Pseudo Lidar ++

- **higher 3D localisation error of stereo is only due to depth estimation inaccuracy**

"we observe that stereo methods do indeed *detect* objects with high reliability, yet they estimate the depth of the *entire* object as either too far or too close. See Figure 1 for an illustration"

=> propose to debias the model (based on ideas of Wang et al. 2019)
- main reason for depth error : not computed directly (but through disparity)
=> predict depth directly (SDN architecture)
- SDN still has error
=> match the few but exact LiDAR measurements first with pixels (irrespective of depth) and then with their corresponding 3D points to obtain accurate depth estimates for several nodes in the graph (using only a 4 beam cheap lidar)

Pseudo Lidar

bounding boxes from the frontal view of a scene, pseudo-LiDAR begins with image-based depth estimation, predicting the depth $Z(u, v)$ of each image pixel (u, v). The resulting depth map Z is then back-projected into a 3D point cloud: a pixel (u, v) will be transformed to (x, y, z) in 3D by

$$z = Z(u, v), \qquad x = \frac{(u - c_U) \times z}{f_U}, \qquad y = \frac{(v - c_V) \times z}{f_V}, \tag{1}$$

A 3D cloud is obtained and can then be processed like a Lidar point cloud.

A disparity estimation algorithm takes a pair of left-right images I_l and I_r as input, captured from a pair of cameras with a horizontal offset (i.e., baseline) b. Without loss of generality, we assume that the algorithm treats the left image, I_l, as reference and outputs a disparity map D recording the horizontal disparity to I_r for each pixel (u, v). Ideally, $I_l(u, v)$ and $I_r(u, v + D(u, v))$ will picture the same 3D location. We can therefore derive the depth map Z via the following transform,

$$Z(u, v) = \frac{f_U \times b}{D(u, v)} \quad (f_U\text{: horizontal focal length}). \tag{2}$$

Pipeline:
1) 4D disparity volume that captures pixel difference
2) Build a 3D cost volume from that
3) PSMNet extracts deep feature map and concatenates features from left and right image followed by 3D convs and computes disparity

of 3D convolutions. The resulting 3D tensor S_{disp}, with the feature channel size ending up being one, is then used to derive the pixel disparity via the following weighted combination,

$$D(u, v) = \sum_d \text{softmax}(-S_{\text{disp}}(u, v, d)) \times d, \tag{3}$$

where softmax is performed along the 3rd dimension of S_{disp}. PSMNet can be learned end-to-end, including the image feature extractor and 3D convolution kernels, to minimize the disparity error

$$\sum_{(u,v)\in\mathcal{A}} \ell(D(u, v) - D^\star(u, v)), \tag{4}$$

4) SDN

A stereo network designed and learned to minimize the disparity error (cf. Equation 4) may over-emphasize nearby objects with smaller depths and therefore perform poorly in estimating depths for faraway objects. To see this, note that Equation 2 implies that for a given error in disparity δD, the error in depth δZ increases *quadratically* with depth:

$$Z \propto \frac{1}{D} \Rightarrow \delta Z \propto \frac{1}{D^2}\delta D \Rightarrow \delta Z \propto Z^2 \delta D. \tag{5}$$

2 Adaptations :

1) Depth loss (instead of disparity loss) => corrects emphasis on smaller errors of nearby objects
2) Depth Cost Volume is changed to operate on grid of depths => better for neighbourhood effects

$$Z(u,v) = \sum_z \text{softmax}(-S_{\text{depth}}(u,v,z)) \times z.$$

We construct the new depth volume, C_{depth}, based on the intuition that $C_{\text{depth}}(u,v,z,:)$ and $C_{\text{disp}}\left(u, v, \frac{f_U \times b}{z}, :\right)$ should lead to equivalent "cost". To this end, we apply a bilinear interpolation

=> much lower errors far away only small increses in near range

Depth Correction

- graph based
- Only low res lidar
- Input Matching : KD-Tree and KNN
- Edge weights : based on manifold learning a quadratic equation is optimised
- Correction : points with lidar measurements are updated to them, remaining depths are updated so they can still be reconstructed as a weighted sum of KNN

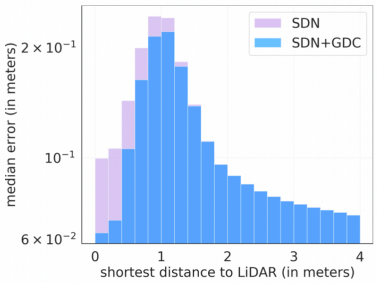

Figure 9: **Median depth estimation errors w.r.t. the shortest distances to 4-beam LiDAR points on KITTI validation set.**

Summary

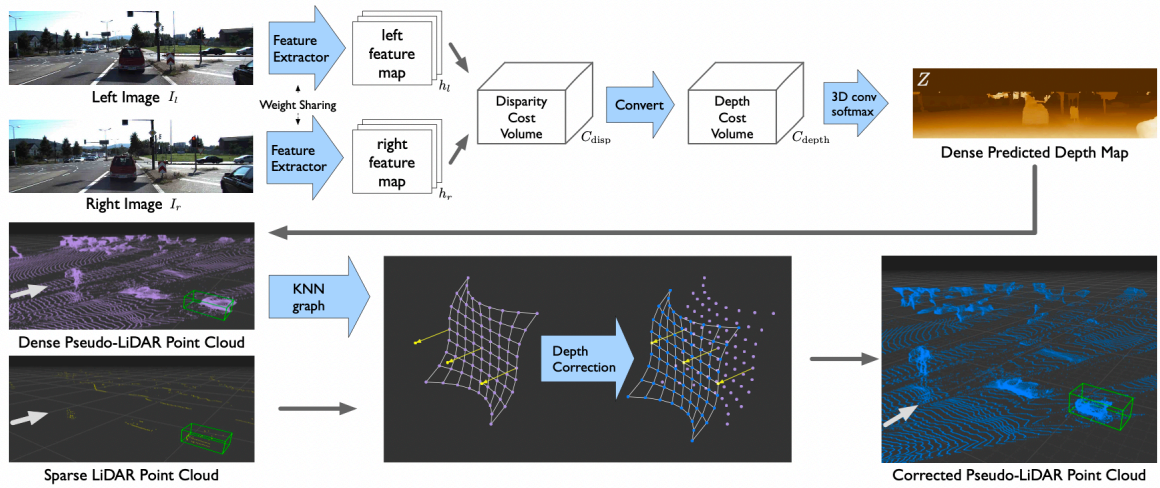

Figure 5: **The whole pipeline of improved stereo depth estimation:** (top) the stereo depth network (SDN) constructs a depth cost volume from left-right images and is optimized for direct depth estimation; (bottom) the graph-based depth correction algorithm (GDC) refines the depth map by leveraging sparser LiDAR signal. The gray arrows indicates the observer's view point. We superimpose the (green) ground-truth 3D box of a car, the same one in Figure 1. The corrected points (blue; bottom right) are perfectly located inside the ground truth box.