DIG

Multi-View Depth Estimation by Fusing Single-View Depth Probability with Multi-View Geometry (CVPR 2022 Oral)

2023.11.2

Contribution & Motivation

1. Probabilistic depth sampling：

2. Depth consistency weighting for multi-view matching

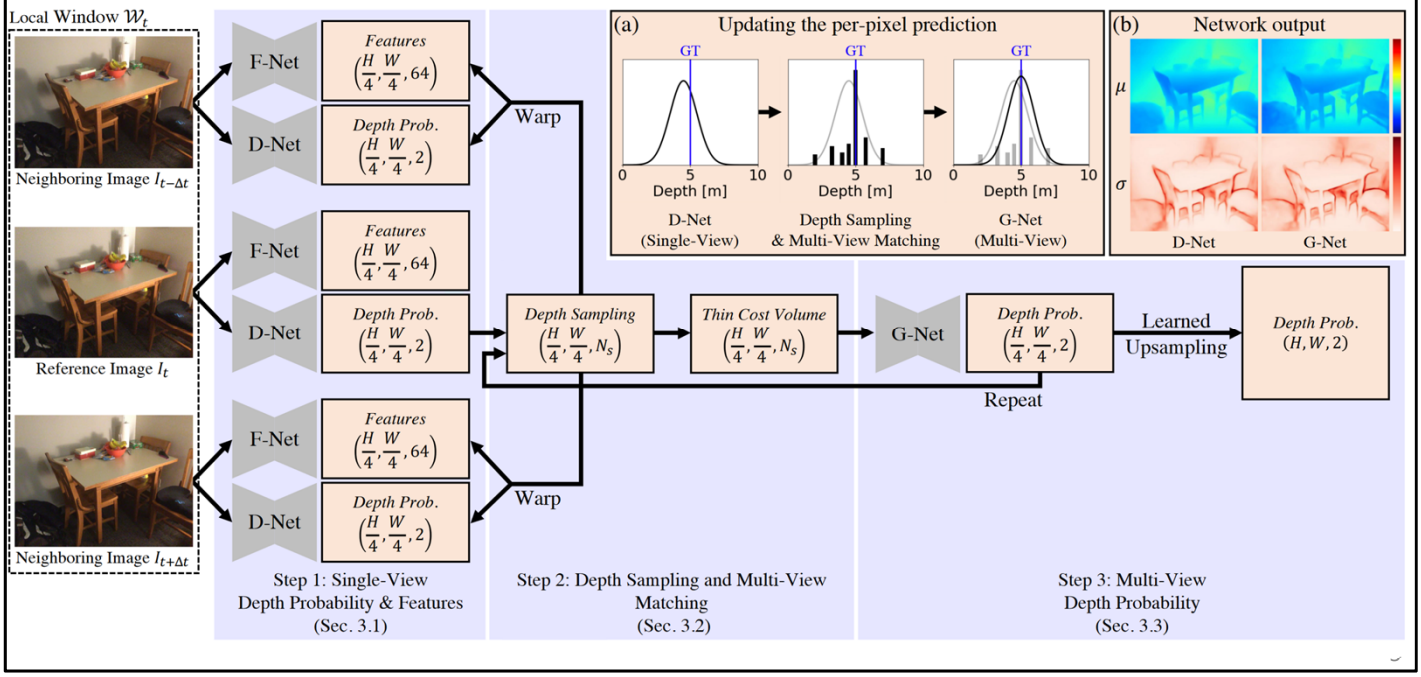
3. Iterative refinement

2

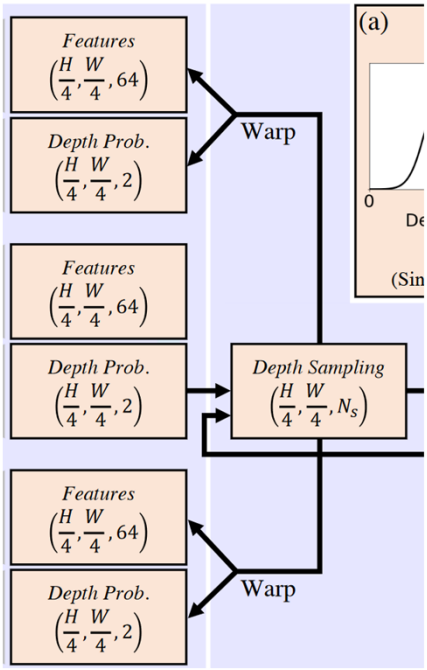
具体而言：沿着cost volume的下降方向循环预测残差索引字段，以检索下一次迭代的成本值。新更新的索引字段用于直接索引（即通过线性插值采样）深度假设来渲染深度图，深度图经过迭代优化以接近地面真实深度，使系统端到端可训练。

感觉心有点虚，很少把自己的contribution写的这么长，而且在第一节就去放一个对比图强调自己创新。

Method: Overview of network



Method: Single-View Depth Probability and Features



我们对 μ 使用线性激活，对 σ^2 使用f (x) = ELU(x) + 1 以确保正方差和平滑梯度。

$$L_{u,v}(d^{\text{gt}}_{u,v}|I_t) = \frac{1}{2}\log\sigma^2_{u,v}(I_t) + \frac{\left(d^{\text{gt}}_{u,v} - \mu_{u,v}(I_t)\right)^2}{2\sigma^2_{u,v}(I_t)}.$$

$$s_{u,v,k}(I_t) = \sum_{i\neq t}\langle \mathbf{f}_{u,v}(I_t), \mathbf{f}_{u_{ik},v_{ik}}(I_i)\rangle$$

任何现有的深度估计网络都可以用作 D-Net。我们使用带有 EfficientNet B5 [43] 主干的轻量级卷积编码器-解码器。

当减少误差 (dgt – μ)2 具有挑战性时，网络会学习估计高 σ2。这通常发生在对象边界附近和远处的点 [22]。相反，当估计的 σ2 较低时，正确的深度可能接近估计的 μ。

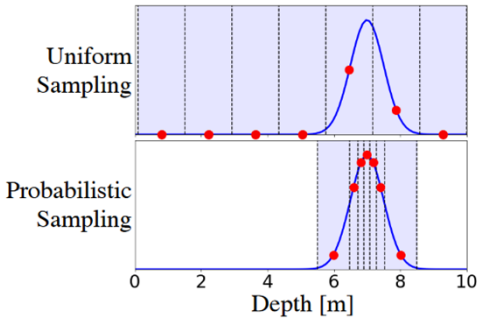
点乘，(uik, vik) 是 (u, v, dk) 定义的 3D 坐标在第 i 个图像上的投影

Method: Fusing Single-View Depth Probability with Multi-View Geometry

$$d_{u,v,k} = \mu_{u,v} + b_k \sigma_{u,v},$$

$$\text{where} \quad b_k = \frac{1}{2}\left[\Phi^{-1}\left(\frac{k-1}{N_s}P^* + \frac{1-P^*}{2}\right) + \Phi^{-1}\left(\frac{k}{N_s}P^* + \frac{1-P^*}{2}\right)\right].$$

N_s bins



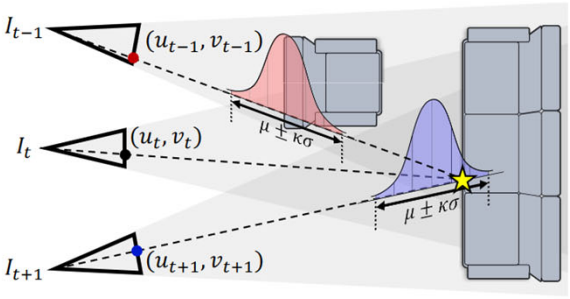
因此我们可以在评估更少的候选者的同时获得更高的准确度。对于具有高不确定性的像素，候选之间的间距增加，以便可以评估更广泛的候选。

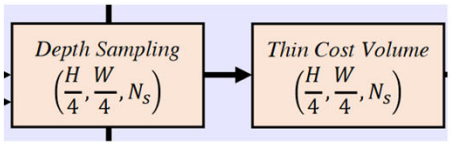
Method: Fusing Single-View Depth Probability with Multi-View Geometry

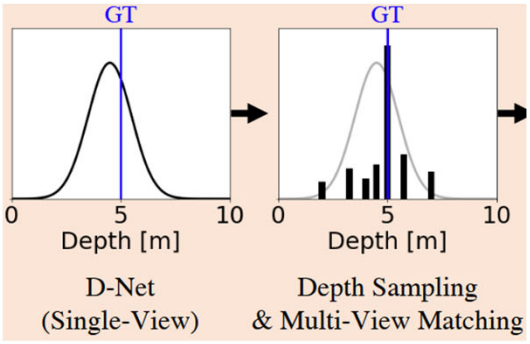
$$s_{u,v,k}(I_t) = \sum_{i \neq t} w^{\text{dc}}_{u_{ik},v_{ik},d_{ik}} \langle \mathbf{f}_{u,v}(I_t), \mathbf{f}_{u_{ik},v_{ik}}(I_i) \rangle$$

$$w^{\text{dc}}_{u_{ik},v_{ik},d_{ik}} = \delta\left(p_{u_{ik},v_{ik}}(d_{ik}|I_i) > p_{\text{thres}}\right).$$

$$p_{\text{thres}} = \exp(-\kappa^2/2)/\sigma_{u_{ik},v_{ik}}\sqrt{2\pi}$$







(uik, vik) 是由 (u, v, dk) 定义的 3D 坐标在第 i 个图像上的投影。

dik 在 κ-sigma 置信区间内，则权重变为 1

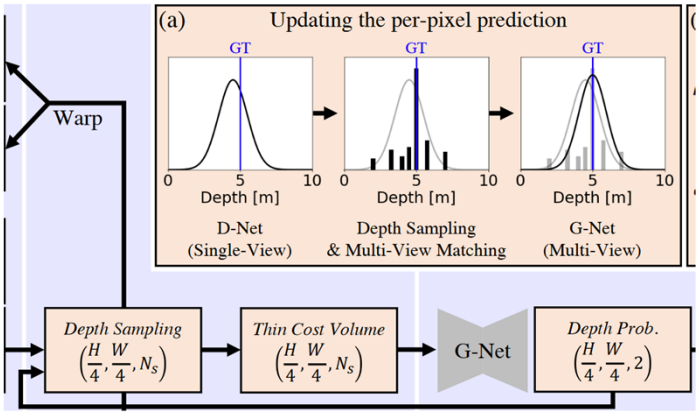
我们设置 pthres = exp(−κ2/2)/σuik,vik √2π，以便如果 dik 在 κ-sigma 置信区间内，则权重变为 1。这意味着 pthres 对于每个像素和每个视图都是自适应的。如果 D-Net 不确定深度（即高 σ），则 pthres 会变低，从而允许考虑更多深度候选。

深度一致性加权丢弃单视图深度概率低的候选者。当多视图匹配不明确或不可靠时，这种加权尤其有用。例如，如果像素位于无纹理表面内，则大范围的深度候选将导致相似的匹配分数。如果场景包含反射表面，则会在反射之间计算匹配分数，从而导致深度估计过高。在这两种情况下，MaGNet 都可以通过支持具有高单视图深度概率的深度候选来做出稳健的预测。

对于参考帧 It 中的像素 (ut, vt)，深度候选定义 3D 点（用 ⋆ 标记）。将该点投影到相邻视图，并评估每个视图中的深度概率。对于 It−1，由于遮挡，⋆ 不在 μ ± κσ 范围内。在这种情况下一致性权重变为0。

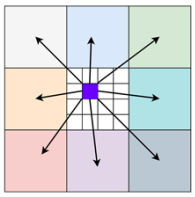
Method: Estimating Multi-View Depth Probability Distribution

$$\mu_{u,v}^{new} = \mu_{u,v} + b_{k'}\sigma_{u,v} \qquad \sigma_{u,v}^{new}/\sigma_{u,v} \qquad \mathcal{N}(\mu_{u,v}^{new}, \sigma_{u,v}^{new})$$



feature-map of D-Net

$H/4 \times W/4 \times (4 \times 4 \times 9)$ mask



7

迭代细化和网络训练。多视图匹配过程（即概率深度采样 → 一致性加权匹配 → G-Net 更新）重复 Niter 次，产生 Niter 预测。对于每个预测，都会计算 NLL 损失（方程 2），并将它们的总和用于训练 G-Net 和上采样层。按照[45]，第 i 个预测由 γNiter−i 加权，其中 0 < γ < 1，以更加重视最终输出。

迭代细化有两个方面的好处。首先，如果其中一个候选者获得了较高的匹配分数，则均值将向该候选者移动，方差将减小，以便在下一次迭代中，网络可以在该候选者附近执行更精细的深度搜索，以找到更好的候选者。具有较高匹配分数的候选人。迭代更新还可以防止D-Net预测不准确的故障模式。例如，如果真实深度不在初始搜索空间 [μu,v − βσu,v, μu,v + βσu,v] 内，则采样的候选者都不会获得高匹配分数。在这种情况下，G-Net将学习增加方差来衰减损失（方程 2），并且网络可以在下一次迭代中执行更广泛的深度搜索。

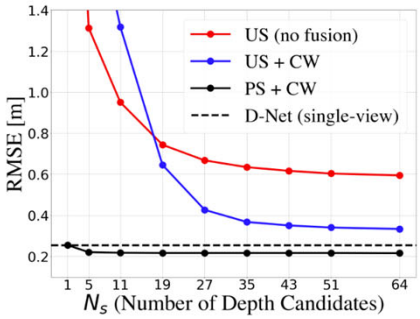
Experiment

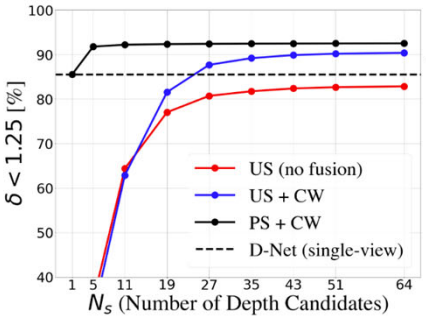
Method	Cap	Train on ScanNet → Test on ScanNet					Train on ScanNet → Test on 7-Scenes				
		abs rel	abs diff	rmse	$\text{rmse}_{\log}$	δ<1.25	abs rel	abs diff	rmse	$\text{rmse}_{\log}$	δ<1.25
MVDepthNet [46]	10m	0.1116	0.2087	0.3143	0.1500	88.04	0.1905	0.3304	0.4260	0.2221	71.93
DPSNet [20]		0.0986	0.1998	0.2840	0.1348	88.80	0.1675	0.2970	0.3905	0.2061	76.03
NAS [24]		0.0941	0.1928	0.2703	0.1269	90.09	0.1631	0.2885	0.3791	0.1997	77.12
CNM-Net [30]		0.1102	0.2129	0.3032	0.1482	86.88	0.1602	0.2751	0.3602	0.2030	76.81
DELTAS [41]		0.0915	0.1710	0.2390	0.1226	91.47	0.1548	0.2671	0.3541	0.1860	79.66
UCS-Net [6]		0.0845	0.1605	0.2335	0.1145	92.22	0.2113	0.3668	0.4683	0.2369	69.31
Long et al. [29]		0.0812	0.1505	0.2199	0.1104	**93.13**	0.1465	0.2528	0.3382	0.1967	80.36
Ours (D-Net)		0.1186	0.2070	0.2708	0.1461	85.46	0.1339	0.2209	**0.2932**	0.1677	83.08
Ours (full)		**0.0810**	**0.1466**	**0.2098**	**0.1101**	92.98	**0.1257**	**0.2133**	0.2957	**0.1639**	**85.52**
NeuralRGBD [27]	5m	0.1013	0.1657	0.2500	0.1315	91.60	0.2334	0.4060	0.5358	0.2516	68.03
Long et al. [29]		0.0805	0.1438	0.2029	**0.1083**	**93.33**	0.1465	0.2528	0.3382	0.1967	80.36
Ours (D-Net)		0.1177	0.1991	0.2526	0.1439	85.70	0.1339	0.2209	**0.2932**	0.1677	83.08
Ours (full)		**0.0804**	**0.1409**	**0.1960**	0.1084	93.13	**0.1257**	**0.2133**	0.2957	**0.1639**	**85.52**

Method	Multi	abs rel	sq rel	rmse	$\text{rmse}_{\log}$	δ<1.25
MonoDepth2 [16]	×	0.106	0.806	4.630	0.193	87.6
FeatDepth [39]	×	0.099	0.697	4.427	0.184	88.9
BTS [26]	×	0.059	0.245	2.756	0.096	95.6
AdaBins [1]	×	0.058	0.190	2.360	0.088	96.4
SC-GAN [47]	✓	0.063	0.178	**2.129**	0.097	96.1
Ours (D-Net)	×	0.061	0.209	2.422	0.092	96.0
Ours (full)	✓	**0.054**	**0.162**	2.158	**0.083**	**97.1**
NeuralRGBD [27]	✓	0.100	0.473	2.829	0.128	93.2
Ours (D-Net)	×	0.063	0.254	2.471	0.102	95.8
Ours (full)	✓	**0.050**	**0.167**	**1.971**	**0.085**	**97.7**

时间一致性：估计深度图的平均绝对误差的标准差进行评估，然后给了张定性的图片。

Ablation





N_{iter}	N_s	abs rel	sq rel	rmse	$\text{rmse}_{\log}$	δ<1.25
1	5	0.097	0.035	0.217	0.121	90.75
	7	0.096	0.035	0.217	0.121	90.81
	9	0.096	0.035	0.217	0.121	90.81
	11	**0.095**	**0.034**	**0.216**	**0.120**	**90.94**
1	5	0.097	0.035	0.217	0.121	90.75
2		0.090	0.032	0.209	0.115	92.15
3		**0.087**	0.031	0.207	**0.113**	92.61
4		**0.087**	**0.030**	**0.206**	**0.113**	**92.73**

Method	N_s	abs rel	sq rel	rmse	$\text{rmse}_{\log}$	δ<1.25
UCS-Net [6]	(64, 32, 8)	0.091	0.156	0.229	0.120	91.49
UCS-Net + PS	(8, 8, 8)	0.092	0.157	0.214	0.118	90.74
Ours	(5×3, 0, 0)	**0.082**	**0.143**	**0.202**	**0.110**	**92.78**

DIG

Thanks