DIG

Multi-view Depth Estimation using Epipolar Spatio-Temporal Networks (CVPR 2021)

2023.10.12

Contribution & Motivation

①Epipolar Spatio-Temporal Transformer：传播时间相干性以执行多个帧的联合深度估计，使估计的深度图在时间上更加相干。（直接把MVSNet那一套搬到视频会有伪影。）

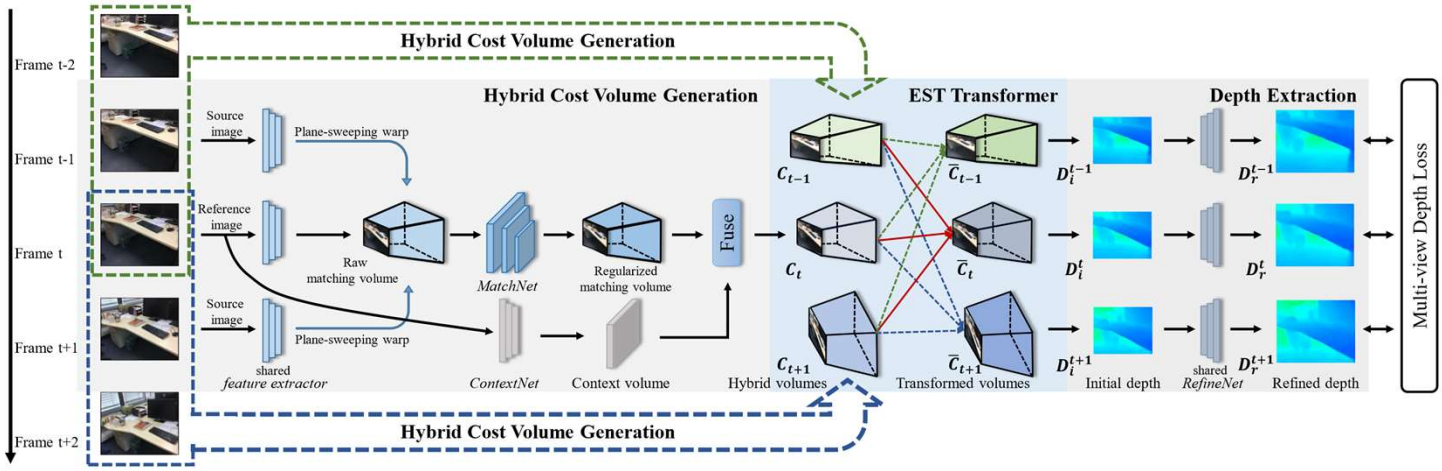
②实现成本正则化的混合网络，它由两个专家网络组成，分别学习 3D 局部匹配信息和 2D 全局上下文信息。

2

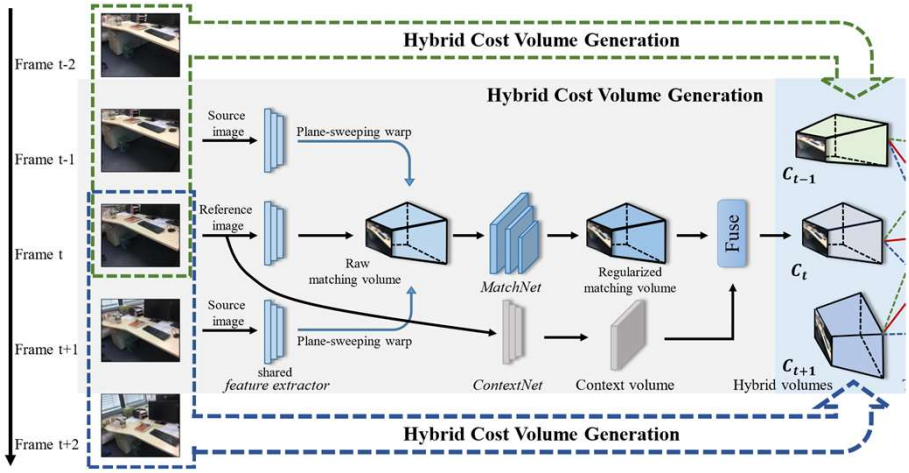
第一个解决了：粗分辨率下的误差难以恢复，容易遗漏快速移动的小物体，多级级联训练通常需要多次训练迭代（通常超过 100 万次）

相比之下，我们的更新算子只有 2.7M 个参数，在 inference 过程中可以执行 100 多次而不会出现divergence。

Approach: Overview of network



Approach: Hybrid (3D matching + 2D context) cost volume generation



1. feature extractor: Spatial Pyramid Pooling
2. raw matching volume: source--- warp --->refernce, 2C(concatenate)*D(depth)*H*W*N(2)
3. *MatchNet:* 3 * 3d conv -> view average pool -> N * 3d conv -> C*D*H*W
4. *ContextNet:* 3 images--- Resnet-50 -----> C'(==D)*H*W --(+reg)-> (C+1)*D*H*W
5. *Hybird Cost Volume:* 每一个都包含matching 和 context

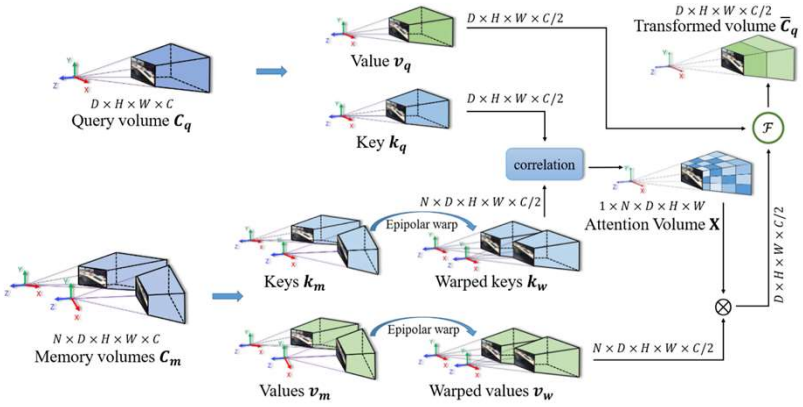
4

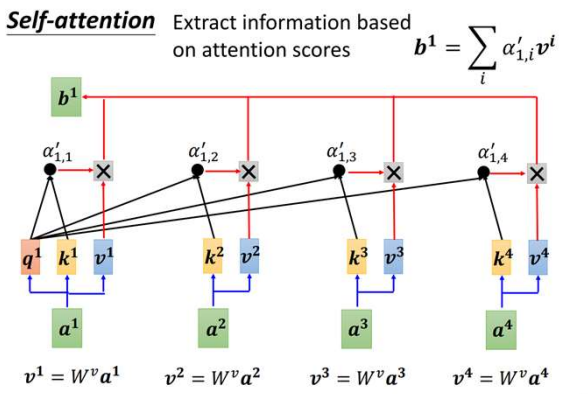
我们观察到全局上下文信息本质上是2D信息？
然后这里的每个

Approach: Epipolar Spatio-Temporal transformer

1. Consistency Constraint: 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。Cwarp t−1 、Cwarp t+1 和 Ct 应在重叠区域的体素中包含相似的特征。
2. EST Transformer: 每次把其中一个当成Query， Ct -> Query, Ct-1 & Ct+1 -> Volume (Query, Volume) -- two identical conv ---> (K, V)

$$f(v_q, y) = w \odot y + (1-w) \odot g(v_q, r \odot y), y = \sum_{i=1}^{N} x_i v_w^i$$





5

xi ∈ R1×1×D×H×W 衡量查询与第 i 个内存卷的扭曲密钥的相似度，N 是内存卷的数量，

内积。

w, r ∈ D×H×W are two learned weight volumes which measure the reliability of the retrieved values

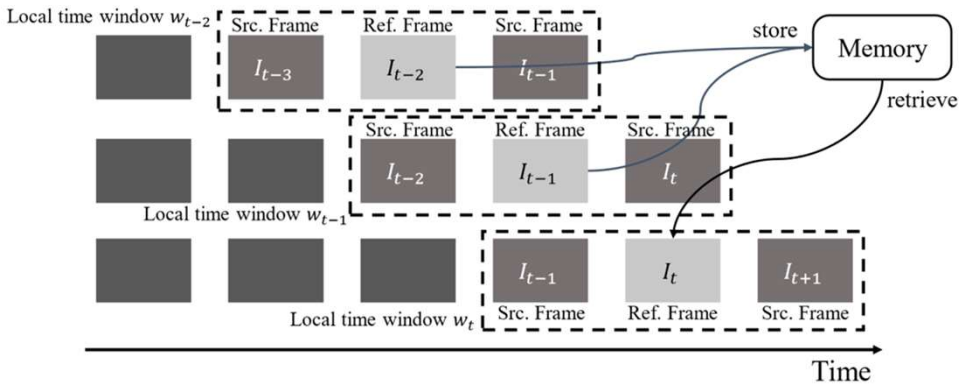
Q*K*V

self-attition的变体少做了一个correlation，把Vq通过后Fusion的形式融入

Approach: RefineNet Depth Regression and Inference

我们将混合成本量、转换成本量和两阶段 RefineNet 的四种类型的深度图表示为 Ds，s = 0, 1, 2, 3。i是指target image。

$$loss = \frac{1}{N}\sum_{s=0}^{3}\sum_{i=1}^{N}\lambda^{s-3}\left\|\mathbf{D}_s^i - \hat{\mathbf{D}}_s^i\right\|_1$$



6

我们从存储过去N帧的键和值对的存储空间中检索相关值

混合成本量

权重为什么是1.95 1.56 1.25 1?

ESTM就是用前两帧，

Experiment

Range	Method	ScanNet					7scenes				
		Abs Rel	Abs	Sq Rel	RMSE	$\sigma < 1.25$	Abs Rel	Abs	Sq Rel	RMSE	$\sigma < 1.25$
10m	MVDepth [32]	0.1167	0.2301	0.0596	0.3236	84.53	0.2213	0.4055	0.2401	0.5154	67.33
	MVDepth-FT	0.1116	0.2087	0.0763	0.3143	88.04	0.1905	0.3304	0.1319	0.4260	71.93
	DPS [17]	0.1200	0.2104	0.0688	0.3139	86.40	0.1963	0.3471	0.1970	0.4625	72.51
	DPS-FT	0.0986	0.1998	0.0459	0.2840	88.80	0.1675	0.2970	0.1071	0.3905	76.03
	NAS [20]	0.0941	0.1928	0.0417	0.2703	90.09	0.1631	0.2885	0.1023	0.3791	77.12
	CNM [23]	0.1102	0.2129	0.0513	0.3032	86.88	0.1602	0.2751	0.0819	0.3602	76.81
	DELTAS [30]	0.0915	0.1710	0.0327	0.2390	91.47	0.1548	0.2671	0.0889	0.3541	79.66
	Ours-EST(concat)	0.0818	0.1536	0.0301	0.2234	92.99	**0.1458**	0.2554	0.0745	0.3436	79.82
	Ours-EST(adaptive)	**0.0812**	**0.1505**	**0.0298**	**0.2199**	**93.13**	0.1465	**0.2528**	**0.0729**	**0.3382**	**80.36**
5m	Neuralrgbd [21]	0.1013	0.1657	0.0502	0.2500	91.60	0.2334	0.4060	0.2163	0.5358	68.03
	Ours-EST(concat)	0.0811	0.1469	0.0279	0.2066	93.19	**0.1458**	0.2554	0.0745	0.3435	79.82
	Ours-EST(adaptive)	**0.0805**	**0.1438**	**0.0275**	**0.2029**	**93.33**	0.1465	**0.2528**	**0.0729**	**0.3382**	**80.36**

Table 3. Memory and computation complexity analysis.

Model	Params	MACs	Memory	Time
DPS [17]	**4.2M**	442.7G	**1595M**	337ms
NAS [20]	18.0M	527.7M	1689G	212ms
Neuralrgbd [21]	5.3M	616.6G	2027M	195ms
DELTAS [30]	124.6M	**98.6G**	2395M	495ms
Ours-ESTM	36.2M	176.9G	1799M	**71ms**

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

Cont.	Trans.	Inference type	Abs	Sq Rel	RMSE	$\sigma < 1.25$
✗	✗	Independent	0.3333	0.0994	0.4897	80.89
✗	✓	Joint	0.3429	0.1291	0.4927	81.36
✗	✓	ESTM	0.3319	0.1073	0.4822	81.43
✓	✗	Independent	0.3220	0.0897	0.4657	82.82
✓	✓	Joint	**0.3133**	**0.0883**	**0.4556**	**83.52**
✓	✓	ESTM	0.3137	0.0884	0.4554	83.43

Memory size	Abs Rel	Abs	Sq Rel	RMSE	$\sigma < 1.25$
1	0.1530	0.2632	0.783	0.3494	79.07
2	0.1465	0.2528	0.0729	0.3382	80.36
3	**0.1460**	**0.2520**	**0.0727**	**0.3376**	**80.44**
4	0.1461	0.2521	0.0728	0.3377	80.44

ESTDepth / eval_hybrid_seq.sh

flamehaze1115 initial code

Code Blame 6 lines (6 loc) · 384 Bytes

```
python eval_hybrid_seq.py --seq_len 5 --summary_freq 10 --ndepths 64 \
--loadckpt ./checkpoint/model_000006.ckpt \
--datapath /userhome/35/xxlong/dataset/scannet_test \
--evalpath ~/workplace/EST/output/hybrid_EST_V4_ndepths64 \
--testlist ./data/scannet_split/test_split.txt --IF_EST_transformer True \
--depth_min 0.1 --depth_max 10. --save_init_prob False --save_refined_prob False
```

ESTDepth / eval_hybrid.sh

flamehaze1115 initial code

Code Blame 9 lines (9 loc) · 415 Bytes

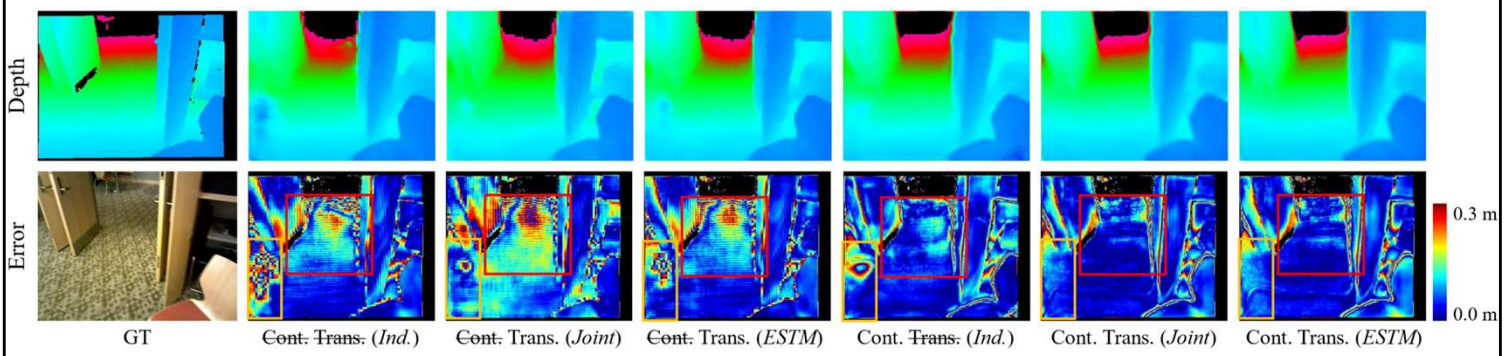
```
python eval_hybrid_seq.py --seq_len 5 --summary_freq 10 --ndepths 64 \
--loadckpt ./checkpoint/model_000006.ckpt  \
--datapath /userhome/35/xxlong/dataset/scannet_test/ \
--evalpath ~/workplace/EST/output/hybrid_EST_V4_ndepths64 \
--testlist ./data/scannet_split/test_split.txt \
--IF_EST_transformer True \
--depth_min 0.1 --depth_max 10. \
--save_init_prob False --save_refined_prob False \
--eval_dataset scannet
```

✓	ResNet-18	✗	Independent	0.1253	0.3213	0.0873	0.4623	0.1759	83.31	95.91	98.49
✓	ResNet-18	✓	Joint	0.1269	0.3180	0.0933	0.4605	0.1758	83.61	95.58	98.25
✓	ResNet-18	✓	ESTM	0.1262	0.3160	0.0897	0.4580	0.1756	83.63	95.68	98.31
✓	ResNet-50	✗	Independent	0.1258	0.3220	0.0897	0.4657	0.1894	82.82	95.55	98.33
✓	ResNet-50	✓	Joint	0.1243	0.3133	0.0883	0.4556	0.1910	83.52	95.60	98.30
✓	ResNet-50	✓	ESTM	0.1254	0.3137	0.0884	0.4554	0.1913	83.43	95.68	98.33

这是因为如果估计的深度不够准确，联合估计将在多个深度图上传播错误的信息。但ESTM受到的影响较小，因为随着更多帧的顺序处理，错误可以逐渐减轻。总体而言，混合正则化网络和 EST 变压器的结合可以提高最佳性能。

当ContextNet采用ResNet-18作为主干时，ESTM深度比Joint深度要好一些。当ContextNet采用ResNet-50作为主干时，联合深度稍微优于ESTM深度。这可以提供额外的证据，表明当估计的深度不够准确时，ESTM 推理操作由于其统一的长期时间相干性而比联合估计表现更好。

Ablation



Tying: 默认情况下，我们会在更新运算符的所有实例中绑定权重。在这里，我们测试了我们方法的一个版本，即每个更新运算符学习一组单独的权重。当权重绑定时，准确率更高，参数数量也明显减少。

DIG

Thanks