," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[89] B. Yang, Y. Zhang, Y. Xu, Y. Li, H. Zhou, H. Bao, G. Zhang, and Z. Cui, "Learning object-compositional neural radiance field for editable scene rendering," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[90] Q. Wu, X. Liu, Y. Chen, K. Li, C. Zheng, J. Cai, and J. Zheng, "Object-compositional neural implicit surfaces," in *Proceedings of European Conference on Computer Vision*, 2022.

[91] Z. Fan, P. Wang, Y. Jiang, X. Gong, D. Xu, and Z. Wang, "Nerf-sos: Any-view self-supervised object segmentation on complex scenes," *arXiv2209.08776*, 2022.

[92] J. Cen, J. Fang, C. Yang, L. Xie, X. Zhang, W. Shen, and Q. Tian, "Segment any 3d gaussians," *arXiv:2312.00860*, 2023.

[93] J. Cen, Z. Zhou, J. Fang, W. Shen, L. Xie, X. Zhang, and Q. Tian, "Segment anything in 3d with nerfs," in *Proceedings of Advances in Neural Information Processing Systems*, 2023.

[94] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *Proceedings of European Conference on Computer Vision*, 2020.

[95] V. Sitzmann, M. Zollhöfer, and G. Wetzstein, "Scene representation networks: Continuous 3d-structure-aware neural scene representations," in *Proceedings of Advances in Neural Information Processing Systems*, 2019.

[96] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[97] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 1–14, 2023.

[98] H. Lee, M. Ra, and W.-Y. Kim, "Nighttime data augmentation using gan for improving blind-spot detection," *IEEE Access*, vol. 8, pp. 48049–48059, 2020.

[99] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, 2020.

[100] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2017.

[101] A. Punnappurath, A. Abuolaim, A. Abdelhamed, A. Levinshtein, and M. S. Brown, "Day-to-night image synthesis for training nighttime neural isps," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[102] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[103] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[104] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[105] D. Dai and L. V. Gool, "Dark model adaptation: Semantic image segmentation from daytime to nighttime," in *IEEE International Conference on Intelligent Transportation Systems*, 2018.

[106] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv:1706.05587*, 2017.

[107] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 41, no. 7, pp. 1655–1668, 2019.

[108] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015.

[109] F. Radenović, G. Tolias, and O. Chum, "Deep shape matching," in *Proceedings of European Conference on Computer Vision*, 2018.

[110] H. Kuehne, H. Jhuang, E. Garrote, T. A. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2011.

[111] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *arXiv:1212.0402*, 2012.

[112] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," *arXiv:1705.06950*, 2017.

[113] M. Monfort, C. Vondrick, A. Oliva, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. M. Brown, Q. Fan, and D. Gutfreund, "Moments in time dataset: One million videos for event understanding," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 42, no. 2, pp. 502–508, 2020.

[114] Y. Xu, J. Yang, H. Cao, K. Mao, J. Yin, and S. See, "Arid: A new dataset for recognizing action in the dark," in *Deep Learning for Human Activity Recognition*, 2021.

[115] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2019.

[116] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[117] F. Zhang, Y. Li, S. You, and Y. Fu, "Learning temporal consistency for low light video enhancement from single images," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[118] S. Zheng and G. Gupta, "Semantic-guided zero-shot learning for low-light image/video enhancement," in *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, 2022.

[119] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv:2304.07193*, 2023.

[120] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2014.

[121] Y. Ge, H.-X. Yu, C. Zhao, Y. Guo, X. Huang, L. Ren, L. Itti, and J. Wu, "3d copy-paste: Physically plausible object insertion for monocular 3d detection," in *Proceedings of Advances in Neural Information Processing Systems*, 2023.

[122] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of European Conference on Computer Vision*, 2016.

[123] G. Wang, Z. Chen, C. C. Loy, and Z. Liu, "Sparsenerf: Distilling depth ranking for few-shot novel view synthesis," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2023.

[124] J. Yang, M. Pavone, and Y. Wang, "Freenerf: Improving few-shot neural rendering with free frequency regularization," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[125] J. Collins, S. Goel, K. Deng, A. Luthra, L. Xu, E. Gundogdu, X. Zhang, T. F. Yago Vicente, T. Dideriksen, H. Arora, M. Guillaumin, and J. Malik, "Abo: Dataset and benchmarks for real-world 3d object understanding," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[126] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang, *et al.*, "Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai," *arXiv:2109.08238*, 2021.

[127] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi, "Objaverse: A universe of annotated 3d objects," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[128] X. Yu, M. Xu, Y. Zhang, H. Liu, C. Ye, Y. Wu, Z. Yan, C. Zhu, Z. Xiong, T. Liang, *et al.*, "Mvimgnet: A large-scale dataset of multi-view images," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

个人简历、在学期间研究成果

作者简介

罗润冬，男，2002 年 1 月出生于上海市，2020 年考入北京大学信息科学技术学院，攻读理学学士学位；2024 年 8 月将赴美国康奈尔大学计算机系攻读博士学位。

发表论文 （* 代表共同第一作者）

会议论文

[1] **Rundong Luo***, Yifei Wang*, and Yisen Wang. “Rethinking the Effect of Data Augmentation in Adversarial Contrastive Learning.” In Proceedings of International Conference on Learning Representations (ICLR), 2023.

[2] **Rundong Luo**, Wenjing Wang, Wenhan Yang, and Jiaying Liu. “Similarity Min-Max: Zero-shot Day-Night Domain Adaptation.” In Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV), 2023. **(Oral), top 2%.**

[3] **Rundong Luo***, Hong-Xing Yu*, and Jiajun Wu. “Unsupervised Discovery of Object-Centric Neural Fields.” Under Review at Neural Information Processing Systems (NeurIPS), 2024.

[4] **Rundong Luo***, Haoran Geng, Congyue Deng, Puhao Li, Zan Wang, Baoxiong Jia, Leonidas Guibas, and Siyuan Huang.“Physically-Plausible Part Completion for Interactable Objects.” Under Review at Neural Information Processing Systems (NeurIPS), 2024.

期刊论文

[1] Wenjing Wang*, **Rundong Luo***, Wenhan Yang, and Jiaying Liu. “Unsupervised Illumination Adaptation for Low-Light Vision.” IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2024.

专利申请

[1] 一种无监督低光照域自适应训练方法及检测方法，专利申请人：刘家瑛，**罗润冬**，汪文靖，专利申请号：202211129606.6，申请日期：2022年9月16日

致谢

四载寒暑，岁月如流。自从我怀揣憧憬踏入燕园的那一刻起，至今日的依依惜别，每一段时光都镌刻着成长的印记。这段旅程，既是学术的积累，也是人生的历练，使我从一个懵懂的学子蜕变为一个对科研充满热情的探索者。

在北大的四年，是我人生中最为宝贵的学习与成长阶段，培养了我自由、自强且务实的底色。自由是不受拘束的探索：北大自由的学术氛围与丰富的教育资源，让我得以自由探索自己热爱的方向；自强是向上攀登的信念，北大聚集了一群能力出众的同学，只有自强才能让我不辜负自己曾经的努力；务实则是做成事情的习惯，北大给予了我们各种人生的选择：学业科研、学生工作、实践实习、甚至谈恋爱，也让我意识到虚度时光不干事是停滞的人生。

北大的四年也是探寻人生目标并为之奋斗的阶段。在这四年里，我有幸遇到了许多优秀的老师和学长。他们不仅在学术上给予我无私的指导和帮助，更在人生道路上为我指引方向。我衷心感谢王奕森老师和王一飞学长，是他们引领我走进了科研的大门，我还要感谢 STRUCT 小组的刘家瑛老师、汪文靖学姐、樊泽嘉学姐、杨文瀚学长、胡煜章学长、黄浩峰学长、林里浪学长、张佳航学长。是你们的帮助和支持让我在科研道路上少走了许多弯路。回想起那些与大家在实验室度过的日夜，那些为了赶稿而熬过的通宵，大家在会议室举起手机自拍，照片里满是洋溢着活力的笑颜。

我的大学生活也因为有了同学和好友们的陪伴而更加丰富多彩。我最棒的室友宁川若、程凯、刘明昊；从高中到现在的朋友张泽楷、周书涵、胡煜骐、杨宇铭、张天宇；以及大学的新朋友们范博皓、耿浩然、周雨扬、吴乐鸿。我无法列出你们每个人的名字，但你们的陪伴让我在困境中找到了力量，在成功时分享喜悦。那一次次出游聚餐、一次次深夜卧谈，每一次，都有你们，都成为了我青春最珍贵的回忆。

特别要感谢我的爸爸妈妈。是你们无条件的爱和支持让我能够勇往直前，无论遇到什么困难和挑战都能够坚持下去。你们的关心和鼓励是我前进的动力源泉，也是我面对挫折时最坚实的后盾。十余载国内求学路接近尾声，但你们的教诲和关爱将永远伴随我前行，也愿我能更多地陪伴在你们身边。

回首本科四年，也有遗憾。大一的我在对绩点的追求中忽略了生活的多彩和科研的探索；后续几年虽然努力科研并矢志申请梦校，但终究未能如愿。然而这些遗憾也让我更加珍惜现在所拥有的一切，也让我更加明确自己未来的方向和目标。

文末搁笔，我也即将跨越大洋开启新的生活和学术旅程。就如同高中时数学联赛失利和高考超常发挥带我走上了如今的学术道路一样，选择与结果往往充斥着机缘巧合，希望我在抵达路的末端时，不会后悔现在的抉择。和大家顶峰相见！

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：

日期：　　年　　月　　日

学位论文使用授权说明

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；

论文作者签名：　　导师签名：

日期：　　年　　月　　日