Generalizing to the Open World: Deep Visual Odometry with Online Adaptation

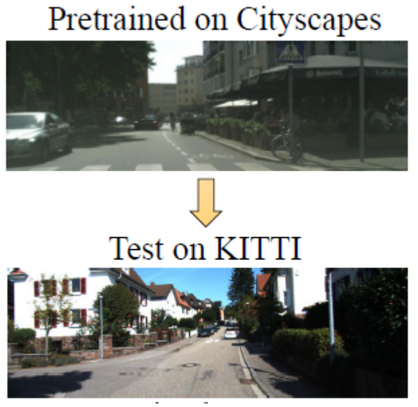
Shunkai Li Xin Wu Yingdian Cao Hongbin Zha
Key Laboratory of Machine Perception (MOE), School of EECS, Peking University
PKU-SenseTime Machine Vision Joint Lab

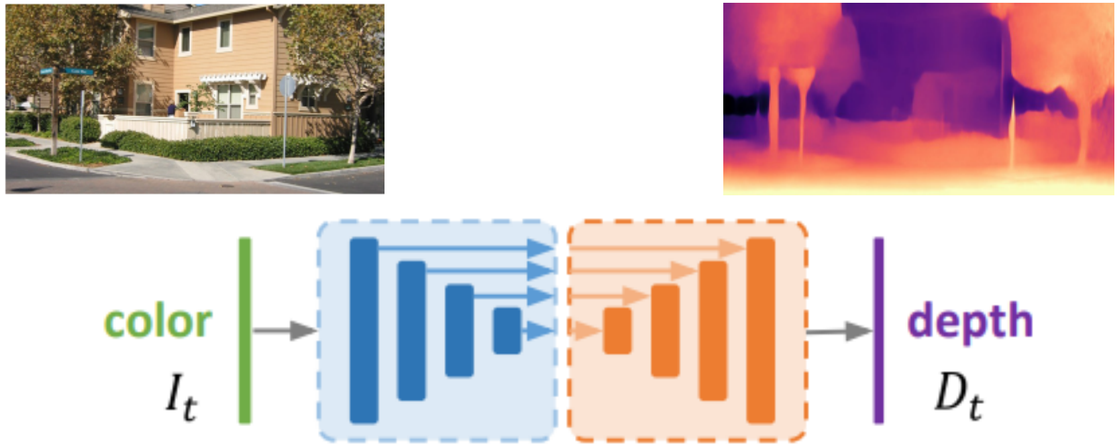
zhuofehuang

2021/06/10

Past Depth-Net structures

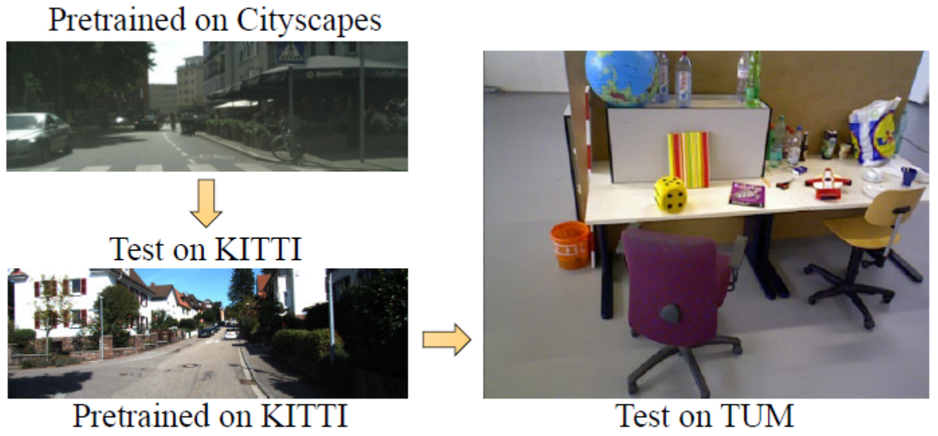
- U-Net like encoder-decoder structure
- Train and test on the same environment(eg. outdoors like KITTI, cityscapes)
- Off-line one-time inference



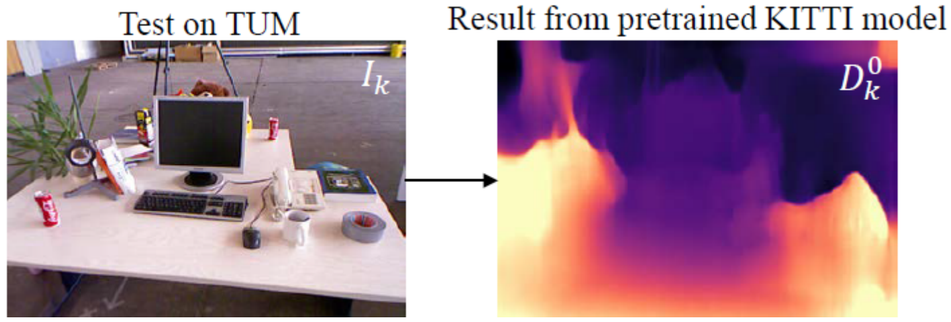


Collapse in unseen environment

- Train a model on outdoor data
- Test on indoor data



- How to solve generalization?



Bad result

Contributions

- We replace Pose-CNN with FlowNet, which provides more robust pose estimation.

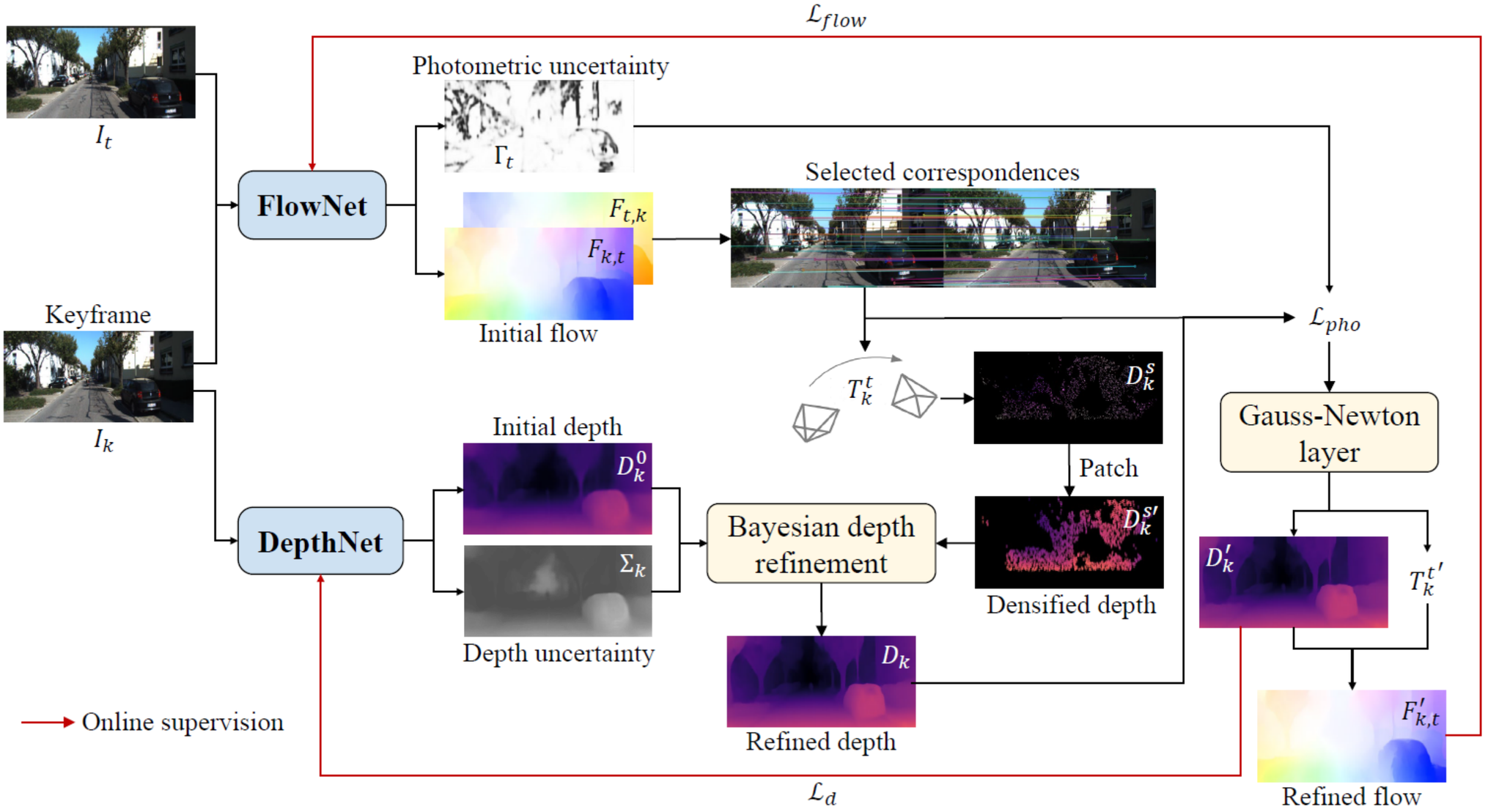
- The predicted depth is continuously refined by a Bayesian fusion framework, which is further used to train depth and optical flow during online learning.

- We introduce online learned depth and photometric uncertainties for better depth refinement and differentiable Gauss-Newton optimization.

Outline

- Depth modeling
- Online depth refinement
- Photometric residuals with learned uncertainty
- Differentiable Gauss-Newton optimization

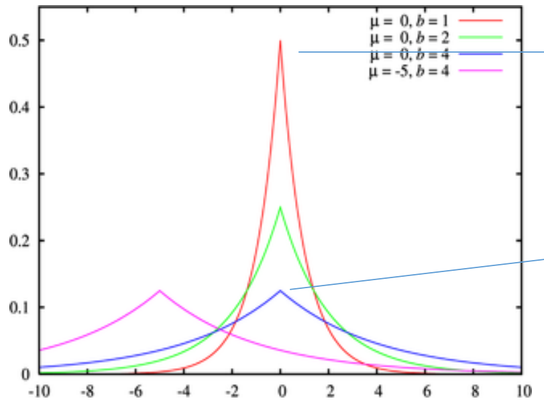
Framework


$\mathcal{L}_{flow}$
I_t
Photometric uncertainty
Γ_t
FlowNet
$F_{t,k}$
$F_{k,t}$
Initial flow
Selected correspondences
$\mathcal{L}_{pho}$
Keyframe
I_k
T_k^t
D_k^s
Gauss-Newton layer
Initial depth
D_k^0
Patch
DepthNet
Σ_k
Bayesian depth refinement
$D_k^{s'}$
Densified depth
D_k'
$T_k^{t'}$
Depth uncertainty
D_k
Refined depth
$F_{k,t}'$
Refined flow
Online supervision
$\mathcal{L}_d$

Review uncertainty

- Key idea: predict a posterior probability distribution for each pixel parameterized with its mean μ as well as its variance σ: $\mathrm{p}(d_{gt}|d,\sigma)$

$$p(d_{gt}|d,\sigma) = \frac{1}{2\sigma}\exp\left(-\frac{|d_{gt}-d|}{\sigma}\right)$$



lower σ, lower uncertainty

higher σ, higher uncertainty

σ : uncertainty

Negative log-likelihood (NLL loss):

$$-\log p(d_{gt}|d,\sigma) = \frac{|d_{gt}-d|}{\sigma} + \log(\sigma) + const$$

Depth modeling (Train)

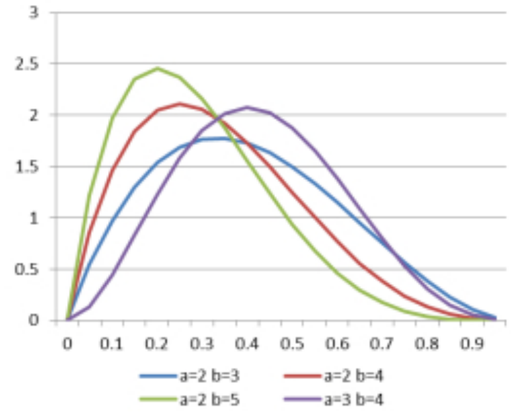
- Consider inverse depth $z_i = \frac{1}{d_i}$

- Model the good measurement as Gaussian distribution around the ground truth z_i, while the bad one is regarded as observation noise which is uniformly distributed

$$p(z_i^t|z_i, \rho_i^t) := \rho_i^t \mathcal{N}(z_i^t|z_i, \tau_i^2) + (1 - \rho_i^t)\mathcal{U}(z_i^t|z_i^{\min}, z_i^{\max})$$

- where the probability of being an inlier ρ_i^t is in Beta distribution:

$$Beta(\rho_i^t|a_i^t, b_i^t)$$



Depth modeling (Inference)

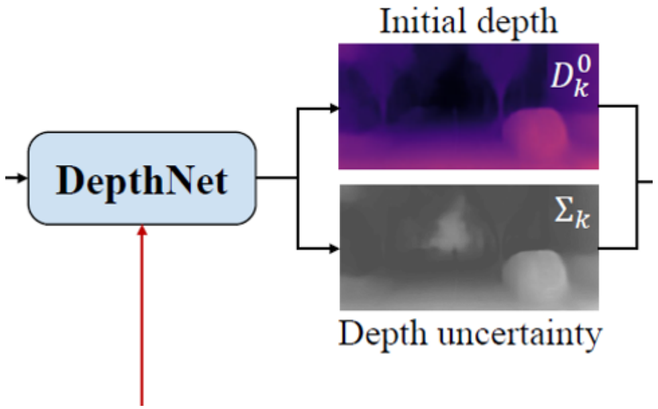
- While inference, we find MAP at each timestamp, and apply Gaussian and Beta distribution for approximation：

$$q(z_i^t, \rho_i^t | a_i^t, b_i^t, \mu_i^t, \sigma_i^{t2}) := Beta(\rho_i^t | a_i^t, b_i^t) \mathcal{N}(z_i^t | \mu_i^t, \sigma_i^{t2})$$

- Initialization：

$$\mu_i^0 = \frac{1}{d_k^0}, \quad \sigma_i^0 \in \Sigma_k, \quad z_i^{\max} = \mu_i^0 + \sigma_i^0,$$

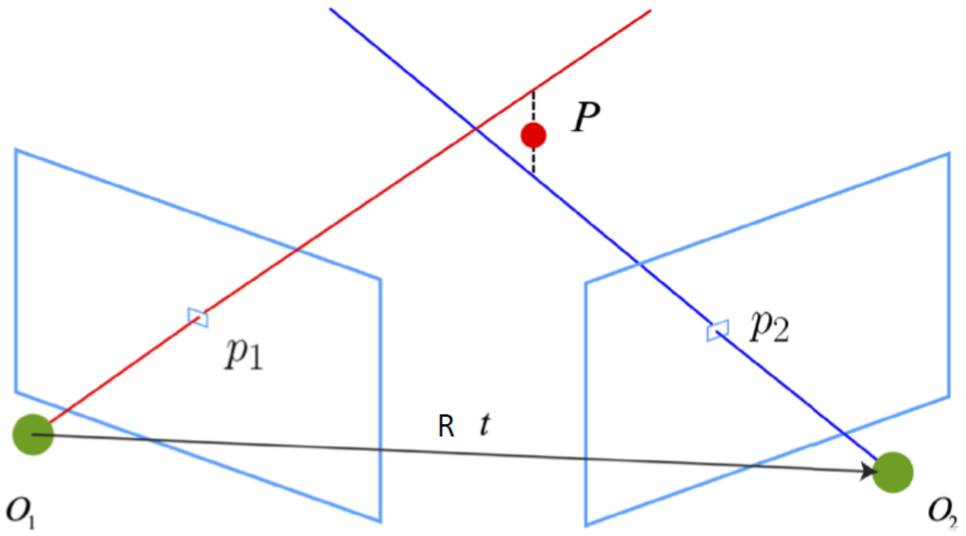
$$z_i^{\min} = \begin{cases} \mu_i^0 - \sigma_i^0, & \text{if } \mu_i^0 - \sigma_i^0 > 0 \\ 1e^{-6}, & \text{else} \end{cases}$$

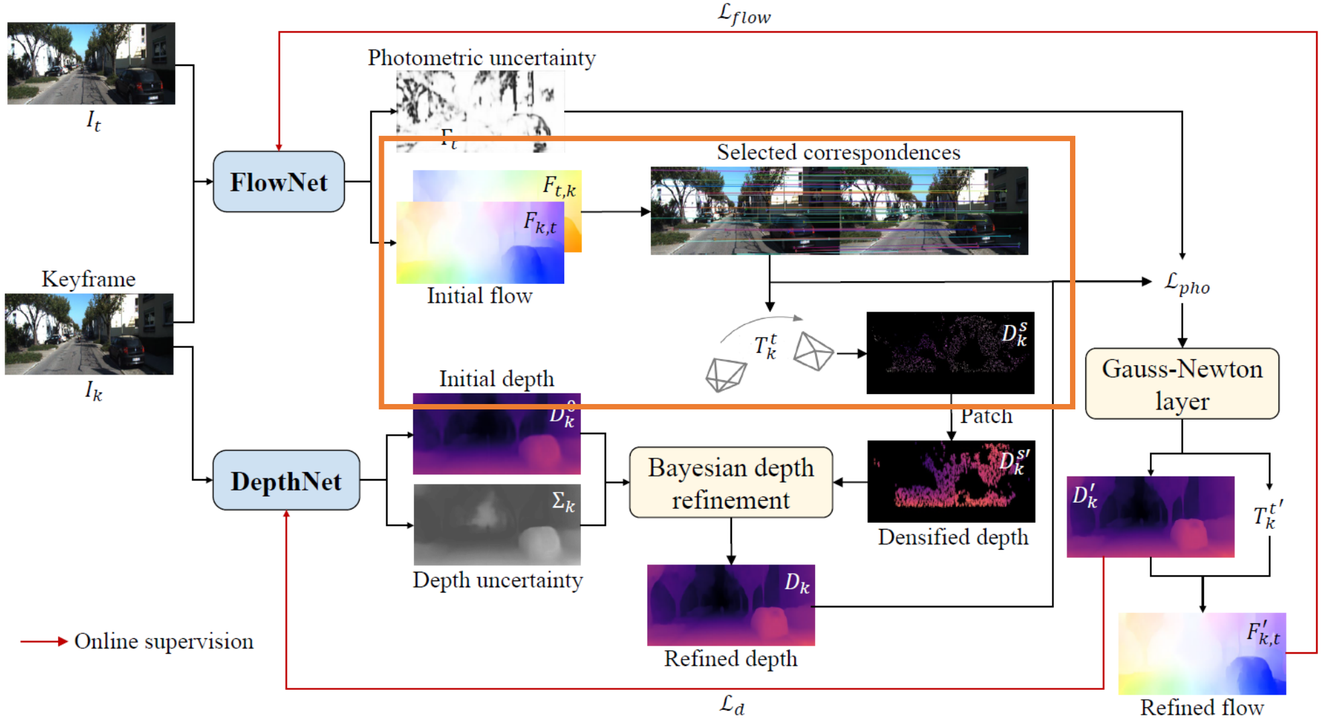


Outline

- Depth modeling
- Online depth refinement
- Photometric residuals with learned uncertainty
- Differentiable Gauss-Newton optimization

Midpoint triangulation





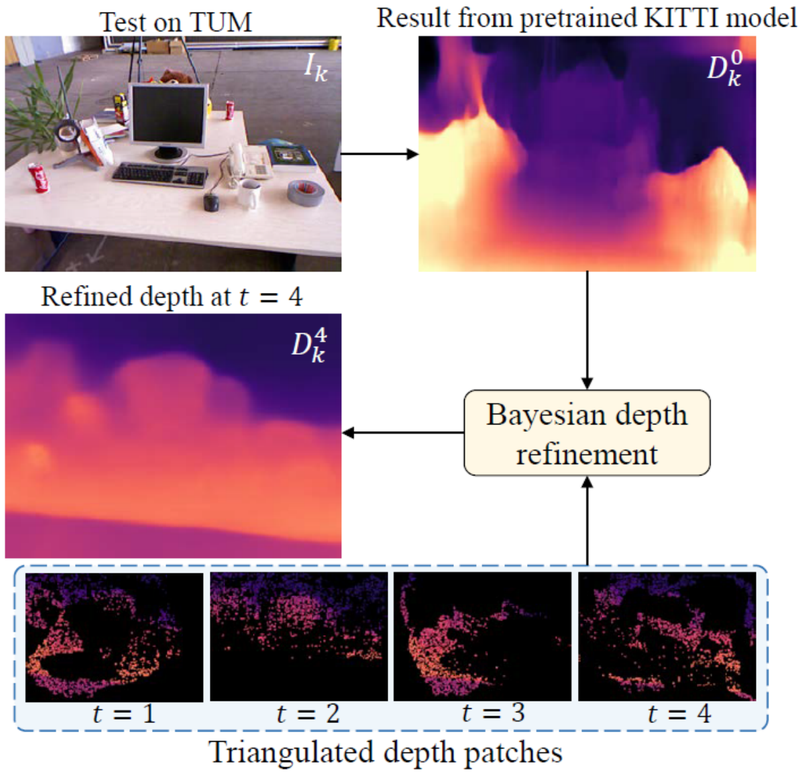
$$d_i^k = \arg\min_{d_i^k} [\mathrm{dis}(L_k, d_i^k)^2 + \mathrm{dis}(L_t, d_i^k)]^2,$$

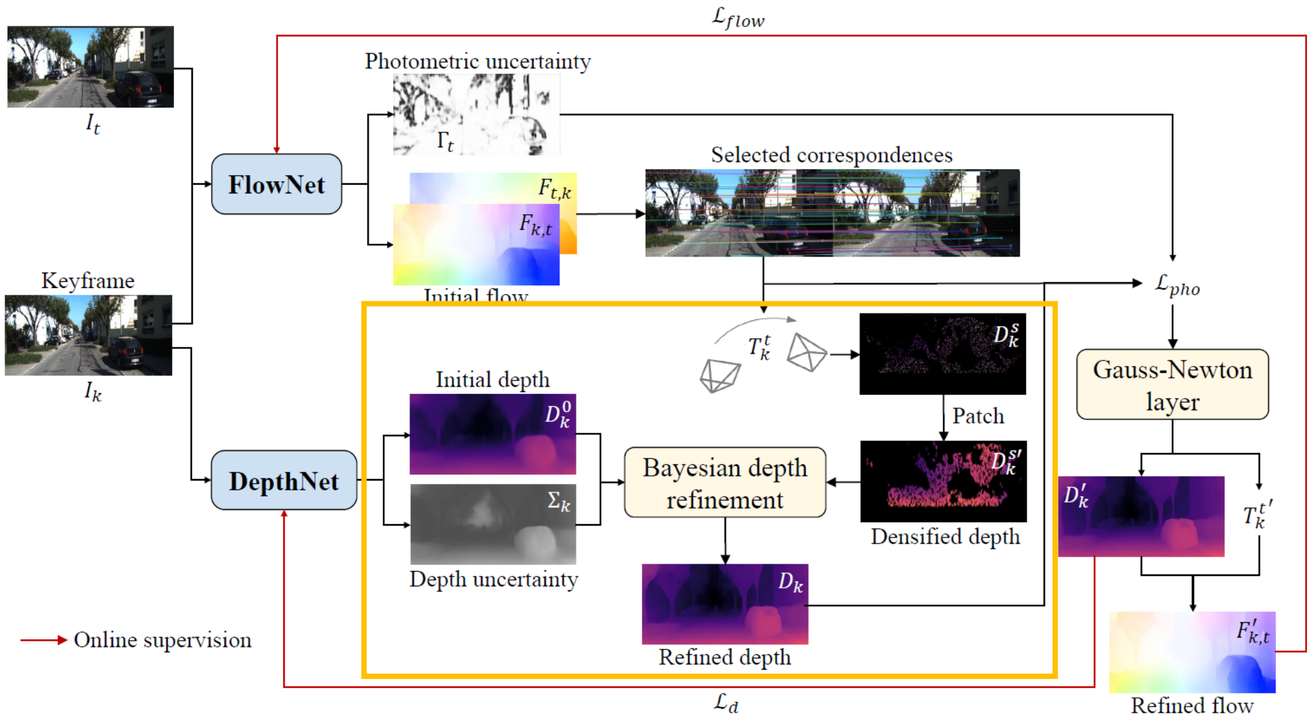
differentiable

Online adaptation

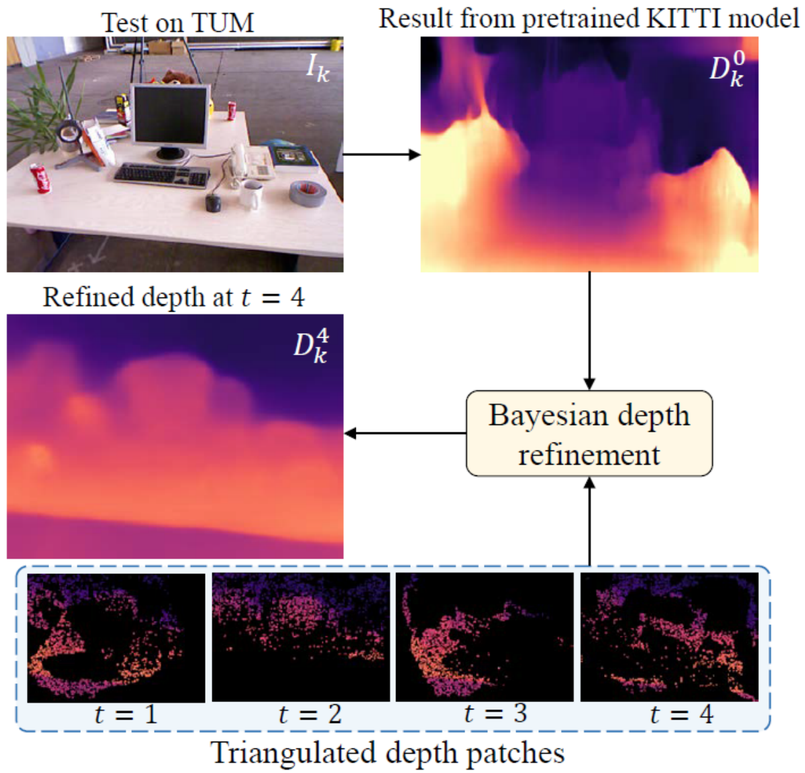
Sparse triangulated depth map D_k^s: ~2000 points

Densify each point with 3×3 window: $D_k^{s'}$





Online adaptation



$D_k^{s'}$: used to update prior inverse depth estimate z_i^t

$$q(z_i^t, \rho_i^t | a_i^t, b_i^t, \mu_i^t, {\sigma_i^t}^2) := Beta(\rho_i^t | a_i^t, b_i^t) \mathcal{N}(z_i^t | \mu_i^t, {\sigma_i^t}^2)$$

Meanwhile, a_i^t, b_i^t, μ_i^t, ${\sigma_i^t}^2$ are incrementally updated by Bayesian formulation.

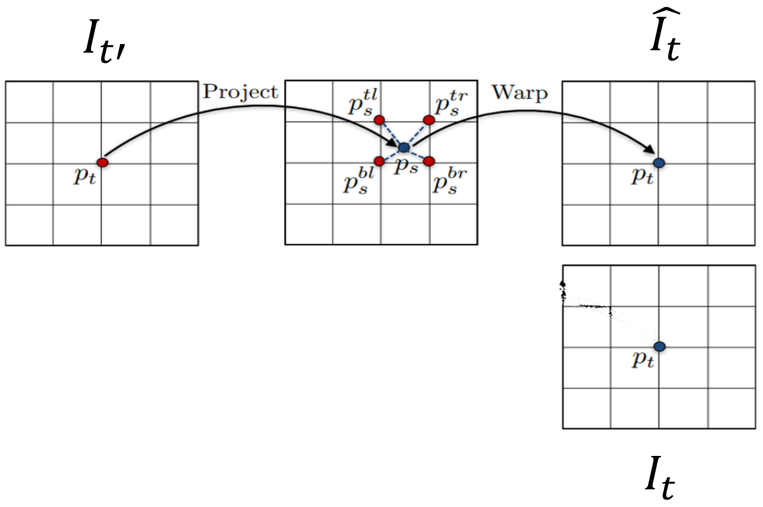
$$p(B_i|A) = \frac{p(B_i)p(A|B_i)}{P(A)} = \frac{p(B_i)p(A|B_i)}{\sum_{i=1}^{n} p(B_i)p(A|B_i)}$$

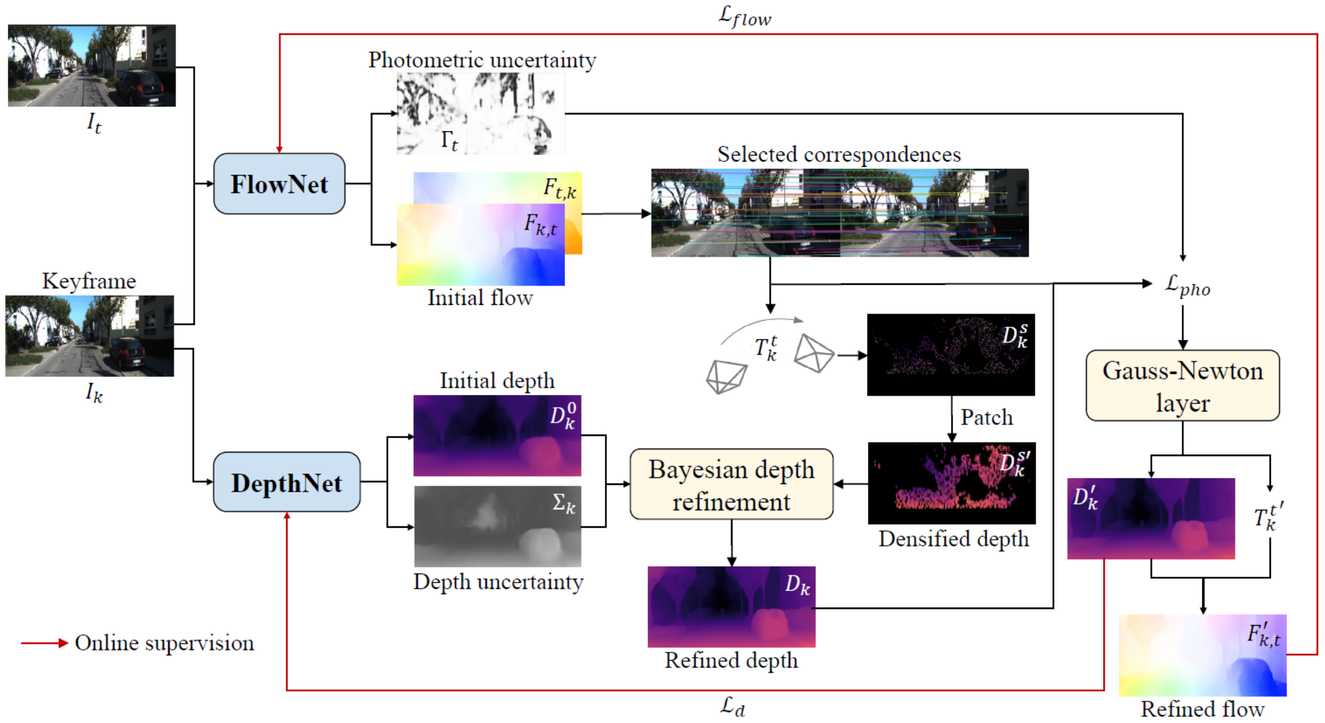
Help generalization: during adaptation, Depth-Net online learns prior knowledge of new scene

Outline

- Depth modeling
- Online depth refinement
- Photometric residuals with learned uncertainty
- Differentiable Gauss-Newton optimization

Photometric loss with uncertainty





Assuming observation noise is Laplacian:

$$\mathcal{L}_{pho} = \sum -\log p(I|\mu_I, \gamma) = \frac{\|\hat{I}_t - I_t\|_1}{\Gamma_t} + \log\Gamma_t$$

Outline

- Depth modeling
- Online depth refinement
- Photometric residuals with learned uncertainty
- Differentiable Gauss-Newton optimization

Gaussian-Newton Optimization

First order Taylor Expansion: $f(x+\Delta x) \approx f(x) + J(x)\Delta \mathrm{x}$

Find Δx minimizing $\|f(x+\Delta x)\|^2$: $\Delta x^* = argmin_{\Delta x} \frac{1}{2}\|f(x) + J(x)\Delta \mathrm{x}\|^2$

Expansion: $\frac{1}{2}\|f(x) + J(x)\Delta \mathrm{x}\|^2 = \frac{1}{2}(f(x) + J(x)\Delta \mathrm{x})^{\mathrm{T}}(f(x) + J(x)\Delta \mathrm{x})$

$$= \frac{1}{2}(\|f(x)\|^2 + 2f(x)^T J(x)\Delta \mathrm{x} + \Delta \mathrm{x}^T J(x)^T J(x)\Delta \mathrm{x})$$

Calculate $\frac{\mathrm{d}\, f^2}{\mathrm{d}\, \Delta x} = 0$: $2J(x)^{\mathrm{T}} f(x) + 2J(x)^{\mathrm{T}} J(x)\Delta \mathrm{x} = 0$

$$J(x)^{\mathrm{T}} J(x)\Delta \mathrm{x} = -J(x)^{\mathrm{T}} f(x)$$

GN layer

Starting with an initial depth and pose: D_k, T_k^t

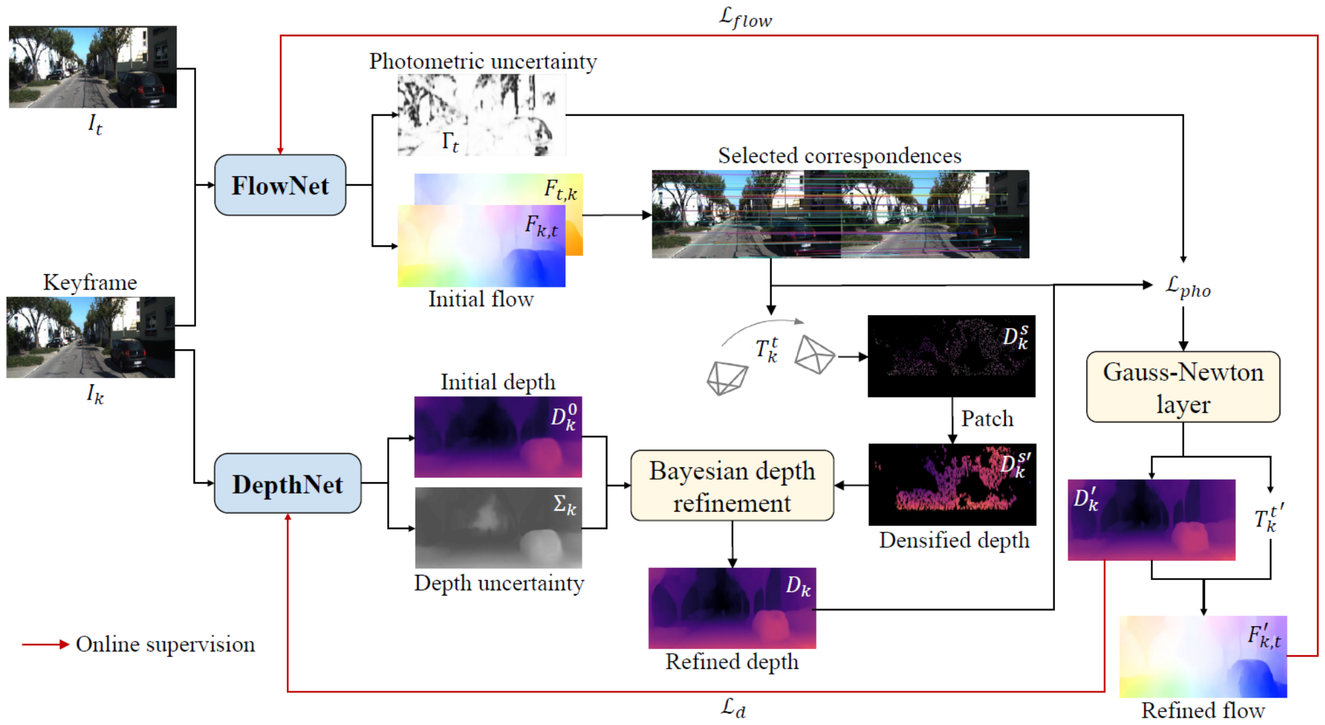
Compute the weighted photometric loss $r_i(p)$:

$$r_i(p) = \frac{\widehat{I_i}(p_i) - I_i(p)}{\gamma_i}, \gamma \in \Gamma_t$$

First order derivatives:

$$J_i^D(p) = \frac{1}{\gamma_i} \frac{\partial \hat{I}_i(p_i)}{\partial p_i} \frac{\partial p_i}{\partial D_k(p)}, \quad J_i^T(p) = \frac{1}{\gamma_i} \frac{\partial \hat{I}_i(p_i)}{\partial p_i} \frac{\partial p_i}{\partial T_k^t}$$

$$\delta = -(J^T J)^{-1} J^T r, \quad J = [J^D \; J^T]$$



Differentiable, designed as a layer

Converge within 3 iterations

Ablation study

Method	Supervision	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
SfMLearner [47]	-	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Garg *et al.* [12]	stereo	0.169	1.080	5.104	0.273	0.740	0.904	0.962
Vid2Depth [24]	-	0.163	1.240	6.220	0.250	0.762	0.916	0.968
GeoNet [39]	-	0.155	1.296	5.857	0.233	0.793	0.931	0.973
Zhan *et al.* [41]	stereo	0.135	1.132	5.585	0.229	0.820	0.933	0.971
Mahjourian *et al.* [25]	-	0.163	1.240	6.220	0.250	0.762	0.916	0.968
SAVO [22]	-	0.150	1.127	5.564	0.229	0.823	0.936	0.974
SC-SfMLearner [1]	-	0.137	1.089	5.439	0.217	0.830	0.942	0.975
Zhao *et al.* [43]	-	0.113	0.704	4.581	**0.184**	0.871	0.961	0.984
Monodepth2 [15] (w/o pretrain)	-	0.132	1.044	5.142	0.210	0.845	0.948	0.977
Monodepth2 (ImageNet pretrain)	-	0.115	0.882	4.701	0.190	0.879	0.961	0.982
Ranjan *et al.* [28]	-	0.148	1.149	5.464	0.226	0.815	0.935	0.973
Ours (w/o RDS)	-	0.136	1.087	5.118	0.210	0.843	0.952	0.980
Ours (w/o PU)	-	0.115	0.799	4.282	0.253	0.882	0.965	0.981
Ours	-	**0.106**	**0.701**	**4.129**	0.210	**0.889**	**0.967**	**0.984**

Table 3. Depth estimation results on KITTI dataset by Eigen *et al.* [6] split. The results are capped at 80 meters. As for error metrics Abs Rel, Seq Rel, RMSE and RMSE log, lower value is better; as for accuracy metrics δ, higher value is better. w/o RDS: without refined depth for online supervision. w/o PU: without online learned photometric uncertainty.

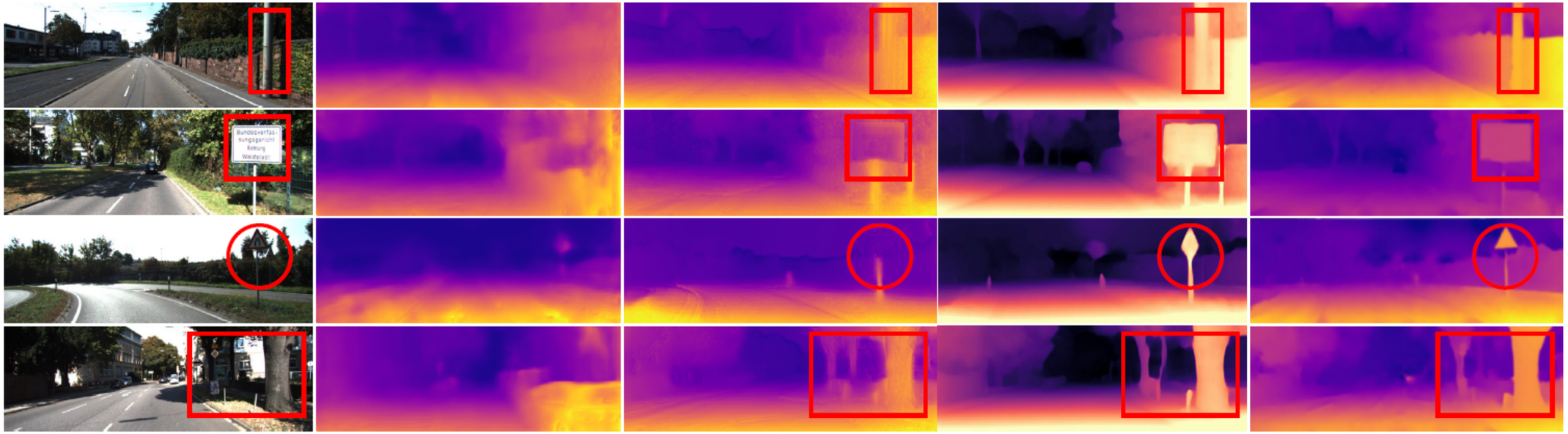


Figure 6. Depth estimation results on KITTI dataset. Thanks to our triangulation process and multi-frame depth refinement, our method shows better predictions and preserves sharp edges while other methods tend to predict vague depth. Best viewed in color.

Generalization (Test on VO)

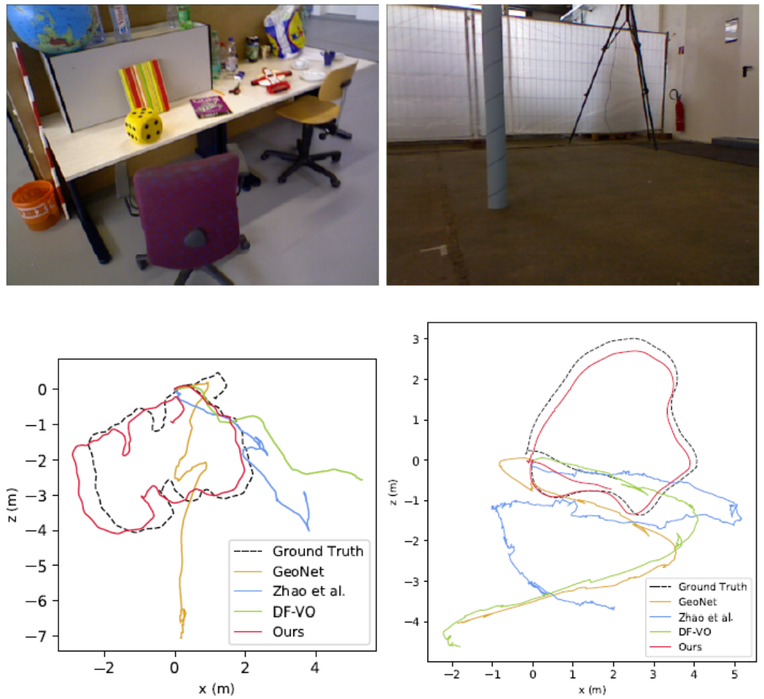
- Train on KITTI, test on RGB-D TUM

Sequence	Vid2Depth [24]	GeoNet [39]	Zhan *et al.* [41]	SAVO [22]	Li *et al.* [21]	DF-VO [42]	Zhao *et al.* [43]	DSO [7]	ORB-SLAM2 (LC) [26]	Ours	Ours (w/o RDS)	Ours (w/o PU)
fr2/desk	0.698	0.462	0.570	0.402	0.214	0.306	0.485	X	X	**0.158**	0.572	0.221
fr2/pioneer_360	0.581	0.662	0.453	0.402	0.218	0.599	0.693	X	X	**0.201**	0.638	0.254
fr2/pioneer_slam	0.367	0.301	0.309	0.338	0.190	0.585	0.354	0.737	X	**0.176**	0.481	0.210
fr2/360_kidnap	0.564	0.579	0.430	0.421	0.357	0.745	0.468	X	0.582	0.384	0.605	**0.371**
fr3/cabinet	0.492	0.282	0.316	0.281	0.272	0.447	0.227	X	X	**0.213**	0.453	0.276
fr3/long_office_hou_valid	0.401	0.316	0.327	0.297	0.237	0.227	0.534	0.327	**0.042**	0.133	0.529	0.168
fr3/nostr_texture_near_loop	0.328	0.277	0.340	0.440	0.255	0.564	0.348	0.093	**0.057**	0.159	0.401	0.186
fr3/str_notexture_far	0.227	0.258	0.235	0.216	0.177	0.505	0.175	0.543	X	**0.104**	0.432	0.201
fr3/str_notexture_near	0.235	0.198	0.217	0.204	0.128	0.603	0.218	0.481	X	**0.207**	0.579	0.224

Table 2. Quantitative evaluation of different methods pretrained on KITTI and tested on TUM-RGBD dataset. We evaluate relative pose error (RPE) which is presented as translational RMSE in [m/s]. LC: loop closure, X: fail. w/o RDS: without refined depth for online supervision. w/o PU: without online learned photometric uncertainty.

Generalization (Test on VO)

- Train on KITTI, test on RGB-D TUM



(c) fr3/long_office_hou_valid (d) fr2/pioneer_360

Figure 5. Visual odometry results pretrained on outdoor KITTI and tested on indoor TUM dataset. All the other learning-based baselines tend to fail when faced with large domain shift. In contrast, our method is still able to recover accurate VO estimation by online adapting to challenging indoor datasets.

Thanks

- Q & A