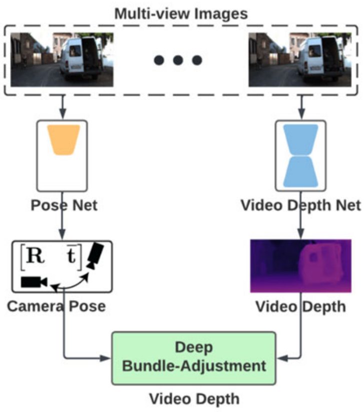
LightedDepth: Video Depth Estimation in light of Limited Inference View Angles (CVPR 2023)

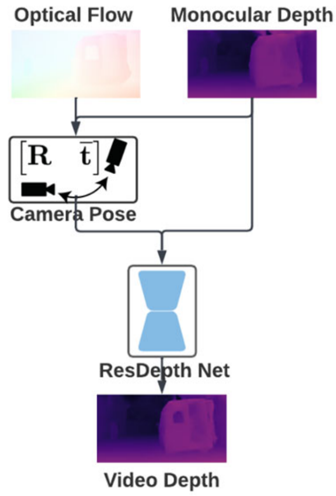
2024.01.11

Contribution & Motivation & Revision

1.we decompose into two sub-tasks that are robust to deficient view angles, and connect them via an efficient scale alignment algorithm.

2. stabilize the indoor normalized pose estimation with the additional projection constraint.





Method

$$\begin{cases} \overline{\mathbf{P}}^{\dagger}, s^{\dagger} = \arg\min_{\overline{\mathbf{P}}, s} \left(h_e \left(\overline{\mathbf{P}}, \mathbf{O} \right) + \right. \\ \qquad\qquad \left. \lambda \cdot h_c \left(f \left(\mathbf{D}^{*}, p \left(\overline{\mathbf{P}}, s \right) \right), \mathbf{O} \right) \right) \\ \mathbf{D}^{\dagger} = \arg\min_{\mathbf{D}} h_p \left(g \left(f \left(\mathbf{D}^{*}, p \left(\overline{\mathbf{P}}, s \right) \right), \mathbf{I}_j \right), \mathbf{I}_i \right). \end{cases}$$

Pose Estimation

Depth Estimation

D∗ and O are initial monodepthmap and flowmap.

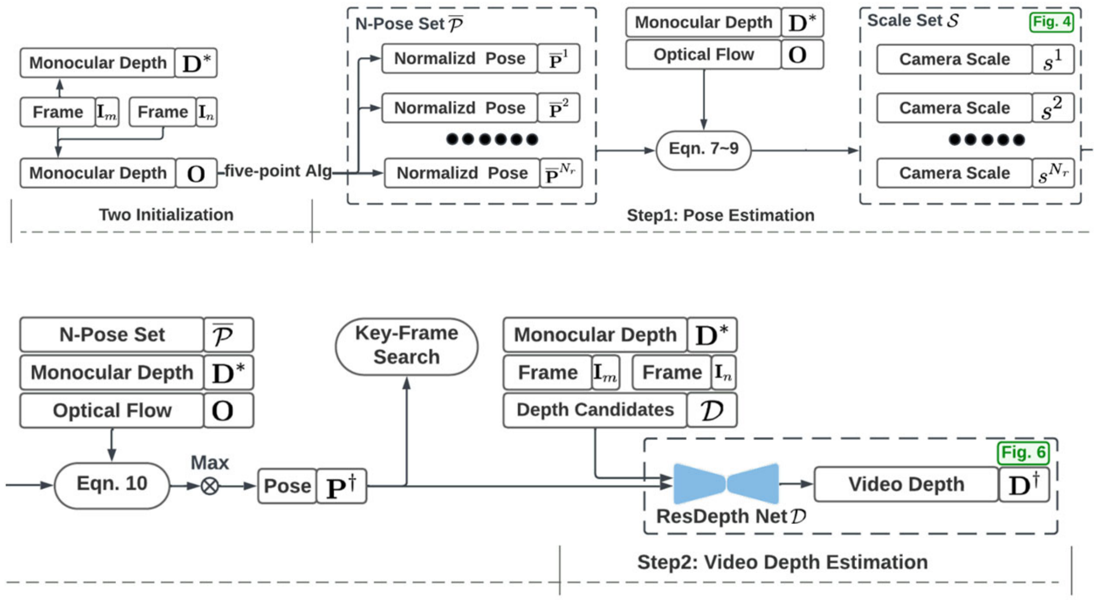
D† and λ are the optimized video depthmap and a predefined weighting parameter

Functions he(·) and hc(·) are epipolar and projection consistency con- straints detailed in Sec. 3.1.

f (·) produces 2D pro- jection locations g(·) applies bilinear sampling to In at 2D locations from f (D, P).

train的时候是在给定Pose和光流的极线约束与，投影约束下训练

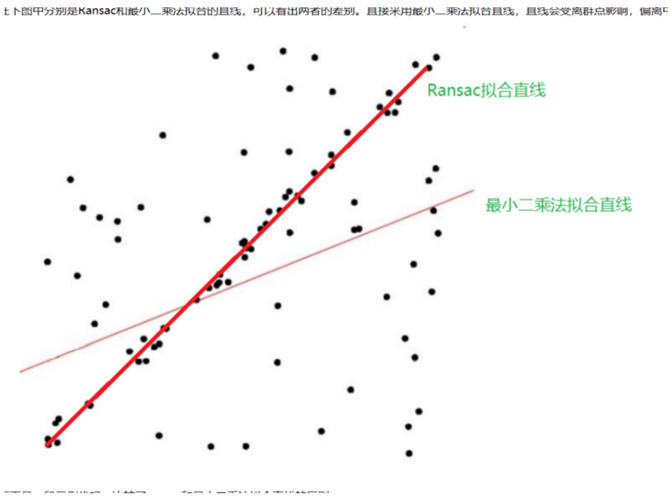
Method: Overview



初始Pose: Essential Matrix 的求解算法--Nister 五点算法

Preliminary: RANSAC

1.在数据中随机选择n个点设定为内群
2.计算适合内群的模型，如线性直线模型
y = ax + b
3.把其它刚才没选到的点带入刚才建立的模型中，计算是否为内群点
4.记下内群数量
5.重复以上步骤, 迭代k次
6.比较哪次计算中内群数量最多,内群最多的那次所建的模型就是我们所要求的解



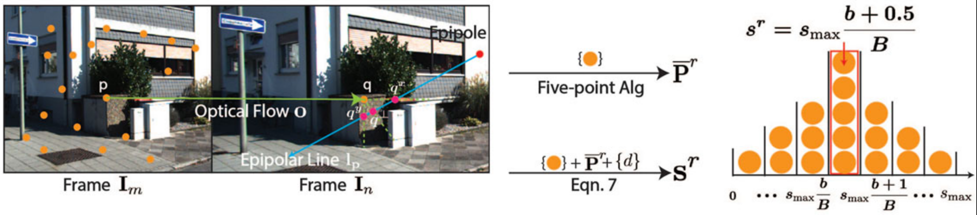
D∗ and O are initial monodepthmap and flowmap.

Functions he(·) and hc(·) are epipolar and projection consistency con- straints detailed in Sec. 3.1.

Method: Pose Estimation

$$\{\mathbf{p}\}, \{\mathbf{o}\} \text{ and } \{d\}$$

$$\mathbf{q}_k = \mathbf{p}_k + \mathbf{o}_k$$



(a) Pixel-wise scale estimation (b) Camera scale estimation

$$d'\mathbf{q} = d'\begin{bmatrix} q^x & q^y & 1 \end{bmatrix}^{\mathsf{T}} = d\mathbf{K}\,\mathbf{R}\,\mathbf{K}^{-1}\,\mathbf{p} + s\mathbf{K}\,\bar{\mathbf{t}}.$$

$$s = \arg\min_{s} \left(d^x - d\right)^2 + \left(d^y - d\right)^2 .$$

$$\log(s) = \log(d) + m, \quad (7)$$

$$m = -\log\frac{1}{2}\left(\frac{x - q_k^x \cdot z}{q_k^x \mathbf{m}_3^{\mathsf{T}}\mathbf{p}_k - \mathbf{m}_1^{\mathsf{T}}\mathbf{p}_k} + \frac{y - q_k^y \cdot z}{q_k^y \mathbf{m}_3^{\mathsf{T}}\mathbf{p}_k - \mathbf{m}_2^{\mathsf{T}}\mathbf{p}_k}\right).$$

$$\begin{cases} h_e(\overline{\mathbf{P}}^r, \{\mathbf{o}\}) = \sum_{k=1}^{N_r}\left(\mathbf{q}_k^{\mathsf{T}}\mathbf{K}^{-\top}\mathbf{E}\mathbf{K}^{\mathsf{T}}\mathbf{p}_k < k_e\right) & (10a) \\ h_c(\overline{\mathbf{P}}^r, s^r, \{\mathbf{p}\}, \{\mathbf{q}\}, \{d\}) = & \\ \quad \sum_{k=1}^{N_r}\left(\|f(d_k, p(\overline{\mathbf{P}}^r, s^r)) - \mathbf{q}_k\|^2 < k_c\right). & (10b) \end{cases}$$

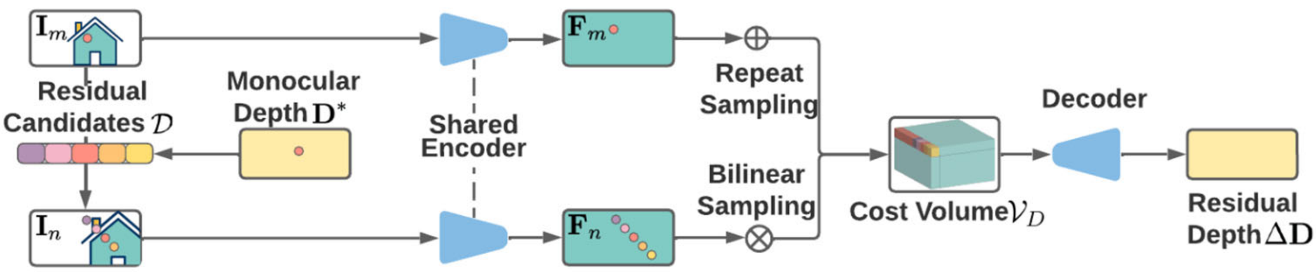
选10000个像素，选R组5个点，算R个pose，根据R个Pose和Depth在10000个像素中投票出R个scale。
然后在这个R个Pose和Scale的约束下，去算 he + hc最小的

随着旋转的累积，流变得与场景深度无关，使得图像线索对于深度来说不太有用。此外，它将非线性投影变换退化为线性仿射变换，破坏了基于极线约束的五点算法。
单应矩阵的应用场景是相机只有旋转而无平移的时候，两视图的对极约束不成立，基础矩阵E为零矩阵，这时候需要使用单应矩阵H
1） 给定一个图像上的一个点，被本质矩阵或基本矩阵相乘，其结果为此点在另一个图像上的对极线，在匹配时，可以大大缩小搜索范围。

Method: Video Depth Estimation

Construct Cost Volume $\mathcal{V}_D$. We sample residual depth candidates $\mathcal{D}$ of size $k_{\mathcal{D}}$ around initial monocular depthmap $\mathbf{D}^*$ with predefined interval Δd as:

$$\mathcal{D} = \left\{ \mathbf{D}_i \parallel \mathbf{D}_i = \exp(\Delta d_i) \cdot \mathbf{D}^* \right\}_{i=1}^{k_{\mathcal{D}}}. \qquad (11)$$



损失函数要求 interval 和 最大的estimation bias 差距不大，如果和gt相比的estimation bias变大，第一项也需要变大，因此保证在两侧。

Experiment

Method	Venue	Frame	Labels	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
DORN [14]	CVPR'18	1	D	0.069	0.300	2.857	0.112	0.945	0.998	0.996
BTS [27]	Arxiv'18	1	D	0.059	0.245	2.756	0.096	0.956	0.993	0.998
AdaBins [2]	CVPR'21	1	D	0.058	0.190	2.360	0.088	0.964	0.995	0.999
NeWCRFs [58]	CVPR'22	1	D	0.052	0.155	2.129	0.079	0.974	0.997	0.999
Ours + BTS [27]	CVPR'23	2	D+F	0.037	0.110	1.809	0.059	0.987	0.998	0.999
Ours + AdaBins [2]		2	D+F	0.045	0.108	1.817	0.064	0.987	0.998	0.999
Ours + NeWCRFs [58]		2	D+F	0.041	0.107	1.748	0.059	0.989	0.998	0.999
BA-Net [40]	ICLR'19	5	D+P	0.083	0.025	3.640	0.134	-	-	-
SfMR [50]	CVPR'21	2	D+F+P	0.055	0.224	2.273	0.091	0.956	0.984	0.993
DeepMLE [8]	Arxiv'22	2	D+F+P	0.060	0.203	2.257	0.089	0.967	0.995	0.999
DRO [20]	Arxiv'21	2	D+P	0.047	0.199	2.629	0.082	0.970	0.994	0.998
MaGNet [1]	CVPR'22	3	D	0.051	0.160	2.077	0.079	0.974	0.995	0.999
DeepV2D [41]	ICLR'20	2	D+P	0.064	0.350	2.964	0.120	0.946	0.982	0.991
		5	D+P	0.037	0.174	2.005	0.074	0.977	0.993	0.997
DeepV2cD [22]	ICPRAI'22	5	D+P	0.037	0.167	1.984	0.073	0.978	0.994	-
Ours + MonoDepth2 [18]	CVPR'23	2	D+F	0.032	0.106	1.889	0.057	0.986	0.998	0.999
Ours + BTS [27]		2	D+F	0.029	0.098	1.729	0.053	0.989	0.998	0.999
Ours + AdaBins [2]		2	D+F	0.030	0.089	1.655	0.052	0.989	0.998	0.999
Ours + NeWCRFs [58]		2	D+F	0.028	0.087	1.597	0.049	0.991	0.998	0.999

$$\hat{s} = median(D_{gt})/median(D_{pred})$$

ScanNet	DeMoN [46]	BA-Net [40]	DSO	DeepV2D-2	DeepV2D-8	FivePoint	Ours
Rotation (degree) ↓	3.791	1.009	0.946	0.806	0.714	0.671	0.621 ± 0.007
Translation (degree) ↓	31.626	14.626	19.238	13.259	12.205	13.878	12.840 ± 0.161
Translation (cm) ↓	15.500	2.365	2.165	1.726	1.514	1.524	1.440 ± 0.011

Table 4. **ScanNet Pose Evaluation.** DeMoN, BA-Net, and DSO are trained on ScanNet. DSO is evaluated only on success cases.

下半表将中值缩放[62]应用于预测深度，以与 SfM 方法进行比较
D=semi-dense depthmap, P=IMU pose, F=synthetic optical flow datasets [4, 33]]

Ablation

Mehod	All		Background	
	F1-epe	F1-a1	F1-epe	F1-a1
RAFT [11]	**1.284**	**4.539**	**1.238**	**4.759**
DeepV2D [10]	9.957	22.610	2.180	9.789
Ours	**9.321**	**20.723**	**1.631**	**7.692**

Table 1. **Flow Performance Comparison on KITTI FLOW15 Dataset [4].** RAFT [11] computes flow via regression while

	ResDepth	PoesEstimation	ScaleNet	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	Seq-00 t_{err}
KITTI		✓		0.070	0.275	2.405	0.093	0.959	1.55
	✓	✓		0.038	**0.110**	**1.821**	0.060	**0.987**	1.55
	✓	✓	✓	**0.037**	0.117	1.841	**0.059**	0.986	**1.24**

Table 5. **Ablation on Outdoor Video Depth Estimation.** [Key: 'ResDepth'= Residual depth learning (Sec. 3.2). 'PoseEstimation'= Proposed Pose Estimation Method (Sec. 3.1). 'ScaleNet'=Further refine pose scale with an additional ScaleNet (detailed in Supplementary).]

	FivePoint	PoesEstimation	KeySearch	Abs Rel	Sc Inv	RMSE	log10	$\delta < 1.25$
NYUv2	✓			0.063	0.087	0.248	0.027	0.964
		✓		0.061	0.083	0.239	0.026	0.968
		✓	✓	**0.057**	**0.080**	**0.230**	**0.025**	**0.971**

Table 6. **Ablation on Indoor Video Depth Estimation.** [Key: 'FivePoint'=Baseline Five-point algorithm with RANSAC. 'PoseEstimation'=Proposed Pose Estimation Method (Sec. 3.1). 'KeySearch'=Keyframe search. Bold marks the best score.]

baseline是BTS、NewCRF

Comparison & Comments

Method	Multi	abs rel	sq rel	rmse	$\text{rmse}_{\log}$	$\delta<1.25$
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**

1. 缺乏不使用光流约束pose和scale的消融实验以及和seperate pose network的对比。

2. motivation：预测深度 + pose -> 光流比直接预测光流差，所以要用直接预测光流来矫正Pose进而矫正深度？

3. 暂时伪开源，缺乏训练代码

Thanks