GroCo: Ground Constraint for Metric Self-Supervised Monocular Depth

Aurélien Cecille, Stefan Duffner, Franck Davoine, Thibault Neveu, Rémi Agier

GroCo: Ground Constraint for Metric Self-Supervised Monocular Depth

Aurélien Cecille[1,2], Stefan Duffner[2], Franck Davoine[2], Thibault Neveu[1], and Rémi Agier[1]

[1] Visual Behavior, Lyon, France
[2] INSA Lyon, CNRS, Ecole Centrale de Lyon, Universite Claude Bernard Lyon 1, Université Lumière Lyon 2, LIRIS, UMR5205, 69621 Villeurbanne, France

Abstract. Monocular depth estimation has greatly improved in the recent years but models predicting metric depth still struggle to generalize across diverse camera poses and datasets. While recent supervised methods mitigate this issue by leveraging ground prior information at inference, their adaptability to self-supervised settings is limited due to the additional challenge of scale recovery. Addressing this gap, we propose in this paper a novel constraint on ground areas designed specifically for the self-supervised paradigm. This mechanism not only allows to accurately recover the scale but also ensures coherence between the depth prediction and the ground prior. Experimental results show that our method surpasses existing scale recovery techniques on the KITTI benchmark and significantly enhances model generalization capabilities. This improvement can be observed by its more robust performance across diverse camera rotations and its adaptability in zero-shot conditions with previously unseen driving datasets such as DDAD.

Keywords: Depth · Monocular · Self-Supervised · Metric · Generalization

1 Introduction

Depth estimation is a fundamental task in computer vision, offering crucial 3D insights for various applications such as robotics, augmented reality and intelligent vehicles. Specifically, in the realm of intelligent vehicles, accurate depth perception is vital for navigating safely by identifying and localizing potential obstacles.

Among the various methods, monocular depth estimation is particularly attractive due to its cost efficiency and broad availability across many systems. It presents a viable alternative to more expensive technologies like Lidar and stereo vision, promising wide applicability in real-world scenarios. The advancement of robust monocular depth models, however, is hampered by the need for diverse, large-scale, annotated datasets, which are costly to create.
In response, there has been an increased interest in self-supervised learning, where models are trained using unlabeled data by leveraging the consistency of

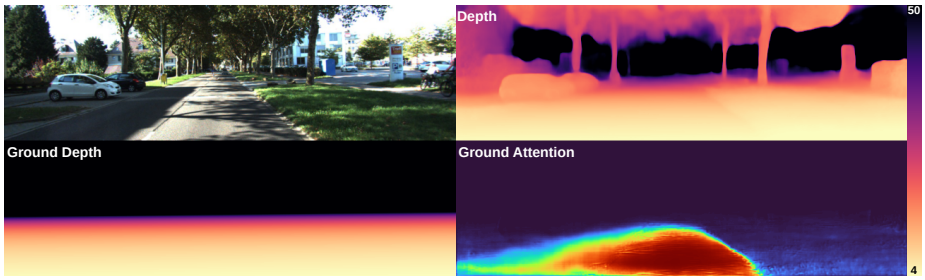

Fig. 1: Example of the models' depth and ground attention prediction. The ground depth is given as input and integrated in the depth prediction using the attention map.

scene geometry across different viewpoints and moments in time. Nonetheless, the self-supervision brings its own set of complications, notably in deriving metric scale information. This problem aligns with the longstanding challenges in the field of monocular visual odometry [1], essential for self-supervised learning of depth. In fact, it is well-documented that the scale of the scene cannot be determined using only monocular images, introducing an inherent ambiguity in the predicted depth and egomotion, which are known only relative to an unknown scale factor. At its core, this issue arises because an image may correspond to various 3D scenes, differentiated only by their scale. This ambiguity is problematic as it obstructs applications that depend on precise distance measurements for decision-making.

Additionally, both self-supervised and supervised depth estimation methods often face the challenge of model overfitting to specific camera parameters [17]. Indeed, within the context of driving, monocular models often infer depth by correlating the vertical position of ground pixels with constant depths, a process heavily relying on unchanging camera intrinsic and extrinsic parameters. Such an approach significantly hampers the models' ability to generalize across different camera setups. Consequently, models require retraining or fine-tuning for each new camera configuration, which significantly limits their practical usability in diverse real-world environments.
In the context of ground-based systems like vehicles or robots, this limitation can be mitigated by using the theoretical flat ground as a reference since it can be deduced from commonly known camera parameters. Two strategies for integrating this ground information exist: the *a posteriori* method, which involves adjusting the scale of depth predictions during the inference phase [21], requiring additional processing steps and necessitating ground segmentation; and the *a priori* method, showcased in recent supervised learning approaches [12, 22], which embeds ground priors into the depth estimation model itself. This latter strategy equips the model with all necessary information for robust scaled depth prediction right from the start, aiming to enhance performance on all types of scenarios.

Despite these advancements, the transition to self-supervised settings remains impeded by the additional scale ambiguity challenge. This gap underscores the need for innovative methodologies capable of leveraging ground plane information effectively in the absence of explicit labels.
Our work directly addresses this challenge by introducing novel loss functions specifically designed for the integration of ground plane priors within a self-supervised learning framework as illustrated in Fig. 1. These innovations significantly enhance the depth estimation models' accuracy and generalizability, facilitating robust performance across diverse camera configurations and environments.
Our four key contributions are the following: **(1)** A self-supervised method for metric depth estimation that enhances generalization across camera poses and datasets. **(2)** Novel loss functions for precise scale recovery. **(3)** A new way of integrating ground attention not requiring any depth annotation. **(4)** An interpretable attention mechanism to accurately localize flat ground areas in images. The source code is available at `https://github.com/Visual-Behavior/GroCo`.

2 Related Work

2.1 Monocular Self-Supervised Depth Estimation

Self-Supervised Depth estimation is a task that has already been widely studied in the past few years. The main idea is to train a model to predict the depth of a scene without labels by exploiting its geometry. The most common way to do this, is to use the simultaneous learning of depth and egomotion [5, 8, 18, 26]. That is, the model is trained to predict the depth of the scene and the motion of the camera at the same time by computing the photometric error between the original image and the reprojected one from the predicted depth and motion. This is a very efficient way to train the model as it does not require any labels but uses the assumption of a static scene and a moving camera. Godart *et al.* [5] proposed a method that is robust when these hypotheses are not satisfied. They manage sequences where there is no egomotion by masking out pixels that do not change between frames and use a minimum loss across adjacent frames to handle dynamic objects.
Model architectures have also been improved, Lyu *et al.* [15] enhanced the quality and sharpness of predicted depth, and recently, Transformer-based architectures have also been used to further increase performance [23, 25].

2.2 Scale Recovery

Metric depth is crucial for downstream tasks. However, since images do not naturally reflect scale changes, incorporating additional information during training is essential for retrieving depth at the correct scale.
Guizilini *et al.* [8], for example, presented a method leveraging vehicle velocity to impose a scale on egomotion estimation, constraining the depth estimation

to be scaled as well. Wagstaff and Kelly [19] proposed to scale the depth using the height of the camera. The process begins by training an up-to-scale model to derive relative depth. Following this, an unsupervised ground segmentation model is developed using the assumption that the bottom middle part of the image is the ground and fitting its relative depth to a plane. All pixels that are close to this plane are then considered as ground. In the subsequent stage, the scale is computed by fitting a plane on the segmented depth and scaling its normal vector with the camera height. New loss terms are then included in the optimization so that the model has to predict depth and egomotion that are equal to their scaled counterpart. By exploiting "off-the-shelf" ground and vehicle segmentation models, Kinoshita and Nishino [11] leverage the assumption of constant camera height to recover the scale of the scene. They especially use the fact that projecting vehicle points to the plane orthogonal to the ground always gives the same height even if the depth of objects changes. Combining this with the prior knowledge of the vehicle height allows to recover the camera height and the scale of the scene. Zhang *et al.* [24] recovers the scale using the IMU sensor combined with an extended Kalman filter (EKF) to provide motion at scale, constraining the depth to adopt the same metric scale. Xiang *et al.* [20] propose to recover the scale using the fact that in the KITTI dataset [16] the rectangular area in the middle bottom part of the image belongs to the ground. Combined with the camera height prior, it allows to determine the scale of the depth.

We notice that all methods that use the camera height rely on some form of ground segmentation, whether model or heuristic-based. This dependency introduces additional challenges in ensuring robustness across diverse scenarios, potentially restricting their usability.

Additionally, most of these models consider the scale as a constant since they only use their prior during inference, and they do not generalize well when a change of camera position should result in scale adjustments.

2.3 Ground Prior

Van Dijk and De Croon *et al.* [17] demonstrated that monocular models estimate depth in two ways: by leveraging the vertical position of the contact point between the object and the ground and through the assimilation of a size prior for objects, with the former having a more significant impact. They further highlighted the sensitivity of these methods to alterations in camera pose, leading to inaccuracies in ground recognition and consequently diminishing overall model efficacy.

To address this limitation, [12, 22] have proposed the integration of a ground prior to provide camera pose information to the model and predict robust metric depth through the use of supervised annotations.

Koledić *et al.* [12] employed a technique that transforms the ground plane into an embedding via a Fourier transform, which is then concatenated with encoder-derived features. This method trains on supervised synthetic data across a wide range of camera poses and thus exhibits robustness to these variations. It can be

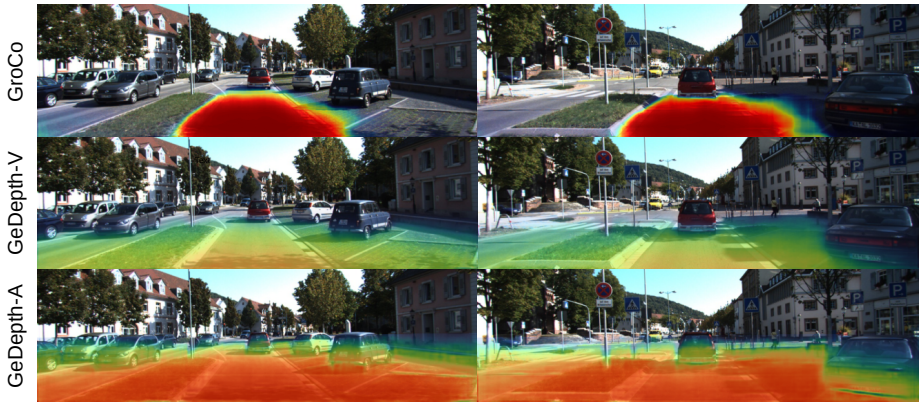

Fig. 2: Result of attention maps compared to GeDepth [22]. While our method outputs very certain and precise ground segmentation, we see that GeDepth tends to have higher recall and uncertainty. We note that although Gedepth attention maps often consider the bottom part of obstacles as ground, it does not impact the end performance because these parts can be compensated by the residual depth or the slope prediction. It also underlines the fact that their adaptive (A) version relies much more on the ground prior compared to the vanilla (V) one, potentially improving robustness.

adapted to real data through a domain adaptation module and the utilization of stereo datasets.

The approach proposed by Yang *et al.* [22], on the other hand, normalizes the ground depth image and directly concatenates it with the input image. Additionally, the authors introduced a ground attention mechanism that works alongside the predicted depth to integrate the ground prior in the final output, termed *Vanilla* version. Subsequently, they presented an *Adaptive* version of their framework, capable of estimating the slope for each pixel within the ground prior, enhancing model accuracy in environments with uneven terrain. Their findings suggest that this method not only generalizes more effectively to unseen datasets but also maintains robustness against changes in image resolution. However, the slope estimation technique shows some limitations. In particular, its inability to adjust the horizon line restricts its applicability to merely offsetting existing ground pixels and consequently introduces issues with positive slopes. Besides, Fig. 2 illustrates a counter intuitive behavior of the ground attention mechanism that considers the bottom part of obstacles as ground.

Despite the promise of these methodologies, they still require the use of stereo cameras or Lidar annotations to circumvent the scale issue inherent to monocular self-supervised settings — an aspect that our work addresses directly.

3 Method

This section outlines our methodology, demonstrating how each component synergistically contributes to solving the scale of the scene and improving gen-

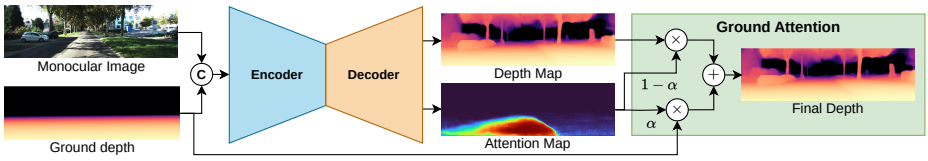

Fig. 3: Illustration of the model architecture, highlighting the integration of ground depth information. The input image and ground depth are concatenated to provide ground aware features. The ground attention mechanism combines the depth map with the ground depth, guided by the attention map, to produce a refined final depth estimation.

eralization across diverse camera setups and datasets, thereby advancing the capabilities of self-supervised learning in depth estimation.

3.1 Ground Plane

To provide the ground prior to the model we use the modifications proposed by [22] since the approach is very flexible and can be integrated with different types of neural network architectures such as CNN or Transformers.
We compute the location of the theoretical ground plane thanks to the camera parameters and height h.

Using the camera intrinsic $K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$ and extrinsic $E = \begin{bmatrix} R \mid t \end{bmatrix}$ such that

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = R^{-1}(K^{-1} d_{u,v} \cdot \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} - t) \tag{1}$$

We can recover the depth of the ground $d_{u,v}$ for each pixels at position (u, v) with height $y = h$ using the following formula:

$$d_{u,v} = \frac{h - t_y}{\frac{R_{1,2}}{f_x}(u - c_x) + \frac{R_{2,2}}{f_y}(v - c_y) + R_{3,2}} \tag{2}$$

Computing the depth for all pixels and keeping only positive values, we obtain a ground depth image representing the distance from the camera to the theoretic ground at each pixel, ignoring obstacles and ground slope variations. This image is then normalized and directly concatenated with the input image as shown is Fig. 3.

We adapt the ground attention scheme from [22], utilizing its vanilla version. The principle is to allow the model to choose between its own predicted depth $\hat{D}$ and the ground prior G as can be seen in Fig. 3. It is done by adding a new attention map α such that the final depth for each pixel i is obtained as follows:

$$D_i = (1 - \alpha_i) \cdot \hat{D}_i + \alpha_i \cdot G_i \tag{3}$$

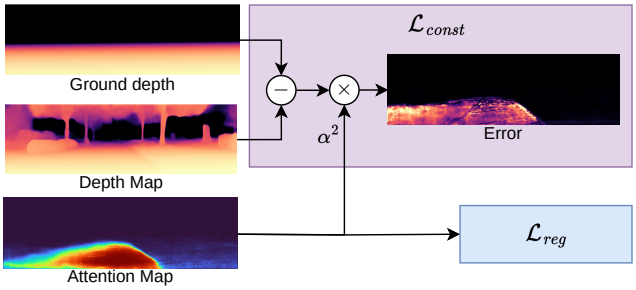

Fig. 4: Overview of the proposed ground constraint loss $\mathcal{L}_{const}$ and attention regularisation $\mathcal{L}_{reg}$. The error image in $\mathcal{L}_{const}$ illustrates how this loss penalizes disagreement between the depth map and ground depth, indirectly ensuring that the scale of depth converges to the one of the ground.

3.2 Scale Constraint

To ensure the accurate scaling of depth predictions, our approach incorporates two novel penalty terms during the training phase, as detailed in Fig. 4. These penalties are designed to more effectively leverage ground prior information, thereby guiding the model to implicitly estimate depth at the correct scale. The first penalty is an activation regularisation on the ground attention that ensures that it segments a minimum proportion of the ground. The second is a constraint loss that solves the scale issue and ensure that the attention correctly segments the ground area.

The regularisation addresses a fundamental challenge: without supervision, models do not automatically align the scale of their depth with the ground prior as they would when trained with labeled data, leading to the dismissal of the ground prior by the attention in favor of maintaining internal consistency in the depth estimates.
To counteract this tendency and promote the integration of the ground prior, we introduce a novel regularization term $\mathcal{L}_{reg}$. This term is designed to encourage the model to incorporate the ground prior into its depth estimation process by penalizing the attention when it does not activate enough, bridging the gap created by the absence of direct scale references from annotations. However, since we do not want the model to take the ground depth everywhere, we only apply this regularisation up to a given threshold τ between 0 and 1, leading to the following formulation:

$$\mathcal{L}_{reg} = \frac{\max(0, \tau - \frac{1}{N}\sum_i^N \alpha_i)^2}{\tau^2} \tag{4}$$

with N the total number of pixels, α_i the i^{th} pixel of the attention map and τ^2 a normalization constant keeping the value in the unit interval.
We found that this formulation is much more robust to hyperparameters than

using a classical regularisation while also being more intuitive to interpret.
Indeed, τ represents the proportion of the ground prior that we are confident at identifying as the ground, typically road surfaces. To prioritize precision and ensure the integrity and scale of depth estimations, it is recommended that τ be set below the proportion of the optimal ground segmentation.
Since τ changes depending on datasets, we propose a rule to compute its value based on the navigable area with respect to image dimensions H and W as well as the expected pathway width P_w and camera height h:

$$\tau = \frac{P_w H}{4hW} \tag{5}$$

Building on this, the constraint loss $\mathcal{L}_{const}$ ensures that the attention correctly activates on ground areas and that in these areas the predicted depth converges to the ground prior. The equality between the predicted depth and the ground prior is necessary so that the scale of the ground is correctly estimated and not degraded by the residual depth. Its effect is twofold: it penalizes the attention on pixels where the ground prior and depth are distant, and, at the same time, penalizes the depth on pixels selected by the attention to make it converge to the ground prior. It can be expressed as:

$$\mathcal{L}_{const} = \frac{1}{N} \sum_{i}^{N} \alpha_i^2 |\hat{D}_i - G_i| \tag{6}$$

We use an absolute distance instead of a relative one to ensure that the attention focuses on closer ground areas. These are more likely to meet the flatness criterion rather than distant areas where this assumption may not hold. The attention is squared to penalize uncertain areas less and allow for a better quality depth prediction as opposed to the raw value that can cause the model to predict more binary attention maps.

We also use the reprojection loss $\mathcal{L}_{reproj}$ and smoothness loss $\mathcal{L}_{smooth}$ from [5] to ensure that the model can estimate the geometry of the scene and correctly propagate the ground scale everywhere, resulting in the final loss:

$$\mathcal{L} = \mathcal{L}_{reproj} + \lambda_{smooth}\mathcal{L}_{smooth} + \lambda_{const}\mathcal{L}_{const} + \lambda_{reg}\mathcal{L}_{reg} \ . \tag{7}$$

Note that, compared to the adaptive method described in [22], we do not let the model predict the slope of the ground. This is for two reasons. The first is that if we give a new degree of freedom to the model, there would be no guarantee that it would converge to a metric depth. And the second is that it is not strictly necessary, since in case the ground is not flat, there is nothing stopping the model from simply discarding the area in the attention and predicting the correct depth.

3.3 Rotation Augmentation

To improve the robustness of the model, rotation augmentation is applied during training to both the images and coherently to the ground. This augmentation

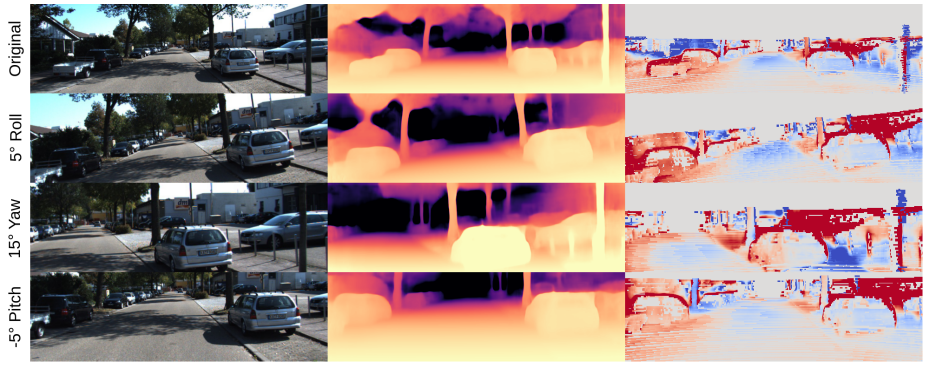

Fig. 5: Visualisation of predictions on the same image with the various rotation augmentations. The last column is the relative per pixel error with the ground truth. The error is between -20% in red and 20% in blue, with 0% or absence of ground truth in white.

simulates camera pose changes and helps the model learn to handle different orientations of the scene. We focus on rotations since they can easily be simulated by warping images, compared to translations that would require to know a dense ground truth depth which is contradictory to the self-supervised setup.
We limit angles amplitudes to 5° for pitch and roll and 15° for yaw to not introduce any black borders or upscaling in images. The ground depth is also augmented to match the rotated images by directly applying the rotation on the camera extrinsic, avoiding interpolation errors. Illustrations of these augmentations can be seen in Fig. 5.
In the same way, we also transform the Lidar depth used to evaluate the model to match the rotated images. This is done by rotating the Lidar point cloud and projecting it to the image plane to obtain the new depth.

3.4 Interpretability

Thanks to the ground attention mechanism, the prediction of the model can be reliably interpreted as seen in Fig. 6. By providing the area where the ground prior and the predicted depth are equal, we can detect failure cases of the model. This could typically be the case with images where the ground is not visible or if it

Fig. 6: Example of segmentation quality of ground attention. Even in cases where there are close obstacles, the attention stays precise.

is very uneven. This information can be used either for humans to perform visual inspection or could for example be used during inference to filter predictions or trigger some sort of warnings in case the model is not able to detect any ground, potentially signaling that the camera has moved.

4 Experiments

4.1 Implementation Details

By default we follow [5] and use a Resnet50 encoder [9] pretrained on the Imagenet dataset [2] and the same decoder coming from [6]. To take the additional inputs from the ground embedding channel, the pretrained weights are kept and the weights of the new channel are initialized with a value of zero. We adapt the outputs of the decoder, replacing the sigmoid by the softplus function to directly predict a strictly positive depth coherent with the ground prior. We also add a new head using the features at all resolutions to predict the attention map similarly to [22].
For the hyper-parameters, we keep the default $\lambda_{smooth} = 10^{-2}$ and set $\lambda_{const} = \lambda_{reg} = 0.1$. τ is set to 0.25 on KITTI and 0.5 on DDAD to reflect the difference in image width and corresponds to a pathway width of two 2.75m wide lanes.
The model is trained using the Adam optimizer [10] with a learning rate of 10^{-4} and a batch size of 12 for 20 epochs on KITTI [4]. On an NVIDIA RTX 3090, it takes about 10 hours for the training to finish.

4.2 Datasets

We use the KITTI dataset extensively since it is the standard for depth estimation in the use case of intelligent vehicles. We report results using the eigen split using both the original Lidar data [4] and the improved depth coming from the KITTI depth benchmark [16]. Unless specified, we will report results on the improved version since it more accurately represents the model performance.
We also use the DDAD dataset [8] to show the generalization of our model to new datasets and cameras. Similarly to [7] we use the front, back, front left and front right cameras to evaluate the model.

4.3 Performance

We first compare our approach to the state-of-the-art methods that only use the camera height to recover the scale of the scene similarly to us. We report standard metrics used for depth evaluation coming from [3]: AbsRel (Absolute Relative Error), SqRel (Squared Relative Error) , RMSE (Root Mean Squared Error), $\text{RMSE}_{\log}$ (Root Mean Squared Log Error), $\delta < 1.25$, $\delta < 1.25^2$ and $\delta < 1.25^3$. δ metrics are accuracy measures and count the proportions of pixels where their ratio with the ground truth is inferior to 1.25^n. In Tab. 1 we see

Table 1: KITTI self-supervised metric depth performance on two versions of labels. Comparison with methods that only use camera height to recover the scale. Results from other works are taken from [11].

Labels	Method	Error (↓)				Accuracy (↑)		
		AbsRel	SqRel	RMSE	$\text{RMSE}_{\log}$	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
[4]	Scale Recovery [19]	0.123	0.996	5.253	0.213	0.840	0.947	0.978
	VADepth [20]	0.120	0.975	4.971	0.203	0.867	0.956	0.979
	Groco	**0.113**	**0.851**	**4.756**	**0.197**	**0.870**	**0.958**	**0.980**
[16]	VADepth [20]	0.091	0.555	3.871	**0.134**	**0.913**	0.983	**0.995**
	Groco	**0.089**	**0.517**	**3.815**	**0.134**	0.910	**0.984**	**0.995**

that our method outperforms the others in the monocular self-supervised metric depth estimation task without additional priors such as segmentation. Fig. 7 shows the performance as a function of τ and its robustness with respect to it.

4.4 Robustness to Camera Position Changes

In order to evaluate our method against a comparable one, we propose a new baseline using the default Monodepth2 [5] pipeline in addition of losses proposed by [19] and leveraging the ground prior to estimate the scale, the method is detailed in the supplementary material. We compare the performance of both models on the KITTI dataset with different camera poses. Both methods were trained with augmentation at training time. Results are reported in Tab. 2. We can see that our method performs better than the baseline for all rotations even though they perform very similarly on original images, demonstrating the gain of using our method to exploit the ground prior. For yaw and roll, we report positive values only since negative ones perform similarly. For the pitch we use the negative one because the positive augmentation leads to the ground not being visible in the image, rendering our method ineffective.
We also report the camera transfer performance against supervised methods in Tab. 3 and show that our method is able to better generalize to new cameras.

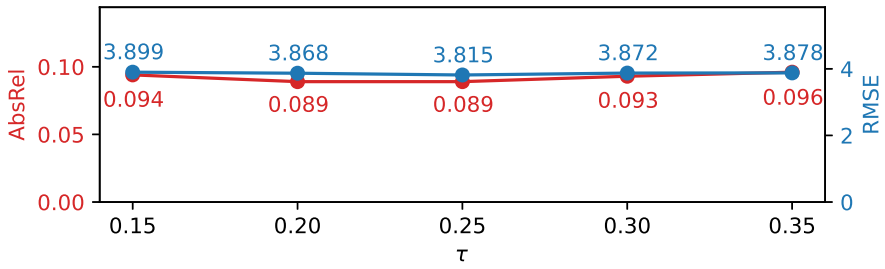

Fig. 7: Model performance for varying τ on KITTI.

Table 2: Performance on KITTI with different camera rotations.

Augment	Method	Error (↓)				Accuracy (↑)		
		AbsRel	SqRel	RMSE	$RMSE_{log}$	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
None	Baseline	0.092	**0.499**	**3.701**	**0.134**	**0.912**	**0.984**	**0.996**
	Groco	**0.089**	0.517	3.815	**0.134**	0.910	**0.984**	0.995
5° **Roll**	Baseline	0.140	0.879	5.083	0.194	0.793	0.960	0.992
	Groco	**0.101**	**0.635**	**4.301**	**0.147**	**0.891**	**0.979**	**0.994**
−5° **Pitch**	Baseline	0.145	0.891	5.341	0.197	0.782	0.952	0.989
	Groco	**0.086**	**0.544**	**3.930**	**0.132**	**0.914**	**0.983**	**0.995**
5° **Yaw**	Baseline	0.107	**0.548**	**3.944**	0.149	0.891	0.980	**0.996**
	Groco	**0.096**	0.552	**3.944**	**0.140**	**0.900**	**0.982**	0.995
15° **Yaw**	Baseline	0.209	1.246	5.919	0.268	0.483	0.932	0.986
	Groco	**0.136**	**0.852**	**4.888**	**0.189**	**0.808**	**0.958**	**0.989**

Table 3: Generalization on new cameras in the same domain compared to supervised methods. The model is trained on the front camera and evaluated on the other ones. Results from other works come from [22], BTS [13] being a CNN-based architecture and DepthFormer [14] a Transformer-based one.

Method	AbsRel (↓)			
	Mean	Back	Left	Right
DepthFormer [14]	0.93	0.83	0.98	0.97
DepthFormer [14] + GeDepth Adaptive [22]	0.66	0.64	0.59	0.75
BTS [13]	0.72	0.82	0.98	0.97
BTS [13]+ GeDepth Adaptive [13, 22]	<u>0.62</u>	<u>0.62</u>	**0.56**	**0.67**
Groco	**0.56**	**0.43**	<u>0.57</u>	<u>0.68</u>

4.5 Generalization to New Datasets

We further evaluated the generalization capacity of our model by training it on the KITTI dataset and measuring its performance on the DDAD dataset. We report the results in Tab. 5. These results are evaluated up to 80m and with an image height of 192 pixels like in the KITTI benchmark. We can see that our model generalizes better to the new dataset than the baseline for all cameras. We also notice that even if our model never saw images of side cameras, its attention is quite robust at segmenting the ground as can be seen in Fig. 8.

Tab. 3 compares our method against the supervised results reported in [22], using the same modalities as them. Point cloud reconstruction of our model are also demonstrated in Fig. 9. We see that despite having close to 8 times less parameters, the performance is quite similar to the supervised methods but vary depending on the metric used. We suspect that this gap comes from the fact

Table 4: Performance when trained on KITTI and evaluated on DDAD compared to supervised methods taken from [22].

Method	AbsRel(↓)	RMSE (↓)	Params
DepthFormer [14]	0.644	17.083	274M
GeDepth-Adaptive [22]	**0.261**	16.132	277M
Groco	0.424	**15.366**	35M

Table 5: Performance when trained on KITTI and evaluated on DDAD for each camera.

Camera	Method	Error (↓)				Accuracy (↑)		
		AbsRel	SqRel	RMSE	$RMSE_{log}$	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Front	Baseline	0.403	5.366	14.364	0.567	0.058	0.283	0.830
	Groco	**0.154**	**1.853**	**8.588**	**0.239**	**0.760**	**0.929**	**0.975**
Back	Baseline	0.272	3.616	10.172	0.335	0.586	0.856	0.937
	Groco	**0.233**	**3.031**	**9.753**	**0.318**	**0.655**	**0.874**	**0.946**
Left	Baseline	0.353	4.031	9.424	0.375	**0.798**	**0.916**	**0.975**
	Groco	**0.256**	**2.875**	**8.656**	**0.321**	0.647	0.861	0.937
Right	Baseline	0.371	4.464	9.737	0.414	0.466	0.752	0.883
	Groco	**0.334**	**3.832**	**9.190**	**0.389**	**0.512**	**0.790**	**0.899**

that in DDAD the "ego-vehicule" is visible in images from the back and side cameras, potentially impacting models performances differently.

5 Limitations and Future Work

Our approach is designed specifically for ground-based vehicles, leveraging the ground as a critical prior. This necessitates the ground's visibility within the field of view and presupposes the existence of at least a partially flat ground, which may limit its effectiveness on uneven terrains. This limit could be alleviated by propagating the scale across time to make sure that even if the flat ground is not visible for some time, the accuracy of depth can be conserved.
Additionally, our model depends on the parameter τ, essential for successful training. Although Fig. 8 indicate that the model can adjust during inference to images with a smaller proportion of flat ground than τ, the parameter stills needs to be set manually for each dataset in the training phase.
Future work could explore strategies to relax this constraint and enhance the ground attention mechanism's recall without sacrificing precision, which is vital for maintaining accurate scale estimation.

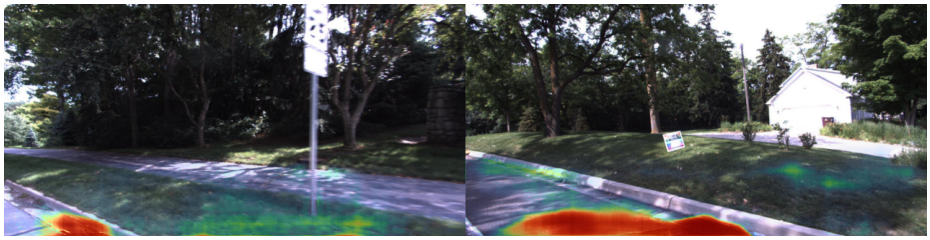

Fig. 8: Example of the attention map for the side camera. Even on an unknown dataset with the road barely visible or sideways, the segmentation of the ground stays precise, ensuring good scale.

Fig. 9: Point cloud reconstructions on the previously unseen DDAD dataset. The viewpoint of the bottom point cloud is translated 2m up and 8m back, while the side one is a top-down view. We can see that even though some artifacts are present, the overall shape and geometry of the scene is preserved.

6 Conclusion

In this study, we introduced a novel self-supervised framework, GroCo, which enhances monocular depth estimation models by leveraging ground plane constraints to address scale ambiguity. Our approach significantly improves generalization across various camera setups and datasets, demonstrating comparable performance to supervised methods. By employing advanced loss functions that facilitate the incorporation of ground attention mechanisms without dependency on annotations, GroCo achieves significant advancements in scale recovery and metric depth estimation accuracy. These results highlight GroCo's potential in advancing the development of self-supervised learning frameworks for real-world applications.

References

1. Aqel, M.O., Marhaban, M.H., Saripan, M.I., Ismail, N.B.: Review of visual odometry: types, approaches, challenges, and applications. SpringerPlus **5**, 1–26 (2016)
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
3. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems **27** (2014)
4. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013)
5. Godard, C., Aodha, O.M., Firman, M., Brostow, G.: Digging into self-supervised monocular depth estimation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3827–3837. IEEE. https://doi.org/10.1109/ICCV.2019.00393, https://ieeexplore.ieee.org/document/9009796/
6. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 270–279 (2017)
7. Guizilini, V., Ambrus, R., Burgard, W., Gaidon, A.: Sparse auxiliary networks for unified monocular depth prediction and completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11078–11088 (2021)
8. Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 3D packing for self-supervised monocular depth estimation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2482–2491. IEEE. https://doi.org/10.1109/CVPR42600.2020.00256, https://ieeexplore.ieee.org/document/9156708/
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Kinoshita, G., Nishino, K.: Camera height doesn't change: Unsupervised monocular scale-aware road-scene depth estimation. https://doi.org/10.48550/arXiv.2312.04530, http://arxiv.org/abs/2312.04530
12. Koledić, K., Petrović, L., Petrović, I., Marković, I.: GenDepth: Generalizing monocular depth estimation for arbitrary camera parameters via ground plane embedding, http://arxiv.org/abs/2312.06021
13. Lee, J.H., Han, M.K., Ko, D.W., Suh, I.H.: From big to small: Multi-scale local planar guidance for monocular depth estimation. arXiv preprint arXiv:1907.10326 (2019)
14. Li, Z., Chen, Z., Liu, X., Jiang, J.: Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. arXiv preprint arXiv:2203.14211 (2022)
15. Lyu, X., Liu, L., Wang, M., Kong, X., Liu, L., Liu, Y., Chen, X., Yuan, Y.: HR-depth: High resolution self-supervised monocular depth estimation **35**(3), 2294–2301. https://doi.org/10.1609/aaai.v35i3.16329, https://ojs.aaai.org/index.php/AAAI/article/view/16329, number: 3

16. Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., Geiger, A.: Sparsity invariant cnns. In: International Conference on 3D Vision (3DV) (2017)
17. Van Dijk, T., De Croon, G.: How do neural networks see depth in single images? In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2183–2191. IEEE. https://doi.org/10.1109/ICCV.2019.00227, https://ieeexplore.ieee.org/document/9009532/
18. Vijayanarasimhan, S., Ricco, S., Schmid, C., Sukthankar, R., Fragkiadaki, K.: SfM-net: Learning of structure and motion from video. https://doi.org/10.48550/arXiv.1704.07804, http://arxiv.org/abs/1704.07804
19. Wagstaff, B., Kelly, J.: Self-supervised scale recovery for monocular depth and egomotion estimation. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2620–2627. https://doi.org/10.1109/IROS51168.2021.9635938, https://ieeexplore.ieee.org/document/9635938, ISSN: 2153-0866
20. Xiang, J., Wang, Y., An, L., Liu, H., Wang, Z., Liu, J.: Visual attention-based self-supervised absolute depth estimation using geometric priors in autonomous driving **7**(4), 11998–12005. https://doi.org/10.1109/LRA.2022.3210298, https://ieeexplore.ieee.org/abstract/document/9904826, conference Name: IEEE Robotics and Automation Letters
21. Xue, F., Zhuo, G., Huang, Z., Fu, W., Wu, Z., Ang, M.H.: Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2330–2337. IEEE (2020)
22. Yang, X., Ma, Z., Ji, Z., Ren, Z.: GEDepth: Ground embedding for monocular depth estimation. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12673–12681. IEEE. https://doi.org/10.1109/ICCV51070.2023.01168, https://ieeexplore.ieee.org/document/10378086/
23. Zhang, N., Nex, F., Vosselman, G., Kerle, N.: Lite-mono: A lightweight CNN and transformer architecture for self-supervised monocular depth estimation. pp. 18537–18546. https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_Lite-Mono_A_Lightweight_CNN_and_Transformer_Architecture_for_Self-Supervised_Monocular_CVPR_2023_paper.html
24. Zhang, S., Zhang, J., Tao, D.: Towards scale-aware, robust, and generalizable unsupervised monocular depth estimation by integrating IMU motion dynamics. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 143–160. Lecture Notes in Computer Science, Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19839-7_9
25. Zhao, C., Zhang, Y., Poggi, M., Tosi, F., Guo, X., Zhu, Z., Huang, G., Tang, Y., Mattoccia, S.: MonoViT: Self-supervised monocular depth estimation with a vision transformer. In: 2022 International Conference on 3D Vision (3DV). pp. 668–678. https://doi.org/10.1109/3DV57658.2022.00077, https://ieeexplore.ieee.org/abstract/document/10044409, ISSN: 2475-7888
26. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. pp. 1851–1858. https://openaccess.thecvf.com/content_cvpr_2017/html/Zhou_Unsupervised_Learning_of_CVPR_2017_paper.html