DIG

RIAV-MVS: Recurrent-Indexing an Asymmetric Volume for Multi-View Stereo (CVPR 2023)

2023.10.12

Review—Cost Volume in MVS

世界空间中的 3D 点将被投影到可见图像 It−1、It、It+1 中，并且它们投影附近的图像纹理应该具有高度相似性。我们将深度估计表述为占用估计问题：如果图像的一个像素（u，v）具有深度值d，则Ct中的体素（u，v，d）被占用，即Ct的学习特征编码每个体素的占用概率。混合成本体积被视为同一 3D 世界空间在不同视点的多个占用测量，即对于世界空间中的 3D 点，其对应的体积 Ct−1、Ct、Ct+1 的体素应保持相似的嵌入向量。

使用此映射，我们将 Ct−1 和 Ct+1 warp到 Ct 的相机坐标空间中，并获得两个扭曲混合体积 Cwarp t−1 、 Cwarp t+1 。两个warp后的volume和 Ct 应在重叠区域的体素中包含相似的特征。

Contribution & Motivation

1. 我们学习通过用于识别深度假设的索引网格接近每个像素的正确深度平面来对cost volume进行索引，所提出的索引字段的循环估计使学习能够锚定在cost volume domain。。

2. 为了促进优化，我们建议分别提升像素和帧级别的cost volume。在像素级别，transformer不对称地应用于参考视图（但不应用于源视图）。通过transformer捕捉远程全局上下文，并通过 CNN 捕获像素级局部特征，我们构建了一个不对称cost volume来存储更准确的匹配相似性线索。在帧级别，我们提出了一个残差姿势网络来纠正位姿。

1. RAFT
2. IterMVS

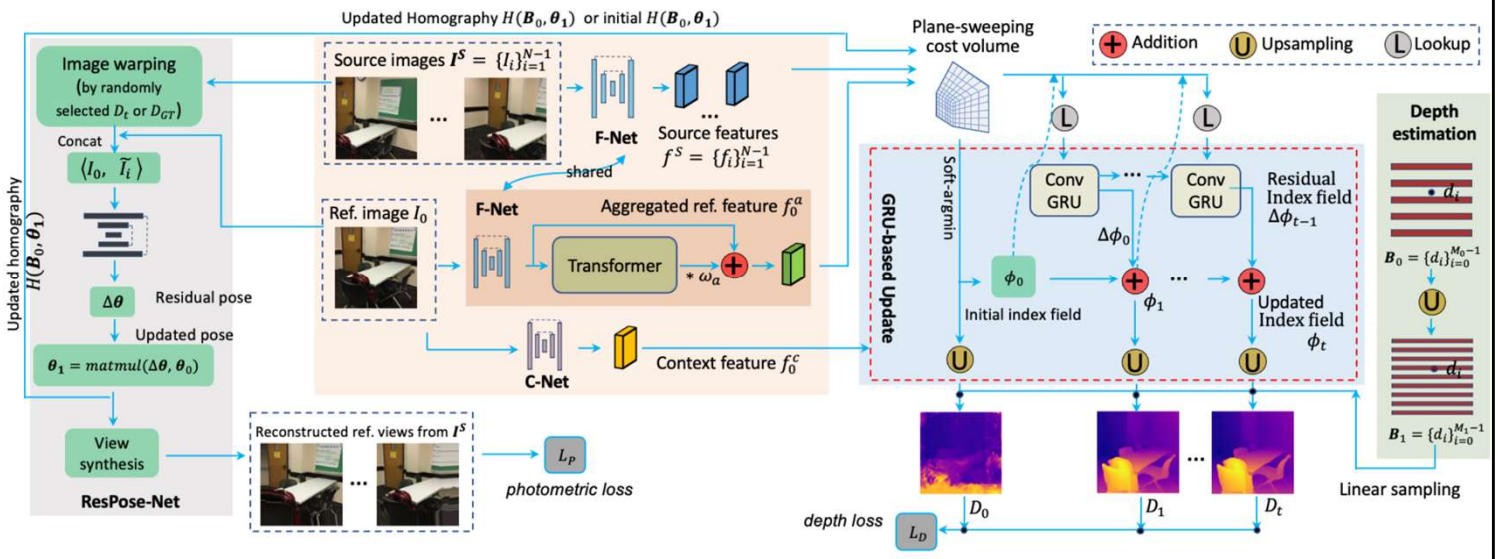
1. DualRefine（同时期）

3

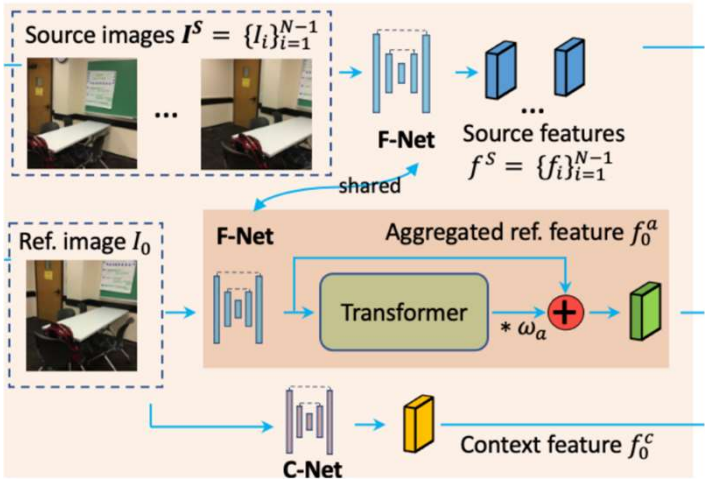
具体而言：沿着cost volume的下降方向循环预测残差索引字段，以检索下一次迭代的成本值。新更新的索引字段用于直接索引（即通过线性插值采样）深度假设来渲染深度图，深度图经过迭代优化以接近地面真实深度，使系统端到端可训练。

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: *Asymmetric* Feature Extraction

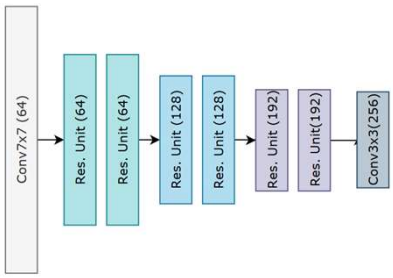


$$\{f_{0,s} \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times F_0}\} \ (s{=}2,4,8,16 \text{ and } F_0{=}32$$

$$f_0 = \mathcal{G}(\langle f_{0,2}\downarrow_2, f_{0,4}, f_{0,8}\uparrow_2, f_{0,16}\uparrow_4\rangle)$$

$\text{Conv}_{3\times3}$, batch normalization, ReLU, and $\text{Conv}_{1\times1}$

$$f_0^a = f_0 + \omega_\alpha \sigma\left(\frac{(f_0 W^Q)(f_0 W^K)^T}{\sqrt{F_1}}(f_0 W^V)\right)$$



5

值得注意的是，这种变换器自注意力仅应用于参考图像，而源特征仍然拥有 CNN 的局部表示。
我们对该变压器层的不对称使用提供了通过自注意力更好地平衡高频特征（高通 CNN）和低频特征的能力 [42,46]。
高频特征有利于局部和结构区域的图像匹配，而低频特征通过变压器的空间平滑（用作低通滤波器）抑制噪声信息，为鲁棒匹配提供更多全局上下文线索，特别是对于充满低纹理、重复模式和遮挡等的区域。
这样，我们的网络可以学习在哪里依赖全局特征而不是局部特征，反之亦然。

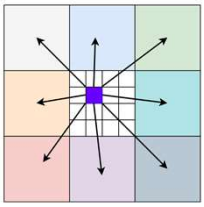
Method: GRU-based Iterative Optimization & Upsampling

$$C_0 \in \mathbb{R}^{H/4 \times W/4 \times M_0} \quad \textit{index fields} \ \{\phi_t \in \mathbb{R}^{H/4 \times W/4}\}_{t=1}^{T}$$

$$\phi_0 = \sum_{i=0}^{M_1-1} i\sigma(C_0) \quad \{C_0^i \in \mathbb{R}^{H/4 \times W/4 \times M_0/2^i}\}_{i=1}^{4}$$

$r = \pm 4$ around the ϕ_t. $\quad C_0^{\phi_t} \in \mathbb{R}^{H/4 \times W/4 \times F_2}$

$$\delta\phi_t, h_{t+1} \Leftarrow \mathrm{GRU}(\langle \phi_t, C_0^{\phi_t}, f_0^c \rangle, h_t); \phi_{t+1} \Leftarrow \phi_t + \delta\phi_t$$



weight mask $W_0 \in \mathbb{R}^{H/4 \times W/4 \times (4 \times 4 \times 9)}$

其中 σ(·) 是沿着成本量 C0 最后一个维度的 softmax 算子，得到的\phi 0就是深度估计，
其实这里的F2应该就是5~9间

Method: Depth Estimation

但是直接上采样的话，会有在不影响量化结果的不连续性。所以把深度假设区间提高一些。

M_1=256 $\quad \mathcal{B}_1 = \{d_i\}_{i=0}^{M_1-1} \quad s_D = \frac{M_1}{M_0} = 4$

mask $W_1 \in \mathbb{R}^{H \times W \times s_D \times M_0}$ reshaped to $H \times W \times M_1$

new depth planes

mask

$\phi_t(p)$ conve

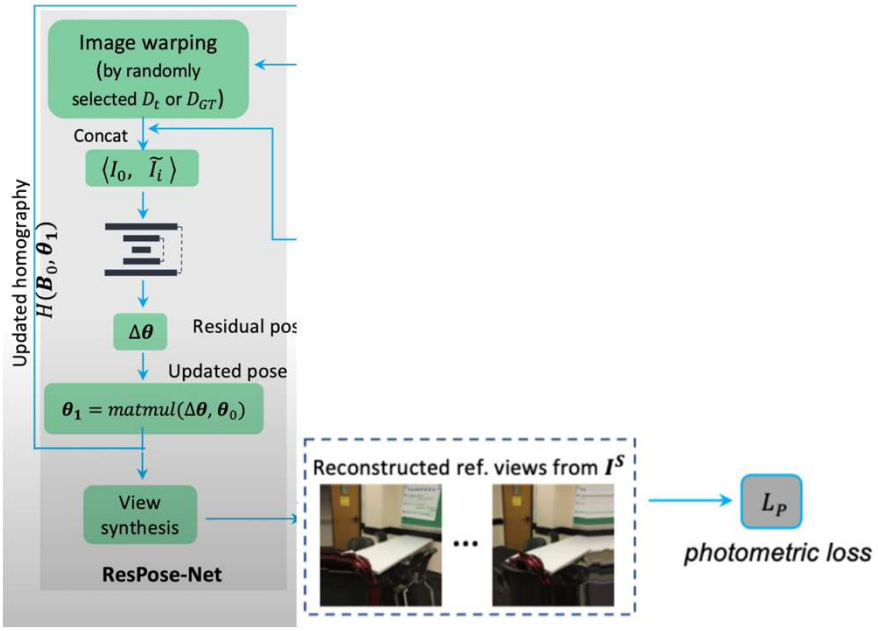
Final depth

$$D_t(p) = \frac{\sum_{i \in \Omega(p)} B[i] \cdot W_1(p, [i])}{\sum_{j \in \Omega(p)} W_1(p, [j])}$$

Neighbor with a radius r=4 centered at upsampled index $\phi_t^u(p)$ for a given pixel p

相当于上采样后，再结合256的深度平面再平均了下，使其更加

Method: Residual Pose Net



Experiment

Method	ScanNet Test-Set (m)							DTU Test-Set (mm)		
	abs-rel	abs	sq-rel	rmse	rmse-log	$\delta < 1.25$	$\delta < 1.25^2$	abs-rel	abs	rmse
MVDepth [54]	0.1167	0.2301	0.0596	0.3236	0.1610	0.8453	0.9639	-	-	-
MVDepth-FT	0.1116	0.2087	0.0763	0.3143	0.1500	0.8804	0.9734	-	-	-
DPSNet [26]	0.1200	0.2104	0.0688	0.3139	0.1604	0.8640	0.9612	-	-	-
DPSNet-FT	0.0986	0.1998	0.0459	0.2840	0.1348	0.8880	0.9785	-	-	-
DELTAS [45]	0.0915	0.1710	0.0327	0.2390	0.1226	0.9147	**0.9872**	-	-	-
NRGBD [37]	0.1013	0.1657	0.0502	0.2500	0.1315	0.9160	0.9790	-	-	-
ESTD [39]	0.0812	0.1505	0.0298	0.2199	0.1104	0.9313	0.9871	-	-	-
MVSNet [59]	0.1032	0.18645	0.0465	0.2743	0.1385	0.8935	0.9775	0.0143	10.7235	25.3989
PairNet [17]	0.0895	0.1709	0.0615	0.2734	0.1208	0.9172	0.9804	0.0129	9.4428	21.4650
IterMVS [52]	0.0991	0.1818	0.0518	0.2733	0.1368	0.8995	0.9741	0.0146	10.6225	28.7009
Ours(base)	0.0885	0.1605	0.0380	0.2347	0.1183	0.9211	0.9810	0.0116	8.2887	21.5806
Ours(+pose,atten)	**0.0734**	**0.1381**	**0.0281**	**0.2080**	**0.1030**	**0.9395**	0.9862	**0.0092**	**6.7771**	**18.5953**

ScanNet ⇒ Others	7-Scenes					RGB-D Scenes V2				
	abs-rel	abs	sq-rel	rmse	$\delta < 1.25$	abs-rel	abs	sq-rel	rmse	$\delta < 1.25$
NRGBD [37]	0.2334	0.4060	0.2163	0.5358	0.6803	-	-	-	-	-
ESTD [39]	0.1465	0.2528	0.0729	0.3382	0.8036	-	-	-	-	-
PairNet [17]	0.1157	0.2086	0.0677	0.2926	0.8768	0.0995	0.1382	0.0279	0.1971	0.9393
IterMVS [52]	0.1336	0.2363	0.1033	0.3425	0.8518	0.0811	0.1245	0.0340	0.2133	0.9496
Ours(base)	0.1148	0.1999	0.0552	0.2857	0.8726	0.0967	0.1336	0.0246	0.1836	0.9427
Ours(+pose,atten)	**0.1000**	**0.1781**	**0.0473**	**0.2664**	**0.8967**	**0.0803**	**0.1168**	**0.0200**	**0.1703**	**0.9632**

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

Variants	ScanNet Test-Set (m)			
	abs-rel	abs	rmse	$\delta < 1.25$
Ours(base)	0.0885	0.1605	0.2347	0.9211
Ours(+pose)	0.0827	0.1523	0.2253	0.9277
Ours(+pose,atten)	**0.0734**	**0.1381**	**0.2080**	**0.9395**

Attention	ScanNet/DTU Test-Set (m)			
	abs-rel	abs	rmse	$\delta < 1.25$
Asym atten (ours)	**0.0734**	**0.1381**	**0.2080**	**0.9395**
Sym. atten	0.0761	0.1496	0.2253	0.9333
MVSNet [59]	0.1032 (0.0143)	0.1865 (10.7235)	0.2743 (25.3989)	0.8935 (0.8936)
MVSNet(+atten)	0.1018 (0.0123)	0.1853 (9.1150)	0.2734 (22.3525)	0.8957 (0.9909)

Itr. T	abs-rel	abs	$\delta < 1.25$
16	0.1413	0.0760	0.9364
24	0.1400	0.0752	0.9375
48	0.1392	0.0747	0.9382
64	0.1392	0.0746	0.9384
96	**0.1392**	**0.0745**	**0.9385**
128	0.1394	0.0745	0.9385

View No.	abs-rel	abs	$\delta < 1.25$
3 (base)	0.1204	0.2121	0.8603
5 (base)	0.1148	0.1999	0.8726
3 (+pose)	0.1162	0.2061	0.8711
5 (+pose)	0.1096	0.1930	0.8840
3 (+pose,atten)	0.1084	0.1923	0.8833
5 (+pose,atten)	**0.1000**	**0.1781**	**0.8967**

Sampling	abs-rel	abs	$\delta < 1.25$
s10 (base)	0.0885	0.1605	0.9211
key (base)	0.0838	0.1598	0.9277
s10 (+pose)	0.0827	0.1523	0.9277
key (+pose)	0.0789	0.1531	0.9339
s10 (+pose,atten)	0.0747	0.1392	0.9382
key (+pose,atten)	**0.0697**	**0.1348**	**0.9472**

View No. on DTU	abs-rel (↓)	abs (mm) (↓)	rmse (↓)
3 (1 ref + 2 source)	0.0149	11.0689	24.8831
5 (1 ref + 4 source)	0.0119	8.8419	21.4327

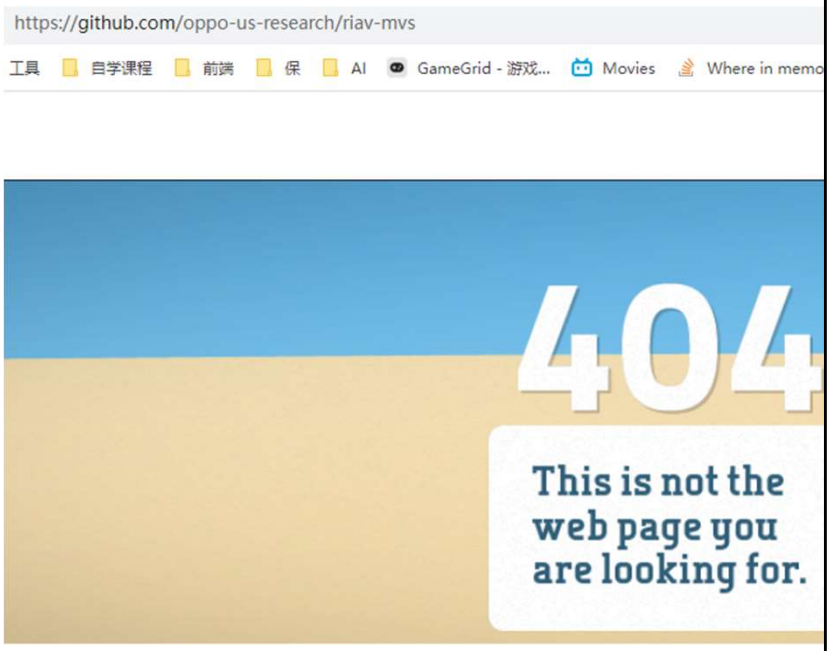
Table 5. 3-view vs. 5-view training and testing on DTU [27].

Methods	Time(fps)	Mem.(MB)	Param.(M)	abs-rel (↓)
Ours(T=8)	6.98	4297	27.6	0.0760
Ours(T=12)	5.91	4297	27.6	0.0752
Ours(T=24)	3.77	4297	27.6	0.0734
IterMVS [52]	22.61	2171	0.34	0.0991
ESTD [39]	14.08	1799	36.2	0.0812

?

Thank You!

Code coming soon
https://github.com/oppo-us-research/riav-mvs



DIG

Thanks