Table 1: Results on the original Eigen test split [5] for other models with our MED volumes and MOM

Ref	Methods	PP	Sup	Data	#Par	abs rel↓	sq rel↓	rmse↓	rmse_{log} ↓	a^1 ↑	a^2 ↑	a^3 ↑
[7]	Monodepth		S	K	32	0.148	1.344	5.927	0.247	0.803	0.922	0.964
[7]	Monodepth-MED		S	K	32	0.112	0.751	4.500	0.196	0.868	0.954	0.980
[7]	Monodepth-MED with MOM fine tune		S	K	32	**0.107**	**0.684**	**4.311**	**0.187**	**0.878**	**0.960**	**0.982**
[20]	SuperDepth		S	K	-	0.112	0.875	4.958	0.207	0.852	0.947	0.977
[20]	SuperDepth-MED		S	K	58	0.111	0.682	4.295	0.190	0.879	0.959	0.982
[20]	SuperDepth-MED with MOM fine tune		S	K	58	**0.108**	**0.647**	**4.180**	**0.184**	**0.886**	**0.962**	**0.983**

Firstly, we thank all reviewers for their detailed reviews. Next, we discuss their comments (R:Reviewer, W:Weakness).

R1.W1: We will clarify in abstract and introduction that our method is self-supervised from stereo pairs. When learning from mono videos, our approach can be easily adopted, since the MOM can be used as long as the network incorporates a disparity prob volume, and the relative camera position is known or estimated. The camera-pose information can be integrated into the warping operation $g(\cdot)$ in Eq. (4) to obtain the mirrored occlusions for the corresponding frame pair. What could be at stake here is the exponential quantization, as inverse depths in SFM are defined up to an unknown and inconsistent scale. The ambiguous scale could prevent the network from taking advantage of all disparity levels. A turn-around for this issue is to incorporate velocity supervision, as introduced in PackNet (Guizilini et al., CVPR2020), or consistent SFM (Tucker and Snavely, CVPR2020) to fully exploit the exponential quantization.

R1.W2: We will follow this suggestion in our paper to convey the main idea clearly and early on the paper.

R1.W3: We will add videos and point-clouds to the supplementary and presentation materials.

R1.W4: Thanks for the advice, it seems that we miss-understood the purpose of the broad impact statement. It will be augmented accordingly. Regarding under/overestimation, it seems we are on the safer side. We measured this by computing the mean median scaling factor [34] between the GT and our depth estimates. We obtained a mean scale factor of 1.016, indicating our network detects objects slightly closer than they are.

R2.W1: We would like to clarify that the occlusion maps are not obtained directly from the depth predictions, but from the information coming from the MED volumes of the stereo pair during training (note the MOM is fed with left and right probability volumes D_L^P and D_R^P). Confidence masks in previous methods [8, 34] either require additional networks or depend on the depth model to generate them, limiting their performance. As can be observed in [8, 34], their generated masks are of low-quality compared to the highly detailed occlusion maps in our MOM.

R2.W2: Please note that the MOM is fed with the MED volumes from the FAL-net under training, not from the fixed FAL-net model from the first stage. This means that the generated occlusion maps in the MOM will get better as the depth estimates from the FAL-net under training get better. We also showed that the improvements just did not come from the additional training schedule, but from the use of the MOM in the ablation section, we further extend on this in Table 1 and the next question.

R2.W3: Our experiments tell us that the effectiveness of the "exponential disparity level" representation is universal. For the sake of completeness, we plugged MED and MOM into the Monodepth (Godard et. al CVPR2017) and the more recent SuperDepth (Pillai et. al, ICRA2019). The latter incorporates ESPCN [24] up-sampling modules in the decoder stage. Incorporating MED volumes and Fine-tuning with MOM showed steady improvements in both networks. This is shown in Table 1 of this rebuttal, which further supports the effectiveness of our overall method.

R3.W1: Surprisingly, we are the first to shed light on the effectiveness of exponential quantization of disparity when learning self-supervised depth via view synthesis. We showed it could improve accuracy from 84% to 93% just by itself.

R3.W2: During training, we monitored the RMSE between the synthetic right view and the GT right view on the KITTI2015 training split for our information. We observed a synthesis performance around 22 in RMSE, which is very good considering that recent works on view synthesis achieve 24 in RMSE [9]. However, since the objective of our method is not view synthesis, discretization and occlusion artifacts are visible in the synthetic images, as naturally expected. To reflect the reviewer's comment, we will add visuals of synthesized images in the supplementary material. The sample of synthetic right view in Fig. 3 of our main paper is actually obtained from the network.

R4.W1: The assumption is correct; we can also borrow a simplified explanation from Reviewer 1: the MOM could be understood as a multi-view occlusion mask generation module. The generated masks are used to filter invalid regions due to parallax.

R4.W2: The summation does not become one, as the planes of the probability distribution are first warped (shifted) to the target view by $g(\cdot)$ in Eq. (4). This shifting not only generates "holes", which are the occluded regions but also areas where the summation is > 1. The latter is the reason why the "max" operator is applied to cap the final occlusion masks O^L and O^R between 0 and 1. We will make sure it is explained more clearly in the final version of the paper.