北京交通大学
BEIJING JIAOTONG UNIVERSITY

北京交通大学软件学院本科毕业设计答辩

联合单目深度估计的深度图像超分辨率重建算法研究

Research on Joint Depth Map Super-Resolution and Monocular Depth Estimation Algorithm

答辩人：唐麒

学号：17301138

指导教师：冯凤娟

答辩时间：2022/6/12

目录

CONTENT

北京交通大学
BEIJING JIAOTONG UNIVERSITY

P 第一部分
Part One

项目简介

- 项目来源
- 研究背景
- 研究意义

1

项目来源

联合单目深度估计的深度图像超分辨率重建算法研究

北京交通大学 信息科学研究所

算法研究岗 2020 年 10 月 29 日 - 至今

探索深度图超分辨率重建任务中颜色引导、细节恢复、模态交互等问题的解决方案。具体地，从多任务学习的角度出发研究一种联合深度估计的深度图超分辨率网络，并探索两个任务之间的交互指导关系，以达到相互促进、互利共赢的效果。

信息科学研究所 Institute of Information Science

MePro 数字媒体信息处理研究中心 Center of Digital Media Information Processing

研究背景

研究背景

1. 深度信息和深度相机

自动驾驶等依赖于高质量的深度信息

便携式消费级深度相机的问世和普及

研究背景

1. 深度信息和深度相机

自动驾驶等依赖于高质量的深度信息

便携式消费级深度相机的问世和普及

2. 成像技术和图像分辨率

成像技术限制导致深度图像分辨率低

硬件设施提高分辨率成本消耗较高等

研究背景

1. 深度信息和深度相机

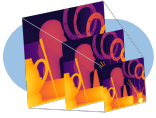
自动驾驶等依赖于高质量的深度信息

便携式消费级深度相机的问世和普及

2. 成像技术和图像分辨率

成像技术限制导致深度图像分辨率低

硬件设施提高分辨率成本消耗较高等



3. 深度图像超分辨率重建

按照是否需要训练分为非学习式超分

辨率重建算法和学习式超分辨率重建

研究背景

1. 深度信息和深度相机

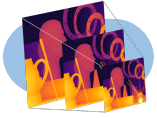
自动驾驶等依赖于高质量的深度信息
便携式消费级深度相机的问世和普及

4. 深度学习

自动驾驶等依赖于高质量的深度信息
便携式消费级深度相机的问世和普及

2. 成像技术和图像分辨率

成像技术限制导致深度图像分辨率低
硬件设施提高分辨率成本消耗较高等



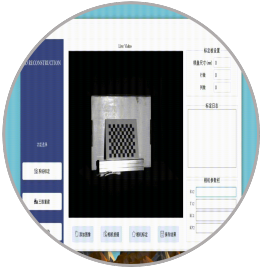
3. 深度图像超分辨率重建

按照是否需要训练分为非学习式超分
辨率重建算法和学习式超分辨率重建

研究意义

游戏领域

获取玩家姿态动作
可提高姿态识别率
提升玩家游戏体验



三维重建

深度相机获取点云
提高密集度和精度
建模真实三维模型

无人驾驶

确定无人车辆位置
获得更高定位精度
环境描述/避障操作

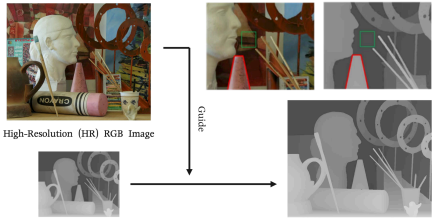
北京交通大学
BEIJING JIAOTONG UNIVERSITY

P 第二部分
Part Two

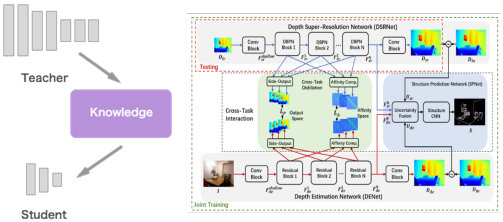
相关工作

- 项目目标

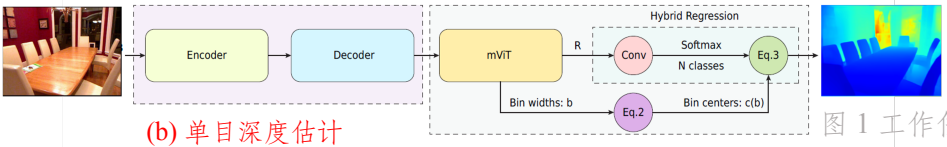
项目目标



(a) 深度图像超分辨率重建算法



(c) 面向深度图像的多任务联合学习



(b) 单目深度估计

图 1 工作任务

北京交通大学
BEIJING JIAOTONG UNIVERSITY

第三部分
Part Three

研究内容

- 网络设计

BridgeNet

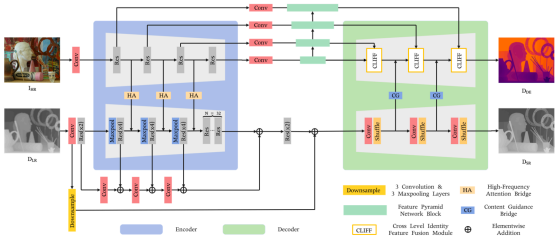


Figure 2: Architecture of the proposed BridgeNet, which consists of a depth super-resolution subnetwork (DSRNet), a monocular depth estimation subnetwork (MDENet), a high-frequency attention bridge (HABdg), and a content guidance bridge (CGBdg). The encoder-decoder structure at the top is the MDENet, and the bottom encoder-decoder structure corresponds to the DSRNet. The HABdg works between the feature encoders of the two subnetworks, focusing on passing the high-frequency color guidance obtained from MDENet to DSRNet. On the contrary, CGBdg works on the decoder side and is used to provide the MDENet with content guidance information learned from the DSRNet.

图 2 单目深度估计和深度图像超分辨率重建联合学习网络架构示意图

- 本文在联合学习网络中将深度图像超分辨率重建任务和单目深度估计任务相关联，以提升深度图像超分辨率重建的性能。本文的整个网络结构具有高度的可移植性，可以为关联深度图像超分辨率重建和单目深度估计任务提供范例。
- 本文提出的联合学习网络包括深度图像超分辨率重建子网络（DSRNet）和单目深度估计子网络（MDENet），以及两个用于联合学习的桥接器，即高频注意力桥（HABdg）和内容引导桥（CGBdg）。
- 在不引入其他监督信息的情况下，本文的方法在多个公开基准数据集上均达到了具有竞争力的性能。

BridgeNet

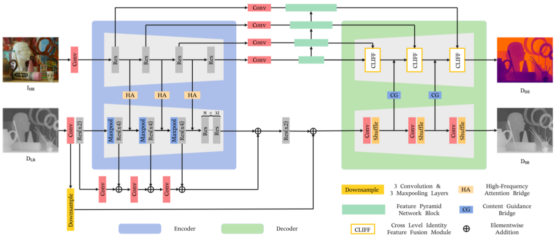


Figure 2: Architecture of the proposed BridgeNet, which consists of a depth super-resolution subnetwork (DSRNet), a monocular depth estimation subnetwork (MDENet), a high-frequency attention bridge (HABdg), and a content guidance bridge (CGBdg). The encoder-decoder structure at the top is the MDENet, and the bottom encoder-decoder structure corresponds to the DSRNet. The HABdg works between the feature encoders of the two subnetworks, focusing on passing the high-frequency color guidance obtained from MDENet to DSRNet. On the contrary, CGBdg works on the decoder side and is used to provide the MDENet with content guidance information learned from the DSRNet.

图 2 单目深度估计和深度图像超分辨率重建联合学习网络架构示意图

给定一组高分辨率的 RGB-D 图像对 $\{I_{HR}^{(n)}, D_{DR}^{(b)}\}_{(n=1)}^{N}$ 和相应的低分辨率深度图像 $\{D_{LR}^{(n)}\}_{(n=1)}$ 作为训练数据，其中 N 是训练图像的数量。此外，低分辨率深度图像在输入网络前被插值到高分辨率深度图像的大小。本文提出的网络以低分辨率深度图像（D_{LR}）和相应的高分辨率彩色图像（I_{HR}）作为输入，同时对深度图像超分辨率重建子网络和单目深度估计子网络进行训练。超分辨的深度图像（D_{SR}）是本文网络的主要输出，此外，估计的深度图像（D_{DE}）也作为辅助输出。

单目深度估计子网络

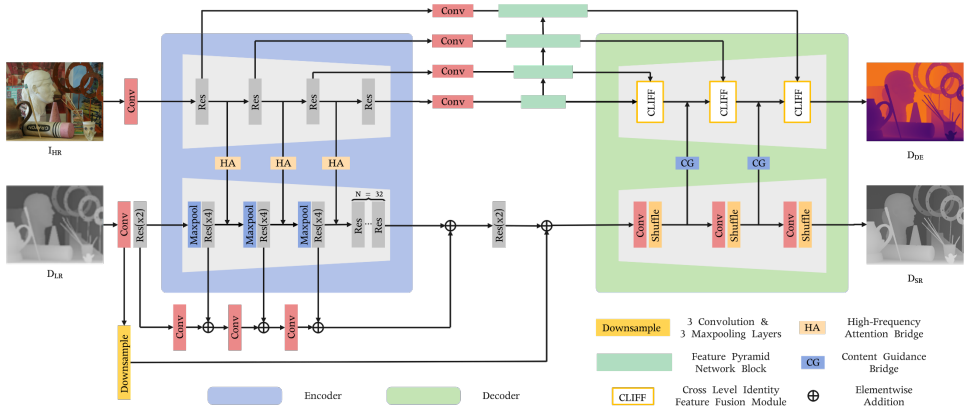


图 3 单目深度估计子网络（MDENet）

单目深度估计子网络

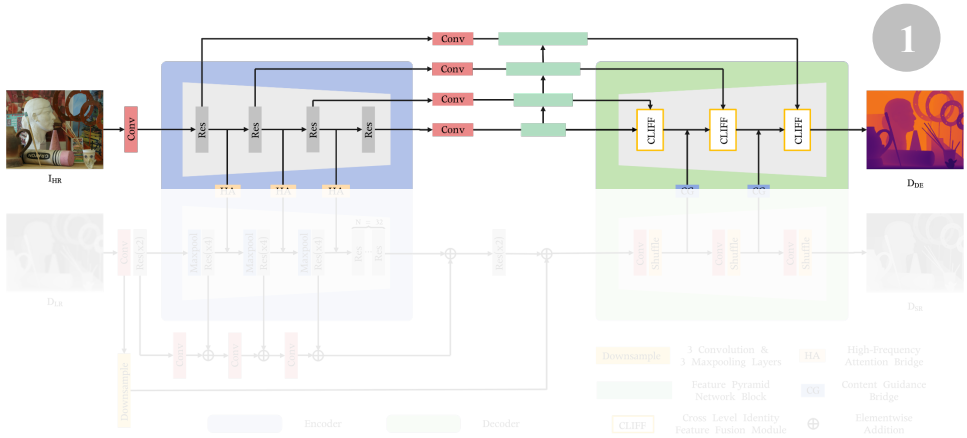


图 3 单目深度估计子网络（MDENet）

单目深度估计子网络

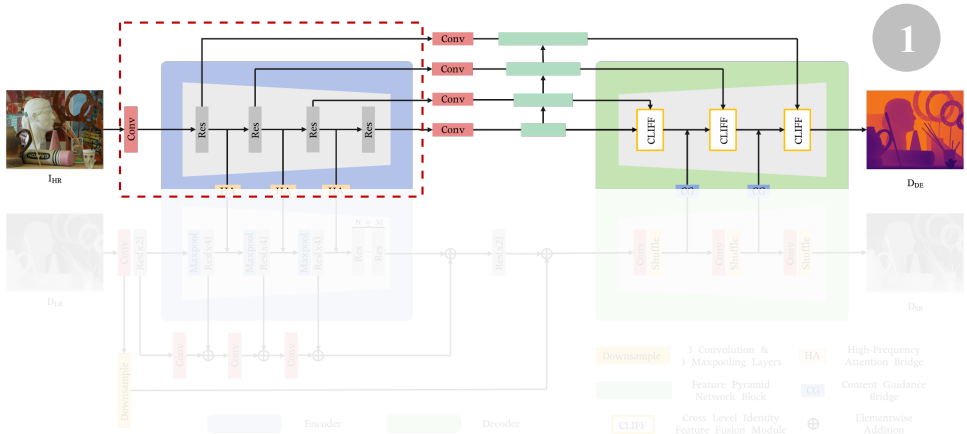


图 3 单目深度估计子网络（MDENet）

单目深度估计子网络

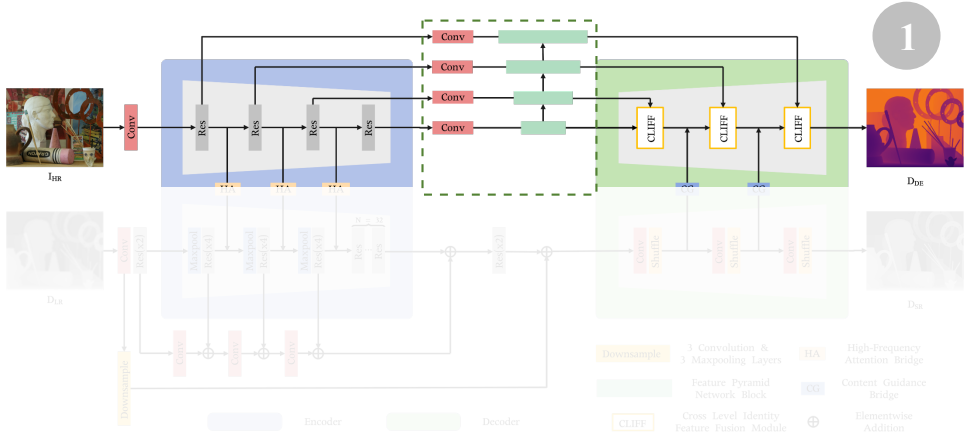


图 3 单目深度估计子网络（MDENet）

单目深度估计子网络

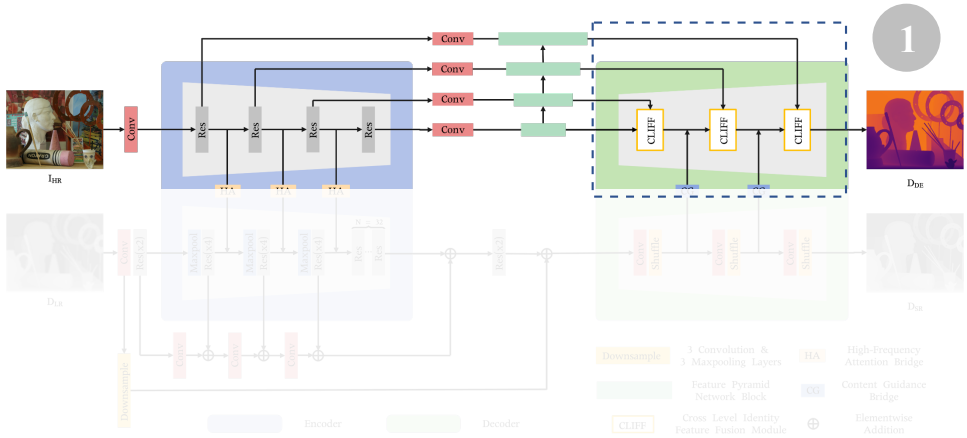


图 3 单目深度估计子网络（MDENet）

深度图像超分辨率重建子网络

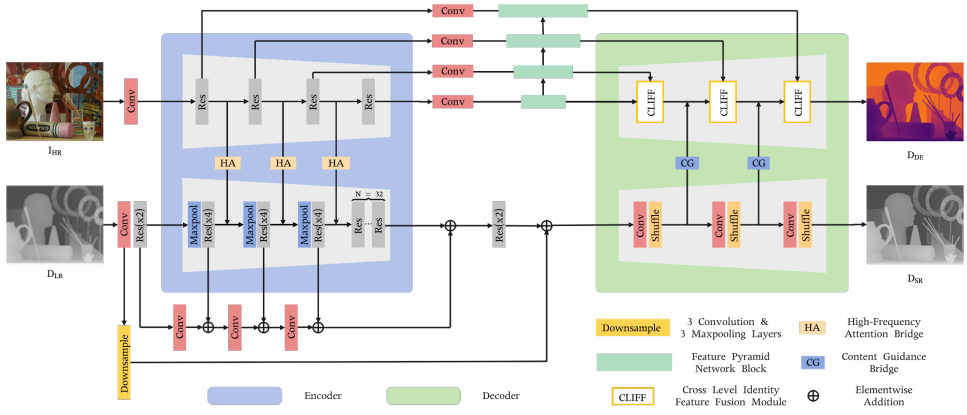


图 4 深度图像超分辨率重建子网络（DSRNet）

深度图像超分辨率重建子网络

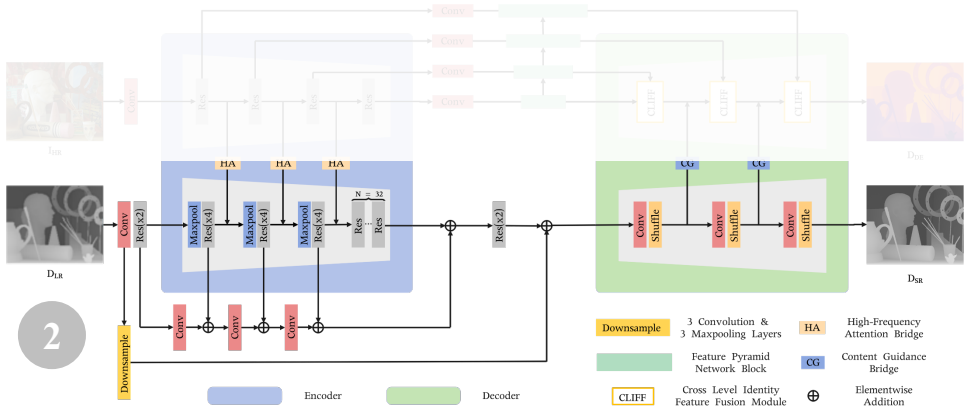


图 4 深度图像超分辨率重建子网络（DSRNet）

深度图像超分辨率重建子网络

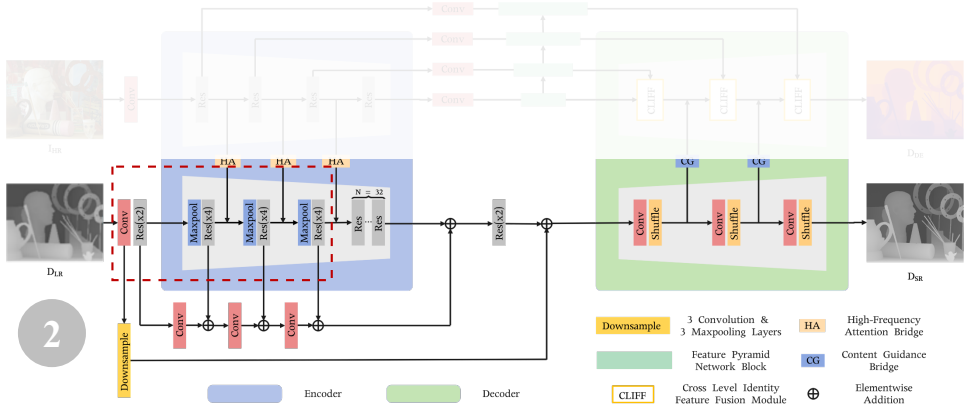


图 4 深度图像超分辨率重建子网络（DSRNet）

深度图像超分辨率重建子网络

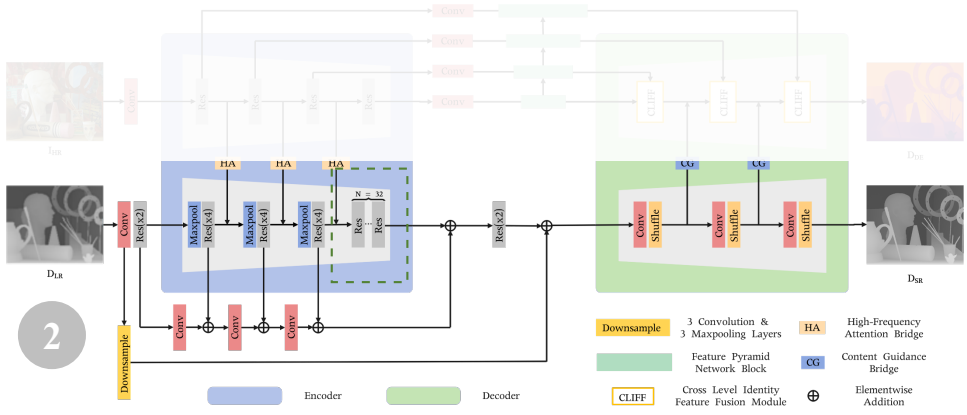


图 4 深度图像超分辨率重建子网络（DSRNet）

深度图像超分辨率重建子网络

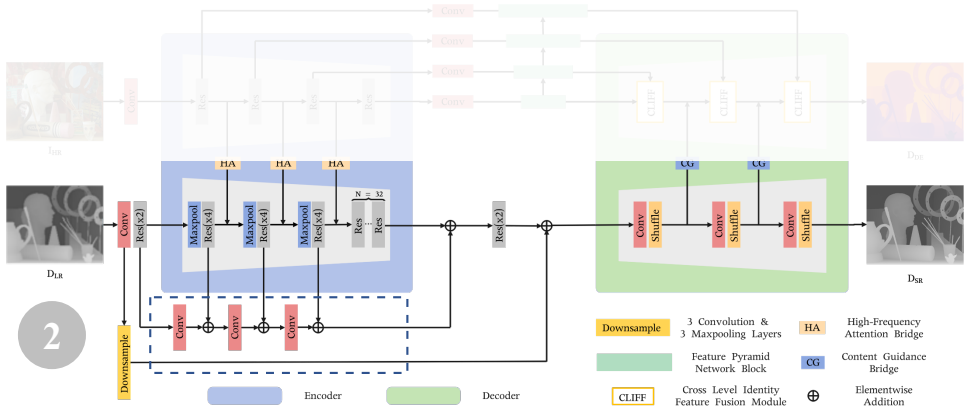


图 4 深度图像超分辨率重建子网络（DSRNet）

深度图像超分辨率重建子网络

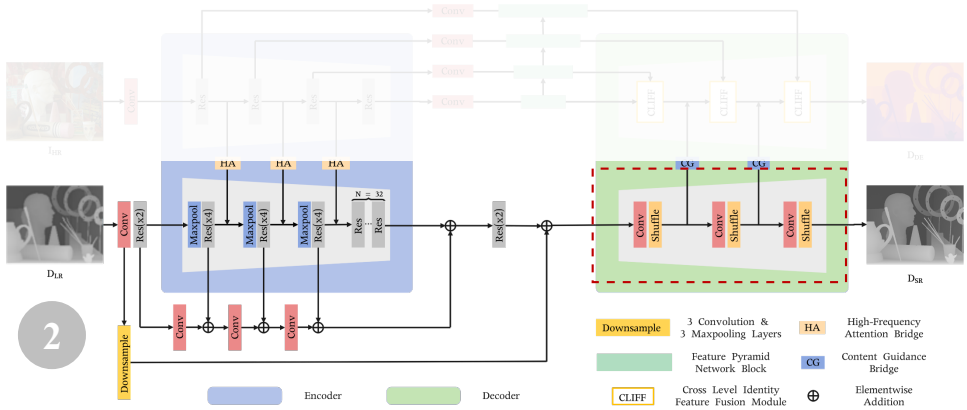


图 4 深度图像超分辨率重建子网络（DSRNet）

深度图像超分辨率重建子网络

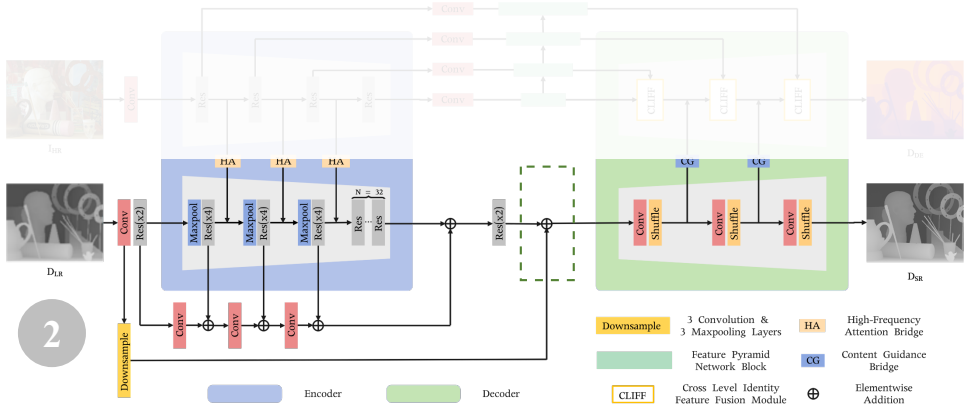


图 4 深度图像超分辨率重建子网络（DSRNet）

深度图像超分辨率重建子网络

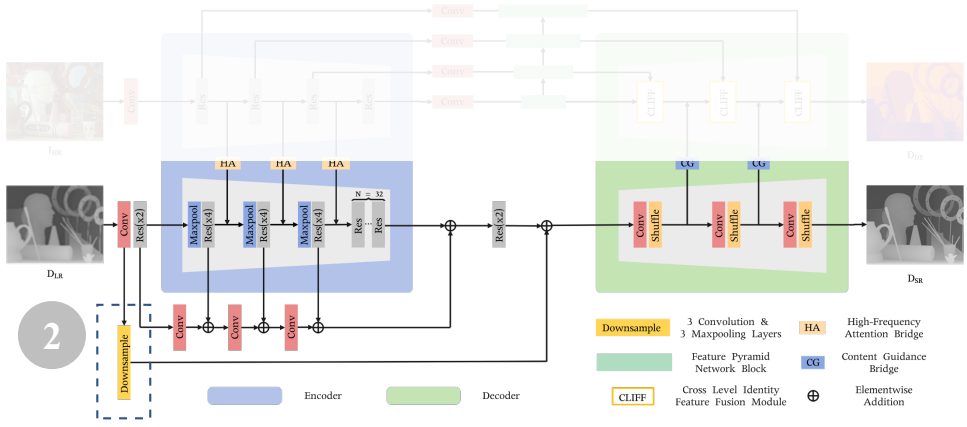


图 4 深度图像超分辨率重建子网络（DSRNet）

高频注意力桥

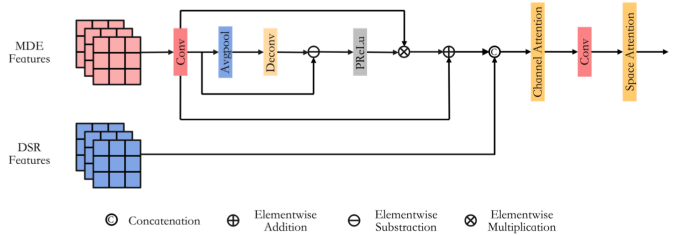


Figure 3: Illustration of HABdg. We first learn the high-frequency attention from the encoder features of the MDENet. Then, it is used to weight the original features to obtain the refined guidance features. After cascading with the features of the corresponding layer of DSRNet, the final output features (*i.e.*, the features of feeding into the next DSRNet encoder layer) are obtained through the CA and SA mechanisms.

图 5 高频注意力桥（HABdg）

高频注意力桥

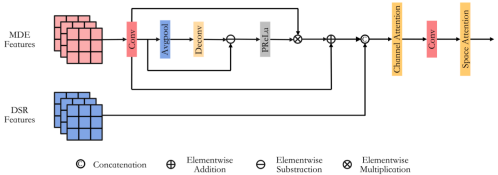


Figure 3: Illustration of HABdg. We first learn the high-frequency attention from the encoder features of the MDENet. Then, it is used to weight the original features to obtain the refined guidance features. After cascading with the features of the corresponding layer of DSRNet, the final output features (*i.e.*, the features of feeding into the next DSRNet encoder layer) are obtained through the CA and SA mechanisms.

$$
\begin{aligned}
F^i_{\text{blurred}} &= \text{deconv}\left(\text{avgpool}\left(F^i_{MDE}\right)\right) \\
A^i_{hf} &= \text{Pelu}\left(F^i_{MDE} - F^i_{\text{blurred}}\right) \\
F^i_{hg} &= F^i_{MDE} + A^i_{hf} \cdot F^i_{MDE} \\
F^i_{\text{comp}} &= \left[F^i_{DSR}, F^i_{hg}\right] \\
F^i_{ha} &= SA\left(\text{conv}_{1\times1}\left(\text{CA}\left(F^i_{\text{comp}}\right)\right)\right)
\end{aligned}
\tag{1}
$$

$F^i_{MDE}, F^i_{\text{blurred}}$ ——单目深度估计子网络第 i 层的特征和获得的模糊特征；

$A^i_{hf}, F^i_{hg}-$ ——获得的第 i 层的高频注意力和优化后的引导特征；

F^i_{DSR}, F^i_{ha} ——深度图像超分辨率重建子网络第 i 层的特征和融合高频信息的特征；

$\text{avgpool}(\cdot)$ ——平均池化操作；$\text{deconv}(\cdot)$ ——反卷积操作；$\text{con}_{1\times1}$ ——卷积核大小为 1×1 的卷积层；

$P\,\text{Relu}(\cdot)$ ——带参数的修正线性单元，即激活函数；

CA, SA ——通道注意力，空间注意力；$[\cdot,\cdot]$ ——通道维度的级联。

内容引导桥

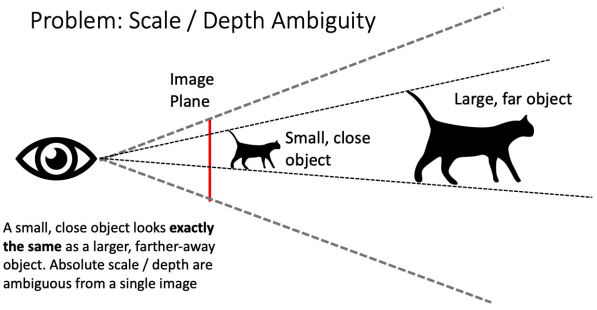


图 6 尺度模糊性

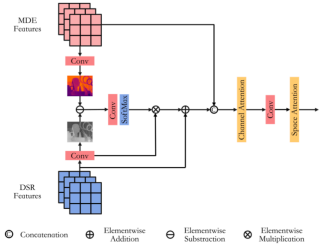


Figure 4: Illustration of CGBdg. We first calculate the difference map between estimated depth map and super-resolved depth map, and then learn the difference weight through a convolution operation and softmax activation. Applying the difference weight to the depth SR encoder features to generate the content guidance for the depth estimation branch. Finally, the depth estimation features and content guidance are concatenated, and fed into the channel attention and spatial attention modules to produce the output features.

图 7 内容引导桥（CGBdg）

内容引导桥

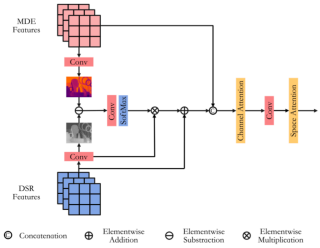


Figure 4: Illustration of CGBdg. We first calculate the difference map between estimated depth map and super-resolved depth map, and then learn the difference weight through a convolution operation and softmax activation. Applying the difference weight to the depth SR encoder features to generate the content guidance for the depth estimation branch. Finally, the depth estimation features and content guidance are concatenated, and fed into the channel attention and spatial attention modules to produce the output features.

$$
\begin{aligned}
M^i_{DSR} &= \mathrm{conv}_{1\times1}\left(Fd^i_{DSR}\right) \\
M^i_{MDE} &= \mathrm{conv}_{1\times1}\left(Fd^i_{MDE}\right) \\
W^i_{diff} &= \mathrm{softmax}\left(\mathrm{conv}_{1\times1}\left(M^i_{DSR} - M^i_{MDE}\right)\right) \\
F^i_{cg} &= Fd^i_{DSR} + W^i_{diff} * Fd^i_{DSR}
\end{aligned}
\tag{2}
$$

M^i_{DSR} ——深度图像超分辨率重建子网络第 i 层生成的深度图像；

M^i_{MDE} ——单目深度估计子网络第 i 层估计的深度图像；

Fd^i_{DSR} ——深度图像超分辨率重建子网络解码器第 i 层的特征；

Fd^i_{MDE} ——单目深度估计子网络解码器第 i 层的特征。

W^i_{diff} ——差异权重；

F^i_{cg} ——第 i 层的内容引导特征。

联合学习策略

Algorithm 1: Joint Learning Strategy

Input: Training data D_{LR}, I_{HR}, D_{HR}

Output: D_{SR}, D_{DE}

Randomly initialize DSRNet and MDENet

for *epoch=1; epoch* $\leq$ *400;* **do**

Step 1

$F_{MDE} = Encoder_{MDE}(I^0_{LR})$

$F^{shallow}_{DSR} = Rec^{(2)}(conv(D_{LR}))$; // shallow feature extraction

for $i=1; i \leq 3$ **do** // i refers to i^{th} layer of encoder

if $i=1$ **then**

$F^i_{DSR} = maxpool(Rec^{(4)}(F^{shallow}_{DSR}))$

else

$F^i_{DSR} = maxpool(Rec^{(4)}(F^{i-1}_{DSR}))$

$F_{channel} = deconv(avgpool(F^3_{MDE}))$

$A^i_{ij} = PRelu(F^i_{DSR} - F_{channel})$

$F^i_{sa} = F^i_{DSR} * A^i_{ij} \cdot F^i_{MDE}$

$F^i_{comp} = [F^i_{DSR}, F^i_{sa}]$

$F^i_{sa} = SA(conv_{1\times1}(CA(F^i_{comp})))$

$F^{inter}_{DSR} = Rec^{(22)}(F^3_{sa})$

$F^{multi-scale}_{DSR} = conv(conv(conv(F^{shallow}_{DSR} + Fe^1_{DSR}) + F^2_{DSR}) + Fe^3_{DSR})$; // multi-scale features fusion

$F^{fusion}_{DSR} = Rec^{(2)}(F^{inter}_{DSR} + F^{multi-scale}_{DSR})$

$F^{low\_res}_{DSR} = Downsample(F^{shallow}_{DSR})$

for $i=1; i \leq 3$ **do** // j refers to j^{th} layer of decoder

if $i=1$ **then**

$Fd^i_{DSR} = pixelshuffle(conv(F^{fusion}_{DSR} + F^{low\_res}_{DSR}))$

else

$Fd^i_{DSR} = pixelshuffle(conv(Fd^{i-1}_{DSR}))$

$D_{SR} - conv_{1\times1}(Fd^3_{SDR})$

Update weights of parts related to DSR with $\mathcal{L}_{DSR} = \|D_{SR} - D_{HR}\|_1$

Step 2

$F_{MDE} = Encoder_{MDE}(I^0_{LR})$

for $k=1k \leq 4$ **do**

if $Fd^1_{MDE} = conv(Fe^1_{MDE})$ **then** k=1

$Fe^1_{MDE} = interpolate(Fe^1_{MDE})$

$Fd^1_{MDE} = conv(Fe^1_{MDE} + Fe^{k-(k+1)}_{MDE})$

for $j=1k \leq 7$ **do**

if $j=1$ **then**

$Fd^j_{MDE} = CLIFF(Fd_{MDE}, Fd^1_{MDE})$

else

$M^j_{SDR} = conv_{1\times1}(Fd^j_{SDR})$

$M^j_{MDE} = conv_{1\times1}(Fd^j_{MDE})$

$W_{Mf} = softmax(conv_{7\times7}(M_{SDR} - M^j_{MDE}))$

$F^j_w = Fd^j_{SDR} * W_{Mf} + Fd^j_{SDR}$

$F^j_{cw} = [Fd^j_{MDE}, F^j_w]$

$F^j_{sa} = SA(conv_{1\times1}(CA(F^j_{cw})))$

$Fd^j_{MDE} = CLIFF(F^j_{sa}, Fd^{j+1}_{MDE})$

$D_{DE} = interpolate(conv_{1\times1}(Fd^7_{MDE}))$

Update weights of parts related to MDE with $\mathcal{L}_{MDE} = \|D_{DE} - D_{HR}\|_1$

本文分别为深度图像超分辨率重建和单目深度估计的损失函数分配了不同的优化器。这是因为深度图像超分辨率重建和单目深度估计的学习难度大不相同，导致两个任务的收敛速度不同，从而很难找到合适的权重设置来确保两个任务都达到最佳性能。因此，在损失函数的设计方面，本文提出分别对深度图像超分辨率重建和单目深度估计相关部分进行优化的策略。其损失函数定义为

$$\begin{aligned} \mathcal{L}_{DSR} &= \|D_{SR} - D_{HR}\|_1 \\ \mathcal{L}_{MDE} &= \|D_{DE} - D_{HR}\|_1 \end{aligned} \tag{3}$$

$\mathcal{L}_{DSR}$ ——深度图像超分辨率重建任务的逐像素 L_1 损失;

$\mathcal{L}_{DSR}$ ——单目深度估计任务的逐像素 L_1 损失。

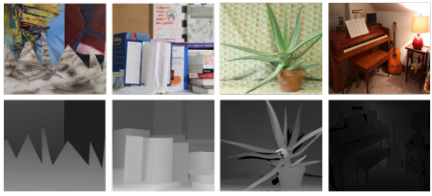
北京交通大学
BEIJING JIAOTONG UNIVERSITY

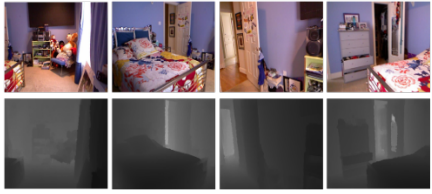
P 第四部分
Part Four

研究成果

- 实现细节
- 实验结果
- 消融实验

实现细节





$$MAD_{\text{Middlebury}} = \frac{1}{N}\sum_{n=1}^{N}\left|D_{HR}^{n} - SR\left(D_{LR}^{n}\right)\right| \quad (4)$$

$$RMSE_{NYUv2} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(D_{HR}^{m} - SR\left(D_{LR}^{m}\right)\right)^{2}} \quad (5)$$

高分辨率的深度图像和彩色图像分别根据 $\times 4$、$\times 8$ 和 $\times 16$ 的上采样因子被裁剪成足够数量的大小为 64^2、128^2 和 256^2 的图像块。为了获得相应的低分辨率深度图像块，使用 Bicubic 插值方法将高分辨率深度图像块下采样为固定大小的 16×16 的图像块。

网络基于 PyTorch 实现，并使用 NVIDIA 2080Ti GPU 加速训练。在训练期间，一次训练所选取的样本数为 8。此外，选取了动量为 0.9，$\beta_1 = 0.9$，$\beta_2 = 0.99$，$\epsilon = 10^{-8}$ 的 ADAM 优化器对训练进行优化。初始学习率设置为 1^{-4}，并且每 100 轮乘以 0.1 以降低学习速率。

实验结果

Table 1: Quantitative depth SR results (in MAD) on Middlebury 2005 dataset. The best performance is displayed in bold, and the second best performance is marked in underline.

	Art			Books			Dolls			Laundry			Mobius			Reindeer		
	×4	×8	×16	×4	×8	×16	×4	×8	×16	×4	×8	×16	×4	×8	×16	×4	×8	×16
CLMF [23]	0.76	1.44	2.87	0.28	0.51	1.02	0.34	0.60	1.01	0.50	0.80	1.67	0.29	0.51	0.97	0.51	0.84	1.55
JGF [22]	0.47	0.78	1.54	0.24	0.43	0.81	0.33	0.59	1.06	0.36	0.64	1.20	0.25	0.46	0.80	0.38	0.64	1.09
TGV [9]	0.65	1.17	2.30	0.27	0.42	0.82	0.33	0.70	2.20	0.55	1.22	3.37	0.29	0.49	0.90	0.49	1.03	3.05
CDLLC [37]	0.53	0.76	1.41	0.19	0.46	0.75	0.31	0.53	0.79	0.30	0.48	0.96	0.27	0.46	0.79	0.43	0.55	0.98
PB [2]	0.79	0.93	1.98	0.16	0.43	0.79	0.53	0.83	0.99	1.13	1.89	2.87	0.17	0.47	0.82	0.56	0.97	1.89
EG [38]	0.48	0.71	1.35	0.15	0.36	0.70	0.27	0.49	0.74	0.28	0.45	0.92	0.23	0.42	0.75	0.36	0.51	0.95
SRCNN [6]	0.63	1.21	2.34	0.25	0.52	0.97	0.29	0.58	1.03	0.40	0.87	1.74	0.25	0.43	0.87	0.35	0.75	1.47
ATGVNet [26]	0.65	0.81	1.42	0.43	0.51	0.79	0.41	0.52	**0.56**	0.37	0.89	0.94	0.38	0.45	0.80	0.41	0.58	1.01
MSG [16]	0.46	0.76	1.53	0.15	0.41	0.76	0.25	0.51	0.87	0.30	0.46	1.12	0.21	0.43	0.76	0.31	0.52	0.99
DGDIE [11]	0.48	1.20	2.44	0.30	0.58	1.02	0.34	0.63	0.93	0.35	0.86	1.56	0.28	0.58	0.98	0.35	0.73	1.29
DEIN [39]	0.40	0.64	**1.34**	0.22	0.37	0.78	0.22	0.38	0.73	0.23	0.36	0.81	0.20	0.35	0.73	0.26	0.40	0.80
CCFN [35]	0.43	0.72	1.50	0.17	0.36	0.69	0.25	0.46	0.75	0.24	0.41	**0.71**	0.23	0.39	0.73	0.29	0.46	0.95
GSRPT [5]	0.48	0.74	1.48	0.21	0.38	0.76	0.28	0.48	0.79	0.33	0.56	1.24	0.24	0.49	0.80	0.31	0.61	1.07
CTKT [32]	**0.25**	**0.53**	1.44	**0.11**	0.26	0.67	**0.16**	0.36	0.65	**0.16**	0.36	0.76	**0.13**	0.27	0.69	**0.17**	0.35	0.77
BridgeNet (Ours)	0.30	0.58	1.49	0.14	**0.24**	**0.51**	0.19	**0.34**	0.64	0.17	**0.34**	**0.71**	0.15	**0.26**	**0.54**	0.19	**0.31**	**0.70**

实验结果

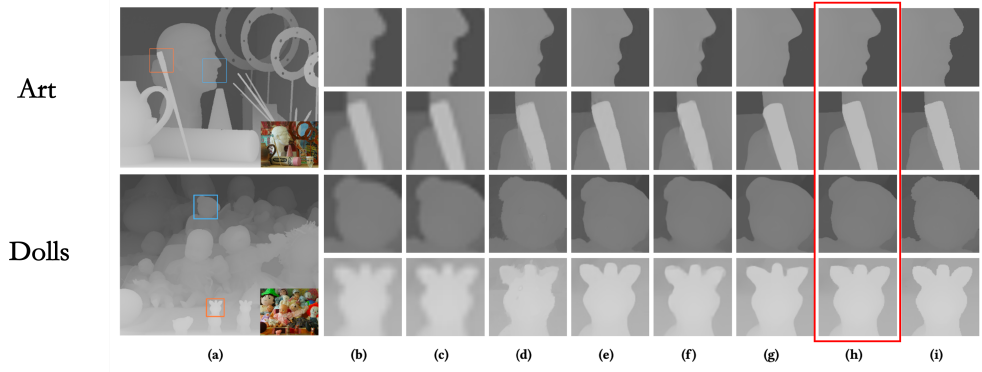


Figure 5: Visual comparisons of ×8 up-sampling results on two examples (*i.e.*, Art in the first row and Dolls in the second row). (a) Ground truth depth maps and color images; (b) LR depth patches; (c)-(h) The super-resolved depth maps generated by Bicubic, TGV [9], MSG [16], DGDIE [11], CTKT [32], and BridgeNet, respectively. (i) Ground truth. Depth patches are enlarged for clear visualization.

消融实验

Table 3: Ablation studies (in MAD) of our BridgeNet on the Middlebury 2005 dataset (×8 case).

	DSRNet	MDENet	HABdg	CGBdg	Middlebury
1	✓				0.366
2		✓			0.472
3	✓	✓			0.363
4	✓	✓	✓		0.355
5	✓	✓		✓	0.361
6	✓	✓	✓	✓	0.343

Table 4: Ablation studies (in MAD) of our HABdg on the Middlebury 2005 dataset (×8 case). 'w/o HABdg' refers to replacing the HABdg by directly propagating features from MDENet to DSRNet.

	Art	Books	Dolls	Laundry	Mobius	Reindeer	Avg.
w/ HABdg	0.58	0.24	0.34	0.34	0.26	0.31	0.343
w/o HABdg	0.65	0.25	0.37	0.38	0.28	0.33	0.376

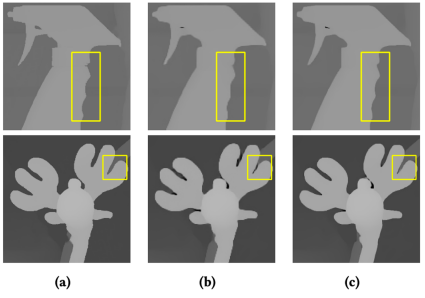


Figure 7: Visual comparisons of different components of our network (×8 case). (a) Ground truth. (b) DSRNet. (c) BridgeNet.

北京交通大学
BEIJING JIAOTONG UNIVERSITY

P 第五部分
Part Five

项目总结

- 总结归纳

论文及专利

- 发明专利：《一种联合单目深度估计的深度图像超分辨率重建方法》
- 2021 ACM MM（CCF A）：BridgeNet: A Joint Learning Network of Depth Map Super-Resolution and Monocular Depth Estimation

BridgeNet: A Joint Learning Network of Depth Map Super-Resolution and Monocular Depth Estimation

北京交通大學
BEIJING JIAOTONG UNIVERSITY

敬请各位老师批评指正

答辩人：唐麒
指导教师：冯凤娟