Supplementary material to "Rebalancing gradient to improve self-supervised co-training of depth, odometry and optical flow predictions"

Marwane Hariat[1], Antoine Manzanera[1], David Filliat [1, 2]

[1]U2IS, ENSTA Paris, Institut Polytechnique de Paris, Palaiseau, France
[2]INRIA FLOWERS
{marwane.hariat, antoine.manzanera, david.filliat}@ensta-paris.fr

1. Divergence tendency of θ_n (Sec. 3.1)

We prove in this section that, under certain reasonable conditions, the random variable Δ diverges almost surely, preventing the weakest networks $(\mathcal{D}_\theta, \mathcal{T}_\alpha)$ to learn anything.

Let us first define:

$$\begin{aligned} \Phi_D &= \Phi\Big(I_t, \hat{I}_s^{\theta,\alpha}\Big) \\ \Phi_F &= \Phi\Big(I_t, \hat{I}_s^{\theta,\alpha}\Big) \end{aligned} \tag{1}$$

Let us denote α_i, the contribution to the stochastic gradient descent (SGD) of training iteration i. This assumes that the sequences $\left(\Phi_D^i\right)_{i\in\mathbb{N}}$ and $\left(\Phi_F^i\right)_{i\in\mathbb{N}}$ satisfies:

$$\begin{aligned} \Phi_D^{i+1} &= \Phi_D^i - \lambda \mathbb{P}\left(\Delta_i < 0\right) \alpha_i \\ \Phi_F^{i+1} &= \Phi_F^i - \lambda \mathbb{P}\left(\Delta_i > 0\right) \alpha_i \end{aligned} \tag{2}$$

where λ quantifies a learning rate. Then:

$$\Delta_{i+1} = \Delta_i + \lambda\alpha_i\Big(\mathbb{P}\left(\Delta_i > 0\right) - \mathbb{P}\left(\Delta_i < 0\right)\Big) \tag{3}$$

$$= \Delta_i + \lambda\alpha_i\Big(2\mathbb{P}\left(\Delta_i > 0\right) - 1\Big) \tag{4}$$

Now let us define the sequence $(\varepsilon_i)_{i\in\mathbb{N}}$ as:

$$\mathbb{P}\left(\Delta_i > 0\right) = \frac{1}{2} + \varepsilon_i \tag{5}$$

Here, ε_i represents the intrinsic bias at iteration i (see Sec. 3.1). Thus,

$$\Delta_{i+1} = \Delta_i + 2\lambda\varepsilon_i\alpha_i \tag{6}$$

and then:

$$\Delta_n = \Delta_0 + 2\lambda\sum_{i=0}^{n}\alpha_i\varepsilon_i \tag{7}$$

The series $\sum \alpha_i\varepsilon_i$ has to be summable, which imposes constraints on the sequence of terms $(\varepsilon_i)_{i\in\mathbb{N}}$. To illustrate this point, let us take the following example:

$$\alpha_i = \frac{(-1)^\beta}{\sqrt{i+1}} \text{ and } \varepsilon_i = \frac{1}{\sqrt{i+1}} \tag{8}$$

With β a random Bernoulli variable of probability parameter $p < 0.5$, to reflect the upward improvement tendency. The choice of denominators reflects the fact that, since the model converges toward an optimum, the increments of the SGD decrease over iterations. Then,

$$\alpha_i\varepsilon_i = \frac{(-1)^\beta}{i+1}, \text{ and } \sum_{i=0}^{n}\alpha_i\varepsilon_i \xrightarrow[n\to\infty]{} +\infty \tag{9}$$

Then,

$$\mathbb{P}\left[0 < \Delta_n\right] \xrightarrow[n\to\infty]{} 1, \text{ and } \theta_n \xrightarrow[n\to\infty]{} +\infty \tag{10}$$

2. Proof of equation 14 (Sec. 4.2)

We show in this section the benefit of taking pixels with values in the neighbourhood $\mathcal{V}_\eta$ of the mode of Δ, compared to pixels with negative Δ.

Let us assume that:

- $\Delta \sim \mathcal{N}\left(\mu, \sigma^2\right)$
- $\exists r > 0,\ \mathbb{E}\left[\Phi_F | \Delta \in \mathcal{V}_\eta\right] < \mathbb{E}\left[\Phi_F | \Delta < 0\right] - r$

The second point illustrates mathematically the fact that moving pixels in the left tail correspond to failed flow predictions because of smoothing issues (see Figure 3). Then,

$$L_1 = \mathbb{E}\left[\Phi_D | \Delta \in \mathcal{V}_\eta\right] \tag{11}$$

$$= \mathbb{E}\left[\Delta | \Delta \in \mathcal{V}_\eta\right] + \mathbb{E}\left[\Phi_F | \Delta \in \mathcal{V}_\eta\right] \tag{12}$$

From our first hypothesis, and using an integral substitution, we have:

$$\mathbb{E}\left[\Delta|\Delta \in \mathcal{V}_\eta\right] = \frac{1}{\sqrt{2\pi}\sigma}\int_{\mu-\eta}^{\mu+\eta} x e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{13}$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\eta}^{\eta} (\sigma x + \mu)\, e^{-\frac{x^2}{2}} \tag{14}$$

$$= \frac{\mu}{\sqrt{2\pi}\sigma}\int_{-\eta}^{\eta} e^{-x^2/2} \tag{15}$$

If η is low enough, a linear approximation gives us:

$$\mathbb{E}\left[\Delta|\Delta \in \mathcal{V}_\eta\right] \approx \frac{2\mu}{\sqrt{2\pi}\sigma}\eta \tag{16}$$

Therefore, from our second hypothesis:

$$L_1 < \frac{2\mu}{\sqrt{2\pi}\sigma}\eta + \mathbb{E}\left[\Phi_F|\Delta < 0\right] - r \tag{17}$$

$$< \frac{2\mu}{\sqrt{2\pi}\sigma}\eta - r - \mathbb{E}\left[\Delta|\Delta < 0\right] + L_2 \tag{18}$$

Let us name:

$$\varepsilon = -\mathbb{E}\left[\Delta|\Delta < 0\right] \tag{19}$$

With an integral substitution, we have:

$$\varepsilon = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{0} -x e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{20}$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\int_{0}^{\infty} x e^{-\frac{(x+\mu)^2}{2\sigma^2}} \tag{21}$$

As,

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{0}^{\infty} x e^{-\frac{(x+\mu)^2}{2\sigma^2}} \xrightarrow[\mu\to\infty]{} 0 \tag{22}$$

Then, if μ is high enough, in other words, if Δ is enough shifted to the right (see Fig. 1), then:

$$0 < r - \varepsilon \tag{23}$$

We have,

$$L_1 < L_2 + \frac{2\mu}{\sqrt{2\pi}\sigma}\eta - (r - \varepsilon) \tag{24}$$

Additionally, if η is low enough, then:

$$\begin{gathered}\frac{2\mu}{\sqrt{2\pi}\sigma}\eta < r - \varepsilon \\ L_1 < L_2\end{gathered} \tag{25}$$

3. Epipolar constraint

Let M the fundamental matrix defined as:

$$\mathrm{M} = \mathrm{K}^{-\mathrm{T}}\left[\mathrm{t}_\alpha\right]_\times \mathrm{R}_\alpha \mathrm{K}^{-1} \tag{26}$$

Let (a, b) such that:

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} = \mathrm{M}p \tag{27}$$

Finally let us define:

$$p_\delta = p + F_\delta\left(p\right) \tag{28}$$

The epipolar loss that we used writes:

$$\mathcal{L}_{\text{epipolar}} = \frac{\mathrm{M}p \cdot p_\delta}{\sqrt{a^2 + b^2}} \tag{29}$$

4. More results

We give in Tab. 1 additional quantitative results of **CoopNet** trained on Cityscapes and evaluated on KITTI to assess the transfer learning capacity of our model. As well as results of **CoopNet** trained on both Cityscapes and KITTI and evaluated on KITTI. In these experiments we show substantial improvements over the benchmarks displayed. More qualitative results are also presented in Fig. 1 and Fig. 2.

Dataset	Method	Train	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
CS+K	*Zhou et al.*[42]	M	0.198	1.836	6.565	0.275	0.718	0.901	0.960
	GeoNet[40]	M	0.153	1.328	5.7370	0.232	0.802	0.934	0.972
	DF-Net[43]	M	0.146	1.182	5.213	0.2080	0.818	0.943	0.978
	Bian et al.[1]	M	0.128	1.047	5.240	0.2080	0.846	0.947	0.976
	CoopNet	M	**0.122**	**0.972**	**5.127**	**0.202**	**0.853**	**0.952**	**0.980**
CS	GeoNet[40]	M	0.210	1.723	6.595	0.281	0.681	0.891	0.960
	Struct2Depth[2]	M+S	0.153	1.109	5.557	0.227	0.796	0.934	0.975
	GLNet[3]	M	0.129	**1.044**	5.361	0.212	0.843	0.938	0.976
	CoopNet	M	**0.125**	1.157	**5.251**	**0.209**	**0.845**	**0.944**	**0.978**

Table 1: **Results of depth estimations**. For each metric the best result is displayed in bold. The depth cutoff is set to 80m. For red metrics, lower is better. For blue metrics, higher is better. A post-processing refinement as done by [2] is performed in these experiments. **Train: M** - Self-supervised methods. **S** - Use of an off-the-shelf semantic algorithms. **CS+K**: Trained on KITTI and Cityscapes combined and evaluated on KITTI. **CS**: Trained on Cityscapes and evaluated on KITTI.

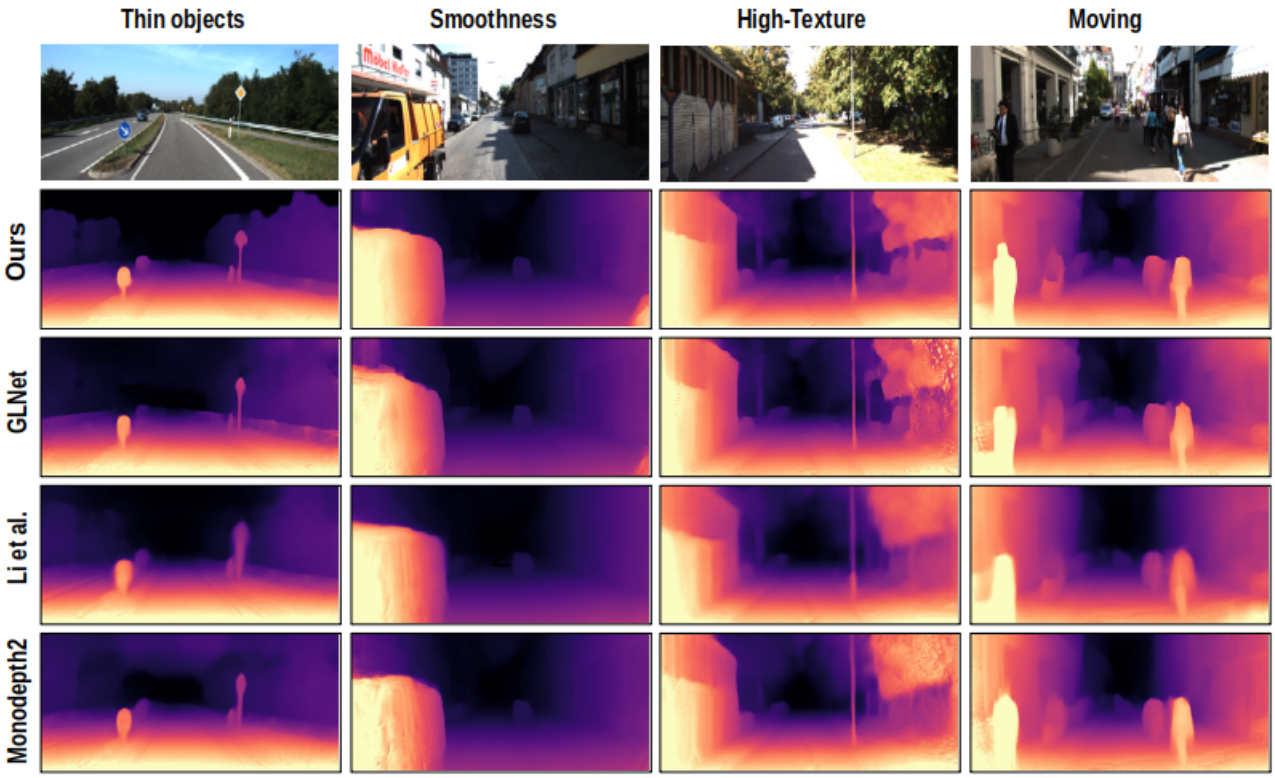

Figure 1: Comparison of depth map estimation algorithms in challenging situations on KITTI.

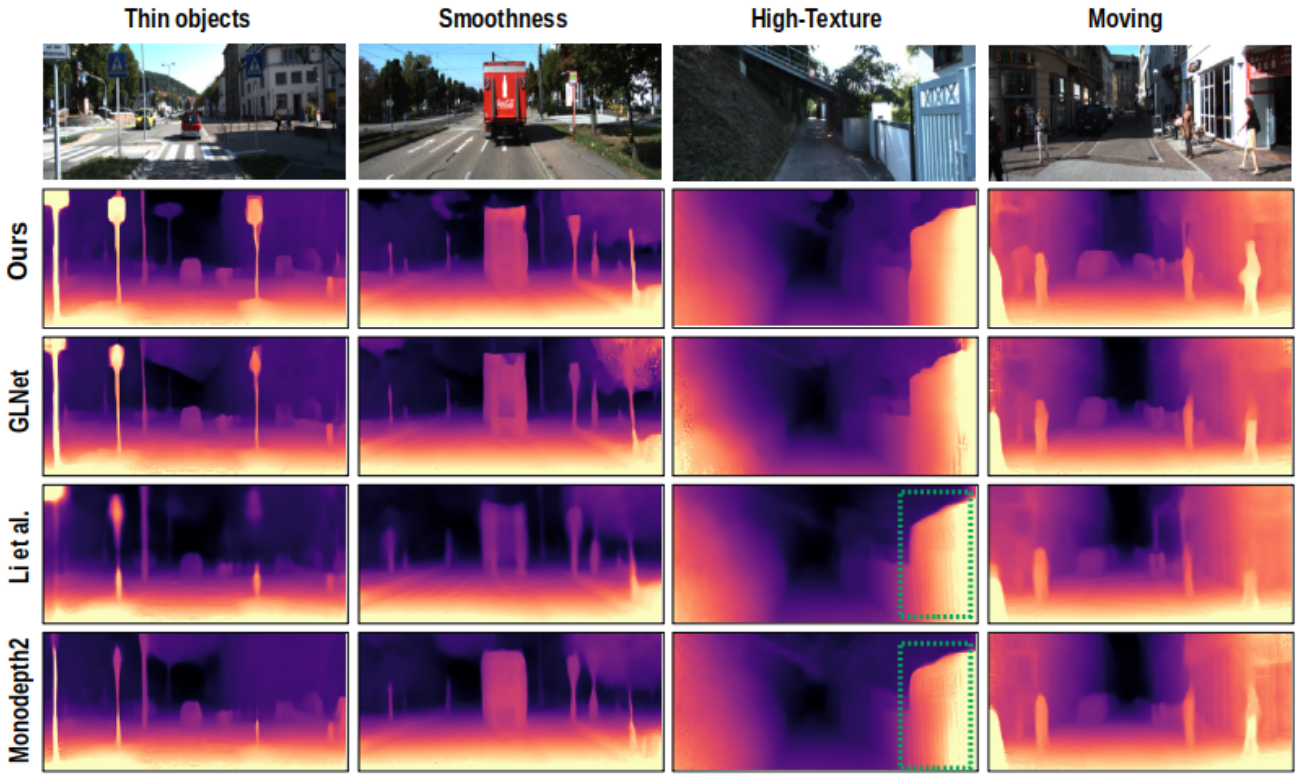

Figure 2: Comparison of depth map estimation algorithms in challenging situations on KITTI. **Note**: For the High-Texture case, both *Monodepth2* and *Li et al.* show copy-artefacts around the grid (see green rectangles).