

For research article

Response to Reviewer 2 Comments

1. Summary

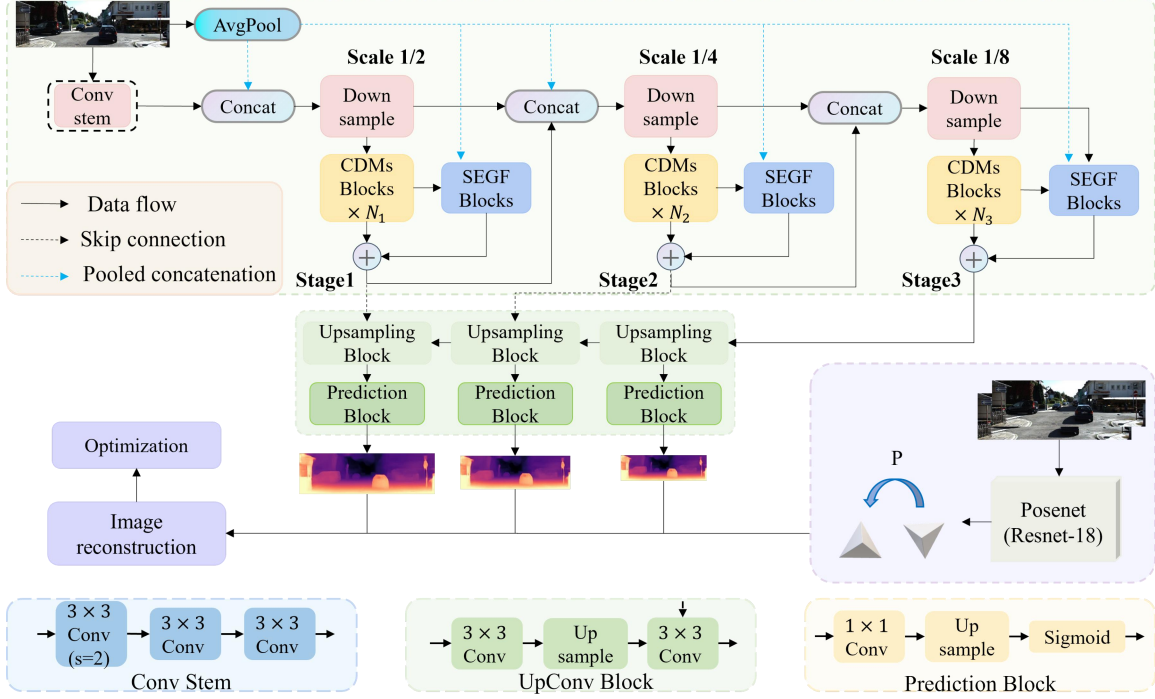
Thank you very much for taking the time to review this manuscript and for providing these meticulous and constructive comments. Your suggestions have been invaluable in helping us improve the technical accuracy, clarity, and overall quality of our work. Please find the detailed point-by-point responses below and the corresponding revisions/corrections highlighted in the re-submitted files.

2. Questions for General Evaluation	Reviewer's Evaluation	Response and Revisions
Does the introduction provide sufficient background and include all relevant references?	Yes	
Are all the cited references relevant to the research?	Yes	
Is the research design appropriate?	Yes	
Are the methods adequately described?	Can be improved	
Are the results clearly presented?	Yes	
Are the conclusions supported by the results?	Can be improved	

3. Point-by-point response to Comments and Suggestions for Authors

Comments 1: Figure 2: Please double-check the inputs of the three upsampling blocks. In the current figure, features from stage 3 and stage 2 are shown as inputs to the right block, while features from stage 1 are connected to the middle block. I believe that features from stage 2 should instead be connected to the middle block, and features from stage 1 should be connected to the left block.

Response 1: Thank you for pointing this out. We carefully examined the model architecture and confirmed that Figure 2 indeed contained a drawing error. We have now corrected it: features from Stage 2 are connected to the intermediate upsampling block, and features from Stage 1 are connected to the left-hand block. The revised Figure 2 is shown below:



Comments 2: If dense depth maps are not available in the KITTI dataset, please clarify how the authors handled pixels in the RGB images without corresponding depth values when training the model and calculating evaluation metrics. Were these pixels simply ignored? Additionally, please clarify whether the sparse ground truth is sufficiently reliable for evaluation.

Response 2: Thank you for pointing this out. The issue of sparse depth ground truth is properly addressed within our self-supervised learning framework, as explained below:

Firstly, regarding the training process: as elaborated in Section 3.4 (Self-supervised Learning) of our paper, our model adopts a self-training approach based on photometric reprojection error. The optimization objective is to minimize the image reconstruction loss between adjacent frames, and the entire training process does not rely on any form of LiDAR ground truth depth data. Therefore, the absence of depth values for certain pixels in the RGB images does not affect the training in any way.

Secondly, during the model performance evaluation stage, we strictly adhere to the standard protocols in the field of self-supervised depth estimation to handle the sparsity of LiDAR point clouds. Specifically, the radar scans produce a point cloud, which is projected onto the image plane for computation at corresponding pixel locations. Since self-supervised methods predict relative depth while LiDAR provides absolute depth values, we apply median scaling to align the scale of the predictions. This is done by computing the median of the ratios between predicted and ground truth values over valid pixels to determine the scaling factor, bringing the predicted depth to the same scale as the ground truth before calculating various error metrics.

Although the ground truth depth data is sparse, this evaluation approach has been widely adopted and validated for reliability by both classical models and novel lightweight models. This reliability stems from two factors: LiDAR point clouds themselves have high measurement accuracy, and all compared methods are evaluated under identical conditions

using the same set of valid pixels and evaluation procedure, ensuring fairness in comparison and credibility of the results.

The revised content can be found in the second and third paragraph of Section 4.1.3 "Evaluation Metrics" on page 9, from line 304 to line 318:

"During the model performance evaluation stage, we strictly adhere to the standard protocols in the field of self-supervised depth estimation to handle the sparsity of LiDAR point clouds. Specifically, the radar scans produce a point cloud, which is projected onto the image plane for computation at corresponding pixel locations. Since self-supervised methods predict relative depth while LiDAR provides absolute depth values, we apply median scaling to align the scale of the predictions. This is done by computing the median of the ratios between predicted and ground truth values over valid pixels to determine the scaling factor, bringing the predicted depth to the same scale as the ground truth before calculating various error metrics.

Although the ground truth depth data is sparse, this evaluation approach has been widely adopted and validated for reliability by both classical models and novel lightweight models [19,30,56,57]. This reliability stems from two factors: LiDAR point clouds themselves have high measurement accuracy, and all compared methods are evaluated under identical conditions using the same set of valid pixels and evaluation procedure, ensuring fairness in comparison and credibility of the results."

Comments 3: Table 1: After including the results of the MViT-Depth series and LiteGfm, the performance improvement appears to be minor, particularly compared to LiteGfm. Please revise Section 4.2.1 (Quantitative Comparison) accordingly and explicitly include the comparison with both the MViT-Depth series and LiteGfm.

Response 3: Thank you for pointing this out. We have revised Section 4.2.1 to include and discuss comparisons with the MViT-Depth family and LiteGfm. The updated analysis shows that, relative to recent lightweight models with state-of-the-art metrics, the principal advantage of MsGf lies in its parameter efficiency. Specifically, MsGf reduces the parameter count by approximately 58% compared with LiteGfm (1.9 M), while maintaining comparable or slightly superior performance on key benchmarks. Against the MViT-Depth suite, MsGf outperforms the best-performing variant, MViT-Depth (6.3 M), yet requires only 13% of its parameters. These results confirm that MsGf achieves an optimal accuracy - efficiency trade-off.

The revised content can be found in the third and fourth paragraph of Section 4.2.1 on page 11, from line 346 to line 354:

"Furthermore, compared to high-performing LiteGfm [57] (1.9M parameters), our model reduces the parameter count by approximately 58% while maintaining comparable performance on several core metrics (e.g., AbsRel, RMSE log, $\delta 1$) and even achieving slight improvements on others (e.g., RMSE, $\delta 1$, $\delta 2$). Compared to the MViT-Depth series[58], our model outperforms their largest variant (MViT-Depth, 6.3M params) across all major metrics, using only 13% of its parameters. These results demonstrate that the MsGf model achieves an optimal trade-off between accuracy and efficiency, offering a superior solution for deploying high-precision monocular depth estimation on resource-constrained devices."