BaseBoostDepth: Exploiting Larger Baselines for Self-Supervised Monocular Depth Estimation

Kieran Saunders, Luis J. Manso, George Vogiatzis

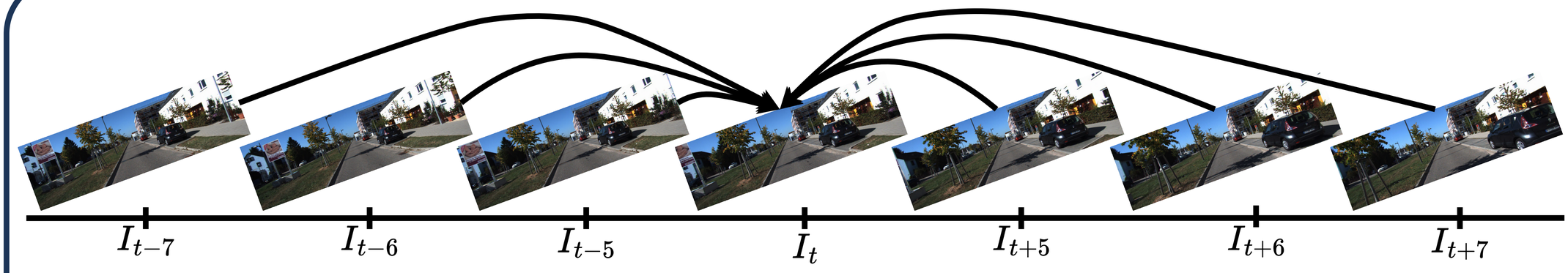
Aston University BIRMINGHAM UK

Loughborough University

AURRIGO

BMVC 2024

Motivation & Overview



In geometric approaches, increasing baseline separation improves depth accuracy, but **most self-supervised models use small separations**. However, using larger separations typically can reduce depth quality due to brightness changes and occlusions. Our method employs curriculum learning to introduce larger baselines progressively, while also incorporating incremental pose estimation to address pose drift, thereby improving accuracy. Additionally, error-induced reconstructions further enhance robustness.

Preliminaries

Image Reconstructions Using Depth And Pose

$$I_{t'\to t} = I_{t'}\langle Proj(D_t, P_{t\to t'}, K)\rangle$$

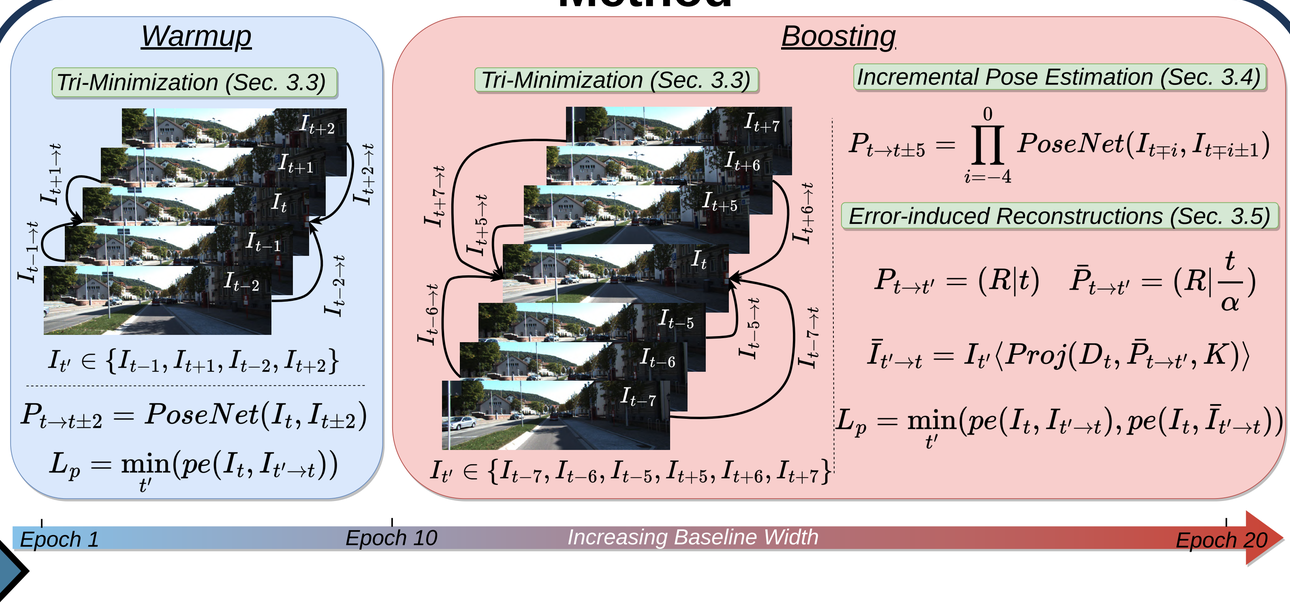
Photometric Loss

$$pe(I_a, I_b) = \frac{0.85}{2}\Big(1 - SSIM(I_a, I_b)\Big) + (1 - 0.85)||I_a - I_b||$$

Minimize Between Reconstructions

$$L_p = \min_{t'}(pe(I_t, I_{t'\to t}))$$

Method



Our method uses a curriculum-learning-inspired optimization strategy that splits training into two stages: warm-up and boosting, where the boosting stage takes full advantage of larger baselines.

- We gradually increase the baseline width during training.
- To reduce the impact of brightness changes and occlusions from larger baselines, we reconstruct the target image using triples of future and past frames, referred to as tri-minimization.
- To address significant drift in pose estimation over larger baselines, we break down the pose estimation process into smaller increments within larger intervals.
- Additionally, we introduce controlled pose errors to add noise to the network.

Quantitative Results

We conduct an extensive ablation study, progressively adding each contribution and comparing the results to the baseline Monodepth2. We also show the effect of using larger frame separations without our curriculum learning scheme.

Ablation	Contributions						KITTI							SYNS	
	Skip	Pre	Tri.	Incri. Pose	Part. Incri.	Err. Rec.	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Acc	Comp
Monodepth2 [9]	1						0.106	0.818	4.750	0.196	0.874	0.957	0.979	2.516	17.193
Monodepth2 [9]	4						0.107	0.832	4.723	0.186	0.887	0.961	0.982	2.512	14.856
Monodepth2 [9]	[4]						0.146	1.164	5.289	0.221	0.813	0.940	0.975	2.465	5.278
BaseBoostDepth	*C*	✗					0.115	0.916	4.856	0.190	0.877	0.960	**0.983**	2.442	5.518
BaseBoostDepth	*C*	✗	✓				0.112	0.867	4.762	0.187	0.879	0.962	**0.983**	**2.417**	**3.433**
BaseBoostDepth	*C*	✗	✓	✓			0.109	0.868	4.767	0.186	0.883	0.961	0.982	2.489	6.547
BaseBoostDepth	*C*	✗	✓	✓	✓		0.107	0.799	4.656	0.184	0.884	**0.963**	**0.983**	2.450	4.290
BaseBoostDepth	*C*	✗	✓	✓	✓	✓	0.106	0.736	4.584	0.184	0.883	**0.963**	**0.983**	2.453	3.810
BaseBoostDepth$_{pre}$	*C*	✓	✓	✓	✓	✓	**0.104**	**0.738**	**4.544**	**0.183**	**0.888**	**0.963**	**0.983**	2.432	4.763

We show state-of-the-art performance on the KITTI dataset using MonoViT's backbone at a resolution of 640×192.

Method	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Monodepth2 [9]	0.106	0.818	4.750	0.196	0.874	0.957	0.979
CADepth [32]	0.102	0.752	4.504	0.181	0.894	0.964	0.983
DIFFNet [36]	0.101	0.749	4.445	0.179	0.898	0.965	0.983
MonoViT [34]	0.098	0.683	4.333	0.174	0.904	0.967	0.984
BaseBoostDepth	0.106	0.736	4.584	0.184	0.883	0.963	0.983
BaseBoostDepth$_{pre}$	0.104	0.738	4.544	0.183	0.888	0.963	0.983
BaseBoostDepth$^{\dagger}_{pre}$	**0.096**	**0.648**	**4.201**	**0.170**	**0.906**	**0.968**	**0.985**

Notably, we achieve state-of-the-art performance on the SYNS dataset across image-based, edge-based, and point cloud-based metrics, showcasing our accuracy in 3D space.

Method	Image-Based					Edge-Based		Point Cloud-Based	
	Abs Rel	MAE	Sq Rel	RMSE	RMSE log	Acc	Comp	F-Score	IoU
Monodepth2 [9]	0.334	6.901	5.285	12.089	0.405	2.516	17.193	0.242	0.149
CADepth [32]	0.363	8.787	5.548	13.512	0.546	2.473	19.045	0.022	0.012
DIFFNet [36]	0.311	6.554	4.690	11.610	0.383	2.411	12.116	0.258	0.161
MonoViT [34]	0.287	6.195	4.399	11.124	0.354	2.443	15.672	0.264	0.164
BaseBoostDepth	0.334	6.878	4.854	11.847	0.409	2.453	**3.810**	0.275	0.174
BaseBoostDepth$_{pre}$	0.328	6.752	4.815	11.752	0.405	2.432	4.763	0.268	0.168
BaseBoostDepth$^{\dagger}_{pre}$	**0.278**	**5.951**	**3.795**	**10.575**	**0.351**	**2.409**	5.314	**0.300**	**0.191**

Qualitative Results

We compare BaseBoostDepth to Monodepth2 (MD2) and observe substantial improvements in edge definition. Notably, this improvement is most pronounced around high-contrast bright pixels.