Movement-induced Priors for Deep Stereo

Yuxin Hou[1] Muhammad Kamran Janjua[2] Juho Kannala[1] Arno Solin[1]

[1]Aalto University, Finland [2]National University of Sciences and Technology, Pakistan

Aalto University

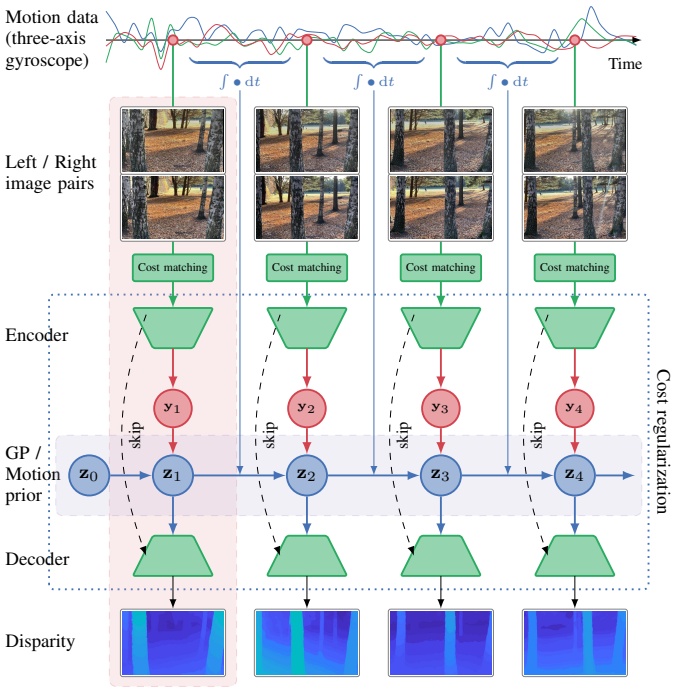


Figure 1. We present a framework for deep stereo inferences with movement-induced priors. The GP inference (blue block) couples the latent-space encodings based on the movement-induced prior that carries over information between stereo pairs. Our proposed prior will not affect computation of cost volume and only incorporate with latent codes of the encoder–decoder part.

Motivation

We introduce movement-induced priors for deep stereo vision by framing the problem as a Gaussian process inference task. Central principles are:

- Solving disparity estimation for image-pair **sequences**
- Fuse information between the latent representations
- The latent representations of pairs with similar scenes should be more correlated

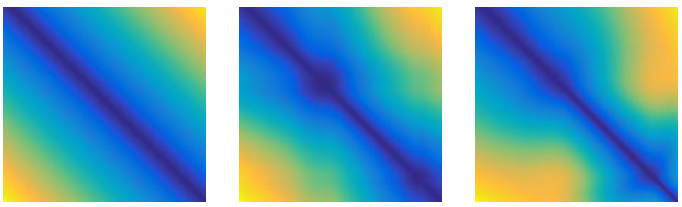
Hierarchy of GP kernels

To inject the prior, we consider three different covariance functions depending on the availability of movement information:

(a) a full pose kernel when full rotations and translations are known

(b) a gyroscope kernel when angular rates of the relative orientation changes are known

(c) a time-decay kernel when movement is unknown



(a) $\mathbf{D}_\mathrm{t}$ (b) $\mathbf{D}_\mathrm{gyro}$ (c) $\mathbf{D}_\mathrm{pose}$

Figure 2. Examples of the different distance matrices. Our Markovian gyroscope distance captures much of the same as the full pose distance, but without access to the pose information.

Probabilistic Gaussian process inference

Though the way of aggregation features can vary a lot, most method use fully-convolutional **encoder–decoder** architectures to regularize cost volumes. We introduce a probabilistic prior to the latent space of encoder–decoders, modify the outputs $\mathbf{y}_i$ from the encoder by a Gaussian process regression model:

$$z_j(t) \sim \mathrm{GP}(0, \kappa(t,t')),$$
$$y_{j,i} = z_j(t_i) + \varepsilon_{j,i}, \quad \varepsilon_{j,i} \sim \mathrm{N}(0, \sigma^2), \tag{1}$$

where the covariance function $\kappa(\cdot,\cdot)$ encodes the movement-induced prior. The encoder output $\mathbf{y}_i$ can now be seen as a 'corrupted' version of the true (unknown) latent encodings $\mathbf{z}_i$.

Movement-induced priors

When we only have observations of angular velocity $\boldsymbol{\omega} = (\omega_\mathrm{x}, \omega_\mathrm{y}, \omega_\mathrm{z})$, we consider the rotational distance metric:

$$d_\mathrm{gyro}(t_i, t_j) = \sqrt{\mathrm{tr}(\mathbf{I}_3 - \textstyle\prod_{k=i+1}^{j} \exp(-[\boldsymbol{\omega}_k]_\times \, \Delta t_k))}, \tag{2}$$

where $\Delta t_k = t_k - t_{k-1}$. To leverage the distance in Markovian fashion, we define the **cumulative pose-to-pose distance** and the kernel:

$$s_i = \sum_{j=1}^{i} d_\mathrm{gyro}(t_{j-1}, t_j) \tag{3}$$

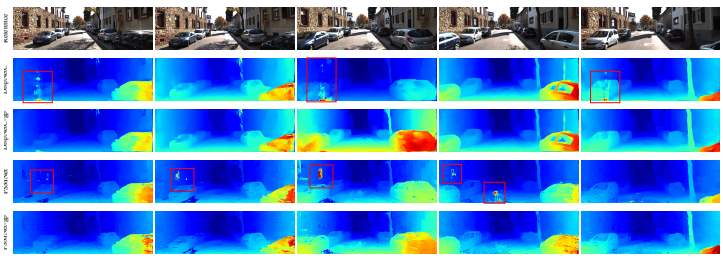
$$\kappa_\mathrm{gyro}(t_i, t_j) = \gamma^2 \left(1 + \frac{\sqrt{3}\,|s_i - s_j|}{\ell}\right) \exp\left(-\frac{\sqrt{3}\,|s_i - s_j|}{\ell}\right) \tag{4}$$

Experiments

We incorporate the proposed movement-induced prior and the GP inference with the two representative models, DispNetC and PSMNet.

Model	Training set			GP used during		KITTI		ZED	
	SceneFlow	KITTI-2015	KITTI Depth	Training	Testing	D1-all	Avg-all	SSIM	PSNR
DispNetC	✓					15.8510	2.3586	0.8446	32.2300
DispNetC-gp	✓				✓	14.5620	1.9730	0.8453	32.2356
PSMNet	✓					60.0773	6.7620	0.8002	31.5341
PSMNet-gp	✓				✓	60.9546	6.2786	0.7966	31.4693
DispNetC-ft	✓	✓				3.8739	0.9600	0.8592	32.3660
DispNetC-ft-gp	✓	✓			✓	3.1305	0.8584	0.8596	32.3726
PSMNet-ft	✓	✓				1.3878	1.1722	0.8536	32.2463
PSMNet-ft-gp	✓	✓			✓	1.3665	1.1698	0.8537	32.2468
DispNetC-ft-seq	✓		✓			1.4134	0.6949	0.8716	32.6382
DispNetC-ft-seq-gp	✓		✓	✓	✓	1.0939	0.6155	0.8797	32.8376
PSMNet-ft-seq	✓		✓			0.5391	0.5890	0.8827	33.0252
PSMNet-ft-seq-gp	✓		✓	✓	✓	0.5350	0.5883	0.8829	33.0280

Both the DispNetC and PSMNet pre-trained from synthetic data show artifacts, and our prior helps to alleviate them.



References

[1] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, pages 5410–5418, 2018.

[2] Yuxin Hou, Juho Kannala, and Arno Solin. Multi-view stereo by temporal nonparametric fusion. In *ICCV*, pages 2651–2660, 2019.

[3] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, pages 4040–4048, 2016.

yuxin.hou@aalto.fi