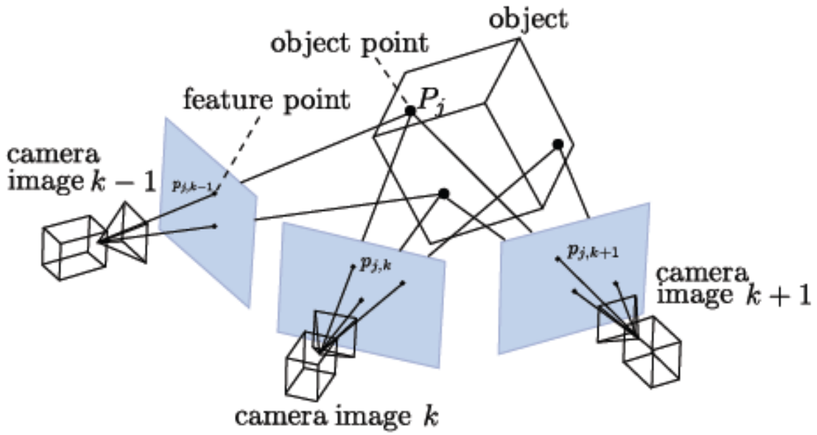
Deep Learning for Depth Estimation

Zhuofei

Monocular Visual Odometry



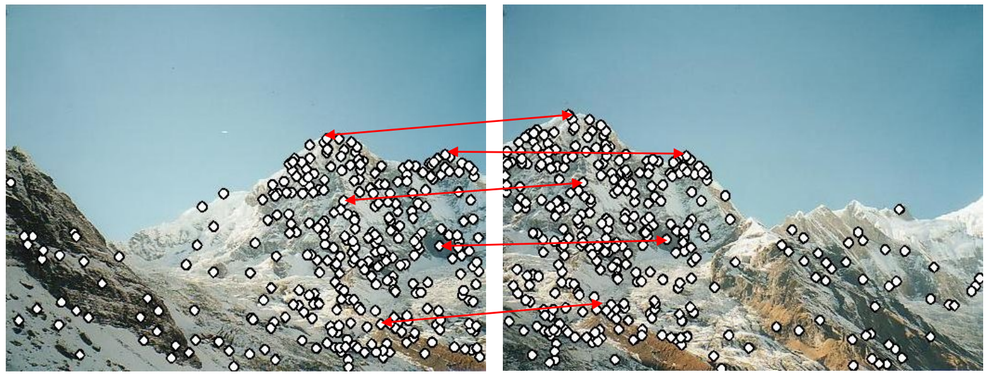
Monocular Image Sequences

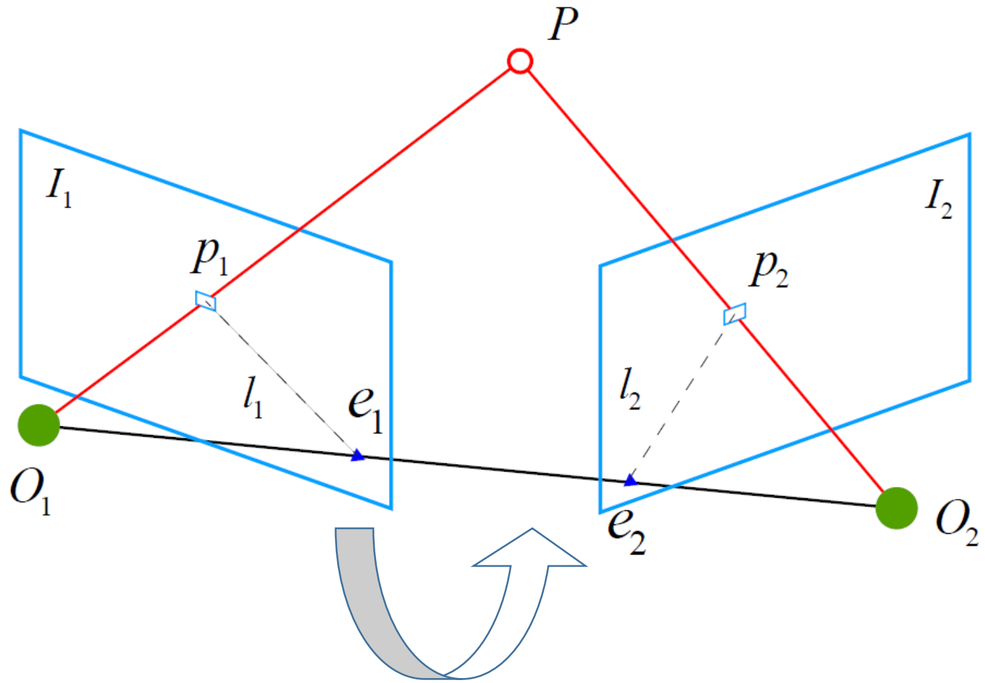
- estimates the ego-motion of a camera
- integrates the relative motion into global poses

$$R^{3\times3}, t^{3\times1}$$

Traditional VO

- Feature-Based Matching + Geometric construction
  - Epipolar geometry: $\mathbf{p}_2^{\mathrm{T}}\mathbf{F}\mathbf{p}_1 = 0$
  - PnP

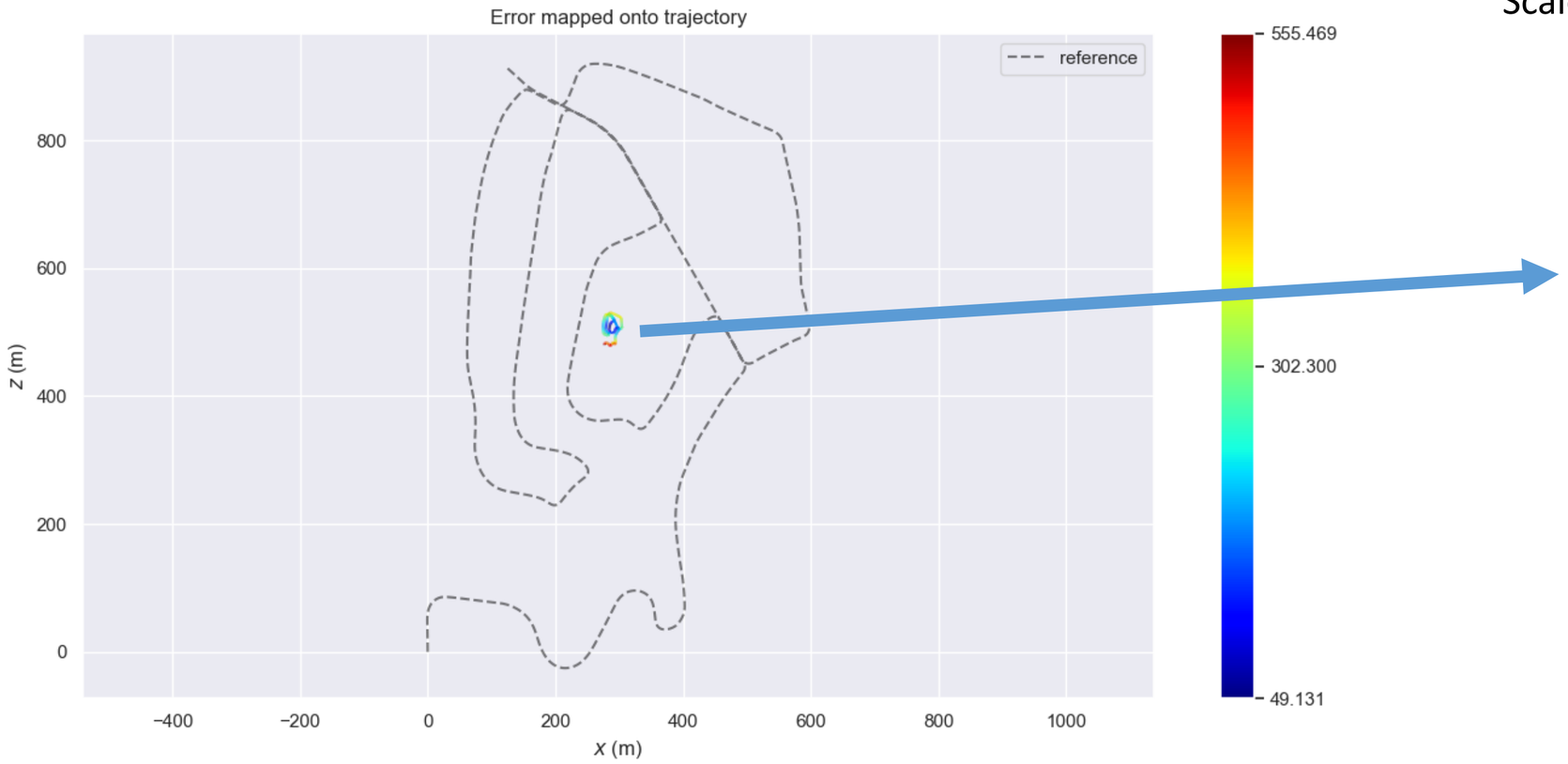




R, t

Limitation 1

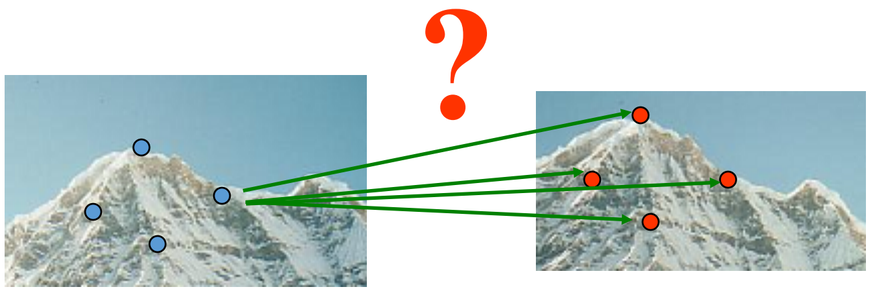
- Scale ambiguity



Scale factor: 17.95

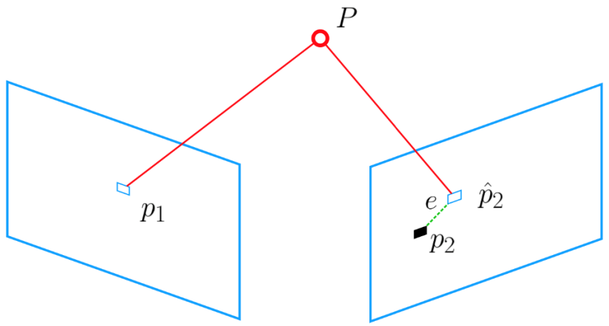
Limitation 2

- Mismatches in repetitive cases
  - For each point hardly correctly recognize the corresponding one



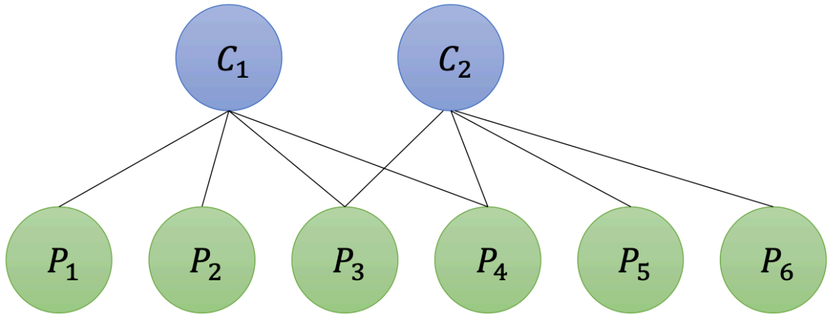
Basic concept: BA vs Direct

Feature-Based BA:

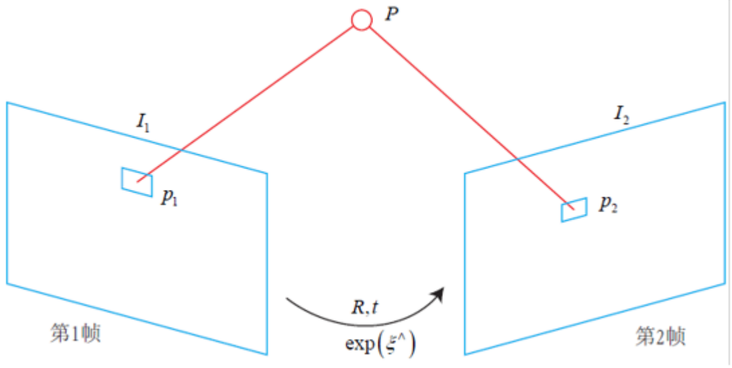


Re-projection error

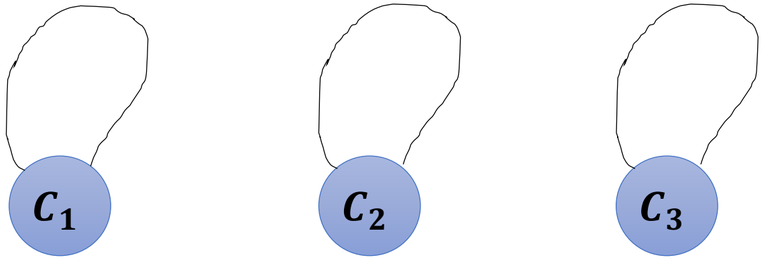
$$e_{ij} = z_{ij} - \pi(\xi_i, \mathbf{p}_j)$$



Direct Method:



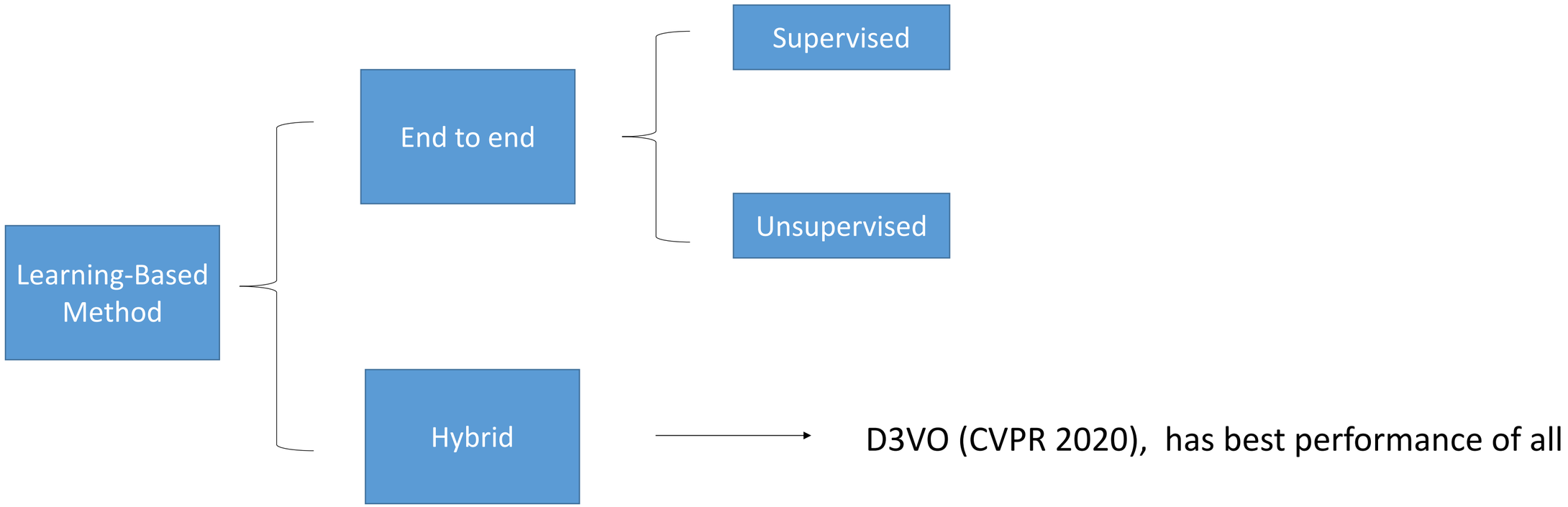
Photometric error



$$e = I_1(p_1) - I_2(p_2)$$

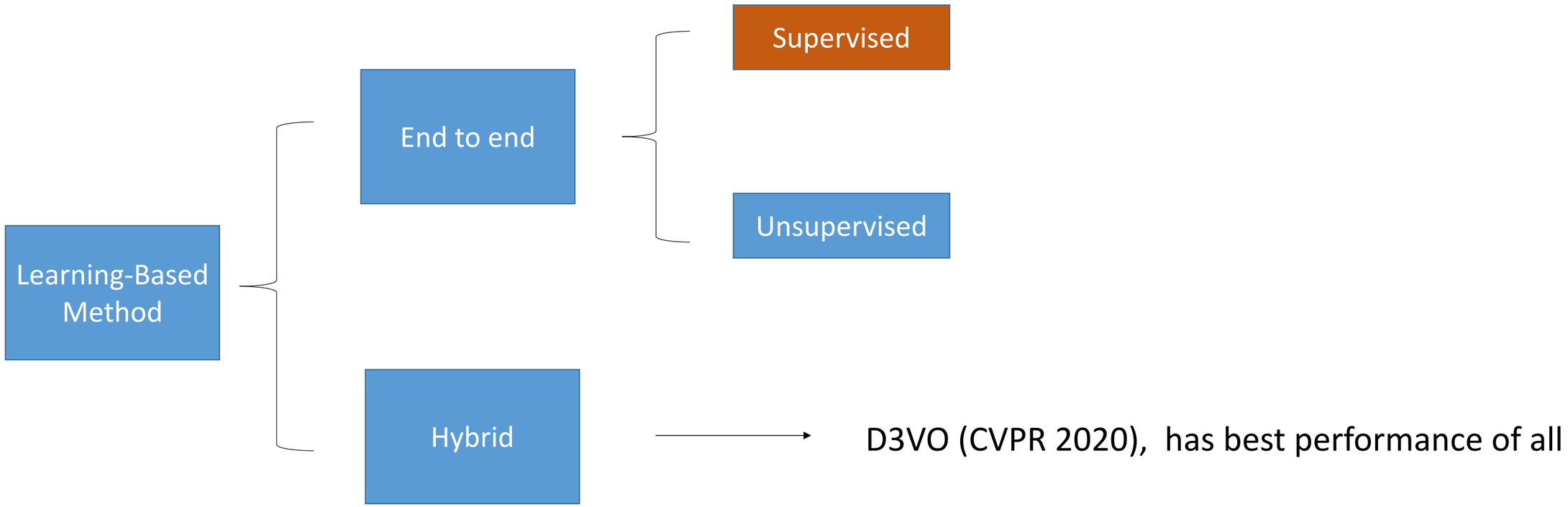
Monocular Visual Odometry

- Learning-Based Classification



Monocular Visual Odometry

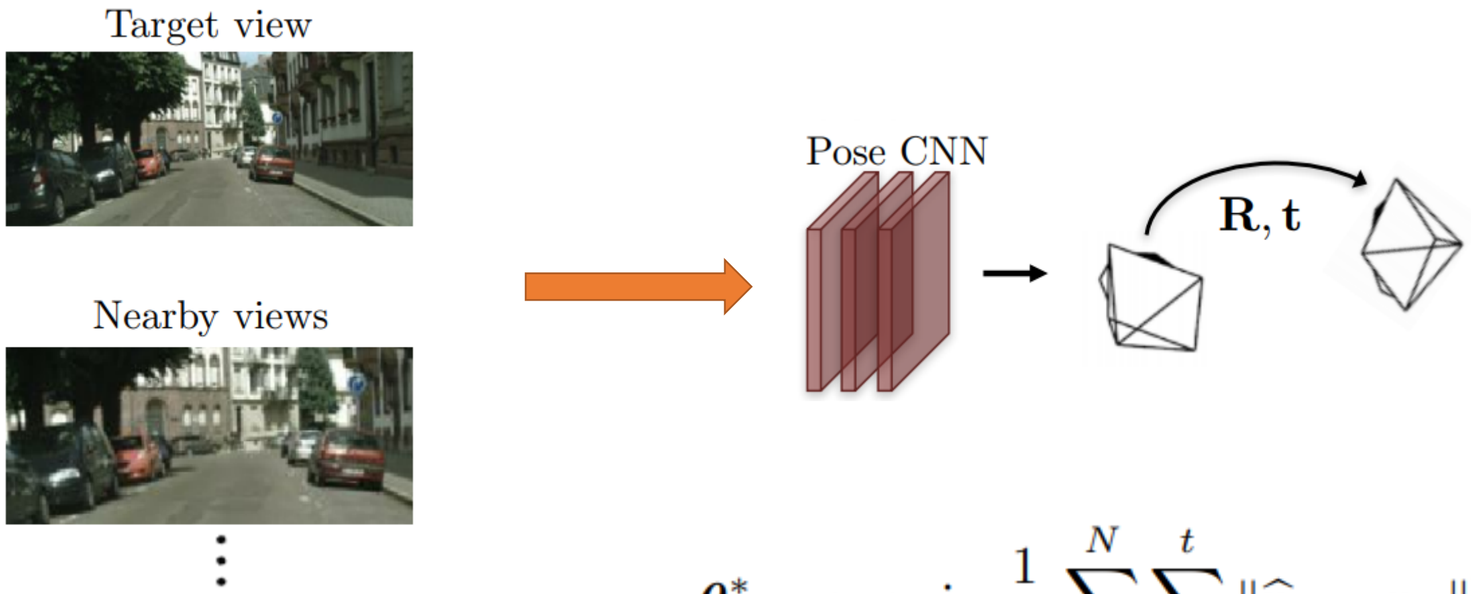
- Learning-Based Classification



Supervised VO

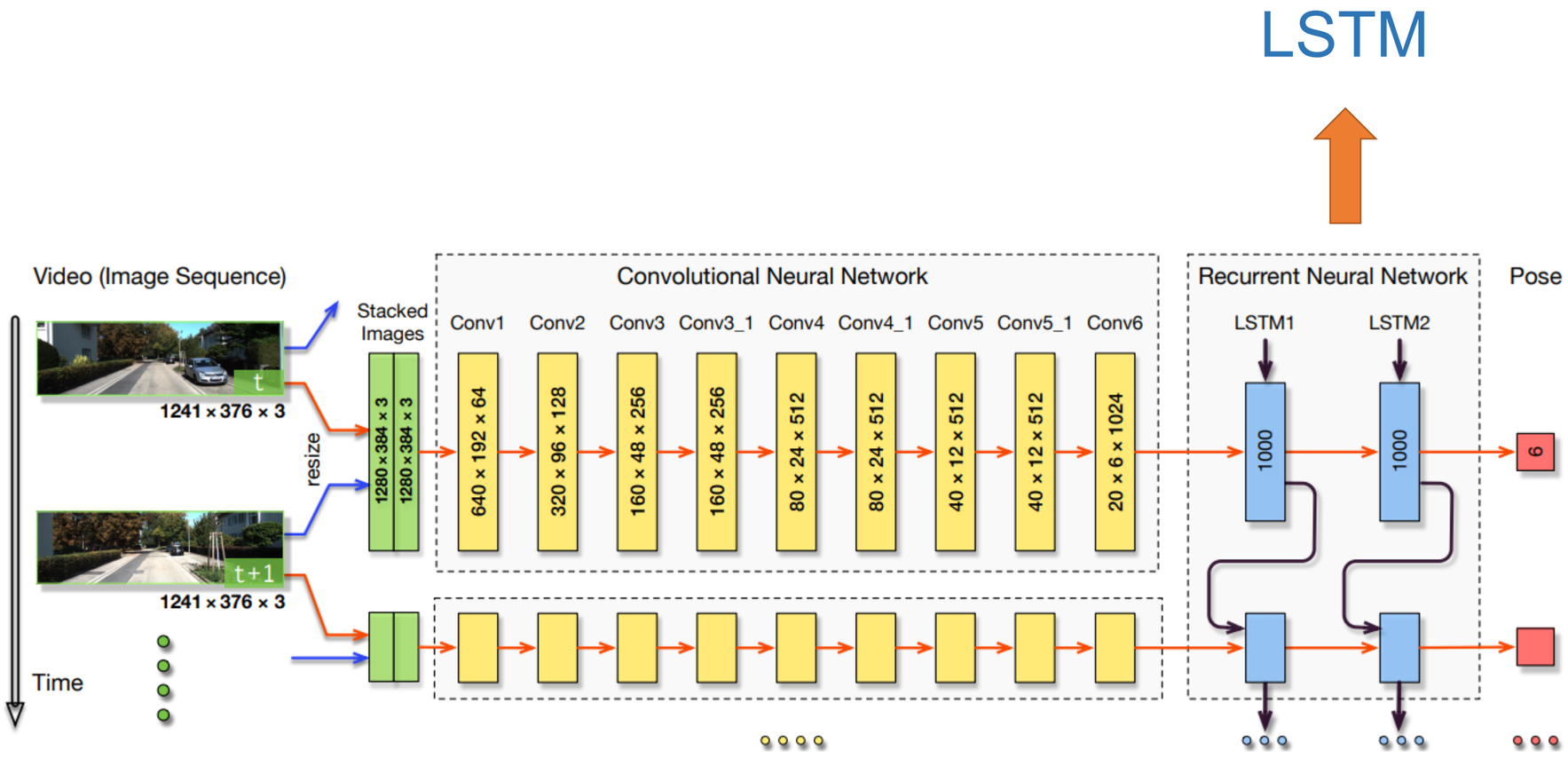
Train a deep end-to-end CNN on annotated dataset:

$$Func\colon I_{1,2,\ldots,N} \rightarrow (R,t)_{1,2,\ldots,N}$$



$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{t} \|\widehat{\mathbf{p}}_k - \mathbf{p}_k\|_2^2 + \kappa \|\widehat{\boldsymbol{\varphi}}_k - \boldsymbol{\varphi}_k\|_2^2$$

Deep VO (ICRA 2017)

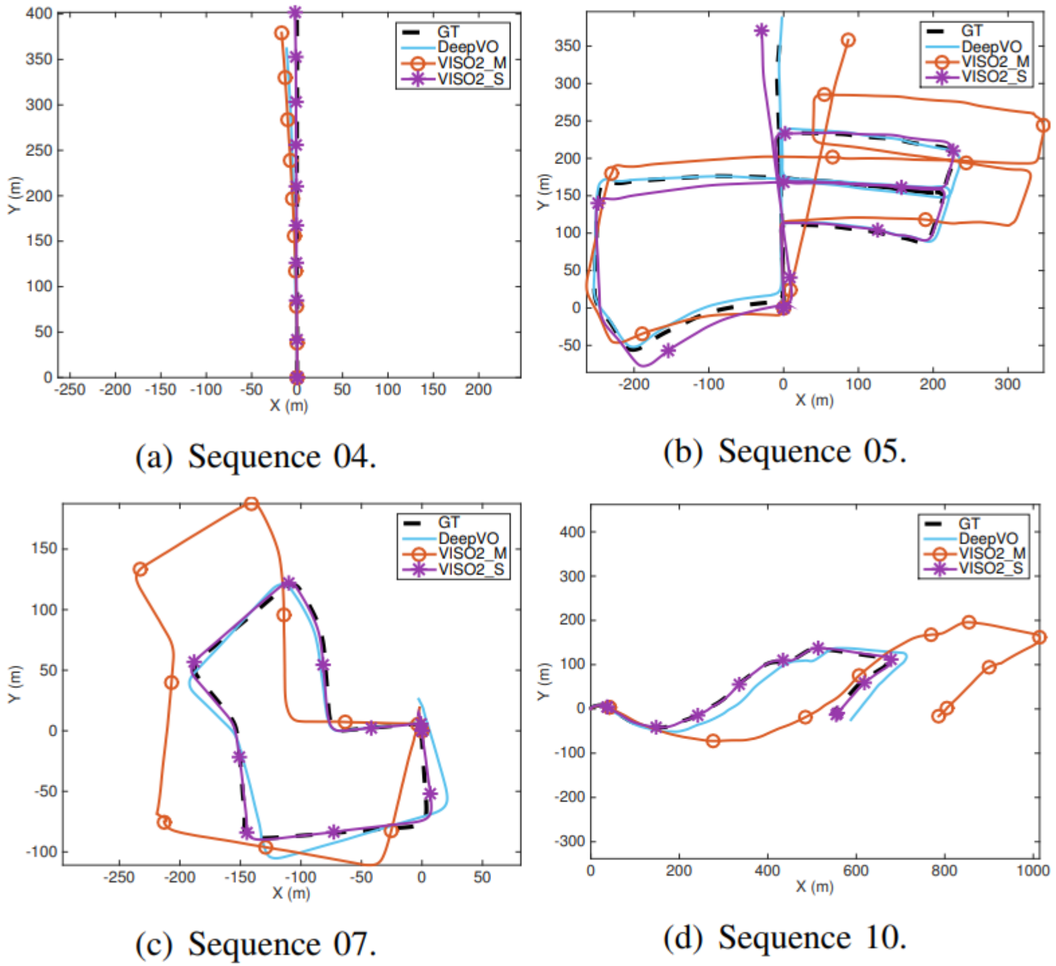


Deep VO (ICRA 2017)

- DeepVO achieves more robust results than the monocular VISO2

Seq.	DeepVO		VISO2_M		VISO2_S	
	$t_{\text{rel}}(\%)$	$r_{\text{rel}}(^\circ)$	$t_{\text{rel}}(\%)$	$r_{\text{rel}}(^\circ)$	$t_{\text{rel}}(\%)$	$r_{\text{rel}}(^\circ)$
03	8.49	6.89	8.47	8.82	3.21	3.25
04	7.19	6.97	4.69	4.49	2.12	2.12
05	2.62	3.61	19.22	17.58	1.53	1.60
06	5.42	5.82	7.30	6.14	1.48	1.58
07	3.91	4.60	23.61	29.11	1.85	1.91
10	8.11	8.83	41.56	32.99	1.17	1.30
mean	5.96	6.12	17.48	16.52	1.89	1.96

- t_{rel}: average translational RMSE drift $(\%)$ on length of 100m-800m.
- r_{rel}: average rotational RMSE drift $(^\circ/100m)$ on length of 100m-800m.
- The DeepVO model used is trained on Sequence 00, 02, 08 and 09. Its performance is expected to improve when it is trained on more data.

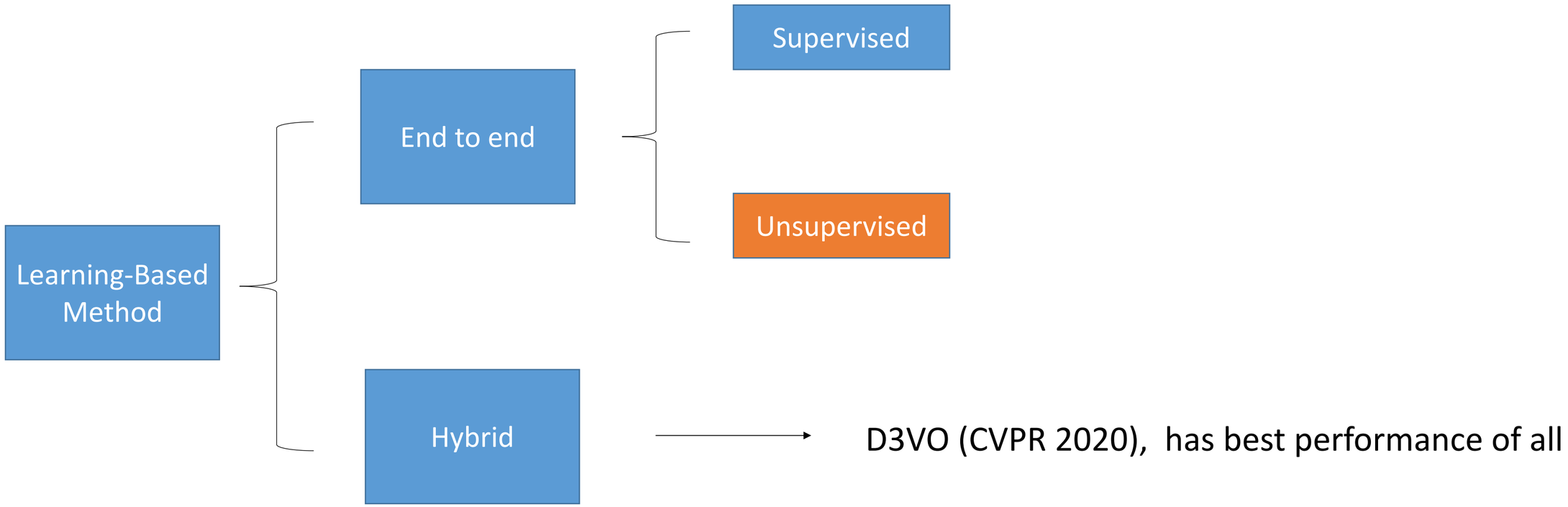


(a) Sequence 04. (b) Sequence 05.

(c) Sequence 07. (d) Sequence 10.

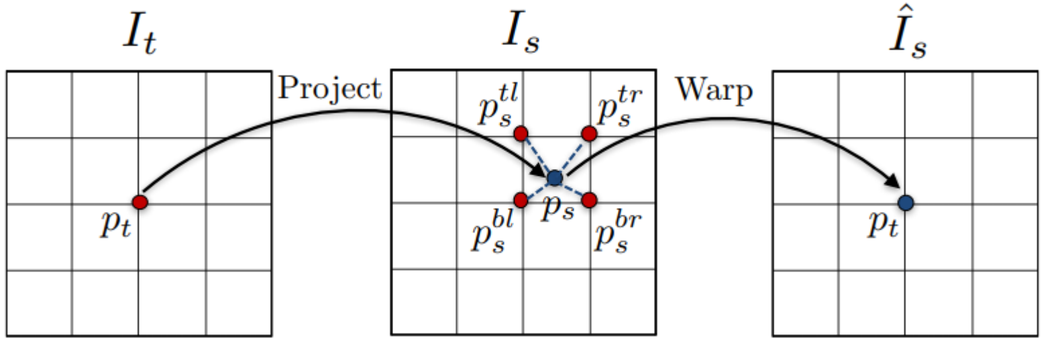
Monocular Visual Odometry

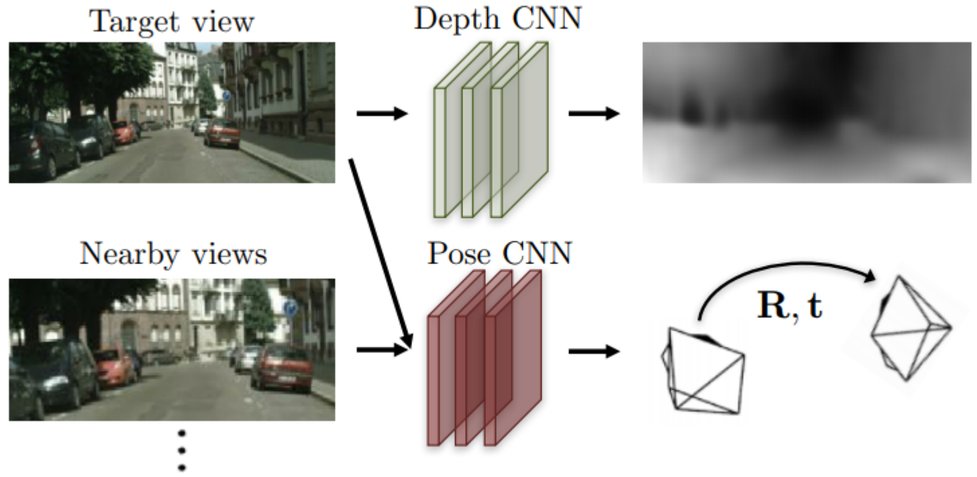
- Learning-Based Classification



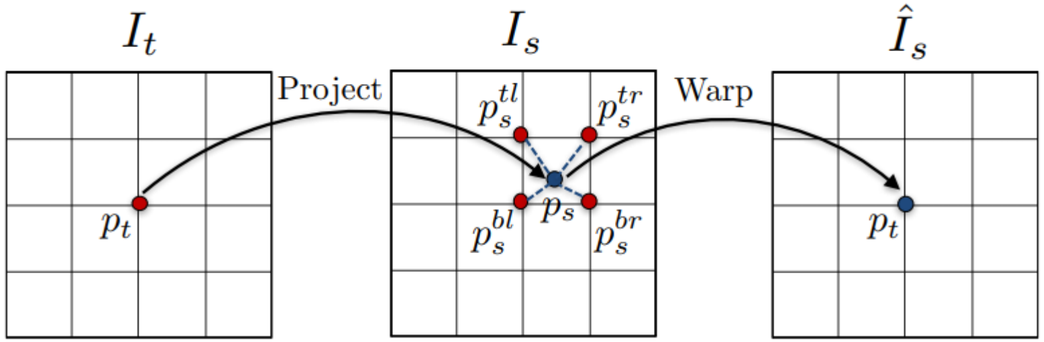
Unsupervised VO

- Utilizing view synthesis as a supervisory signal
- Jointly learns depth and camera ego-motion from video sequences
- Beyond Photometric Loss for Self-Supervised Ego-Motion Estimation (DeepMatchVO, ICRA 2019)
  - SfMLearner (CVPR 2017 oral)
  - GeoNet (CVPR 2018)





Reconstruction Loss

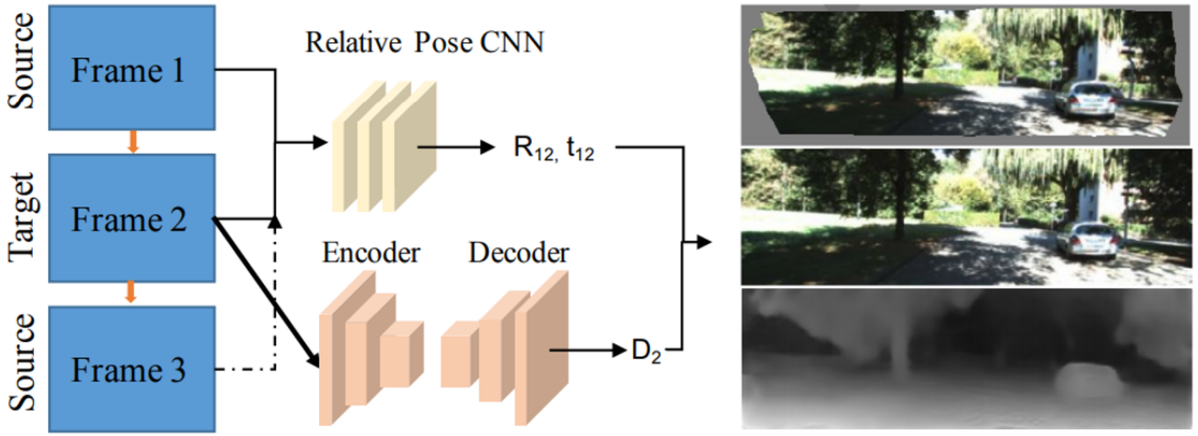


- Image synthesis:

$$p_1 \sim K_1[\hat{R_{12}}|\hat{t_{12}}]\widehat{D_2}(p_2)K_2^{-1}p_2$$

- Add structured similarity (SSIM):

$$\mathcal{L}_{img} = (1-\alpha)||\mathcal{I}_2 - \widetilde{\mathcal{I}_2^1}||_1 + \alpha\frac{1-SSIM(\mathcal{I}_2 - \widetilde{\mathcal{I}_2^1})}{2}$$



- Fix photometrc error:

$$\mathcal{M}(P_M) = \{ \begin{matrix} 1 & Percentile(\mathcal{L}_{img}(i,j)) \leq P_M \\ 0 & otherwise \end{matrix}$$

Fix Photometric Error

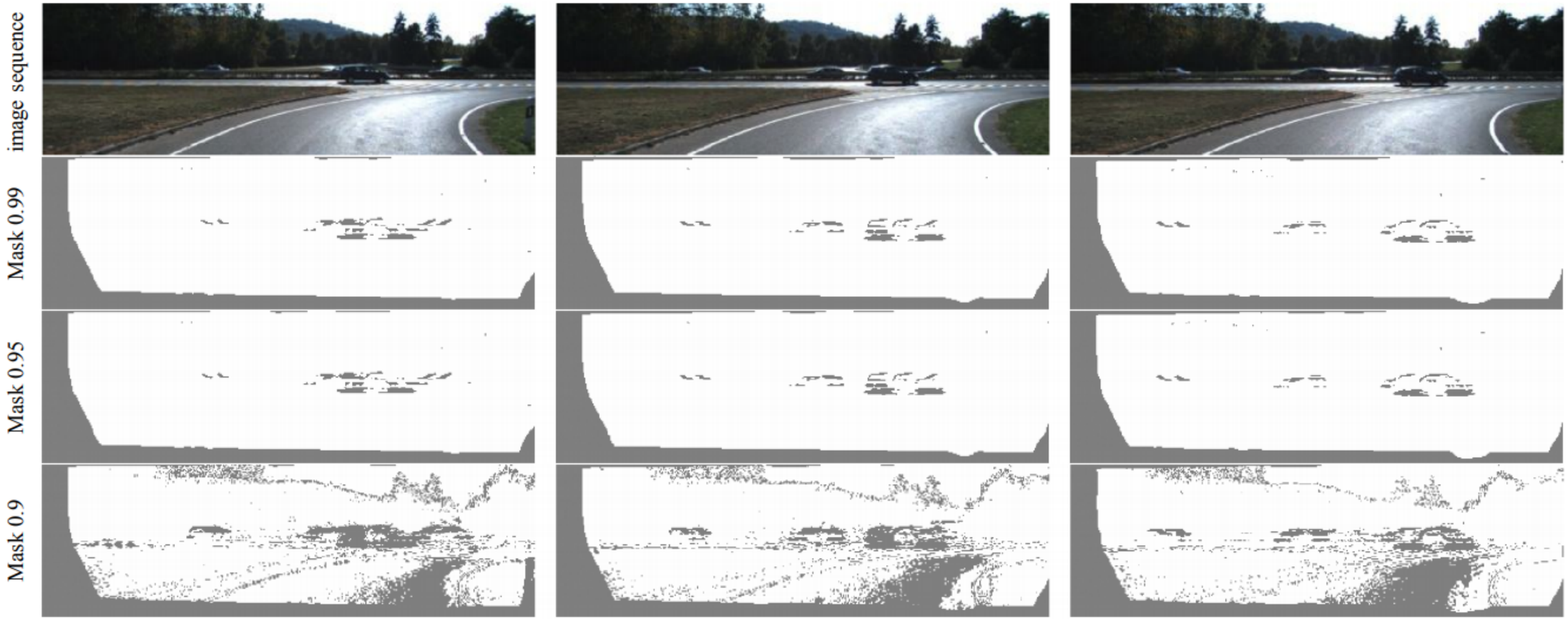


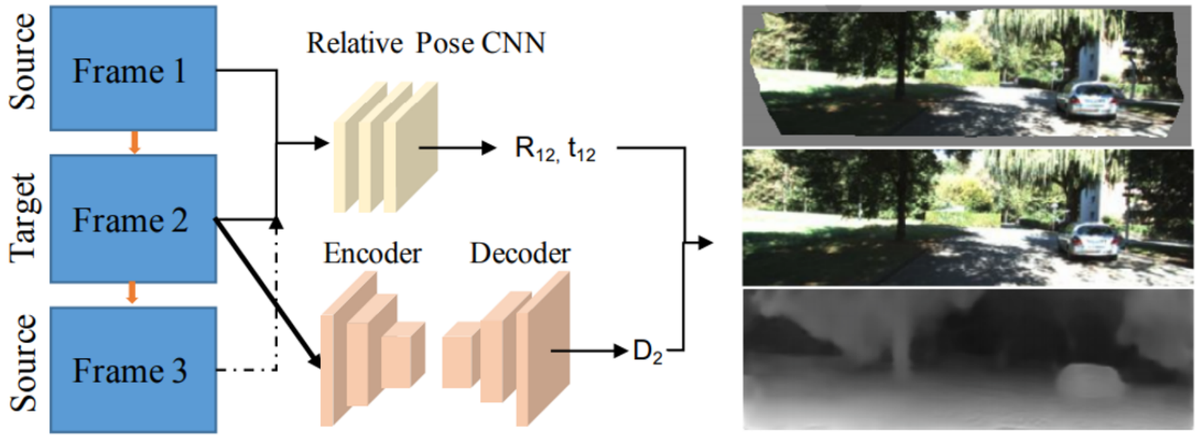
Fig. 2. Threshold masks with P_M from 0.9 to 0.99. We set $P_M (= 0.99)$ modestly so that the loss formulation does not lose too much information.

Smoothness Loss

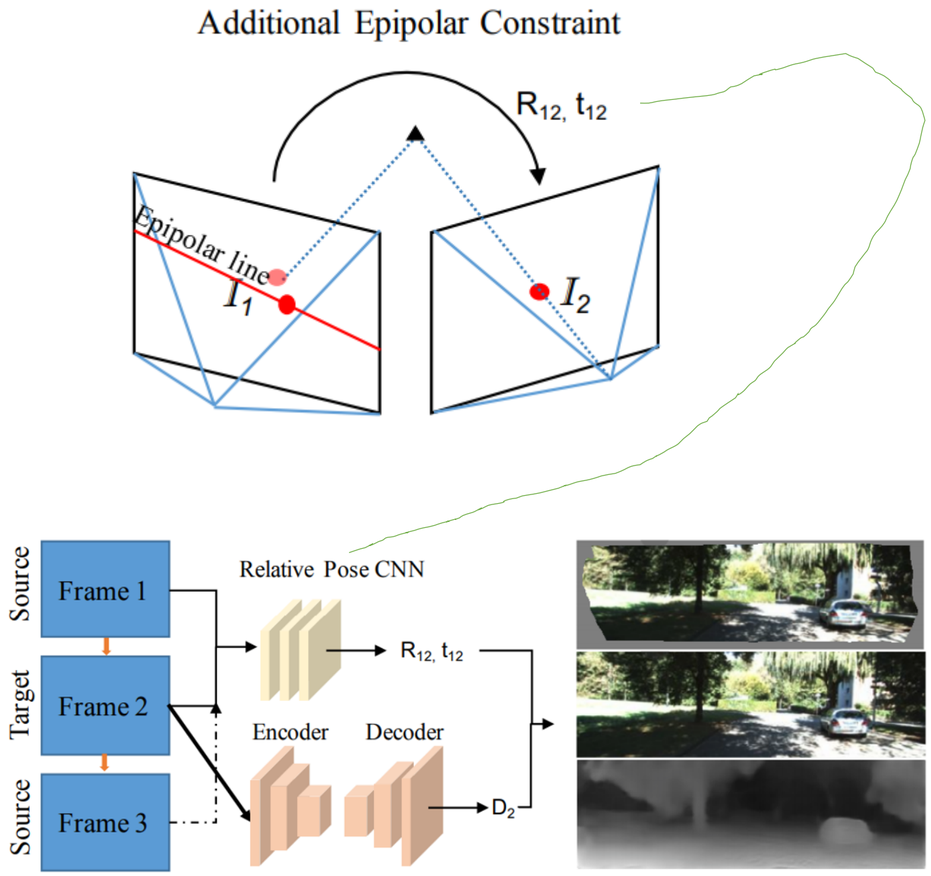
- Resolve the gradient-locality issue in motion estimation
- Remove discontinuity of the learned depth in low-texture regions

- Edge-aware smoothness term:

$$\mathcal{L}_{smooth} = \sum_{p} |\nabla D(p)|^{T} \cdot e^{-|\nabla I(p)|}$$



Pairwise Matching Loss



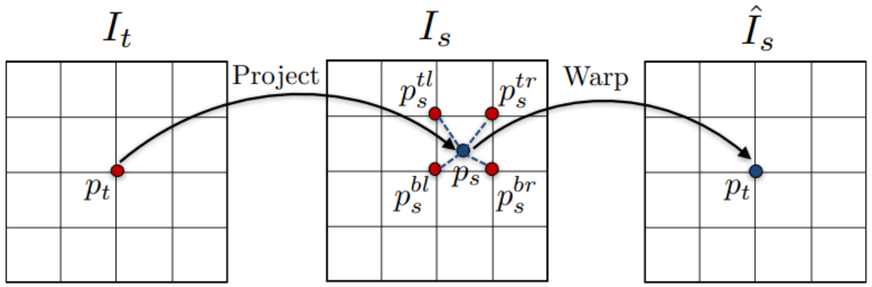
- Epipolar constraint

$$q_i^T F_{12} p_i = (K_2^{-1} q_i')^T R_{12} [t_{12}]_\times (K_1^{-1} p_i') = 0$$

- Epipolar line $l_{12}^{(i)}$:

$$\mathcal{L}_{geo} = \sum_i dist(l_{12}^{(i)}, q_i)$$

Direct Weak Pose

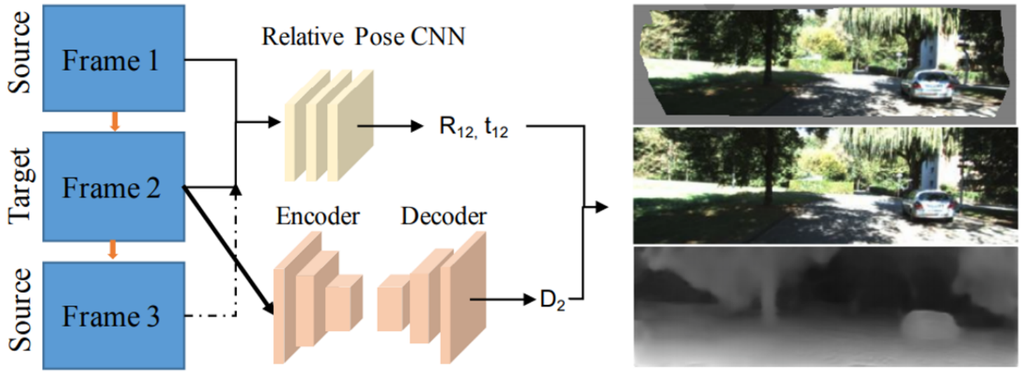


Recall Image synthesis:

$$p_1 \sim K_1[\hat{R_{12}}|\hat{t_{12}}]\widehat{D_2}(p_2)K_2^{-1}p_2$$

3D-2D: PnP → weak pose

Depth



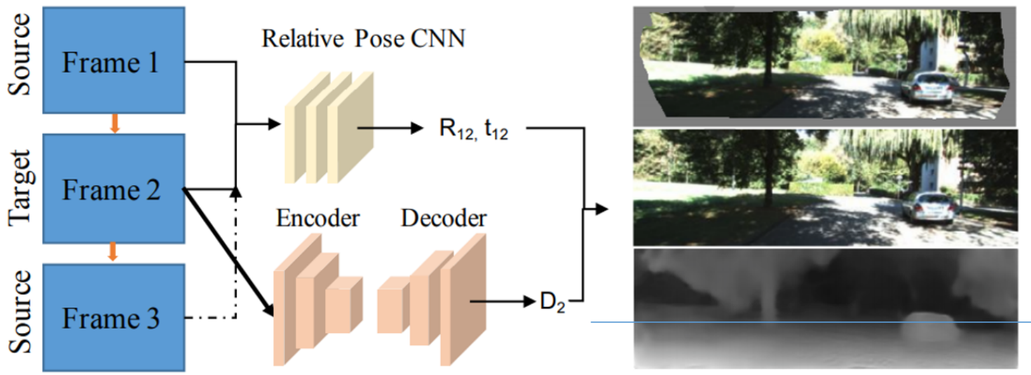
$$\begin{aligned} p_1 &\sim K_1 P_1 P_2^{-1} \widehat{D_2}(p_2) K_2^{-1} p_2 \\ &\sim K_1[R_1|T_1][R_2^T| - R_2^T T_2]\widehat{D_2}(p_2)K_2^{-1}p_2 \\ &\sim K_1[R_1 R_2^T|T_1 - R_1 R_2^T T_2]\widehat{D_2}(p_2)K_2^{-1}p_2 \end{aligned}$$

Use this PnP-pose instead of CNN pose

Prior Weak Pose

CNN: relative pose

3D-2D: PnP → prior pose



Pose error:

$$\mathcal{L}_{pose} = \mathcal{L}_{rotation} + \mathcal{L}_{translation}$$
$$= w_r ||\hat{r_{ij}} - \overline{r_{ij}}||_2 + w_t ||\hat{t_{ij}} - \overline{t_{ij}}||_2$$

Total Error

$$\mathcal{L}_{total} = \mathcal{M}(P_M) \odot \mathcal{L}_{img} + w_s \mathcal{L}_{smooth} + [w_g \mathcal{L}_{geo}] + [w_p \mathcal{L}_{pose}]$$

TABLE I. Three ways to incorporate geometric constraints, compared with baseline method with and without mask. The columns that are marked with red means 'the lower the better', and the columns with purple means 'the higher the better'.

Method	Geometric Info	Cap (m)	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
*Baseline (**w/o** Mask)*	No	80	0.171	1.512	6.332	0.250	0.764	0.918	0.967
*Baseline (**w** Mask)*	No	80	0.163	1.370	6.397	0.258	0.758	0.910	0.962
Pairwise-Matching	Self-generated Matches	80	**0.156**	**1.357**	**6.139**	**0.247**	**0.781**	**0.920**	0.965
Prior-Weak-Pose [21]	Self-generated Pose	80	0.163	1.371	6.275	0.249	0.773	0.918	**0.967**
Direct-Weak-Pose	Self-generated Pose	80	0.162	1.46	6.27	0.249	0.775	0.919	0.965

$w_s = 0.1$: for the whole experiment

$w_g = 0.001, w_p = 0$: Pairwise-Matching

$w_g = 0, w_p = 0.1$: Prior-Weak-Pose

$w_g = 0, w_p = 0$: Direct-Weak-Pose, use for Depth-Net training

TABLE II. Single-view depth estimation performance. The statistics for the compared methods are excerpted from the original papers. 'K' represents KITTI raw dataset (Eigen split) and CS represents cityscapes training dataset. The best results with capped 80m are **bolded**.

Method	Supervision	Dataset	Cap (m)	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen et al. [5] Fine	Depth	K	80	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Liu et al. [26]	Depth	K	80	0.202	1.614	6.523	0.275	0.678	0.895	0.965
Godard et al. [13]	Pose	K	80	**0.148**	1.344	5.927	0.247	0.803	0.922	0.964
Zhou et al. [50] updated	No	K	80	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Mahjourian et al. [29]	No	K	80	0.163	1.24	6.22	0.25	0.762	0.916	0.968
Yin et al. [47]	No	K	80	0.155	1.296	5.857	0.233	0.793	0.931	**0.973**
Yin et al. [47]	No	K + CS	80	0.153	1.328	5.737	0.232	**0.802**	0.934	0.972
Ours	No	K	80	0.156	1.309	5.73	0.236	0.797	0.929	0.969
Ours	No	K + CS	80	0.152	**1.205**	**5.564**	**0.227**	0.8	**0.935**	**0.973**
Garg et al. [11]	Stereo (Pose)	K	50	0.169	1.080	5.104	0.273	0.740	0.904	0.962
Zhou et al. [50]	No	K	50	0.201	1.391	5.181	0.264	0.696	0.900	0.966
Ours	No	K	50	0.149	1.01	4.36	0.222	0.812	0.937	0.973

TABLE III. Visual odometry performance. Learning-based methods use 128×416 images while ORB-SLAM2 uses original images. The pose snippet data is not available for [29] so it is not compared for full pose.

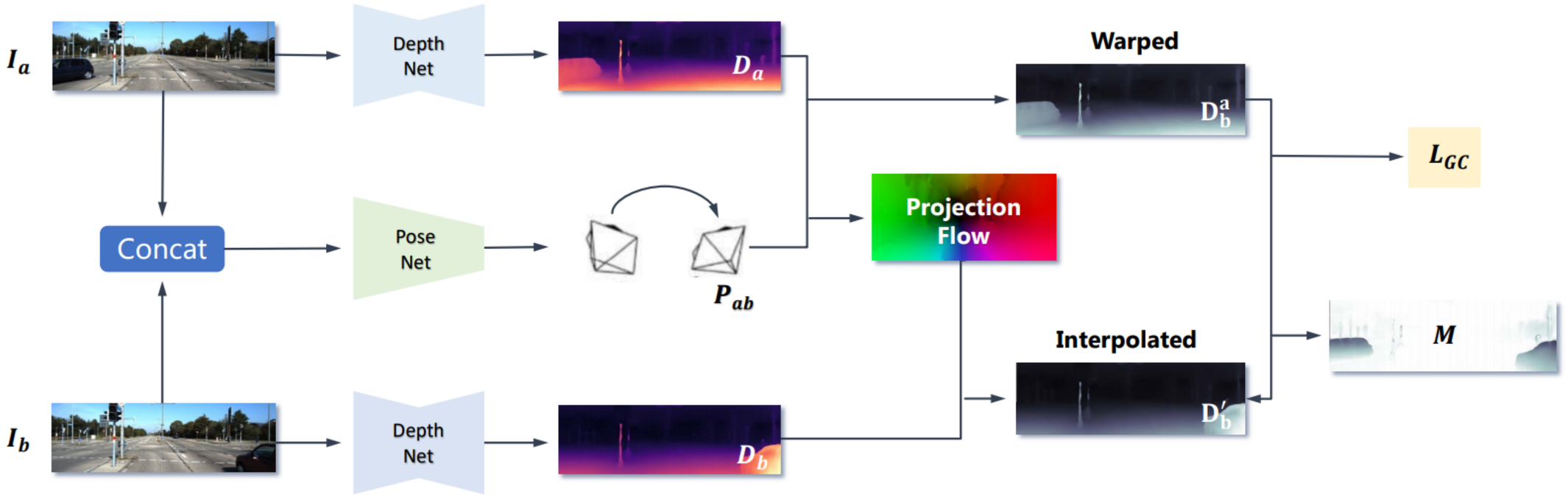
Method	Seq.09		Seq.10 (no loop)	
	Snippet	Full (m)	Snippet	Full (m)
ORB-SLAM2 (full, w LC)	0.014 ± 0.008	7.08	0.012 ± 0.011	5.74
ORB-SLAM2 (full, w/o LC)	-	38.56	-	5.74
Zhou et al. [50] updated (5-frame)	0.016 ± 0.009	41.79	0.013 ± 0.009	23.78
Yin et al. [47] (5-frame)	0.012 ± 0.007	152.68	0.012 ± 0.009	48.19
Mahjourian et al. [29] , no ICP (3-frame)	0.014 ± 0.010	-	0.013 ± 0.011	-
Mahjourian et al. [29] , with ICP (3-frame)	0.013 ± 0.010	-	0.012 ± 0.011	-
Ours et al. (3-frame)	$\mathbf{0.0089 \pm 0.0054}$	18.36	$\mathbf{0.0084 \pm 0.0071}$	16.15

Limitations

- Hard to compete with supervised VO in performance
- Still not able to provide pose estimates in a consistent global scale
  - *Unsupervised Scale-consistent Depth and Ego-motion Learning from Monocular* Video *(NeurIPS 2019):* geometric consistency loss, consistence on global track
  - *UnDeepVO (ICRA 2018):* utilize stereo image pairs to recover the absolute scale

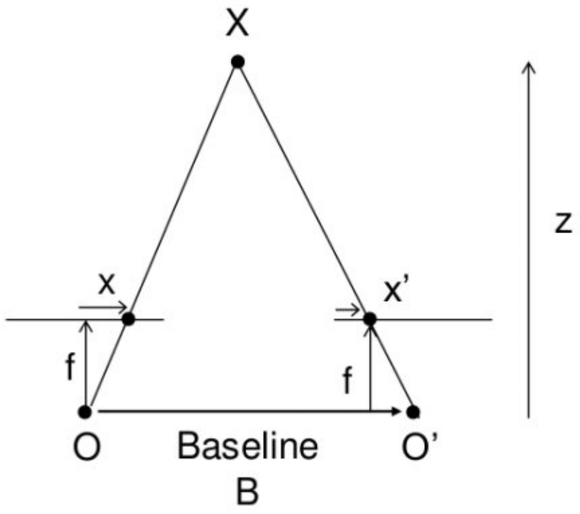
Geo Consistency (NeurIPS 2019)

- enforces the consistency between predicted depth maps and reconstructed depth maps.



UnDeepVO (ICRA 2018)

- Stereo baseline is fixed and known



- Train with stereo, test with mono

- Unsupervised: no access to ground-truth

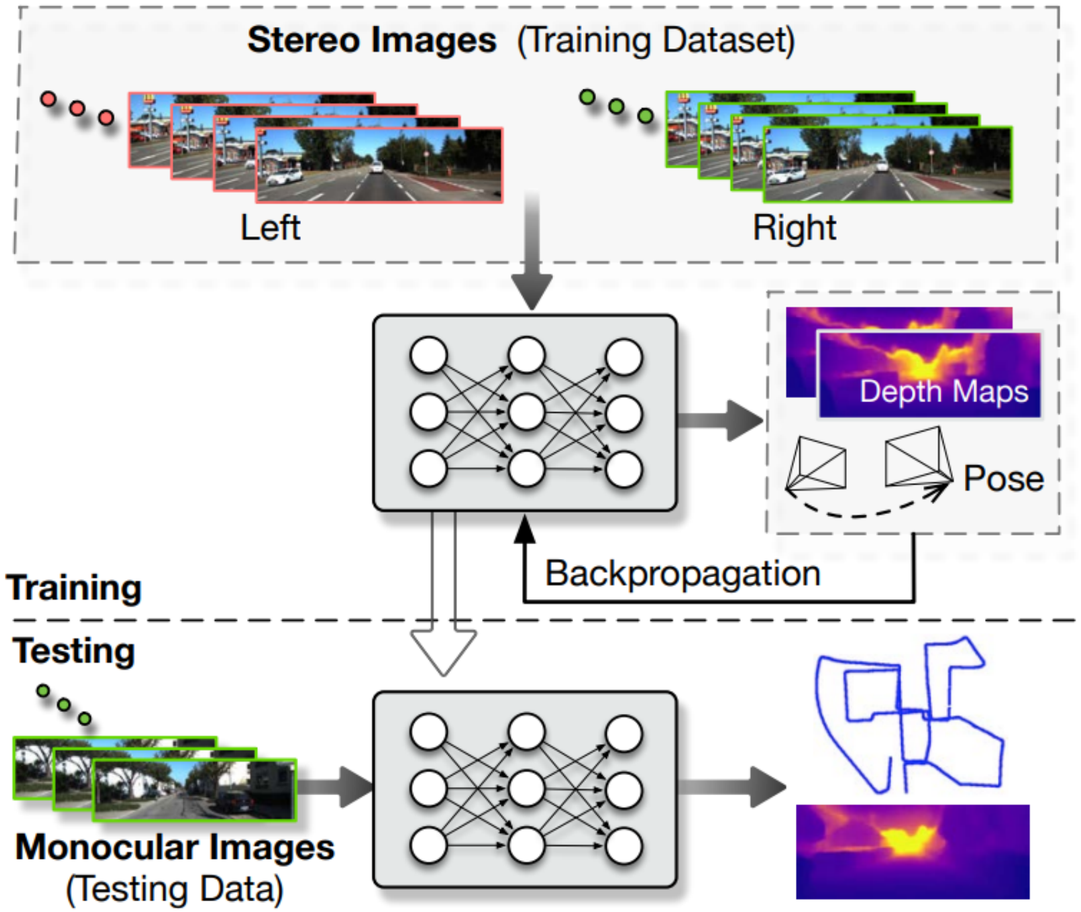
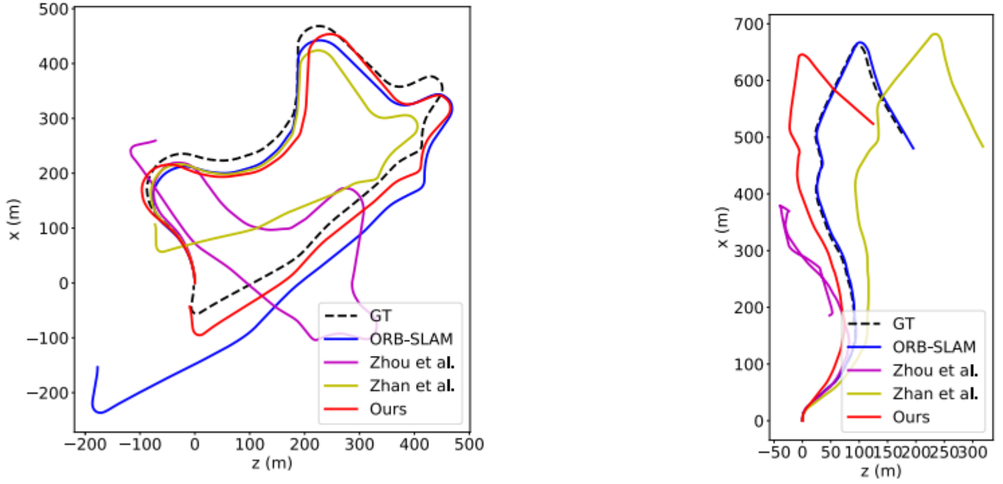


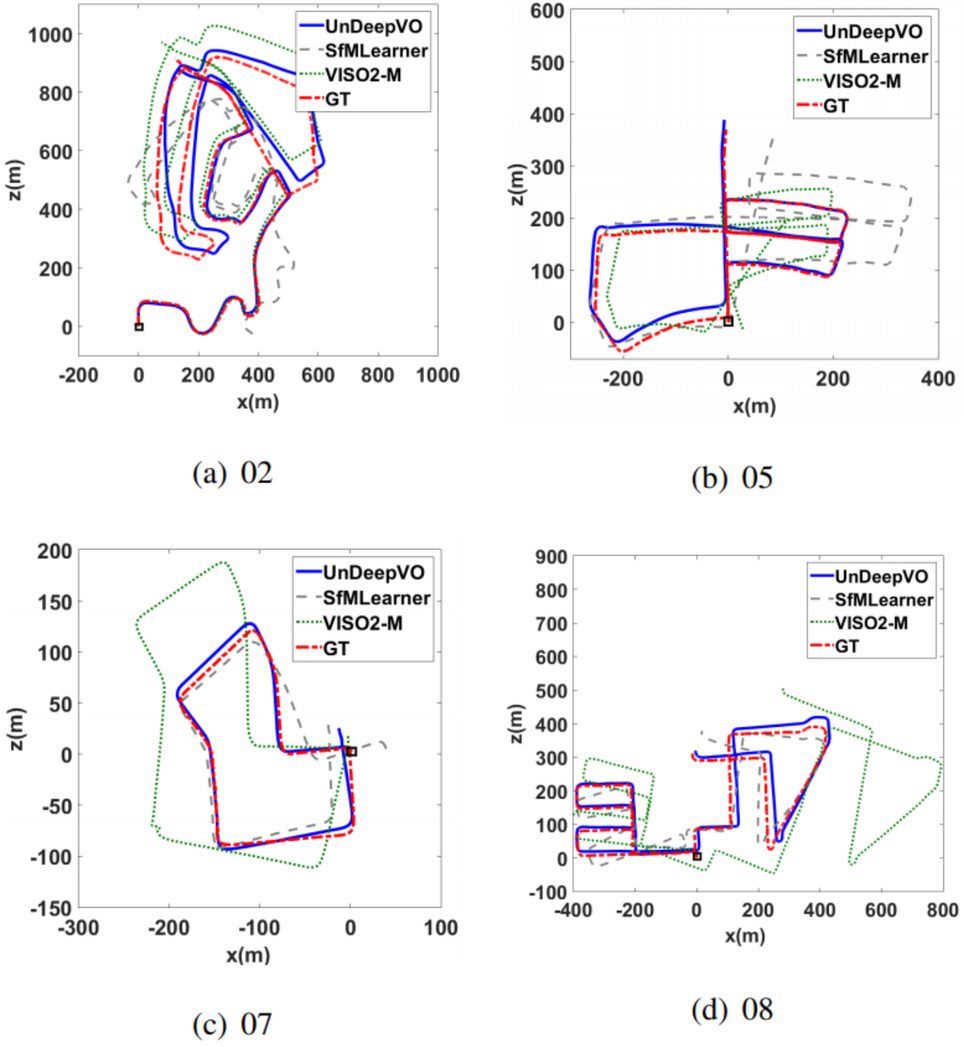
Table 2: Visual odometry results on KITTI odometry dataset [15]. We report the performance of ORB-SLAM [12] as a reference and compare with recent deep methods. K denotes the model trained on KITTI, and CS+K denotes the model with pre-training on Cityscapes [30].

Methods	Seq. 09		Seq. 10	
	t_{err} (%)	r_{err} (°/100m)	t_{err} (%)	r_{err} (°/100m)
ORB-SLAM [12]	15.30	0.26	3.68	0.48
Zhou et al. [7]	17.84	6.78	37.91	17.78
Zhan et al. [17]	11.93	3.91	12.45	**3.46**
Ours (K)	11.2	3.35	**10.1**	4.96
Ours (CS+K)	**8.24**	**2.19**	10.7	4.58



(a) sequence 09 (b) sequence 10

Figure 3: Qualitative results on the testing sequences of KITTI odometry dataset [15].



(a) 02 (b) 05

(c) 07 (d) 08