PHYSICS-BASED DEPTH: A CAMERA MODEL-INFORMED MONOCULAR DEPTH ESTIMATION FOR GROUND-BASED AGENTS

by

PRAVEEN KUMAR REDDY KAMASANI

(Under the direction of Dr. Guoyu Lu)

ABSTRACT

Self-supervised monocular depth estimation, a challenging yet promising area in computer vision due to its independence from labelled data for training, traditionally grapples with the challenge of depth prediction scaled by an unknown factor due to the inherent limitations of monocular vision. Commonly, LiDAR ground truth is used to scale to absolute depth during inference, a method that indirectly relies on labeled data, limiting its practical applications. Our research introduces a novel approach to calculate the absolute depth of flat ground surfaces in images, by utilizing camera model parameters to determine the scaling factor, thereby bypassing the need for LiDAR. This approach, rigorously tested on the KITTI dataset, has shown promising results. By extending from flat ground surfaces to other image regions, our sophisticated yet computational efficiency method can significantly augment the capabilities of both self-supervised and supervised monocular depth estimation depth estimation techniques.

INDEX WORDS: Self-supervised Learning, Monocular Depth Estimation, Camera Model Parameters, Sensor Fusion, Computer Vision, Supervised Learning

PHYSICS-BASED DEPTH: A CAMERA MODEL-INFORMED MONOCULAR DEPTH ESTIMATION FOR GROUND-BASED AGENTS

by

PRAVEEN KUMAR REDDY KAMASANI

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2023

PHYSICS-BASED DEPTH: A CAMERA MODEL-INFORMED MONOCULAR DEPTH ESTIMATION FOR GROUND-BASED AGENTS

by

PRAVEEN KUMAR REDDY KAMASANI

Approved:

Major Professor:	Dr. Guoyu Lu
Committee:	Dr. Xue Iuan Wong Dr. Tianming Liu Dr. Khaled M Rasheed

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
December 2023

Physics-Based Depth: A Camera Model-Informed Monocular Depth Estimation for Ground-Based Agents

Praveen Kumar Reddy Kamasani

November 27, 2023

Acknowledgments

In the completion of this thesis, my heartfelt gratitude first and foremost goes to my advisor, Dr. Guoyu Lu, whose incredible guidance, support, and encouragement have been pivotal in this first step of my research journey. Equally, I am deeply indebted to Dr. Xue Iuan Wong, whose invaluable guidance and knowledge have been crucial in the successful execution of this project. My sincere thanks also extend to my committee members, Dr. Tianming Liu and Dr. Rasheed Khaleed, for their invaluable support and guidance throughout my master's research and coursework. I am profoundly grateful to my labmates, colleagues, and fellow MSAI students, whose camaraderie, insights, and shared experiences have not only made this journey deeply memorable but also immensely enriching. Additionally, I extend my appreciation to the Institute for Artificial Intelligence at the University of Georgia for their comprehensive support in every aspect of my graduate studies. Last but not least, my wholehearted gratitude goes to my friends and family, whose unwavering support and presence have been my strength through thick and thin.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Monocular Depth Estimation (MDE) plays a crucial role in real-world Artificial Intelligence applications by significantly enhancing AI's ability to interpret and interact with the physical environment. MDE serves as a cornerstone in robotics Wang et al. (2019); Luo et al. (2021); Tateno et al. (2017), scene understandingHazirbas et al. (2017), autonomous drivingLi et al. (2023b), augmented realityTang et al. (2022) and 3D reconstructionNewcombe et al. (2011). At its core, Monocular Depth Estimation fundamentally seeks to generate a depth map from a single RGB image by assigning precise true depth values to each pixel. But MDE is an ill-posed problem due to the inherent ambiguity of its scale because the same 2D image can be projected from infinitely many 3D scenes.

Even though, the advent of convolutional neural networks (CNN) has propelled the field to new heights Simonyan and Zisserman (2014); Szegedy et al. (2015); He et al. (2016) by exhibited prowess in MDE Eigen et al. (2014); Fu et al. (2018); Lee et al. (2019), most of the existing state-of-the-art MDE methods are based on supervised training Eigen et al. (2014); Fu et al. (2018); Ranftl et al. (2020); Bhat et al. (2021) and for supervised methods, sparse labelled ground truth collected by sensors such as LiDAR is required. Expensive data gathering and labeling processes limits data scale in supervised methods Fu et al. (2018); Bhat et al. (2021, 2022); Li et al. (2022);

Luo et al. (2018). To avoid the cost of depth labeling, researchers explored various self-supervised MDE frameworks to add the unsupervised constraints. Most of the early works in self-supervised MDEs use a regression module to estimate a pixel-wise depth map Godard et al. (2019); Gordon et al. (2019); Peng et al. (2021); Watson et al. (2019) and uses photo-metric consistency loss to train the model. However, self-supervised learning in monocular depth estimation is primarily hindered by a significant limitation: the depth predicted by the model during inference is subject to an unknown scaling factor, a consequence of the intrinsic constraints of monocular vision. Typically, this scaling factor is determined using LiDAR ground truth data during testing and inference phases to convert the depth into an absolute measure. However, this dependency on LiDAR for absolute scale calibration considerably restricts the practical utility of these methods, as it necessitates reliance on an additional sensor for labeled data during inference. This bottleneck challenges the fundamental principles of self-supervised learning and underscores the justification for adopting supervised models, which generally demonstrate superior performance, particularly given the unavoidable dependence on labeled data in the inference process.

Meanwhile, due to the convenience brought by deep neural networks, extensive information from the available sensors has been ignored. In this work, we propose a novel monocular depth estimation methodology to compute the absolute depth of the flat ground surface of the image, using camera intrinsic and extrinsic parameters, to derive scaling factor, in the place of LiDAR labelled ground truth, to convert the self-supervised predicted depth maps to absolute scale. By deriving the scaling factor from a computed ground surface depth map, self-supervised monocular depth estimation becomes fully self-reliant, relying exclusively on camera imagery and independent of LiDAR data. We name this depth estimated from the camera physics model parameters as *physics depth*. In this work, we also present a method for calculating the depth of all the objects within an image, such as vehicles, pedestrians and buildings, to create a complete dense depth map. This technique builds upon our initial depth map of the flat ground surface, derived using camera model parameters. This approach offers a robust depth prior, enhancing depth estimation accuracy both

during the training and inference phases of neural networks. Notably, our method enables effective training of both self-supervised and supervised networks without relying on specialized ground truthing equipment, thereby reducing costs and relying solely on the camera for depth supervision.

Furthermore, with the rise of ViT Dosovitskiy et al. (2020), transformer-based approaches Ranftl et al. (2021); Yang et al. (2021); Yuan et al. (2022a) are gaining traction. Notably, these methodologies predominantly focus on training models to predict depths from single images utilizing ground truth. However, the inherent ambiguity of MDE poses challenges, as MDE is theoretically an ill-posed problem, which can mainly estimate depth based on learned scenes. Due to the missing second camera or active sensors, MDE systems trained on one type of scene (e.g., outdoor) typically do not perform well on other types of scenes (e.g., indoor). Current deep learning paradigms, despite their advancements, have not addressed this underlying issue, leading to limitations in their generalization capabilities. By incorporating this dense physics-based depth, we devise unique fusion modules to amalgamate physics depth with RGB imagery, serving as inputs that synergize with networks. More crucially, this strategy can seamlessly integrate with any supervised depth estimation framework.

Even though, solving the problem of scaling factor for self-supervised monocular depth estimation techniques helps solve a big bottleneck of converting predicted depth maps to absolute scale during test and inference stages for practical applicability of self-supervised techniques, MDE by self-supervised learning still suffers from high errors: the model estimates the depth of the whole object either too far or too close, as photometric consistency and cross-frame consistency are still not direct constraints. With the computed dense physics depth map, we specifically design the training framework to interact and support unsupervised networks. More critically, this algorithm can be applied as an extension component for any unsupervised depth estimation networks.

In summary, our contributions in this work are multifold and include the following aspects: 1) We introduce a novel methodology that harnesses the camera model parameters to compute the dense depth map of flat ground surface. 2) We solved the uncertain scale issue in self-supervised

monocular depth estimation and its dependence on LiDAR ground truth by providing scaling factor from the ground surface dense physics depth map to predict absolute depth map during inference. 3) We propose a mechanism that extends the foundational flat ground surface physics depth map to full image of the scene, using semantic segmentation information, to supervise the depth estimation network while training. 4) We proposed an information fusion module that adeptly integrates physics depth into image data, yielding multi-modal features. This enriched output can subsequently feed into any supervised model, markedly enhancing depth prediction accuracy. 5) Targeting at the physics depth calculated from the camera model, we designed a neural network training framework to effectively integrate the physics depth supervision with self-supervised methods. 6) We also present a methodology to both validate and rectify the calibration results of camera orientation relative to the ground.

Chapter 2

Motivation

Depth perception through cameras is a crucial aspect of modern imaging and computer vision technologies, significantly impacting various fields such as autonomous vehicles, robotics, augmented reality, and medical imaging. This importance stems from the camera's ability to mimic the human eye's depth perception, which is vital for understanding and interacting with the three-dimensional world. In autonomous vehicles, for instance, accurate depth perception is essential for safe navigation and obstacle avoidance. It allows the vehicle to accurately gauge distances to other objects, enabling timely and appropriate responses to dynamic road situations. In robotics, depth perception enhances a robot's ability to interact with its environment, perform complex tasks, and navigate through varied terrains. For augmented reality applications, depth perception is key to creating immersive experiences, as it helps in accurately overlaying virtual objects onto the real world in a way that they appear to exist within the same space. In the medical field, depth perception in cameras aids in precise imaging, enabling better diagnosis and treatment planning. The challenge in achieving accurate depth perception lies in the complexity of replicating the human eye's ability to perceive depth, which involves interpreting various visual cues and processing them in real-time. This makes it a significant area of research and development, with advancements potentially leading to more sophisticated, safe, and interactive technology applications.

2.1 Human Depth Perception

Human depth perception is a complex process that allows us to perceive the three-dimensional structure of the world around us. This ability primarily stems from binocular vision, where our two eyes, positioned slightly apart, capture slightly different images. The brain then merges these images, a process known as binocular disparity, to create a perception of depth. This stereoscopic vision is crucial for judging distances accurately. Additionally, our depth perception is enhanced by several monocular cues. These include relative size (objects appearing smaller when they are farther away), texture gradient (the density of texture increasing with distance), interposition (objects blocking part of another object are perceived as closer), linear perspective (parallel lines appearing to converge in the distance), and motion parallax (objects closer to us moving faster across our field of vision than those further away). Another important monocular cue is relative height, where objects positioned higher in our field of vision are perceived as being farther away, while those lower appear closer. This cue is particularly effective in landscapes where the ground provides a consistent reference level. Our ability to perceive depth is also influenced by our past experiences and knowledge of the size of familiar objects, which helps the brain interpret size and distance. The integration of these binocular and monocular cues, along with cognitive processing, enables us to navigate and interact effectively with our environment.

2.2 Relative Height Monocular Depth Estimation Cue

The relative height cue, a monocular depth cue, is based on the observation that objects positioned closer to the horizon line in our visual field appear farther away, while objects farther from the horizon appear closer Dunn et al. (1965). The relative height cue can be seen demonstrated in Fig. 2.1, where we can state that point *a* is the closest, point *b* is the second closest and point *c* is the farthest on the ground, and similarly point *1* is the closest, point *2* is the second closest and point *3* is the farthest on the sky based on their distance from the person who have taken the picture.

Humans and animals use relative height to navigate through environments, avoiding obstacles, and determining safe paths. In various fields such as painting, drawing, photography, virtual reality environments, and video gaming, the concept of relative height cue is utilized extensively to create a convincing three-dimensional space on a two-dimensional screen. It plays a significant role in how we perceive the layout of the terrain and can be a critical factor in judging distances in both urban and rural landscapes. Algorithms designed to interpret visual data can use the relative height cue to estimate the distance of objects within an image, which is crucial for applications like autonomous vehicles and robotic navigation. While relative height is a powerful depth cue, it's not without limitations. It relies heavily on the presence of a ground plane and a clear horizon line to be effective. In the absence of these, or if the ground plane is tilted or the horizon is obscured Gardner et al. (2010), the cue becomes less reliable. Furthermore, the effectiveness of this cue tends to decrease as the viewing distance increasesSurdick et al. (1997). Despite its simplicity, the cue is an integral part of a sophisticated system that humans use to interact with complex environments.

2.3 Why Not Use Additional Sensors for Depth Estimation?

Opting for cameras over additional sensors like LiDAR for depth estimation presents a more advantageous alternative, especially in the context of mass-produced products. This preference stems from several key factors. Firstly, LiDAR, though reliable, incurs a significantly higher cost than camera systems, making it less viable for mass production. Cameras, utilizing advanced techniques like stereo vision, can generate denser and more detailed depth maps, offering a more comprehensive environmental understanding than the sparser data produced by LiDAR. Furthermore, incorporating multiple sensor types increases the complexity of both physical integration and data processing, with each sensor introducing unique data characteristics and formats. This necessitates more complex algorithms for effective data fusion and interpretation. The rapid advancements in camera technology, coupled with breakthroughs in computer vision algorithms, have significantly

Figure 2.1: Picture Demonstrating the Relative Height Depth Cue:

A scenic view of Starship preparing to launch at Boca Chica,TX. Source: @SpaceX on Twitter

enhanced the capability of cameras in accurate depth estimation. Modern techniques leveraging machine learning and neural networks have further bolstered camera performance, rendering them a technologically sound and cost-effective option. Additional sensors can also lead to increased power consumption and device bulkiness, a critical consideration in portable or compact devices. In conclusion, while LiDAR and similar sensors have their benefits, cameras are often the more suitable choice for depth estimation in a wide range of applications due to their cost-efficiency, data richness, reduced complexity, and technological advancements.

Chapter 3

Related Work

3.1 Monocular Depth Estimation Network Architecture

Monocular depth estimation performance varies significantly across different architectures. Yin and Shi (2018) transitioned from the VGG encoder to a ResNet encoder. Guizilini et al. (2020) introduced 3D convolutions in PackNet, aiming to efficiently compress and decompress features while preserving details. To fuse multi-scale features, Wang et al. (2020) employed attention mechanisms. Recognizing the inherent limitations of CNNs, Zhou et al. (2021) integrated HRNet for self-supervised monocular depth estimation, capitalizing on HRNet's Wang et al. (2020) prowess in modeling multiscale features. In the latest time, Transformer is also developed for depth estimation Li et al. (2023c).

3.2 Supervised Depth Estimation

In the field of Monocular Depth Estimation (MDE) utilizing neural networks, the groundbreaking research by Eigen et al. (2014) stands as a cornerstone contribution. Their pioneering work introduced a coarse-to-fine convolution neural network alongside a scale-invariant loss function. Sub-

sequently, MDE has garnered considerable attention, focusing on enhancing performance through increasingly sophisticated architectures and loss functions Laina et al. (2016); Lee et al. (2018); Liu et al. (2015); Miangoleh et al. (2021). Some scholars have redefined the challenge as an ordinal regression task. Currently, supervised learning in monocular depth estimation is primarily divided into two methodologies: a pixel-wise regression approach Eigen et al. (2014); Zhao et al. (2021); Ranftl et al. (2021); Huynh et al. (2020) and a pixel-wise classification framework Fu et al. (2018); Diaz and Marathe (2019). While regression methods facilitate the prediction of continuous depths, they pose optimization challenges. In contrast, classification methods enable discrete depth predictions and are comparatively more straightforward to optimize. In the decoding stage, Lee et al. (2019) introduced a local planar guidance layer to infer the plane coefficients, which were used to recover the full depth of resolution of the map. More recently, Adabins Bhat et al. (2021) has applied transformer to estimate the depth.

3.3 Self-supervised Depth Estimation

Self-supervised depth estimation from monocular videos or stereo image pairs is gaining prominence, particularly due to the challenges in obtaining accurate ground truth. In the domain of monocular depth, Zhou et al. (2017) spearheaded a self-supervised framework by jointly training depth and pose networks based on image reconstruction loss. Godard et al. (2019) further advanced this field by introducing a minimum re-projection loss and auto-masking loss, setting a new benchmark.

Building upon the groundwork of Newcombe Newcombe et al. (2011), subsequent studies like those by Guizilini et al. (2020) and Chawla et al. (2021) addressed the scale ambiguity inherent in monocular Structure-from-Motion (SfM) methods. They achieved this by integrating real-time data sources such as GPS or camera velocity. These self-supervised models hinge on the photometric consistency of the re-projection, as noted by Wang Wang et al. (2004). In stereo training

contexts, models use synchronous stereo image pairs to predict disparity Scharstein and Szeliski (2002), which inversely relates to depth. Since the relative camera pose is known in stereo setups, the primary task of these models is disparity map prediction. Garg et al. (2016) pioneered this approach, training a self-supervised monodepth model using stereo pairs and a photometric consistency loss. This methodology was further refined by Godard et al. (2017) with additional constraints like left-right consistency. Moreover, Garg et al. (2020) extended this to predict continuous disparity values. While stereo views inherently provide an absolute depth scale, current monocular self-supervised models predict only relative depth scales, requiring ground truth for scale calibration. By leveraging physics depth data from ground surfaces and road-connecting regions, these unsupervised models can be enhanced to predict absolute depth scales, thus increasing accuracy for datasets like KITTI.

3.4 Geometric Priors

Geometric priors are gaining increasing prominence in the realm of monocular depth estimation. Traditional multi-view stereo methods, as highlighted by Gallup et al. (2010), predominantly emphasize optimization efficiency. In the domain of self-supervised learning for monocular depth estimation, multi-view geometry plays a crucial role. It facilitates the warping of images from the source to the target viewpoint, thus creating the reprojection error, which serves as the loss function for the depth estimation network Godard et al. (2019). In addition to photometric consistency, geometric consistency is also being increasingly leveraged, particularly in the context of point cloud comparisons Mahjourian et al. (2018); Hirose et al. (2021).

Another widely utilized geometric prior is the surface normal constraint, as discussed in works like those by Kusupati Kusupati et al. (2020) and Long Long et al. (2021). This constraint ensures the alignment of normal vectors deduced from both estimated and ground-truth depths. However, it's important to acknowledge the limitations of the planarity prior, notably its tendency to yield

inaccurate depth estimations in regions exhibiting high curvature. The piecewise planarity prior Chauve et al. (2010); Gallup et al. (2010); Bódis-Szomorú et al. (2014) offers a tangible approximation to real-world scenarios. This prior segments the scene into 3D planes Yang and Zhou (2018); Zhang et al. (2020), aiming to categorize the scene into dominant depth planes.

Despite the inherent ambiguity in monocular depth estimation, contemporary supervised learning paradigms remain predominantly grounded on truth labels. Even as novel architectures like the Transformer enhance prediction accuracy, they do not address the foundational challenges associated with monocular depth estimation errors. While geometric priors can mitigate some uncertainty, their contributions to the overarching problem remain marginal. Diverging from traditional geometric priors, we leverage camera model parameters to compute scene depth. This approach furnishes more precise and generalizable depth predictions, largely bolstering model performance.

Chapter 4

KITTI Dataset

The KITTI dataset(Geiger et al., 2013), short for the Karlsruhe Institute of Technology and Toyota Technological Institute dataset, is a well-known benchmark dataset in the field of computer vision, particularly focusing on applications in autonomous driving. It was created by researchers from the Karlsruhe Institute of Technology and the Toyota Technological Institute in Chicago. KITTI dataset was captured using a VW station wagon, which can be seen in Fig. 4.1, equipped with various sensor modalities, including high-resolution color and grayscale stereo cameras, a Velodyne 3D laser scanner and a high-precision GPS/IMU inertial navigation system. The KITTI Dataset includes two color cameras with a resolution of 1.4 megapixels and consisting of approximately 50,000 color images from each camera. The data encompasses a wide range of real-world traffic scenarios, from freeways and rural areas to inner-city scenes, featuring both static and dynamic objects. The dataset is notable for its calibration, synchronization, and timestamping, providing both rectified and raw image sequences. The KITTI dataset was recorded while driving in and around Karlsruhe, Germany, Germany in 2011. The dataset aims to advance the development of computer vision and robotic algorithms for autonomous driving. The sensor setup includes two stereo camera rigs (one for grayscale and one for color) with a baseline of approximately 54 cm.

KITTI data is collected with diverse set of sensors, which can be seen in Fig. 4.2 , with the

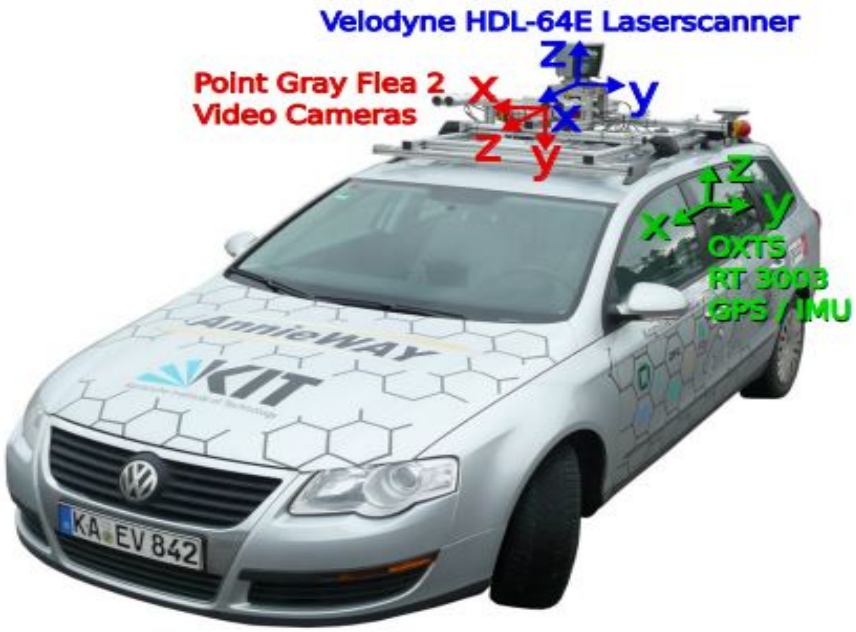

Figure 4.1: KITTI Recording Platform:
VW Passat station wagon is equipped with four video cameras (two color and two grayscale cameras), a rotating 3D laser scanner and a combined GPS/IMU inertial navigation system.

following capabilities: a) 2 × PointGray Flea2 grayscale cameras (FL2-14S3M-C), 1.4 Megapixels, 1/2" Sony ICX267 CCD, global shutter. b) 2 × PointGray Flea2 color cameras (FL2-14S3C-C), 1.4 Megapixels, 1/2" Sony ICX267 CCD, global shutter. c) 4 × Edmund Optics lenses, 4mm, opening angle $\sim 90\circ$ vertical opening angle of region of interest (ROI) $\sim 35\circ$. d) 1 × Velodyne HDL-64E rotating 3D laser scanner, 10 Hz, 64 beams, 0.09-degree angular resolution, 2 cm distance accuracy, collecting $\sim$ 1.3 million points/second, field of view: $360\circ$ horizontal, $26.8\circ$ vertical, range: 120 m

The KITTI dataset is widely acclaimed for its diverse range of data, as seen in 4.3. In the fig. 4.4, you'll find a sample scene presented in three distinct imaging formats: RGB, which offers full-color imagery from right and left RGB camera; grayscale, which provides a monochromatic view from the left and right grayscale camera; and in fig. 4.5, you will find a Velodyne image, which offering a unique 3D perspective of the scene's depth and structure.

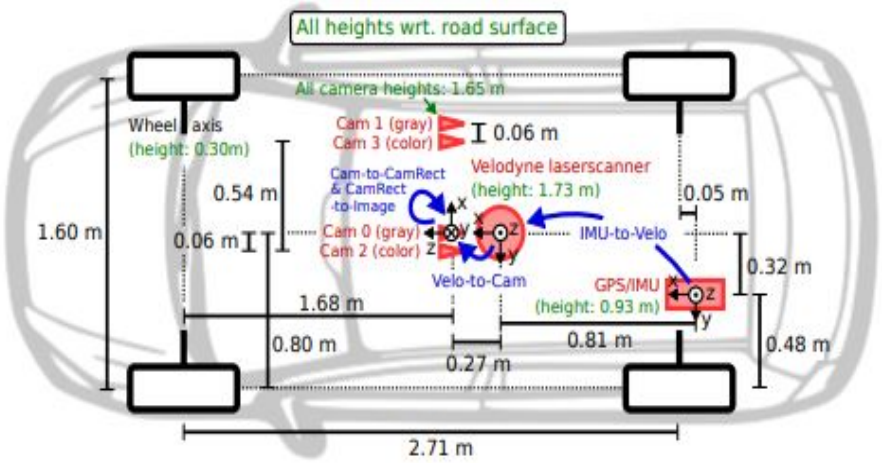

Figure 4.2: KITTI Sensor Setup:

This figure illustrates the dimensions and mounting positions of the sensors (red) with respect to the vehicle body. Heights above ground are marked in green and measured with respect to the road surface. Transformations between sensors are shown in blue.

Figure 4.3: Sample RGB Images from KITTI:

This figure demonstrates the diversity of KITTI dataset. The left color camera image is shown.

Figure 4.4: Sample Camera Images from KITTI:

0000000000.png of 2011_09_26_drive_0001_sync

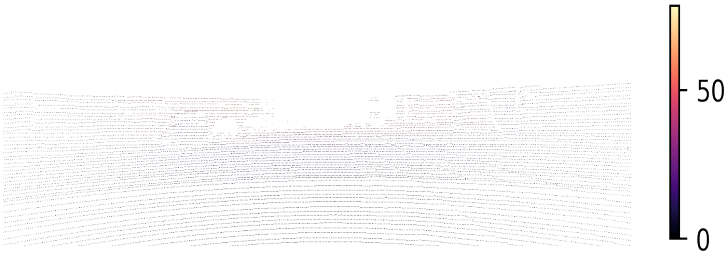

Figure 4.5: Sample Velodyne Depth Map from KITTI:

Sparse LiDAR depth map of the above RGB images

Chapter 5

Physics Depth Methodology

In this work, we introduce a novel monocular depth estimation methodology to compute the absolute depth using camera intrinsic and extrinsic parameters, accompanied by semantic segmentation information of the image. We termed it *physics depth*, as we are using the fundamental physics principles to calculate the depth of the flat surface in the camera's field of view by adopting the biological monocular relative height depth cue [1] on to camera systems. In this methodology, the procedure begins by creating a physics-based depth map across the camera's entire field of vision, based on the assumption that camera's field of view is on a perfectly flat surface, as depicted in 5.1, we name the depth map generated *full physics depth map*. Subsequently, we identify the actual flat surface areas in the image using semantic segmentation information to segment the region of full physics depth map, where the actual flat surface is present. We use this flat surface physics depth map to calculate the *scaling factor* for converting self-supervised monocular depth maps to absolute scale during inference. The depth data from these flat surfaces is then extrapolated to adjacent ground and vertical surfaces. Finally, we employ image inpainting techniques to address any remaining gaps in the depth map. The efficacy of our method was validated using the KITTI Geiger et al. (2013) dataset, with results showing a close alignment in accuracy with LiDAR-derived depth measurements, especially for proximal flat surfaces.

Figure 5.1: Sample Scene with Perfectly Flat Surface:

Source: DALL·E 2023-11-16 05.08.36 - An image showcasing a flat surface on the bottom half, with a road leading towards the horizon at sunset.

5.1 Physics Depth Map for Full Field of view

The transformation of a three-dimensional point from the world coordinate system (x_w, y_w, z_w) to the camera coordinate system (x_c, y_c, z_c), followed by its projection from the camera coordinate system to the two-dimensional image plane (u, v), can be accurately described using the following linear camera model equation 5.1:

$$z_c \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} K & 0 \end{bmatrix} \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{5.1}$$

In this context, $\mathbf{K}$ denotes the camera's intrinsic matrix, while $\mathbf{R}$ represents the rotation ma-

trix, and $\mathbf{T}$ is the translation vector, collectively comprising the camera's extrinsic parameters. Substituting $\mathbf{K}$, $\mathbf{R}$, and $\mathbf{T}$ into Equation 5.1 yields the following equation:

$$Z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & O_x \\ 0 & f_y & O_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} & x_{13} & t_x \\ x_{21} & x_{22} & x_{23} & t_y \\ x_{31} & x_{32} & x_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \tag{5.2}$$

In this framework, u, v represent the pixel coordinates on the image plane, where the origin of the coordinate system is located at the optical center (O_x, O_y) of the image, often referred to as the principal point. The terms f_x, f_y denote the camera's focal lengths along the x and y axes, respectively.

Lets say the product of camera's intrinsic and extrinsic matrices are represented by A as shown below:

$$A = \begin{bmatrix} K & 0 \end{bmatrix} \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix} \tag{5.3}$$

By substituting A in Equation 5.2,

$$\begin{aligned} z_c u &= a_{11}x_w + a_{12}y_w + a_{13}z_w + a_{14} \\ z_c v &= a_{21}x_w + a_{22}y_w + a_{23}z_w + a_{24} \\ 1 &= a_{31}x_w + a_{32}y_w + a_{33}z_w + a_{34} \end{aligned} \tag{5.4}$$

If we know that the height of the world coordinate system above the ground is h, and assuming that the Y-axis of the world coordinate system points towards the ground, then $y_w = h$. By Substituting $y_w = h$ in equation 5.4, we solve for x_w, z_w, and z_c. Utilizing the world coordinate system (x_w, y_w, z_w), we can derive the precise camera coordinates (x_c, y_c, z_c) using the following equation:

$$\begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} = \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \tag{5.5}$$

From equation 5.5, we can derive camera coordinates (x_c, y_c, z_c) for pixel (u, v).

Also by substituting equation 5.5 in equation 5.1, we get

$$Z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & O_x \\ 0 & f_y & O_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} \tag{5.6}$$

From equation 5.6, if the height of the camera in the camera coordinate system, denoted as $y_c = h$, is known, we can directly solve for x_c and z_c. Utilizing the coordinates (x_c, y_c, z_c), we generate a 3D point cloud representing a flat surface and calculate the point-to-point distance from the camera's center, (0,0,0), to a point on the ground. This process enables us to compute the complete physics depth map, assuming that the entire field of view of the camera encompasses a perfectly flat surface as seen in 5.2. We then identify the actual flat surface areas in the image via semantic segmentation, isolating the regions where the flat surface is present, roads in the case of KITTI, and we named it *Road Physics Depth Map* as demonstrated in 5.3. Our method was evaluated using the KITTI Geiger et al. (2013) with detailed results presented in the 'Experiments' section of this paper.

5.2 Extension of Ground Physics Depth

In our evaluations,the *Road Physics Depth Map* aligns closely with LiDAR measurements for flat surfaces in front of the camera, providing a dense depth map as opposed to the sparser LiDAR counterpart.

Figure 5.2: Sample Full Physics Depth Map:

A physics-based depth map of a sample KITTI image, calculated under the assumption that the camera's entire field of view is focused on a perfectly flat surface.

However, this approach predominantly targets the flat surface in front of the camera, which could lead to overfitting to road regions when training a depth prediction model, restricting its applicability during training. To mitigate this overfitting risk, we broadened the scope of our physics-based depth method to cover the entire image by beginning with applying the physics depth logic to surfaces that are almost at the level of camera base and flat —encompassing road, sidewalks, parking lots, rail tracks, and more—by presuming a uniform flatness over these terrains and we call it *Ground Physics Depth Map* as demonstrated in 5.4.

Furthermore, we extrapolated physics depth to vertical entities that are in contact with the flat surface like vehicles, pedestrians, and buildings by propagating depth values vertically from the intersection of the horizontal and vertical structures and we call it *Edge Extended Physics Depth Map* as demonstrated in 5.5.

After vertical extension of the physics depth, some objects have partial depth maps as only part of the structure was in contact with the horizontal surface, we filled the missing depth map of those objects using Telea Inpainting Technique Telea (2004) by making use of available depth maps. In this context, Telea Inpainting proved to be the optimal choice due to its incorporation of

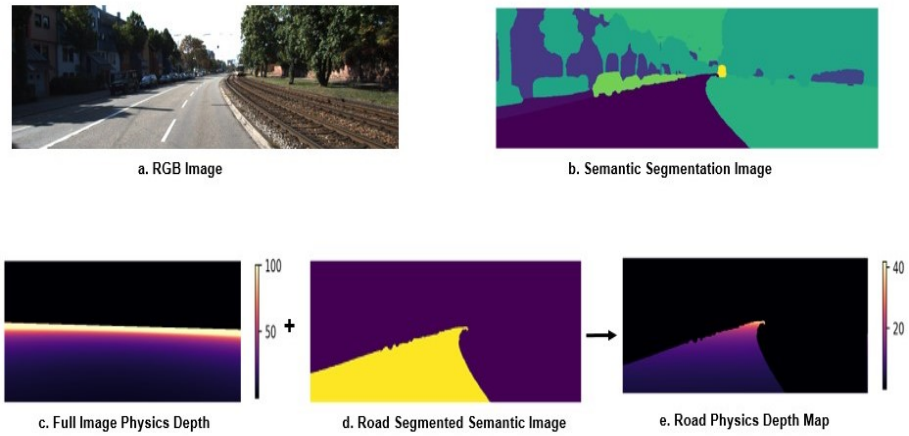

Figure 5.3: End-to-end Road Physics Depth methodology demonstrated on a sample image:

(a) RGB image (b) semantic segmented image (c) full physics depth map along with scale (d) road segmented from semantic segmented image (e) physics depth map of road, along with scale

both the directional rate of change and the geometric distance from adjacent pixels during depth propagation. For objects that are not in contact with ground, we filled their depth by extending the depth of the objects that are in-between the object and the ground. Finally, we fill the sky with 1.5 times the maximum depth of the in pained depth map to create *Dense Physics Depth Map* without any gaps as demonstrated in 5.6. The detailed results of these experiments are presented in the Experiments section of this paper.

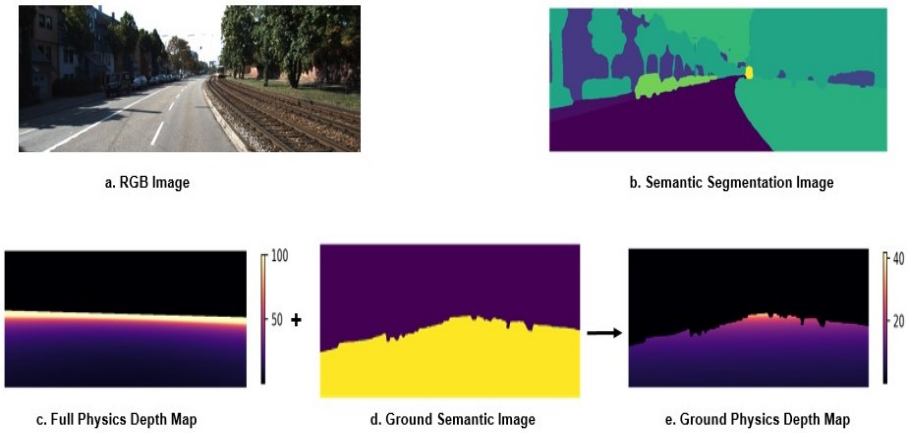

Figure 5.4: End-to-end Ground Physics Depth methodology demonstrated on a sample image:

(a) RGB image (b) semantic segmented image (c) full physics depth map along with scale (d) Ground segmented from semantic segmented image (e) physics depth map of Ground, along with scale

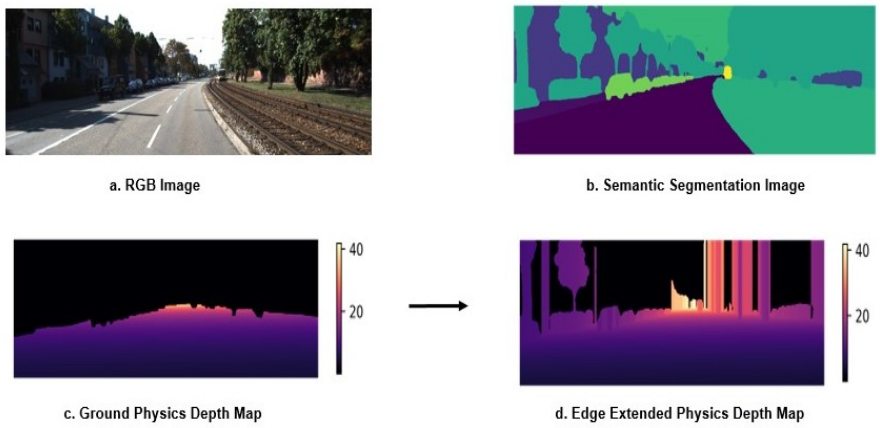

Figure 5.5: End-to-end Edge Extended Physics Depth methodology demonstrated on a sample image:

(a) RGB image (b) semantic segmented image (c) ground physics depth map along with scale (d) edge extended physics depth map, along with scale

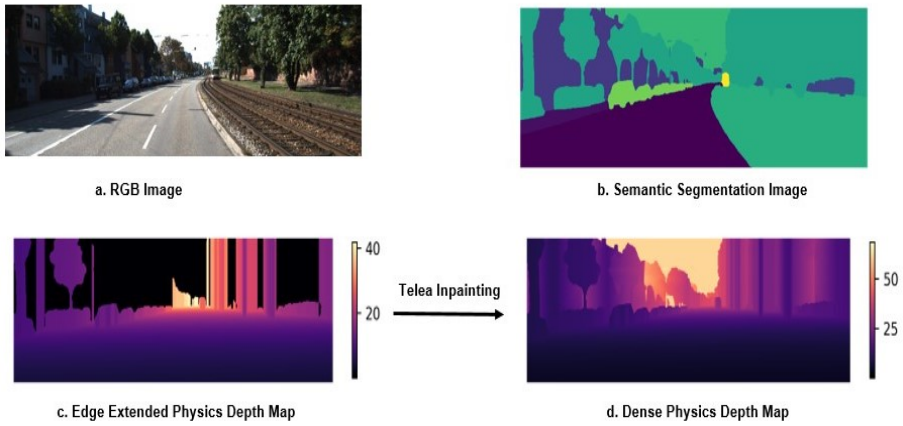

Figure 5.6: End-to-end Dense Physics Depth methodology demonstrated on a sample image:

(a) RGB image (b) semantic segmented image (c) edge extended physics depth map along with scale (d) physics depth map of complete image, along with scale

Chapter 6

Physics Depth in Supervised Monocular Depth Estimation

6.1 Fusion Unified Framework

In traditional supervised learning for depth estimation, RGB images serve as the sole input, with models predicting depth for each pixel. Monocular depth estimation faces a fundamental challenge: a single point may correspond to an indeterminate number of depths, making accurate depth prediction inherently difficult. Current models focus on enhancing network architectures and increasing model complexity, yet these improvements do not fundamentally resolve the accuracy limitations of monocular depth estimation. Our approach introduces the concept of 'physics depth,' calculating precise depths for the road and its immediate surroundings. By integrating physics depth as prior knowledge into the model, it not only provides accurate depth for road sections but also aids the network in learning depths for other areas through physical depth cues.

This chapter presents an advanced Information Fusion module, synergistically combining physics-based depth data with RGB imagery, as depicted in Fig. 6.1. Comprising Multi-Head Relational Attention (MHRA), Depth Information Selection, and a Supervision Model, this module leverages

the complementary strengths of both data types to enhance our supervised learning framework for monocular depth estimation.

Inspired by the UniFormer's spatio-temporal network structure Li et al. (2023a), as shown in Fig. 6.1, our Fusion Unified framework extracts and integrates features from both RGB and depth images. We acknowledge that while physics-based depth offers partial information, it is interconnected with other image regions. Learning this mapping allows our network to more effectively address monocular depth estimation tasks, benefiting from the physics-based depth.

Our Fusion Unified framework consists of five key modules: Dynamic Position Embedding (DPE), global and local Multi-Head Relation Aggregator (MHRA), Physics Depth Selection (PDS), and the Supervision model. The PDS module, given the accuracy of road depth data, incorporates an adaptive weighting mechanism for physics depth features, focusing on feature emphasis and overfitting mitigation.

Multi-Head Relational Attention

The MHRA module aims to capture both global and local details from the physics depth and RGB images. To balance the precision of road depth data and mitigate over-reliance on physics depth, we introduce a weight matrix in the PDS module. This matrix discerns which depth aspects to emphasize or downplay. As the network learns these weight matrices, it can adaptively integrate depth information, avoiding dependency solely on depth features.

Global and Local Multi-Head Relational Attention

MHRA is designed to extract comprehensive information from the physics depth and RGB images. CNN structures capture local details, while Transformer architectures, thanks to their attention mechanism, extract global information. In our framework, local MHRA focuses on extracting details from RGB images, which are rich in nuances. In contrast, global MHRA captures broad trends within the physics depth data, revealing scene structures.

The mathematical formulation is as follows:

$$X_{RGB} = DPE(x_{rgb}) + x_{rgb}, \quad X_{Depth} = DPE(x_{depth}) + x_{depth} \tag{6.1}$$

$$Y_{RGB} = MHRA_{local}(\mathrm{Norm}_{\mathrm{BN}}(X_{RGB})) + X_{RGB} \tag{6.2}$$

$$Y_{Depth} = MHRA_{global}(\mathrm{Norm}_{\mathrm{LN}}(X_{Depth})) + X_{Depth} \tag{6.3}$$

MHRA evaluates token relationships in a multi-head format:

$$R_n(X) = A_n V_n(X) \tag{6.4}$$

$$MHRA(X) = \mathrm{Concat}(R_1(X), R_2(X), \ldots, R_N(X))U \tag{6.5}$$

Here, $R_n(\cdot)$ represents the n^{th} head, and U is a learnable parameter matrix for integrating N heads. $V_n(X)$ is a linear transformation of original tokens. A_n learns token affinity with local and global modes. For RGB images, local information extraction is prioritized:

$$A_n^{\mathrm{local}}\left(X_i, X_j\right) = a_n^{i-j} \tag{6.6}$$

a_n is a learnable parameter, and $(i-j)$ indicates the relative position between tokens i and j. The depth image is processed to extract the global relationship. $Q_n\left(\cdot\right)$ and $K_n\left(\cdot\right)$ are distinct linear transformations.

$$A_n^{global}\left(X_i, X_j\right) = \frac{e^{Q_n(X_i)^T K_n\left(X_j\right)}}{\sum_{j' \in \Omega_{H \times W}} e^{Q_n(X_i)^T K_n\left(X_{j'}\right)}} \tag{6.7}$$

”

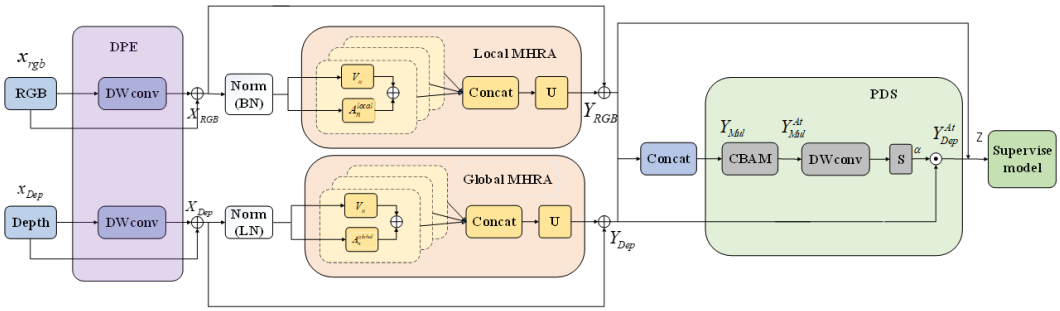

Figure 6.1: Information Fusion module consists of four key modules, Dynamic Position Embedding (DPE), global and local Multi-Head Relation Aggregator (MHRA), Physics Depth Selection, and the Supervision Model.

6.2 Physics Depth Selection

In this study, we address the precision of road data depth information derived from meticulously calculated camera parameters. The input for our module includes the road data, the filled depth information representing physical depth, and the original RGB image. Although road data contributes to the depth values' precision, the process to fill the environment with this data exhibits certain inaccuracies. To mitigate these, we utilize a deep neural network to refine and enhance the depth information, which largely retains its accuracy, serving as a global reference.

Our approach prioritizes RGB information to augment the physical depth data, addressing any inconsistencies during the filling process. Emphasis is placed on extracting RGB texture details and analyzing color space distribution to refine the output.

In the proposed Perceptual Depth Synthesis (PDS) module, depth features Y_{Dep} and image features Y_{RGB} are concatenated to form multimodal features Y_{Mul}. These features are then further refined using the Convolutional Block Attention Module (CBAM), which focuses on the "what" and "where" aspects in the channel and spatial axes. The CBAM module, as described by Woo et al. (2018) Woo et al. (2018), sequentially applies channel and spatial attention modules, thereby efficiently adjusting network features.

The mathematical representation of the PDS module's process is as follows:

$$Y_{Mul} = Y_{RGB} + Y_{Dep} \tag{6.8}$$

The CBAM's functionality is detailed further:

$$F' = M_c(F) \otimes F, \quad F'' = M_s(F') \otimes F' \tag{6.9}$$

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{6.10}$$

$$M_s(F) = \sigma(f^{7\times7}(AvgPool(F); MaxPool(F))) \tag{6.11}$$

In these equations, $\otimes$ denotes element-wise multiplication, and σ is the sigmoid function. The MLP represents a multi-layer perceptron with shared weights W_0 and W_1. A convolution operation with a 7×7 filter size is denoted by $f^{7\times7}$.

The output, Y_{Mul}^{At}, significantly impacts the extent to which physical depth information is integrated into the network. We utilize depthwise separable convolution, coupled with the sigmoid function, to learn the weight matrix associated with the physical depth feature α. This process selectively filters physical depth features and adapts the integration level into the network. The relevant formulas are:

$$Y_{Mul}^{At} = CBAM(Y_{Mul}), \quad \alpha = \sigma(DW(Y_{Mul}^{At})) \tag{6.12}$$

$$CBAM(F) = M_s(M_c(F) \otimes F) \otimes (M_c(F) \otimes F) \tag{6.13}$$

Finally, the multimodal features Z, enriched with key physical depth information Y_{Dep}^{At}, are generated by combining Y_{Dep}^{At} with Y_{RGB}:

$$Y_{Dep}^{At} = \alpha \cdot Y_{Dep}, \quad Z = Y_{Dep}^{At} + Y_{RGB} \tag{6.14}$$

These multimodal features Z serve as inputs for contemporary monocular depth estimation models to enhance image depth estimation.

6.3 Supervision module

Our framework facilitates the integration of any existing supervised learning model by utilizing the output features from the Information Fusion Module, thus obviating the need for structural alterations. Furthermore, we introduce a novel smoothing loss function aimed at bolstering the accuracy of monocular depth prediction while circumventing the introduction of new errors.

RGB images function as a global prior in this study, applying smoothing constraints to the depth map and ensuring uniformity in depth or surface normals. These constraints are crucial for addressing the noise and discontinuities commonly present in depth maps or surface normals, which may arise from variable illumination, material differences, occlusions, motion blur, among other factors. The application of a smoothing constraint serves to attenuate these imperfections, maintaining the integrity of fine details.

The smoothing loss function is delineated as follows:

$$L_s = \frac{1}{N} \sum_{i,j} \left(|\partial_x d_{i,j}| \, e^{-\|\partial_x I_{i,j}\|} + |\partial_y d_{i,j}| \, e^{-\|\partial_y I_{i,j}\|} \right) \tag{6.15}$$

By implementing an L1 penalty on the depth gradients, denoted by ∂d, we promote the depth map's local smoothness. We enhance this loss function by introducing an edge-aware weighting factor that considers the RGB image gradients, ∂I, to effectively manage the depth discontinuities that typically correspond with image gradients, as observed by Godard et al. (2017).

6.4 End Note

In this chapter, we introduce a novel physics-based supervised learning approach for depth estimation. Contrary to existing supervised methods that primarily rely on complex network architectures, unique geometric priors, and extensive data augmentations for incremental improvements in performance, our method strategically employs physics-based scene depth analysis for more precise depth prediction. This advancement significantly enhances the evaluation metrics in monocular depth estimation tasks.

Our approach capitalizes on the potential of physics-based depth estimation, which is derived from detailed camera modeling. By doing so, it notably improves the performance of supervised learning models, particularly in accurately discerning ground and environmental features. The core objective is to empower these models to inherently predict depth values with high accuracy, utilizing insights gained from physics-based depth analysis.

This methodology is not only straightforward in its deployment but also serves as a fundamental tool to substantially boost the depth prediction capabilities of various supervised models. It represents a shift from conventional techniques, offering a more direct and effective path to enhanced depth estimation in a wide range of applications.

Chapter 7

Physics Depth in Self-supervised Monocular Depth Estimation

7.1 Physics Depth Scaling Factor

Our model innovatively commences with physics depth to address the pervasive scale challenges in single-view depth estimation. Typically, single-view models grapple with scale accuracy, constrained to estimating depth in relation to other points within the scene. In contrast, our approach capitalizes on physics depth as a foundational element during the training phase. This inclusion allows our model to autonomously adjust and correct depth prediction errors, crucially preserving the integrity of scale measurements.

By integrating physics depth, our model gains a significant advantage: it anchors depth estimations to real-world measurements, rather than relying solely on relative depth cues. This results in a more reliable and accurate depth perception, particularly important in applications where scale precision is paramount. The ability of our model to maintain consistent scale across various depths and viewpoints marks a substantial improvement over traditional single-view methods. Our approach, therefore, not only addresses but effectively overcomes the inherent limitations of scale

inconsistency typically associated with single-view depth estimation.

This advancement in handling scale issues opens new possibilities for single-view depth estimation applications. It allows for more accurate 3D reconstructions, enhanced object detection, and improved navigation capabilities in autonomous systems. By leveraging physics depth, our model sets a new standard in depth estimation accuracy, proving especially beneficial in scenarios where precise depth information is critical for decision-making or analysis.

7.2 Self-supervised Network Architecture

In the realm of self-supervised learning for depth estimation, traditional methods typically employ the source image to reconstruct the target image. The model's training hinges on the discrepancy between the target and reconstructed images, utilizing this difference as the loss function. A pivotal challenge in this context is the integration of physical depth data as a prior, particularly given the limited extent of physics depth data, which only represents a small segment of the overall image.

Our research addresses this challenge by adopting physics depth as a foundational prior. This approach aids the self-supervised model in refining estimations of ground and environmental depths. This methodology stands in stark contrast to conventional self-supervised models, which often commence with arbitrary initial values. By initiating with physics depth, our model gains a more accurate starting point, making the correction of depth inaccuracies during self-supervised training markedly more efficient compared to training from scratch.

The entire framework for physics depth computation and training with self-supervised network is shown in Fig. 7.1.

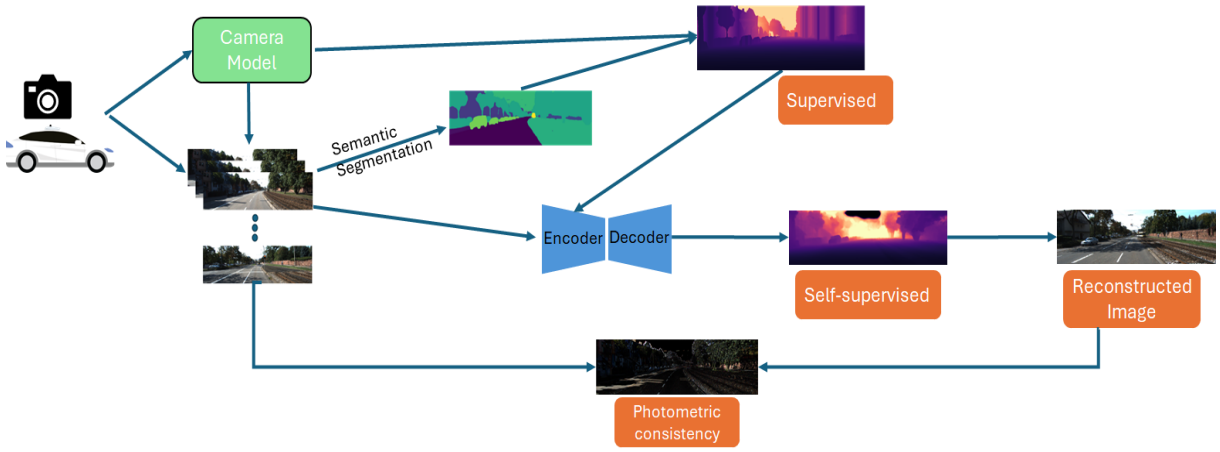

Figure 7.1: The entire framework for the computation of physics depth based on camera model and the application of physics depth for neural network supervision through interaction with general self-supervised neural networks.

7.3 Physics-Depth Supervision

Our methodology initiates the depth network training by utilizing physics depth as a foundational guide. This strategy equips the model with an initial understanding of various areas' depths, thus improving its self-driven depth prediction capabilities, particularly in regions with sparse or uncertain data.

We begin by employing physics depth as a reference during the training phase. This initial step grants the model a fundamental grasp of depth across different areas, enhancing its autonomous prediction ability. Notably, the accuracy of physics depth varies across regions; it tends to be more reliable in and around roadways than in other areas. Consequently, we assess the reliability of physics depth in diverse locations and prioritize regions with higher depth certainty during the training process. This approach bolsters the precision of our self-learning methods by leveraging more dependable starting data.

A crucial aspect of our methodology is evaluating the trustworthiness of physics depth data through self-learning. This involves utilizing a network designed to discern scene positional changes between consecutive frames, aiding in the more effective application of physics depth.

The mathematical representation is as follows:

$$\begin{aligned} I_{t-1\to t} &= I_{t-1}\left\langle proj\left(D_t, T_{t\to t-1}\right)\right\rangle, \\ con_{i,j} &= \text{confidence}\left(I_t, I_{t-1\to t}\right), \end{aligned} \tag{7.1}$$

where $I_{t-1\to t}$ is the reconstructed image at time t using the image at time $t-1$, with $\langle proj\left(\cdot\right)\rangle$ depicting the reconstruction function based on depth and pose networks. The function confidence $(\cdot)$ computes the confidence level for each physics depth, using I_t and the reconstructed $I_{t-1\to t}$.

In our study, the model prioritizes pixels with higher reliability in depth data, thereby enhancing depth estimation accuracy. Conversely, it allocates less emphasis to areas with lower confidence depths. This approach minimizes errors in high-confidence areas while accepting some imprecision in less reliable regions. To quantify the discrepancy between the model's predicted depths and actual depths, we utilize the L2 loss function, also known as Mean Squared Error (MSE). Incorporating the confidence level into the loss computation ensures the model's performance aligns with the reliability of the depth data.

The corresponding formula is:

$$L_{phy} = \sum_{i=1}^{M}\sum_{j=1}^{N} con_{ij}\cdot\left(d_{ij}^{phy} - \hat{d}_{ij}\right)^2, \tag{7.2}$$

where con_{ij} represents the confidence level of physics depth at pixel (i,j), d_{ij}^{phy} is the labeled physics depth, and $\hat{d}_{ij}$ is the depth predicted by the model.

7.4 Self-supervised Training

In the self-supervised training paradigm, depth estimation is framed as an image reconstruction challenge. This approach eschews traditional ground truth labels in favor of utilizing unlabeled monocular videos during training. Our methodology hinges on leveraging both photometric and geometric consistencies as dual pillars to optimize image reconstruction jointly. This dual-

consistency approach allows for a more robust and accurate estimation of depth by capitalizing on the complementary strengths of both photometric and geometric information.

Photometric Consistency

For consecutive frames I_{t-1} and I_t, our model independently estimates their corresponding depths, D_{t-1} and D_t. As outlined in Equation 7.3, frames I_{t-1} and I_t can be projected into structured 3D point clouds Q_{t-1} and Q_t, respectively. Utilizing the pose network, we estimate the camera's movement from time $t-1$ to t. Through the application of the transformation matrix $T_{t-1\to t}$ and the point cloud Q_t, an estimated version of Q_{t-1}, denoted as $\hat{Q}_{t-1}$, is obtained: $\hat{Q}_{t-1} = T_{t-1\to t}Q_t$. Subsequently, frame I_t is reconstructed by warping I_{t-1} using the principles detailed in Equation 5.1 and Equation 7.4. The photometric loss is computed by Equation 7.5 using the reconstructed target image $I_{t-1\to t}$ and the target image I_t.

$$Q_{t-1}^{xy} = D_{t-1}^{xy} \cdot K^{-1} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{7.3}$$

$$I_{t-1\to t}\left[u\right] = I_{t-1}\left\langle u' \right\rangle \tag{7.4}$$

$$L_{ph} = ph\left(I_t, I_{t-1\to t}\right) \tag{7.5}$$

$$\begin{aligned} &ph\left(I_t, I_{t-1\to t}\right) = \tfrac{\alpha}{2}\left(1 - SSIM\left(I_t, I_{t-1\to t}\right)\right) \\ &+ \left(1-\alpha\right)\left\|\left(I_t, I_{t-1\to t}\right)\right\|_1 \end{aligned} \tag{7.6}$$

with α commonly set to 0.85 Godard et al. (2019), ph is a photometric reconstruction error. Furthermore, for each pixel p, the minimum of the losses computed from forward and backward neighboring frames allows to mitigate the effect of occlusions Godard et al. (2019) on the reprojection

process.

$$L_{ph}\left(p\right) = \min_{s \in [-1,1]} pe\left(I_{t-1}\left(p\right), I_{t-1 \to t}\left(p\right)\right) \tag{7.7}$$

1 stands for forward, 2 stands for backward.

$$L_s = \left|\partial_x d_t^*\right| e^{-\left|\partial_x I_t\right|} + \left|\partial_y d_t^*\right| e^{-\left|\partial_y I_t\right|} \tag{7.8}$$

As in previous works Godard et al. (2019), the edge-aware smoothness loss is used to improve the depth map.

Chapter 8

Experiments

8.1 Physics Depth

Below are the results of evaluations conducted on road physics depth maps, against LiDAR depth maps, for a single sample image and the entire KITTI dataset:

	Lower Better					Higher Better		
	AbsRel	Sq Rel	RMSE	RMSE log	Abs %Error	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Sample Image	0.0472	0.0367	0.8260	0.0513	4.7210	0.9999	0.9999	0.9999
Whole KITTI Dataset	0.0675	0.3328	2.6092	0.0907	6.7467	0.9583	0.9937	0.9989

Table 8.1: Results of evaluation of Road Physics Depth on single sample image vs complete KITTI dataset

From the Table 8.1,it can be seen that there is a noticeable decrease in the effectiveness of road physics depth when evaluating on entire KITTI dataset. This reduction in effectiveness may stem from various factors: flatness quality of road surfaces, quality of KITTI camera calibrations and accuracy of semantic segmentation information. Our observations specifically highlight considerable quality issues with the camera calibration details in the KITTI dataset. Camera calibration issues may arise due to the calibration being performed only once at the start of the day, as is the case for the sensor setup shown in 4.1. This approach to calibration is prone to minor variations

throughout the day. These variations can be attributed to fluctuations in the sensor setup due to vibrations caused by changes in road surface conditions and driving dynamics.

Below are the results of evaluations conducted on the physics depth methodology at different stages of the extension against LiDAR depth map:

	Lower is Better					Higher is Better		
	AbsRel	Sq Rel	RMSE	RMSE log	Abs %Error	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Road Physics Depth Map	0.0472	0.03675	0.8261	0.05134	4.7211	0.9999	0.9999	0.9999
Ground Physics Depth Map	0.1051	0.2741	1.5604	0.1498	10.5194	0.8855	0.9694	0.997
Edge Extended Physics Depth Map	0.1708	0.7108	3.5310	0.2221	17.0814	0.7460	0.9384	0.9823
Dense Physics Depth Map	0.1933	0.9537	4.0243	0.2436	19.3293	0.7173	0.9169	0.9760

Table 8.2: Results of evaluation of different stages of expansion of physics depth methodology

In Table 8.1, it is evident that there is a decrease in the effectiveness of depth maps when the physics-based depth analysis is extended from the road surface to other parts of the image, as demonstrated in 8.1. This decline in quality can be attributed to the absence of perfectly flat surfaces in the remaining ground areas and the techniques employed to extend the physics-based depth map from the ground to vertical structures.

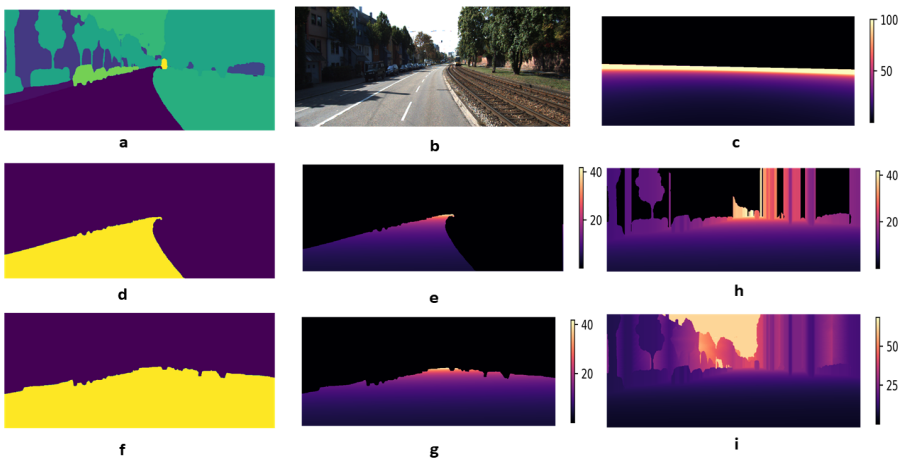

Figure 8.1: End-to-end Physics Depth Methodology demonstrated on a sample KITTI image:

(a) semantic segmented image (b) RGB image (c) full physics depth map along with scale (d) road segmented from semantic segmented image (e) physics depth map of road, along with its scale (f) ground surface segmented from semantic segmented image (g) physics depth map of ground surface, along with its scale (h) edge extended physics depth map (i) dense physics depth maps

8.2 Physics Depth as Scaling Factor

Below are the results of a comparative study on two of the most prominent self-supervised architectures: MonoDepth2 Godard et al. (2019) and MonoVIT Zhao et al. (2022), focusing on scenarios where the scaling factor is calculated using LiDAR depth versus physics-based depth.

		Lower Better				Higher Better		
MonoDepth2	Median Scaling Factor	AbsRel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
LiDAR Scale	32.260	0.159	1.689	5.168	0.238	0.830	0.931	0.967
Physics Depth Scale	32.487	0.158	1.968	5.287	0.242	0.842	0.930	0.966

Table 8.3: Evaluation of MonoDepth2 with LiDAR Depth Scaling Factor and Physics Depth Scaling Factor

MonoVIT	Median Scaling Factor	Lower Better				Higher Better		
		AbsRel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
LiDAR Scale	28.354	0.110	0.759	4.248	0.199	0.872	0.954	0.979
Physics Depth Scale	28.096	0.108	0.743	4.241	0.200	0.874	0.955	0.979

Table 8.4: Evaluation of MonoVIT with LiDAR Depth Scaling Factor and Physics Depth Scaling Factor

Despite the challenges posed by camera calibration issues in the KITTI dataset, substituting LiDAR with physics-based depth maps for determining the scaling factor does not notably diminish the model's performance. The evaluations affirm that physics-based depth maps can be considered a dependable substitute for LiDAR Depth Maps in calculating the scaling factor, thereby enhancing the autonomy of self-supervised models.

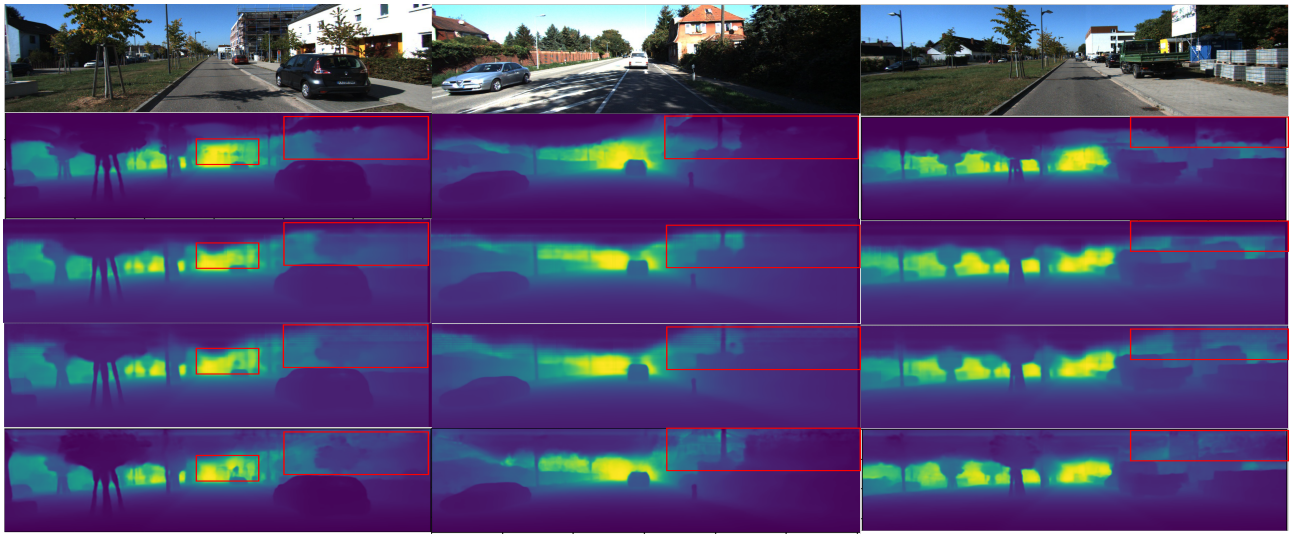

Figure 8.2: **Visual results on KITTI:** From top to bottom, the models are AdaBinsBhat et al. (2021), SwinV2-L 1K-MIMXie et al. (2023), NeWCRFsYuan et al. (2022b), our models.

8.3 Physics Depth in Supervised Depth Estimation

KITTI Monodepth Evaluation:

Using the standard KITTI Eigen split, which comprises 697 images, we conducted an evaluation of our model. Table 8.6 presents a summary of the performance of state-of-the-art (SoTA) supervised methods on the KITTI dataset, clearly indicating that our method outperforms previous approaches significantly. Even in comparison to MIM using the same model architecture, the introduction of physics depth information led to substantial improvements. For the `swin_v2_base` structure, the RMSE (Root Mean Square Error) improved from 2.05 to 1.2301, and for the `swin_v2_large` structure, it improved from 1.96 to 1.1652. In Figure 8.3, we observe that when comparing the prediction results of AdaBins, SwinV2-L, 1K-MIM, and NeWCRFs models, our model excels in capturing intricate scene details and demonstrates superior scene recovery capabilities.

Ablation study:

To thoroughly assess the impact of the proposed components in our methods on performance, we conducted detailed ablation studies on the KITTI, presented in Table 8.5.

Physics Depth: We observe from the comparison between row 1 and row 2 that the impact is substantially enhanced when the physics depth information is directly fused with the RGB information. This underscores the significant potential of physics depth in improving the predictive capabilities of supervised learning models for monocular depth estimation.

Information Fusion: In the comparison between Row 3 and Row 4, we observe that the information fusion module, which combines features from depth and RGB images, substantially enhances the model's predictive capacity for depth estimation. Moving on to Row 4 vs. Row 5, we note that, even after employing the same information fusion module, utilizing all the data produces superior results compared to using only the top 80% of the data. Data with higher errors can still provide valuable insights to the model, whereas excessively clean data may lead to model overfitting. To address this challenge, we introduce a physics depth selection module within the information fusion module. This module intelligently highlights physics depth features that enhance model predictions while effectively filtering out features that could hinder the model's performance. This adaptive approach enables us to harness the full potential of physics depth, leading to a significant enhancement in the model's predictive capabilities.

80% Physics Depth: We calculate the depth of each pixel along with its ground truth error and select the top 80% of the depth data, ordered from the smallest to the largest error, as the input to our model. The results are presented in Row 2 vs. 3. It is evident that both using all the data and using the top 80% of the data enhance the predictive capabilities of the model. However, leveraging the physics depth information within the top 80% of the error range yields greater improvements. This outcome can be attributed to the fact that not all physics depth measurements are absolutely accurate, and significant errors can adversely impact the model's performance.

ID	PD	80%PD	FM	AbsRel ↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta < 1.25 \uparrow$
1	✗	✗	✗	0.050	0.139	1.966	0.078	0.976
2	✗	✓	✗	0.0310	0.0620	1.3839	0.0499	0.9943
3	✓	✗	✗	0.0314	0.0648	1.4330	0.0509	0.9941
4	✗	✓	✓	0.0258	0.0453	1.1978	0.0424	0.9964
5	✓	✗	✓	0.0251	0.0428	1.1652	0.0415	0.9966

Table 8.5: A study of our methods on the KITTI dataset: PD: Physics Depth. IF: Information Fusion Module. 80%PD: Utilizing 80% of physics depth data

Method	year	AbsRel ↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
Eigen.Eigen and Fergus (2015)	2015	0.203	1.548	6.307	0.282	0.702	0.898	0.967
DORN.Fu et al. (2018)	2018	0.071	0.268	2.271	0.116	0.936	0.985	0.995
VNL.Yin et al. (2019)	2019	0.072	-	3.258	0.117	0.938	0.990	0.998
BTS Lee et al. (2019)	2019	0.061	0.261	2.834	0.099	0.954	0.992	0.998
Adabins Bhat et al. (2021)	2021	0.058	0.190	2.360	0.088	0.964	0.995	0.999
P3Depth Patil et al. (2022)	2022	0.071	0.270	2.842	0.103	0.953	0.993	0.998
DepthFormer Li et al. (2023c)	2023	0.052	0.158	2.143	0.079	0.975	0.997	0.999
NeWCRFsYuan et al. (2022b)	2022	0.052	0.155	2.129	0.079	0.974	0.997	0.999
iDiscPiccinelli et al. (2023)	2023	0.050	0.145	2.067	0.077	0.977	0.997	0.999
URCDCShao et al. (2023)	2023	0.050	0.142	2.032	0.076	0.977	0.997	0.999
MiM(base) Xie et al. (2023)	2023	0.052	0.141	2.050	0.078	0.976	0.998	0.999
MiM(large)Xie et al. (2023)	2023	0.050	0.139	1.966	0.075	0.977	0.998	0.999
Trap AttentionNing and Gan (2023)	2023	0.050	0.128	1.869	0.074	0.980	0.998	0.999
LightedDepth Zhu and Liu (2023)	2023	0.028	0.087	1.597	0.049	0.991	0.998	0.999
Ours (MiM base)	2023	0.0271	0.0483	1.2301	0.0442	0.9959	0.9993	0.9998
Ours (MiM large)	2023	0.0251	0.0428	1.1652	0.0415	0.9966	0.9994	0.9998

Table 8.6: For a quantitative depth comparison using the Eigen split of the KITTI dataset, we employ MIM as our supervised model. Specifically, we utilize MIM with the following configurations: MIM Base: Swin_v2_base, MIM Large: Swin_v2_large.

8.4 Physics Depth in Self-supervised Depth Estimation

KITTI Monodepth Evaluation:

Using the standard KITTI Eigen split, which comprises 697 images, we conducted an evaluation of our model. Table **??** presents a summary of the performance of state-of-the-art (SoTA) self-supervised methods on the KITTI dataset, clearly indicating that our method outperforms previous approaches significantly. Compared to our backbone model, SQLDepth, the integration of physical depth, associated confidence measures, and the consistency checks in both two-dimensional and three-dimensional spaces has yielded substantial improvements. This is reflected in the Root

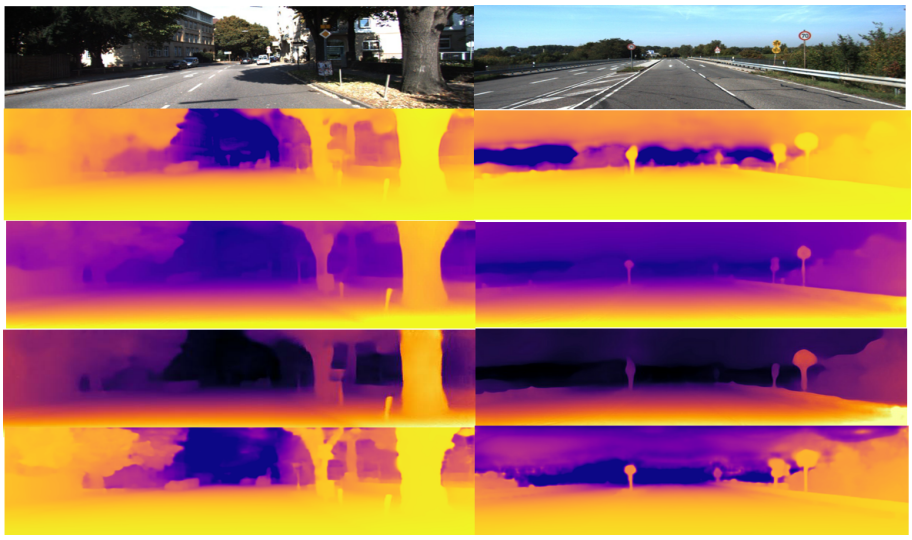

Figure 8.3: **Qualitative results on KITTI:**.From top to bottom the models are MonoVitZhao et al. (2022), RA-Depth He et al. (2022), ManyDepthWatson et al. (2021), our models.

Method	Year	Test frames	AbsRel ↓	Sq Rel↓	RMSE↓	RMSElog ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
Pilzer et alPilzer et al. (2018)	512×256	1	0.240	4.264	8.049	0.334	0.710	0.871	0.937
Struct2DepthCasser et al. (2019)	416×128	1	0.145	1.737	7.280	0.205	0.813	0.942	0.976
Monodepth2 Godard et al. (2019)	416×128	1	0.129	1.569	6.876	0.187	0.849	0.957	0.983
Lee et al.Lee et al. (2021b)	832×256	1	0.111	1.158	6.437	0.182	0.868	0.961	0.983
InstaDMLee et al. (2021a)	832×256	1	0.111	1.158	6.437	0.182	0.868	0.961	0.983
ManyDepth Watson et al. (2021)	416×128	$2\ (-1, 0)$	0.114	1.193	6.223	0.170	0.875	0.967	0.989
SQLDepth Wang et al. (2023)	416×128	1	0.110	1.130	6.264	0.165	0.881	0.971	0.991
Backbone (SQLDepth)	416×128	1	0.103	1.090	5.937	0.157	0.895	0.974	0.991

Table 8.7: For a quantitative depth comparison of the Cityscape dataset Cordts et al. (2016), we employ SQLDepth Wang et al. (2023) as our supervised model.

Mean Square Error (RMSE) outcomes achieved by our model. Figure 8.3 illustrates that, when evaluating against the predictions of MonoVit, SQLDepth, 1K-MIM, and NeWCRFs models, our model excels at capturing intricate details in complex scenes and demonstrates superior scene reconstruction capabilities.

Ablation Study:

To elucidate the individual contributions of various components in our model during monocular training, we conducted an ablation study, the results of which are presented in Table 8.8. This study involved modifying different elements of our model. We observed that the baseline model,

Back-bone	Physics Depth	confidence	AbsRel ↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta < 1.25 \uparrow$
✓			0.087	0.659	4.096	0.165	0.920
✓	✓		0.086	0.621	3.912	0.161	0.921
✓	✓	✓	0.086	0.594	3.886	0.159	0.918
✓	✓	✓	0.085	0.583	3.785	0.158	0.922

Table 8.8: **Ablation study on KITTI.** Input is 1024×320. PD: Physics Depth. L_{con}: Loss of Physics-Depth Supervision. L_{2D}: Loss of 2D Spatial Consistency. L_{3D}: Loss of 3D Spatial Consistency

devoid of our enhancements, exhibited the lowest performance. Conversely, integrating all our components resulted in substantial improvements, as demonstrated in the enhanced version.

Physics Depth: As reported in Table 8.8, We find that Providing models with accurate ground depth provides the strongest improvement in the model's ability to predict depth. This improvement is due to the intrinsic role of depth estimation in deciphering the relative position and dimensions of objects within a scene. As a central reference point, ground depth provides critical contextual information. It establishes a benchmark that aids in the accurate interpretation of the spatial relationships between various objects and the ground. This input is effective in resolving ambiguities related to object size and distance perception. In addition, during the training phase, known ground depth facilitates faster adaptation to the geometric structure of the scene.

Confidence of Physics Depth: As reported in Table 8.8. Compared to the mere incorporation of physics depth, our approach yields better results by computing a confidence score for the physics depth estimates. Since the physics depth is not always exact—being highly accurate in some regions while erroneous in others—direct usage could disrupt the self-supervised learning process. Specifically, regions with small errors may impede the optimization of self-supervision due to their misleading influence. By integrating confidence scores during training, our model prioritizes areas with higher accuracy, reducing the emphasis on erroneous regions. This strategy allows subsequent self-supervision to refine these areas further.

Chapter 9

Conclusion

In this work, we introduce an innovative approach to calculate the absolute depth of flat ground surfaces in images, utilizing camera model parameters to determine the scaling factor, thereby bypassing the need for LiDAR and introduce a physics-based supervised learning approach for depth estimation. While existing supervised techniques often rely on advanced network architectures, unique geometric priors, and diverse data augmentations to marginally enhance model performance, our method leverages physic depth for precise depth prediction. This significantly elevates the evaluation metrics of monocular depth estimation. By tapping into the potential of physics depth estimation calculated through camera model, we seek to offering a better alternative to LiDAR for ground surface to calculate the scaling factor for self-supervised learning models, enhance the performance of supervised models, especially in discerning ground and environmental details. Our ultimate aim is to enable these models to intrinsically predict accurate depth values leveraging insights from the physics depth. This approach is straightforward to deploy and offers a foundational mechanism to augment the depth prediction capabilities of various supervised and unsupervised models.

Bibliography

Bhat, S. F., Alhashim, I., and Wonka, P. (2021). Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018.

Bhat, S. F., Alhashim, I., and Wonka, P. (2022). Localbins: Improving depth estimation by learning local distributions. In *European Conference on Computer Vision*, pages 480–496. Springer.

Bódis-Szomorú, A., Riemenschneider, H., and Van Gool, L. (2014). Fast, approximate piecewise-planar modeling based on sparse structure-from-motion and superpixels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 469–476.

Casser, V., Pirk, S., Mahjourian, R., and Angelova, A. (2019). Unsupervised monocular depth and ego-motion learning with structure and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0.

Chauve, A.-L., Labatut, P., and Pons, J.-P. (2010). Robust piecewise-planar 3d reconstruction and completion from large-scale unstructured point data. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 1261–1268. IEEE.

Chawla, H., Varma, A., Arani, E., and Zonooz, B. (2021). Multimodal scale consistency and awareness for monocular self-supervised depth estimation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5140–5146. IEEE.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223.

Diaz, R. and Marathe, A. (2019). Soft labels for ordinal regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4738–4747.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dunn, B. E., Gray, G. C., and Thompson, D. (1965). Relative height on the picture-plane and depth perception. *Perceptual and Motor Skills*, 21(1):227–236.

Eigen, D. and Fergus, R. (2015). Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658.

Eigen, D., Puhrsch, C., and Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27.

Fu, H., Gong, M., Wang, C., Batmanghelich, K., and Tao, D. (2018). Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011.

Gallup, D., Frahm, J.-M., and Pollefeys, M. (2010). Piecewise planar and non-planar stereo for urban scene reconstruction. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 1418–1425. IEEE.

Gardner, J. S., Austerweil, J. L., and Palmer, S. E. (2010). Vertical position as a cue to pictorial depth: Height in the picture plane versus distance to the horizon. *Attention, Perception, & Psychophysics*, 72:445–453.

Garg, D., Wang, Y., Hariharan, B., Campbell, M., Weinberger, K. Q., and Chao, W.-L. (2020). Wasserstein distances for stereo disparity estimation. *Advances in Neural Information Processing Systems*, 33:22517–22529.

Garg, R., Bg, V. K., Carneiro, G., and Reid, I. (2016). Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 740–756. Springer.

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237.

Godard, C., Mac Aodha, O., and Brostow, G. J. (2017). Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279.

Godard, C., Mac Aodha, O., Firman, M., and Brostow, G. J. (2019). Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838.

Gordon, A., Li, H., Jonschkowski, R., and Angelova, A. (2019). Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8977–8986.

Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., and Gaidon, A. (2020). 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2485–2494.

Hazirbas, C., Ma, L., Domokos, C., and Cremers, D. (2017). Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part I 13*, pages 213–228. Springer.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778.

He, M., Hui, L., Bian, Y., Ren, J., Xie, J., and Yang, J. (2022). Ra-depth: Resolution adaptive self-supervised monocular depth estimation. In *European Conference on Computer Vision*, pages 565–581. Springer.

Hirose, N., Koide, S., Kawano, K., and Kondo, R. (2021). Plg-in: Pluggable geometric consistency loss with wasserstein distance in monocular depth estimation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12868–12874. IEEE.

Huynh, L., Nguyen-Ha, P., Matas, J., Rahtu, E., and Heikkilä, J. (2020). Guiding monocular depth estimation using depth-attention volume. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 581–597. Springer.

Kusupati, U., Cheng, S., Chen, R., and Su, H. (2020). Normal assisted stereo depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2189–2199.

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., and Navab, N. (2016). Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE.

Lee, J. H., Han, M.-K., Ko, D. W., and Suh, I. H. (2019). From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*.

Lee, J.-H., Heo, M., Kim, K.-R., and Kim, C.-S. (2018). Single-image depth estimation based on fourier domain analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 330–339.

Lee, S., Im, S., Lin, S., and Kweon, I. S. (2021a). Learning monocular depth in dynamic scenes via instance-aware projection consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1863–1872.

Lee, S., Rameau, F., Pan, F., and Kweon, I. S. (2021b). Attentive and contrastive learning for joint depth and motion field estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4862–4871.

Li, K., Wang, Y., Zhang, J., Gao, P., Song, G., Liu, Y., Li, H., and Qiao, Y. (2023a). Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Li, Y. et al. (2023b). Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1477–1485.

Li, Z., Chen, Z., Liu, X., and Jiang, J. (2023c). Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *Machine Intelligence Research*, pages 1–18.

Li, Z., Wang, X., Liu, X., and Jiang, J. (2022). Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint arXiv:2204.00987*.

Liu, F., Shen, C., Lin, G., and Reid, I. (2015). Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2024–2039.

Long, X., Lin, C., Liu, L., Li, W., Theobalt, C., Yang, R., and Wang, W. (2021). Adaptive surface normal constraint for depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12849–12858.

Luo, C., Yang, X., and Yuille, A. (2021). Exploring simple 3d multi-object tracking for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10488–10497.

Luo, Y., Ren, J., Lin, M., Pang, J., Sun, W., Li, H., and Lin, L. (2018). Single view stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 155–163.

Mahjourian, R., Wicke, M., and Angelova, A. (2018). Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5667–5675.

Miangoleh, S. M. H., Dille, S., Mai, L., Paris, S., and Aksoy, Y. (2021). Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9685–9694.

Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., Kohi, P., Shotton, J., Hodges, S., and Fitzgibbon, A. (2011). Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pages 127–136. Ieee.

Ning, C. and Gan, H. (2023). Trap attention: Monocular depth estimation with manual traps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5033–5043.

Patil, V., Sakaridis, C., Liniger, A., and Van Gool, L. (2022). P3depth: Monocular depth estimation with a piecewise planarity prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1610–1621.

Peng, R., Wang, R., Lai, Y., Tang, L., and Cai, Y. (2021). Excavating the potential capacity of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15560–15569.

Piccinelli, L., Sakaridis, C., and Yu, F. (2023). idisc: Internal discretization for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21477–21487.

Pilzer, A., Xu, D., Puscas, M., Ricci, E., and Sebe, N. (2018). Unsupervised adversarial depth estimation using cycled generative networks. In *2018 international conference on 3D vision (3DV)*, pages 587–595. IEEE.

Ranftl, R., Bochkovskiy, A., and Koltun, V. (2021). Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., and Koltun, V. (2020). Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637.

Scharstein, D. and Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47:7–42.

Shao, S., Pei, Z., Chen, W., Li, R., Liu, Z., and Li, Z. (2023). Urcdc-depth: Uncertainty rectified cross-distillation with cutflip for monocular depth estimation. *arXiv preprint arXiv:2302.08149*.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Surdick, R. T., Davis, E. T., King, R. A., and Hodges, L. F. (1997). The perception of distance in simulated visual displays: A comparison of the effectiveness and accuracy of multiple depth cues across viewing distances. *Presence: Teleoperators & Virtual Environments*, 6(5):513–531.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Tang, Y., Zhao, C., Wang, J., Zhang, C., Sun, Q., Zheng, W. X., Du, W., Qian, F., and Kurths, J. (2022). Perception and navigation in autonomous systems in the era of learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Tateno, K., Tombari, F., Laina, I., and Navab, N. (2017). Cnn-slam: Real-time dense monocular slam with learned depth prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6243–6252.

Telea, A. (2004). An image inpainting technique based on the fast marching method. *Journal of graphics tools*, 9(1):23–34.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al. (2020). Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364.

Wang, Y., Chao, W.-L., Garg, D., Hariharan, B., Campbell, M., and Weinberger, K. Q. (2019). Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8445–8453.

Wang, Y., Liang, Y., Xu, H., Jiao, S., and Yu, H. (2023). Sqldepth: Generalizable self-supervised fine-structured monocular depth estimation. *arXiv preprint arXiv:2309.00526*.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Watson, J., Firman, M., Brostow, G. J., and Turmukhambetov, D. (2019). Self-supervised monocular depth hints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2162–2171.

Watson, J., Mac Aodha, O., Prisacariu, V., Brostow, G., and Firman, M. (2021). The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1164–1174.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.

Xie, Z., Geng, Z., Hu, J., Zhang, Z., Hu, H., and Cao, Y. (2023). Revealing the dark secrets of masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14475–14485.

Yang, F. and Zhou, Z. (2018). Recovering 3d planes from a single image via convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100.

Yang, G., Tang, H., Ding, M., Sebe, N., and Ricci, E. (2021). Transformer-based attention networks for continuous pixel-wise prediction. In *Proceedings of the IEEE/CVF International Conference on Computer vision*, pages 16269–16279.

Yin, W., Liu, Y., Shen, C., and Yan, Y. (2019). Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5684–5693.

Yin, Z. and Shi, J. (2018). Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yuan, W., Gu, X., Dai, Z., Zhu, S., and Tan, P. (2022a). Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3916–3925.

Yuan, W., Gu, X., Dai, Z., Zhu, S., and Tan, P. (2022b). New crfs: Neural window fully-connected crfs for monocular depth estimation. *arXiv preprint arXiv:2203.01502*.

Zhang, W., Zhang, W., and Zhang, Y. (2020). Geolayout: Geometry driven room layout estimation based on depth maps of planes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 632–648. Springer.

Zhao, C., Zhang, Y., Poggi, M., Tosi, F., Guo, X., Zhu, Z., Huang, G., Tang, Y., and Mattoccia, S. (2022). Monovit: Self-supervised monocular depth estimation with a vision transformer. In *2022 International Conference on 3D Vision (3DV)*, pages 668–678. IEEE.

Zhao, J., Yan, K., Zhao, Y., Guo, X., Huang, F., and Li, J. (2021). Transformer-based dual relation graph for multi-label image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 163–172.

Zhou, H., Greenwood, D., and Taylor, S. (2021). Self-supervised monocular depth estimation with internal feature fusion. *arXiv preprint arXiv:2110.09482*.

Zhou, T., Brown, M., Snavely, N., and Lowe, D. G. (2017). Unsupervised learning of depth

and ego-motion from video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1851–1858.

Zhu, S. and Liu, X. (2023). Lighteddepth: Video depth estimation in light of limited inference view angles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5003–5012.