We thank the reviewers for their valuable feedback. We are encouraged they found our method well-motivated (R1, R2, R3), rigorous (R1), novel (R2, R4), simply reproducible (R1) and effective (R3), compatible with other algorithms (R1, R3), and well-validated by experiments (All). All the reviewers found our paper well-written and clear to follow. Given the time and page limit, we respond to the major comments and will incorporate all feedback.

@**R1- GANet:** Given the limit time, we only managed to train and evaluate **CDN**-GANET Deep MM on Scene Flow: 0.68/7.7/2.97 (EPE/1PE/3PE) (cf. Table 1). We will include other results on GANET in the final version.

@**R1- Why multi-modal ground-truths (GTs)?:** There are three reasons. First, pixels are discrete: a single pixel may capture different depths. Second, real datasets need to project signals from a depth sensor (e.g., LiDAR) to a depth map. As pixels are discrete and the cameras and LiDAR might be placed differently, multiple LiDAR points of different depths may be projected to the same pixel. Third, for stereo estimation, pixels along boundaries or occluded regions cause ambiguity to the model; multi-modal GTs offer better supervision for training, especially in early epochs.

@**R1- Why not MM in Table 1?** We want to emphasize the gain by our algorithm design. We report the non-MM results for a fair comparison with baselines which are mostly trained with uni-modal depths. We will specify this.

@**R1- MM in Table 3 & 4:** Conceptually our approach should improve, but we still evaluate using the (likely noisy) uni-modal GTs. We conduct an analysis as in Table 6: w/ MM achieves 2.08/13.2/8.65, better than w/o MM.

@**R1- α & k:** Table B shows the errors with varying α & k on Scene Flow using **CDN**-SDN-MM (cf. Table 4). A smaller α leads to a larger error, which makes sense as it relies less on the GTs. After all, we attribute the small gain in Table 4 to evaluation using uni-modal GTs. MM does improve convergence and depth on boundaries (see above).

@**R1, R2, R4- Offset network:** We will add details. It has 30K parameters, only 0.3% w.r.t. PSMNET. The novelty is in a single loss to jointly learn the offset and the main network.

@**R1- Bin sizes:** Our method outputs modes and needs (a) the bin containing the correct depth to have the highest probability and (b) the offset to be accurate. A smaller bin size makes (a) harder. A larger bin size makes (a) easier but makes (b) harder as the range of offsets gets larger.

Table A: MM ablation.

α	k	RMSE	ABSR
0.8	3	**1.80**	**0.028**
0.8	5	1.82	0.029
0.8	7	1.88	0.035
0.5	3	1.81	0.029
0.2	3	2.20	0.062

@**R1- Kendall [12]:** 3D Convs smooth the estimation but cannot guarantee uni-modal distributions. [12] employs pre-scaling to sharpen the probability (in their Fig 2), which might resolve the issue but makes the prediction concentrate on discrete disparity values. We do not prevent predicting a multi-modal distribution, especially for pixels whose disparities are inherently multi-modal. We output argmin (after an offset), which is what [12] hopes to achieve.

@**R1, R3- All Areas on KITTI:** There are two possible reasons. First, **CDN**-GANET overly focuses on foreground pixels that contain more ambiguities and discontinuities. Second, we used the same hyper-parameters as the original GANET and did not specifically tune it for **CDN**. We note that, # foreground:# background pixels is $\sim 0.15/0.85$; the degradation on background is ~ 0.16 3PE for both non occlusion and all, smaller than the gain on foreground.

@**R2- Learned offsets, explanations, insights:** Fig 3 shows how the offsets shift the distribution on a pixel and we will add more qualitative results. The offset network learns to produce the sub-grid disparity at each grid disparity values. The bin size balances the difficulty of predicting the bin location and the offset (please see @R1- Bin sizes) and we found $s = 2$ to perform well. It is the only hyper-parameter to tune and only integral values are considered.

@**R2- KL divergence (KLD):** We apply the Wasserstein distance (WD) to overcome non-overlapped supports in measuring divergences, which occur even if the target $p^\star$ is Dirac. Thus, using the WD is valid and more preferable than manually adding a smoothing Gaussian/Laplacian to the KLD. While in Eq. (10) one can pair the offset with either $\tilde{p}$ or $p^\star$, it makes more sense to view the offset as a way to improve the prediction $\tilde{p}$ rather than to adjust the target $p^\star$.

@**R2- Literature survey:** We will include more papers, especially those that discuss mean/mode and KLD.

@**R2- Ablation (cf. Table 5):** We use the mode for the WD-only model. Using a bin size $s = 2$ w/o offsets, the mode is restricted to integers and EPE suffers. Using mean has 1.26/13.5/4.18, worse than mode since the WD does not align the mean to the GT. Using mode for PSMNET has 1.57/39.7/4.40, worse than mean with a similar reason.

@**R3- CDN-SDN on KITTI:** We showed it in Table S3 (Suppl.). **CDN**-SDN is for depth estimation and we trained it on KITTI detection following [43] (L244). See also Table B for the results on KITTI detection Val using other metrics.

@**R3- L272-276:** Our approach has advantage on hard pixels whose disparity is ambiguous. We see (a) a gain on the foreground and (b) that foreground has a higher error than All. We thus argue that most of these hard pixels are on the foreground.

@**R3- L299-300:** We visualized the depth results w/ and w/o MM at early epochs and observed this. We will include both qualitative and quantitative results (like Table 6).

Table B: CDN-SDN on KITTI.

Method	RMSE	ABSR
SDN	3.08	0.044
CDN-SDN	3.00	0.042
CDN-SDN-MM	**2.99**	**0.040**

@**R4- Semantic segmentation:** Thanks for pointing out these papers that use semantic labels to guide the model to resolve depth discontinuities (i.e., predict uni-modal distributions). Our method, in contrast, does not prevent predicting multi-modal distributions along depth discontinuities, but changes the outputting rule (i.e., argmin with a predicted offset). Our method can also capture depth discontinuities within an object or an object class.

@**R4- Modeling the offsets:** While the *learned* offsets may lead to common supports between the predicted and GT distributions, we have to first come up with a loss to *learn* the offsets. Concretely, to learn b in Eq. (10), we need a loss that can measure the divergence between $\tilde{p}$ and $p^\star$. The WD offers a principled loss to learn the two networks jointly.

@**R4- Others:** Thanks for the great suggestions on analyses and we will try to include them in the final version.