**Computer Science and Application** 计算机科学与应用, 2023, 13(4), 708-719
Published Online April 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.134070

Hans汉斯

# 单帧非自然图像深度估计与动态合成

**余林江，杨伊欣**

合肥工业大学计算机与信息学院，安徽 合肥

收稿日期：2023年3月13日；录用日期：2023年4月10日；发布日期：2023年4月17日

## 摘 要

**深度学习在单目深度估计任务上表现优异，通过学习单帧图像与深度图像之间存在的映射关系来估计图像的深度。但是，目前单目深度估计的研究仅关注于自然图像，当把它应用于非自然图像，如绘画图像时，相对于自然图像，它们有着低纹理、切边锐利、平滑过渡相对少的特点，会出现深度估计中前后不同物体的层次感不明显，以及同一物体上出现深度不一致的问题。本文根据这类图像设计了一个由单目深度估计模块和RGB图像指导的精细化模块构成的精细单目深度估计网络RefineDepth来改善以上问题。同时，由于绘画图像缺乏对应深度信息，本文通过三维场景卡通风格渲染图像来模拟绘画类非自然图像的方式，制作了两个绘画图像数据集SSMO和SU3D，并建立了一个真实的山水画测试集。实验结果表明，模型在测试的数据集上都取得了出色的结果。最后，将绘画图像进行基于深度图像渲染，动态合成立体效果。**

## 关键词

**单目深度估计，非自然图像，精细化，绘画图像数据集，基于深度图像渲染**

# Depth Estimation and Dynamic Synthesis of Single Frame Unnatural Images

**Linjiang Yu, Yixin Yang**

School of Computer and Information, Hefei University of Technology, Hefei Anhui

Received: Mar. 13th, 2023; accepted: Apr. 10th, 2023; published: Apr. 17th, 2023

## Abstract

**Deep learning performs well in monocular depth estimation tasks, estimating the depth of an image by learning the mapping relationship between a single image and a depth image. However, the current research on monocular depth estimation only focuses on natural images. When it is applied**

文章引用: 余林江, 杨伊欣. 单帧非自然图像深度估计与动态合成[J]. 计算机科学与应用, 2023, 13(4): 708-719.
DOI: 10.12677/csa.2023.134070

**to unnatural images, such as painting images, they have low texture, sharp cutting edges, and relatively few smooth transitions. In depth estimation, the layering of different objects before and after is not obvious, and the depth of the same object is inconsistent. Based on such images, this paper designs a refined monocular depth estimation network RefineDepth, consisting of a monocular depth estimation module and a RGB image-guided refinement module to improve the above problems. Meanwhile, due to the lack of corresponding depth information in the painting image, we render images in a cartoon style of 3D scenes to simulate the way of painting unnatural images to make two painting image datasets SSMO and SU3D, and build a real landscape painting test set. The experimental results show that the model has achieved excellent results on the tested datasets. Finally, the painting image is rendered based on the depth image, and the three-dimensional effect is dynamically synthesized.**

## Keywords

**Monocular Depth Estimation, Unnatural Images, Refinement, Painting Image Datasets, Depth Image Based Rendering**



## 1. 引言

随着 VR、AR 等技术的出现，图像的立体化展示逐渐成为对影像等媒介的需求。除了常见的 3D 电影[1]等方面的娱乐应用，图像立体化还可以应用于广告等商业应用，比如通过立体化的海报来实现更好的宣传效果。常用的立体化实现方式有建模、全息投影、神经辐射场(Neural Radiance Field, NeRF) [2]等，其中单目深度估计(Monocular Depth Estimation, MDE)是从单幅图像中估计深度，通过预测平面图像中每个像素的深度值来实现从平面到立体的投影。其研究方法不管是传统的物理方法，还是概率图模型，或者是现如今主流的深度学习，都仅关注于自然图像。

非自然图像，比如手工绘制的绘画图像，不同于建模场景中真实计算的景物深度，没有深度的地面真实值(Ground Truth, GT)。同时，绘画图像有着天生的低纹理、切边锐利、平滑过渡较少等特点，在使用关注于自然图像的模型估计深度时，往往会出现深度估计中前后不同物体的层次感不明显，以及同一物体上出现深度不一致的问题。

针对上述问题，本文提出了一个针对非自然图像的基于 Transformer 网络[3]的单目深度估计网络。首先，通过一个单目深度估计网络估计出场景深度分布，可获得该场景下的深度图像；其次，在后处理中引入了 RGB 图像指导的精细化(RGB-Guided Refinement, RGR)模块，能更好地恢复在深度估计过程中缺失的纹理和边缘等细节：利用原图作为指导，以深度图的地面真实值作为目标进行精细化，增加深度图的细节部分，以提升整体效果。另外，使用基于深度图像的渲染(Depth Image Based Rendering, DIBR)技术[4]实现从原始视图和深度帧生成虚拟视图，实现图像的动态合成。

对于绘画图像而言，由于其主观创造性和抽象性，不存在可以直接使用的成对 RGB 图像和对应的深度地面真实值，往往需要手动去绘制深度，消耗大量人力物力。针对这一问题，本文使用三维场景的卡通风格渲染得到的图像来模拟绘画类非自然图像，以及场景中自带的深度信息作为深度值来设计并制作了两个数据集。模型通过学习这些数据集中非自然图像和对应深度的映射关系，来估计绘画图像的深度

值。同时，本文中制作了一个手工绘制的山水画测试集来测试模型。实验结果表明，模型在测试集上达到了满意的结果。

## 2. 相关工作

单目深度估计在过去通常根据阴影[5]、消失点[6]等图像特征进行计算。然而，这些中的大多数应用于特定的约束场景，方法复杂且低效，缺乏可行性。随着计算机视觉的发展，提出了许多手工特征和概率图模型，例如尺度不变特征变换(Scale-Invariant Feature Transform, SIFT) [7]、条件随机场(Conditional Random Field, CRF) [8]和马尔可夫随机场(Markov Random Field, MRF) [9]，然而这些传统的机器学习方法严重依赖于图像纹理信息，预测结果缺乏精度。

在深度学习的主流框架中，卷积神经网络(Convolutional Neural Networks, CNN)可以自动提取表示场景中深度的空间特征，作为一种前馈神经网络，与传统方法相比，它以更少的参数同时提取深度特征和重建深度图，利用彩色图像和深度图像的映射关系来训练模型，实验证明，单帧图像可通过模型获得较为准确的深度图。Eigen 等人[10]首次使用 CNN 来处理单目深度估计。在此基础上，Wang 等人[11]应用了多尺度融合卷积框架来生成高质量的深度预测。后来，Godard 等人[12]提出的基于双目图像的无监督单目深度估计网络，通过估计出的双目视差图和原双目图像交叉重建，得到重建的双目图像，并且证明利用成对的立体图像和立体视频可以实现单目深度估计的模型训练。最近，Transformer 网络以其强大的性能在计算机视觉领域收到了广泛关注，其中就包括单目深度估计。Yang 等人[13]通过跨深度在瓶颈处嵌入视觉 Transformer (Vision Transformer, ViT)，避免 Transformer 丢失局部信息，以及使用一个注意门解码器融合多级特征。Bhat 等人[14]将自适应容器根据输入场景的表示动态改变，并提出在解码器之后以高分辨率嵌入。Ranftl 等人[15]从密集预测的角度出发，结合 Transformer，利用大规模高分辨率数据的跨数据集方式预训练模型，在单目深度估计领域取得突破。以上的这些算法聚焦于自然图像的深度估计，在非自然图像中的效果却不够信服。

针对图像的细节恢复，多数的深度学习网络选择密集预测来实现。RefineNet 网络[16]，利用不同层特征来完成语义分割。其递归的方式，使用低水平的特征来生成高分辨率的图像。Zhang 等人[17]选择按图像序列预测每个任务，利用上个阶段预测一个任务的信息，在每次迭代中细化另一个任务的特征。Zhou 等[18]通过分离任务间和任务内模式，完善了像素间依赖的使用，从而提取信息。在最新的细节恢复方法中，Hu 等人[19]利用了超分辨率的网络，以掩膜为指导，通过局部精细化进行全局复原。

## 3. 算法介绍

RefineDepth 模型由一个单目深度估计模块(MDE module)以及一个 RGB 图像指导的精细化模块(RGR module)组成的架构(图 1)。最后，通过 DIBR，将原图和获得的最终深度图渲染成 3D 立体图像。单目深度估计模块，负责从彩色绘画图像中估计深度并预测出粗略的深度图；RGB 图像指导的精细化模块，以原有的 RGB 绘画图像作为指导，将粗略的深度图精细化处理，恢复更多的图像深度细节，获得更好的深度估计效果。

**单目深度估计模块(图 2)：**灵感来自 DPT，本文从单目深度估计的视角，探索了使用单个 ViT 结构 ViT-base 和单个头部通过轻量化来实现单目深度估计的可能性。在模型中把输入原图像分割为 16 × 16 的若干块，块的数量为 $H/16\times W/16$，$H$、$W$ 为 RGB 图像的高和宽。将这些块经过特征提取块得到 $H/16\times W/16$ 个 1 × 1 的 768 维标记，这里还添加了一个独立的标记，用于后续将以上标记整合为一个类似图像的像素级特征表达。由于标记特征维度大于输入块中的像素数量，这说明可以完整地学习保留信息。实际上，输入的标记正是以像素级的精度对块中的特征进行解析。在 ViT-base 提取模块中将这些标记送入在

ImageNet [20]上预训练过的视觉 Transformer 中。在网络中使用了 12 层的 Transformer，并在第{3, 6, 9, 12}层增加了一个分支，将不同阶段提取的特征表达送入重组(Recast)模块。

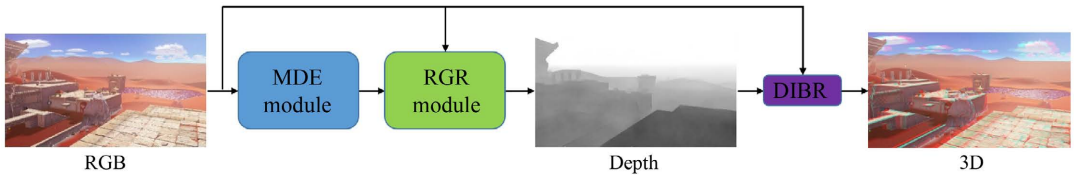


**Figure 1.** Overall framework of RefineDepth model
**图 1.** RefineDepth 模型整体框图

这些特征表达通过重组块进行组合，获得类似图像的像素级输出，这里利用了 CNN 的思想，保留不同阶段的 Transformer 的特征表达，保留不同程度的图像细节，以此更好地估计场景深度。这些特征表达可以在后续的迭代中不断融合上级特征。

在重组模块中，将尺度 $(H/16\times W/16+1)\times D$ 首先映射到尺度为 $H/16\times W/16\times D$ ($D$ 为标记的维度)；然后将这些标记组合起来，使每个特征表达对应一个块，再按照块的顺序连接每一个特征表达，这就产生了中间特征图这一类似于图像的特征表达；再将输入的特征表达经过一个 1 × 1 的卷积投影到 256 维的空间中。除此之外，另加了一个的卷积，它的内部参数是根据上一层重组块输入的尺度而变化；最后经过卷积的特征表达 $s$ 和上一个阶段的重组块作为输入，在求和之后使用两个连续的卷积单元，并对预测的结果进行 2 倍上采样，自上而下进行融合各级特征，最终将尺寸为输入图像一半的特征表达输出到深度重建(Depth Reconstruction)头模块中。头模块由一个带有上采样模块的反卷积组成，输出的深度图与 RGB 原图同尺度同维度。

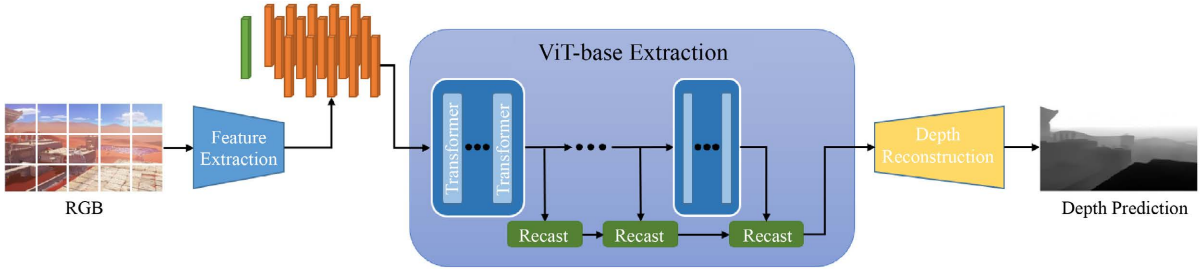


**Figure 2.** MDE module
**图 2.** MDE 模块

**RGB 图像指导的精细化模块(图 3)：**使用单目深度估计网络估计深度的时候，由于没有明确的边缘信息作为指引，往往会使同一平面的被测物体的深度估计不一致，导致深度的离散化。绘画图像，由于其简单的纹理、锐利的切边和较少的平滑过渡，相对于自然图像来说，利用其边缘和纹理等信息可能更好地恢复原有的图像细节，从而获得更加准确的场景深度。因此，提出利用输入的 RGB 原图作为指导，通过增加特征图支路的方式，在深度图像精细化的过程中改善边缘细节。

模块由两部分组成，第一部分是 RGB 原图的支路，将原图灰度化之后经过一个 3 × 3 卷积得到特征图，再将特征加入后续的网络中；第二部分中，设计了一个图到图的回归模型来作为网络主干来精细图像细节，该模型基于常见的图像重建任务，如超分辨率。本文使用 Wang 等人[21]提出的残差密度块(Residualin Residual Dense Block, RRDB)来搭建网络。网络中共有 24 个 RRDB 块，分别在第 5、10、15、

20 个块后面对上个分支的特征融合，并将输出特征送入下个块中。最后融合特征信息，具体来说，先将这两个特征合并起来，然后使用另一个 RRDB 块和两个卷积来重建最终的 SR 特征。

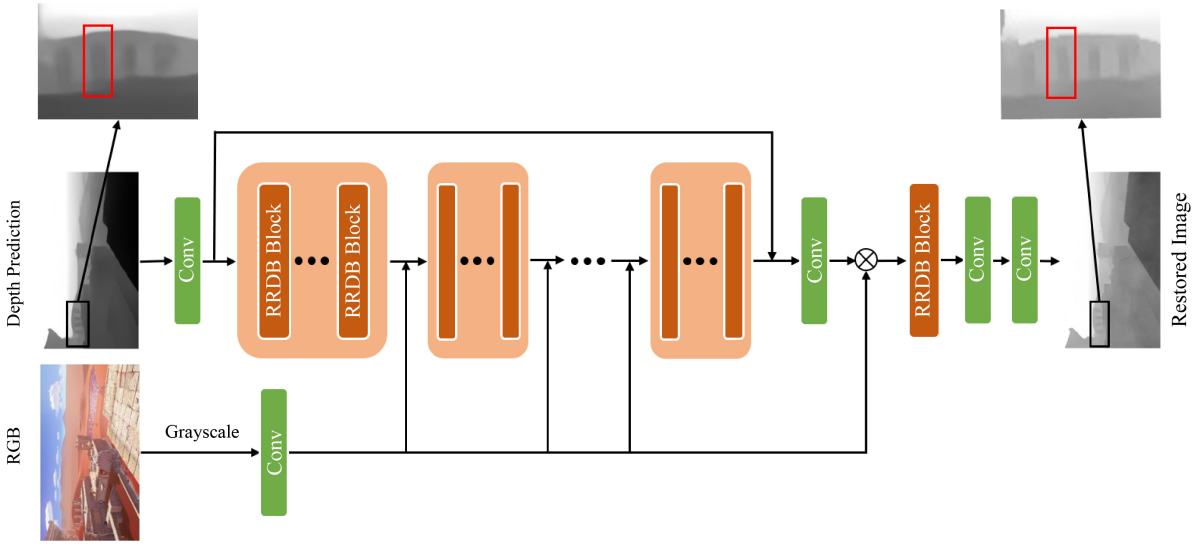


**Figure 3.** RGR module
**图 3.** RGR 模块

## 4. 数据集制作

**SSMO (Shot on Super Mario Odyssey)数据集(图 4)：**本文在《超级马里奥奥德赛》这款游戏中选取了 13 种场景，包括沙漠、城市、海滩等。控制游戏角色在不同场景中以不同视角截取场景图和漫画滤镜图、线稿以及对应的游戏中建模自带的深度图共 1344 对，图像分辨率为 1280 × 720 像素。为了平衡每种场景下的学习效果，提升模型的泛化性，平均地截取了不同场景中的相同数量的图片对。

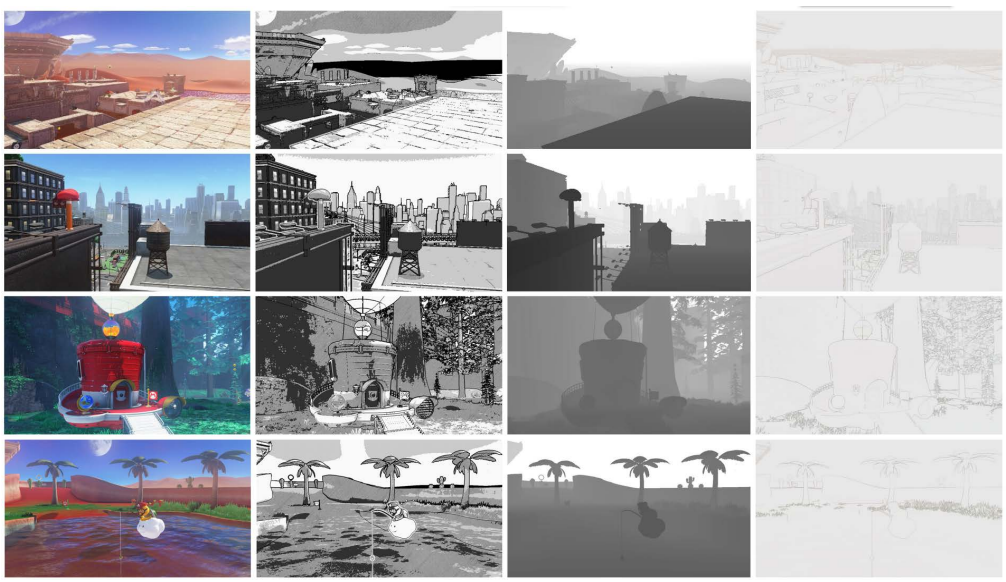

**Figure 4.** SSMO dataset (from left to right): RGB, manga, corresponding GT depth, line draft
**图 4.** SSMO 数据集示例(从左到右)：RGB、漫画、对应 GT 深度、线稿

**SU3D (Shot on Unity3D)数据集(图 5)：**本文利用 Unity3D 构建了大量的三维场景模型并开发了一套轨道相机深度图采集系统，使用该系统采集数据集。与从真实场景中采集深度数据相比，该方法具有易于调整、易于采集、场景类型多样、降低人工成本等诸多优点。根据需要构建不同类型的场景，例如室内布局、城市立面、街景、自然景观等。同时，可以调整相机角度，从不同的视角获取场景，或者对场景中的物体进行增删改查，从不同的场景中获取大量对应的深度数据。在系统中可以根据需要选择场景、在场景中切换相机运行轨道、设置拍摄速度，以及选取不同的相机高度及视角进行拍摄，以此获取 RGB 图。同时，绑定深度计算脚本，确定好场景、轨道和视角后，首先获取 RGB 图像，着色器通过纹理采样得到深度值后赋予深度计算脚本，即可得到 RGB 图像对应的场景深度图。在本文的实验中，我们共采集了 RGB 图像和深度图像共 1000 对，图像分辨率为 640 × 480 像素。在场景方面，本文共选取了 10 个虚拟场景用于图片对的采集，场景类型包括城市街道、乡野田园和森林景象等。与采集 SSMO 数据集相似，本文在采集 SU3D 数据集过程中亦进行了数据的平衡。

为了测试本文的模型在绘画图像上的效果，以验证泛化性性能表现，本文收集了 165 张绘画作品制作了山水画测试集(图 6)，图像分辨率为 1280 × 720 像素。

除此之外，本文还使用了自然图像 NYUv2 数据集[22]共 1449 张，分辨率为 640 × 480 像素，它包含地面真实深度值 GT，以测试本文的模型在自然图像上的表现。


**Figure 5.** (a) The maininterface of the track camera depth map acquisition system; (b) Example of indoor scene tracks; (c) Examples of a SU3D dataset: the first and third columns are the original image, and the second and fourth columns are the corresponding GT depth

**图 5.** (a) 轨道相机深度图采集系统主界面；(b) 室内场景轨道示例；(c) SU3D 数据集示例：第 1、3 列为原图，第 2、4 列分别为对应的 GT 深度

**Figure 6.** Painting image test set
**图 6.** 绘画图像测试集

## 5. 实验

### 5.1. 模型训练

由于单目深度估计和精细化后处理两个任务互相独立，为了更灵活地使用，本文将两个模块分开训练。数据方面把 SSMO 和 SU3D 数据集混合，其中保留 70%的数据集作为训练集，10%作为验证集，其余 20%作为测试集。

在 DPT 中，作者已验证，使用在大型数据集上预训练的 ViT 模块有助于提升网络的学习能力，因此选取经过 ImageNet 预训练的 ViT 作为基础块应用在网络里，由于非自然图像与预训练的数据集有较大区别，因此将该基础块在我们的训练集上再次训练。本文调整批次大小 batch_size 为 8，使用 2 个不同的优化器和不同的学习率来训练模型。具体来说，本文使用优化器 $O_s$ 来更新解码器从头训练的参数，使用另一个优化器 $O_b$ 来更新。对于 $O_s$，本文使用 Adam [23]优化器，学习率为 $3 \times 10^{-5}$；对于 $O_b$，本文同样使用 Adam，学习率设置为 $1 \times 10^{-5}$，在 300 个 epoch 中训练本文的模型。对于单目深度估计来说，是旋转裁切翻转不相关的，本文将输入图像进行 50%随机水平翻转、30%随机裁剪和 20%随机旋转(最大角度 10 度)进行数据增强，增加学习样本的数量，以提升模型的鲁棒性。

RGB 指导的精细化模块中，本文使用目前性能较好的 RRDB 块，采用更加密集的连接进行迭代，并使用 L1 loss 训练模型。由于输入图像和目标图像都是同分辨率，本文将这两者都裁剪至 $64 \times 64$，batchsize 设置为 8 送入网络，使用预训练的面向 PSNR 模型的参数初始化生成器。学习率初始设置为 $2 \times 10^{-4}$，最小为 $1 \times 10^{-7}$，每过 $2.5 \times 10^4$ 次迭代衰减 2 倍，共迭代 $1.0 \times 10^5$ 次结束。本文使用 Adam 优化器，权重衰减为 0，控制参数 $\beta_1$ 设置为 0.9，每迭代 1000 次记录数据。

所有的训练过程均在 Windows 11 系统搭载的 12 GB NVIDIARTX 3060 和 2.5 GH zInteli7-11700 CPU 上完成，在 Pytorch-GPU 11.6.0 环境下使用 Python 3.6 实现了模型。

### 5.2. 结果与分析

#### 5.2.1. 在绘画数据集上对比实验

为了进行直观的对比，我们将本文所提模型与 CNN 为主干网络的 Monodepth2 (mono)和以 Transformer 为主干，在当今单目深度估计方面表现出色的 DPT、GCN 网络模型，在本文提出的绘画图像数据集上训练后的表现进行对比，体现所提出模型的学习能力。

第 1、2 行、第 3、4 行分别是不同模型在 SSMO 数据集、SU3D 数据集预测结果对比(图 7)。我们观察到，monodepth2 由于卷积主干网络更关注于像素级密集预测任务，在学习过程中对于图像纹理把握较差，其结果表现不够出色；GCN 和 DPT 借助 Transformer 架构提升了对图像的全局性和边缘纹理细节的

把握，但是对绘画图像这种过渡平滑少的场景，在估计景物时还是会缺失深度一致性。相较于文中的其他模型而言，本文提出的模型可以从原图中学到更好的深度信息并更好地恢复深度图像的纹理和边缘信息，从而达到准确预测场景深度。

此外，还将 monodepth2、GCN 和 DPT 在绘画图像数据集训练前和训练后的推理结果进行了对比(图 8)。在深度估计中本文更关注结果中场景边缘和结构性的提升，所以使用 SSIM 这一指标(数值越大表示越接近地面真实值)对推理结果进行量化(表 1)。可以看出，在本文的绘画图像数据集上训练之后，不同模型对绘画场景的预测能力都有提升。

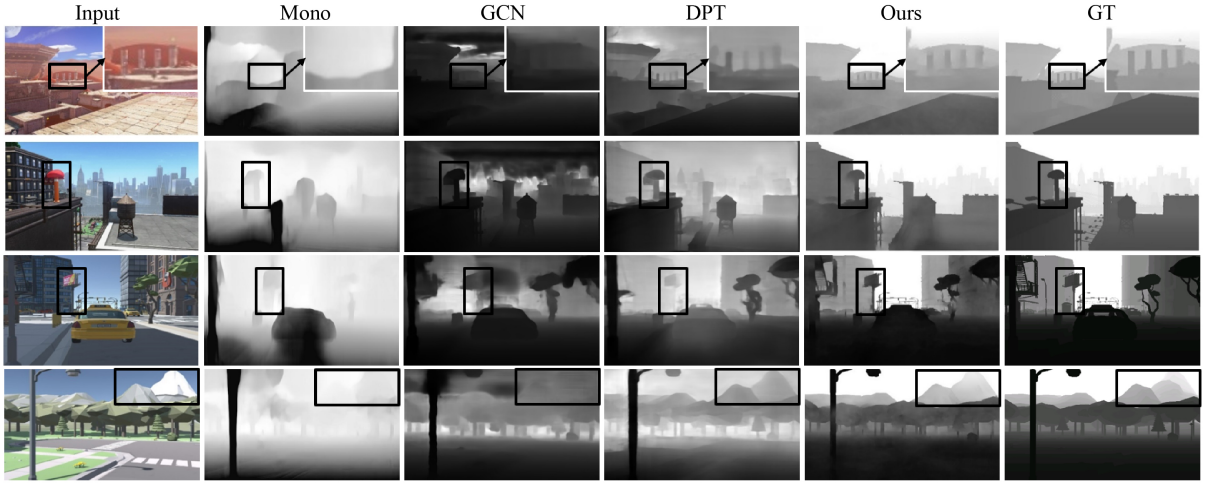


**Figure 7.** Performance of different models after training on the painting dataset
**图 7.** 不同模型在绘画数据集上训练后的表现

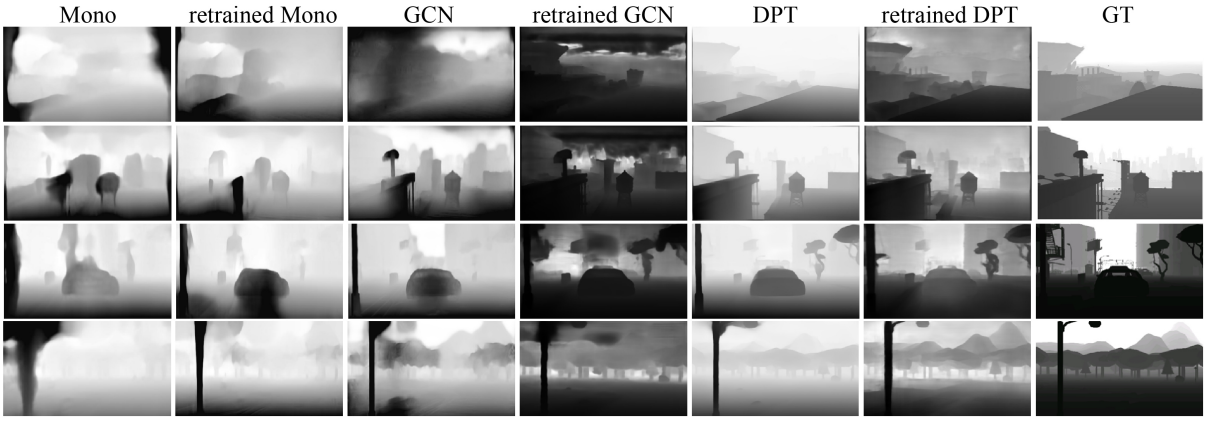


**Figure 8.** Comparison of different models on the painting dataset after training and before training
**图 8.** 不同模型在绘画数据集上训练后与训练前对比

**Table 1.** Average SSIM metrics of different models on the painting dataset
**表 1.** 不同模型在绘画数据集上的平均 SSIM 指标

| | Mono | Retrained mono | GCN | Retrained GCN | DPT | Retrained DPT | Ours |
|---|---|---|---|---|---|---|---|
| SSMO & SU3D | 0.60 | 0.65 | 0.63 | 0.70 | 0.66 | 0.71 | 0.92 |

在消融实验中，本文探索了 RGB 指导的精细化(RGR)模块在深度估计中的作用(图 9)。同样以 SSIM 作为指标，分别计算精细化前和精细化后的结果。经过精细化后的图像质量均有不同程度的提升，SSIM 的数值也从精细化前的 0.70 提升至精细化后的 0.91。

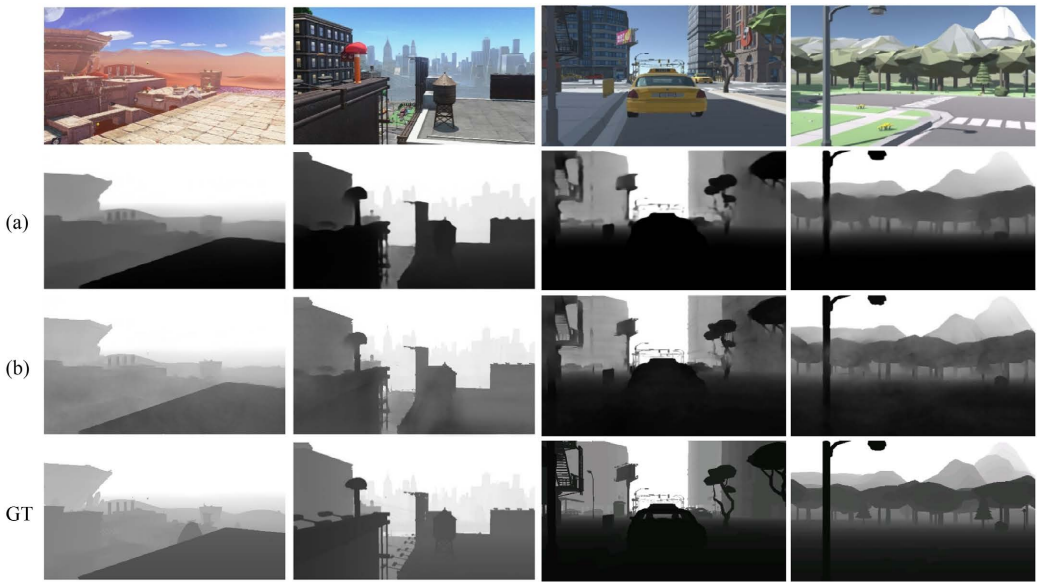


**Figure 9.** Visual comparison of images before and after the RGR module: (a) before passing through the RGR module; (b) after passing through the RGR module
**图 9.** RGR 模块的前后图像直观对比：(a) 经过 RGR 模块前；(b) 经过 RGR 模块后

### 5.2.2. 在山水画测试集上对比实验

本文在山水画测试集上测试了在绘画图像数据集上重新训练后各个模型的预测结果(图 10)，并且与没有重新训练的模型进行对比(图 11)。结果表明，本文模型在其他绘画图像上依然可以更好地估计深度并恢复深度图像的纹理及边缘信息，准确预测场景深度。同时，也说明本文提出的绘画数据集在非自然图像中具有很好的泛化性。

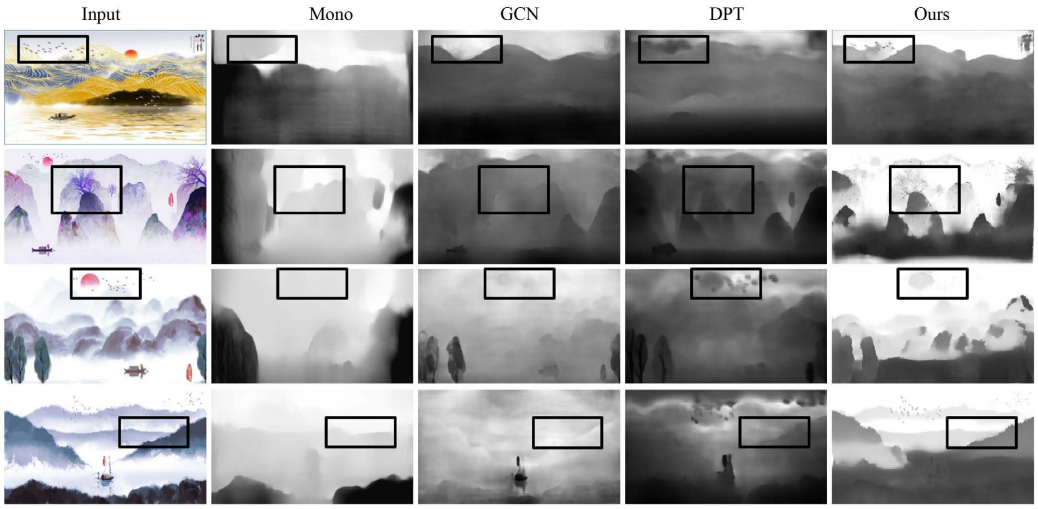


**Figure 10.** Comparison results of different models on the landscape painting test set
**图 10.** 不同模型在山水画测试集上对比结果

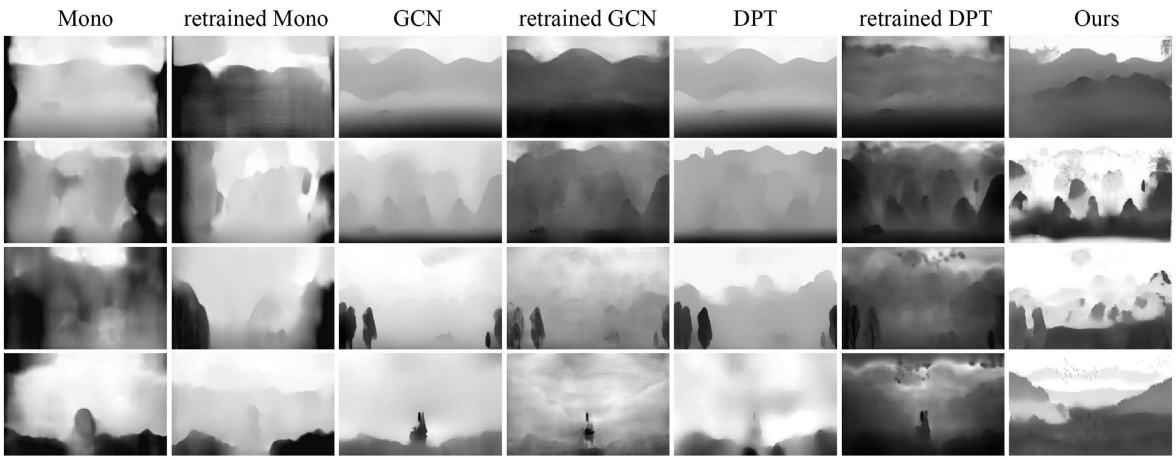


**Figure 11.** Comparison of different models on the landscape painting test set after training and before training
**图 11.** 不同模型在山水画测试集训练后与训练前对比

### 5.2.3. 在自然图像上的拓展实验

为了测试本文模型在自然图像上的表现，将模型在 NYUv2 数据集上重新训练并测试(图 12)。结果显示，在自然图像上也可以获得接近地面真实值的深度估计结果，测试 SSIM 为 0.88，说明本文模型在自然图像上也能很好的恢复图像细节，泛化性出色。

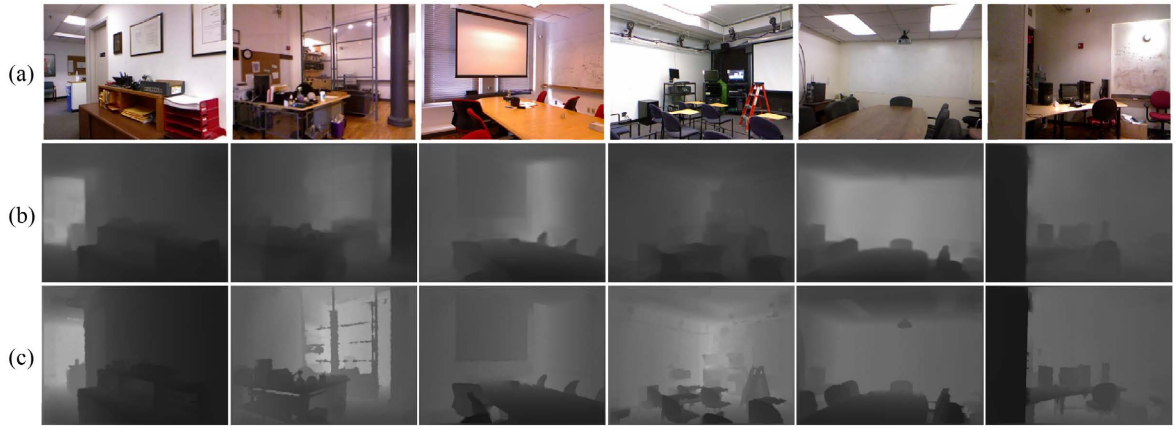


**Figure 12.** The effect of model in this paper on the NYUv2 dataset: (a) original image; (b) depth estimation image; (c) GT
**图 12.** 本文模型在 NYUv2 数据集上的效果：(a) 原图；(b) 深度估计图；(c) GT

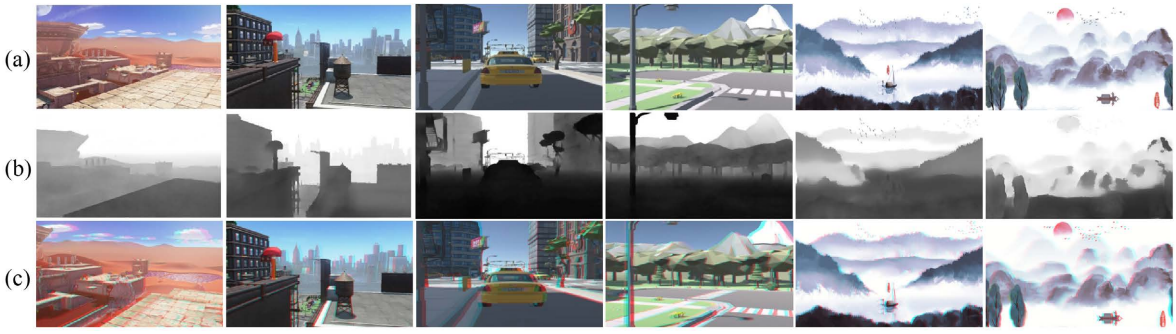


**Figure 13.** Dynamic effect display: (a) RGB image; (b) depth image; (c) red and blue image
**图 13.** 动态效果展示：(a) RGB 图；(b) 深度图；(c) 红蓝图

#### 5.2.4. 基于 DIBR 的动态效果制作

最后，利用 DIBR 技术，将 RGB 图与本文模型估计的深度图进行动态生成，来展示图像的立体化结果。由于文章不能展示动态效果，本文使用深度图进行双向 warp 到虚拟视角，获得左右视图，并通过左右视图偏移模拟视差来表现动态效果，如红蓝图所示(图 13)。本文模型合成结果可清楚展示图中红蓝偏移量，说明模型能够成功胜任动态合成任务，具体动态效果可访问文末链接查看。

## 6. 总结

本文提出 RefineDepth 网络来学习非自然图像在单目深度估计上的表现，并根据绘画图像的特点创新地设计了 RGB 指导的精细化模块对单目深度估计的结果进行后处理。实验结果表明，与以往的单目深度估计模型相比，本文提出的模型在绘画作品上得到了更细粒度和全局一致的预测，经过 DIBR 之后，本文的预测结果也实现了很好的图像立体感。同时，本文的模型在自然图像上也获得了出色的表现。除此之外，设计制作的绘画图像数据集，包括绘画数据集 SSMO 和 SU3D 以及山水画测试集，除了单目深度估计任务之外，同样也可以应用于风格迁移、边缘检测等领域，具有很高的应用价值。立体效果展示及数据集公开详见 https://github.com/PapillonYu/Depth-Estimation-and-Dynamic-Synthesis。

## 参考文献


[1] Xie, J., Girshick, R. and Farhadi, A. (2016) Deep3D: Fully Automatic 2D-to-3D Video Conversion with Deep Convolutional Neural Networks. In: Leibe, B., Matas, J., Sebe, N. and Welling, M., Eds., *Computer Vision—ECCV* 2016, *Lecture Notes in Computer Science*, Springer, New York, 842-857. https://doi.org/10.1007/978-3-319-46493-0_51

[2] Wang, Z., Wu, S., Xie, W., *et al*. (2021) NeRF−−: Neural Radiance Fields without Known Camera Parameters. Preprint. ArXiv: 2102.07064.

[3] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You need. *Advances in Neural Information Processing Systems*, 5998-6008.

[4] 王震. 自由视点立体电视系统的虚拟视点合成技术研究[D]: [硕士学位论文]. 上海: 上海交通大学, 2012.

[5] Prados, E. and Faugeras, O. (2006) Shape from Shading. In: Paragios, N., Chen, Y. and Faugeras, O., Eds., *Handbook of Mathematical Models in Computer Vision*, Springer, Boston, MA, 375-388. https://doi.org/10.1007/0-387-28831-7_23

[6] Tsai, Y.M., Chang, Y.L. and Chen, L.G. (2006) Block-Based Vanishing Line and Vanishing Point Detection for 3D Scene Reconstruction. 2006 *International Symposium on Intelligent Signal Processing and Communications*, Yonago, 12-15 December 2006, 586-589. https://doi.org/10.1109/ISPACS.2006.364726

[7] Lowe, D.G. (1999) Object Recognition from Local Scale-Invariant Features. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Kerkyra, 20-27 September 1999, 1150-1157. https://doi.org/10.1109/ICCV.1999.790410

[8] Lafferty, J., McCallum, A. and Pereira, F.C. (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the Eighteenth International Conference on Machine Learning*, San Francisco, CA, 28 June 2001-1 July 2001, 282-289.

[9] Cross, G.R. and Jain, A.K. (1983) Markov Random Field Texture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-5**, 25-39. https://doi.org/10.1109/TPAMI.1983.4767341

[10] Eigen, D., Puhrsch, C. and Fergus, R. (2014) Depth Map Prediction from a Single Image Using a Multi-Scale Deep Network. *Proceedings of the* 28*th Annual Conference on Neural Information Processing Systems*, Montreal, 8-13 December 2014, 2366-2374.

[11] Wang, L., Zhang, J., Wang, Y., Lu, H. and Ruan, X. (2020) Cliffnet for Monocular Depth Estimation with Hierarchical Embedding Loss. In: Vedaldi, A., Bischof, H., Brox, T. and Frahm, J.M., Eds., *Computer Vision—ECCV* 2020, *Lecture Notes in Computer Science*, Springer, Cham, 316-331. https://doi.org/10.1007/978-3-030-58558-7_19

[12] Godard, C., Aodha, O.M. and Brostow, G.J. (2017) Unsupervised Monocular Depth Estimation with Left-Right Consistency. 2017 *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 21-26 July 2017, 6602-6611. https://doi.org/10.1109/CVPR.2017.699

[13] Yang, G., Tang, H., Ding, M., Sebe, N. and Ricci, E. (2021) Transformers Solve the Limited Receptive Field for Mo-

nocular Depth Prediction. ArXiv: 2103.12091. https://arxiv.org/abs/2103.12091

[14] Bhat, S.F., Alhashim, I. and Wonka, P. (2021) Adabins: Depth Estimation Using Adaptive Bins. ArXiv: 2011.14141. https://arxiv.org/abs/2011.14141

[15] Ranftl, R., Bochkovskiy, A. and Koltun, V. (2021) Vision Transformers for Dense Prediction. 2021 *IEEE/CVF International Conference on Computer Vision*, Montreal, QC, 10-17 October 2021, 12179-12188. https://doi.org/10.1109/ICCV48922.2021.01196

[16] Lin, G., Milan, A., Shen, C., *et al*. (2017) Refinenet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 21-26 July 2017, 5168-5177. https://doi.org/10.1109/CVPR.2017.549

[17] Zhang, Z., Cui, Z., Xu, C., Jie, Z., Li, X. and Yang, J. (2018) Joint Task-Recursive Learning for Semantic Segmentation and Depth Estimation. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *Computer Vision—ECCV* 2018, *Lecture Notes in Computer Science*, Springer, Cham, 238-255. https://doi.org/10.1007/978-3-030-01249-6_15

[18] Zhou, L., *et al*. (2020) Pattern-Structure Diffusion for Multi-Task Learning. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognitio*n, Seattle, WA, 13-19 June 2020, 4513-4522. https://doi.org/10.1109/CVPR42600.2020.00457

[19] Hu, X.T., *et al*. (2022) Restore Globally, Refine Locally: A Mask-Guided Scheme to Accelerate Super-Resolution Networks. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M. and Hassner, T., Eds., *Computer Vision—ECCV* 2022, *Lecture Notes in Computer Science*, Springer, Cham, 74-91. https://doi.org/10.1007/978-3-031-19800-7_5

[20] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Li, F.F. (2009) Imagenet: A Large-Scale Hierarchical Image Database. 2009 *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, 20-25 June 2009, 248-255.

[21] Wang, X., Yu, K., Wu, S., *et al*. (2018) Esrgan: Enhanced Super-Resolution Generative Adversarial Networks. In: Leal-Taixé, L. and Roth, S., Eds., *Computer Vision—ECCV* 2018, *Lecture Notes in Computer Science*, Springer, Cham, 63-79. https://doi.org/10.1007/978-3-030-11021-5_5

[22] Silberman, N., Hoiem, D., Kohli, P., *et al*. (2012) Indoor Segmentation and Support Inference from RGBD Images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y. and Schmid, C., Eds., *Computer Vision—ECCV* 2012, *Lecture Notes in Computer Science*, Springer, Berlin, 746-760. https://doi.org/10.1007/978-3-642-33715-4_54

[23] Kingma, D.P. and Jimmy, B. (2017) Adam: A Method for Stochastic Optimization. https://arxiv.org/abs/1412.6980