UNIVERSITY OF CALIFORNIA
RIVERSIDE

Monocular Depth Estimation Using Attention Guidance

A Thesis submitted in partial satisfaction
of the requirements for the degree of

Master of Science

in

Electrical Engineering

by

Chetan Reddy Mudireddy

June 2021

Thesis Committee:
Dr. M. Salman Asif, Chairperson
Dr. Nanpeng Yu
Dr. Hyoseung Kim

The Thesis of Chetan Reddy Mudireddy is approved:

Committee Chairperson

University of California, Riverside

ABSTRACT OF THE THESIS

Monocular Depth Estimation Using Attention Guidance

by

Chetan Reddy Mudireddy

Master of Science, Graduate Program in Electrical Engineering
University of California, Riverside, June 2021
Dr. Salman Asif, Chairperson

Depth estimation from a single image represents a very exciting challenge in computer vision. In this regards Self-supervised monocular depth estimation has gained immense popularity recently because they dont require groundtruth depth during training. Instead of the groundtruth depth map, the current methods rely on the view synthesis as a supervision for depth prediction. Recently there have been works that leverage the semantic cues while training in a multitask setup. But these methods cause some inherent problem while learning task-specific and task-sharing features which result in less accurate depth features. In this work, we propose to explicitly apply a mechanism by which network can weigh features for different tasks and avoid the interference between tasks of depth estimation and semantic sementation. In other words we employ the attention guided encoder network to learn both the task-specific and task-sharing depth features. Experiments on KITTI dataset demonstrate that our methods compete with the state-of-the-art methods.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Measuring distance relative to a camera remains difficult but absolutely key to unlocking exciting applications such as autonomous driving, 3D scene reconstruction and AR. In robotics, depth is a key prerequisite to perform multiple tasks such as perception, navigation, and planning. Creating a 3D map would be another interesting application, computing depth allows us to back project images captured from multiple views into 3D.

As human beings, we reliaze that how important of a role does vision play in our daily life. The depth estimation being one of the major activities that our eyes perform. Ever since we were born, we start to learn how to gauge how far objects around are present and this enables us to interact with the world around us. This capability plays a major role in our interaction process; how much do I need to move my hand to grab an object infront of me? How should I react to avoid hitting something that is coming my way? How fast

cars are movin on a road? All of these activities depend on our ability to understand and estimate the distance of the objects accurately and faster. This makes the task of teaching machines to learn depth prediction an extremely important one. If our end goal is to make fully autonomous vehicles and robots we need to put a lot of significance on depth estimation as prerequisite for the ability to reason about the world around us.

How do we humans estimate depth:

Lets first understand how we humsns perceive the depth of our sorroundins. This will provide some insignt on how the current depth estimation menthis are inspired from our human cvision system. Theoretically, when light rays from a source hit surfaces, it reflects off and directs towards the back of our retina, projecting them and our eye processes them as 2D just like how an image is formed on an image plane. So how do we actually measure distance and understand our environment in 3D when the projected scene is in 2D? The mechanism at work here is our brain starts to reason about the incoming visual signals by recognizing patterns such as the size, texture and motion about the scene known as **Depth Cues**. There are basically 4 categories of depth cues: Static monocular, depth from motion, binocular and physiological cues [1]. We subconsciously take advantage of these signals to perceive depth remarkably well.

Pictorial depth cues: We humans perceive deoth from a single image depends on the spatial location and arrangemenrs of the things in a given scene. The below table

summarizes some of the cues that we use to reason about the distance of different objedcts which occur to us naturally in our daily interaction with our surroundings.

Figure 1: Figure shows how humans perceive depth of objects using spatial cues. Source[39]

Table 1: Depth cues that enable us to reason about the distance of different objects.

Monoscopic Depth Cues	Examples	Appear Nearer	Appear Fatrther
Size of Objects	Tree	Larger	Smaller
Texture	Grass Patch	High Quality texture	Low Quality, blurry
Linear Perspective	Curb Line	-	Converge to horizon

Depth Cues from Motion (Motion Parallax):

When we observe outside from a moving vehicle, things that are close to us pass faster than the one that is farthest away. The farther something appears, the slower it seems to pass away from the observer.

Depth Cues from Stereo Vision:

Like human eyes, two cameras are displaced horizontally from each other on the same plane to obtain two views of the scene. By comparing these two images, the relative depth information of the scene can be obtained as a disparity map, which measures the horizontal displacement in pixels of the corresponding image points.

1.1 Challenges

Monocular depth estimation is a fundamentally ill-posed problem in Computer Vision. Two significant reasons are projection ambiguity and scale ambiguity. On the application of geometric transformation on a scene, the points in the two different scenes may map to the same location on the plane, see figure 2. As such, many 3D scenes can explain a single 2D image. Humans possess the ability to perceive depth from one image using monocular depth cues like the size of objects, texture, and linear perspective as discussed above. Figure 1 shows how the spatial arrangement of the trees and other objects develops our intuition of the depth associated with every object present in the scene.

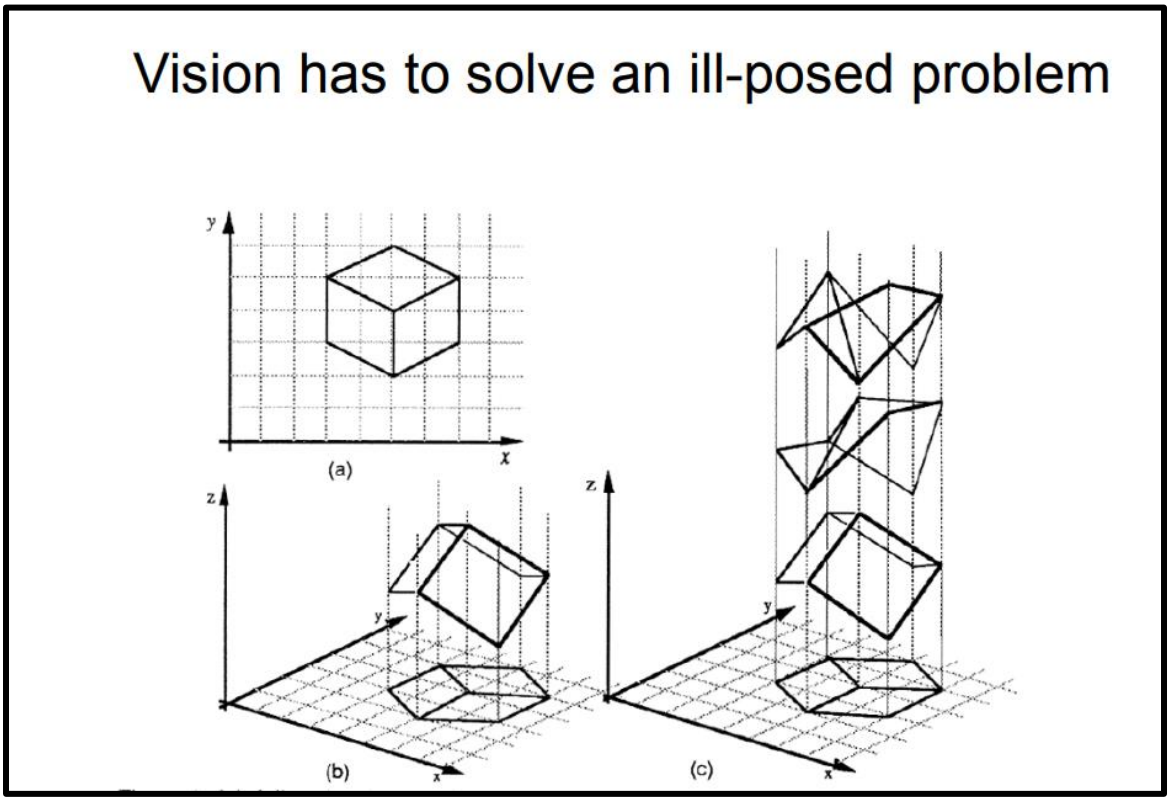

Figure 2: (a) A line drawing provides information only about the x,y coordinates of points lying along the object contours. (b) The human visual system is usually able to reconstruct an object in three dimensions given only a single 2D projection. (c) Any planar line-drawing is geometrically consistent with infinitely many 3D structures. Source[38]

Depth estimation from images is an important tool in a variety of applications, especially in Autonomous systems and robotics. While several dedicated ranging sensors, such as LIDAR provide superior depth accuracy compared to visual methods they are very expensive. Monocular depth estimation is an active area of study in machine learning. While initial supervised learning models for this task have enjoyed success, they require pixel-wise labelled ground truth for training which is very difficult to obtain and turn out

to be very costly to prepare them. Recently, unsupervised methods have surged in usage and they employ novel reconstruction loss terms to avoid the need for ground truth depth, but utilize only one view of the scene during evaluation.

Moving Objects violate the static assumption for SFM Method:

Dynamic objects in the scene further complicate the estimation process. Depth estimation via structure from motion involves a moving camera and consecutive static scenes. **This assumption must hold for matching and aligning pixels**. This assumption breaks when there are moving objects in the scene. To this end, many researchers have looked into several methods to model moving objects in the scene by incorporating velocity information using optical flow [2] or by using instance segmentation mask to model the object‘s motion from one frame to another [3].

1.2 Contributions

This thesis makes the following contributions:

- We propose an attention mechanism to enhance the quality of depth features learned in self-supervised depth estimation networks.
- We demonstrate that the obtained attention guided semantically global context aware depth features can perform better compare to few of the previous methods.

The rest of the thesis is organized as follows. Chapter 2 gives an overview of the Related Work in supervised, unsupervised, and semi-supervised Monocular Depth Estimation. The methodology, i.e., the different types of architecture and loss functions used, details on training, etc. are described in Chapter 3. Chapter 4 contains the results, experiments, and subsequent discussions.

Chapter 2

Related Work

2.1 Monocular Depth Estimation: In Detail

Estimation depth information from images is one of the basic and important tasks in computer vision. To bradly classify depth prediction from images into three different methods depending on what kind of data and methods are used.

Geometry-based methods: These methods are the first of its kind that have been widely investigated to perceive depth for the last forty years. These methods recover 3D structures from a given set of images using the geometric constraints. *Structure from Motion (SfM)* [5] is the primary methods for estimation 3D structures from a series of 2D images. The depth is perceived by applying feature correspondences and geometric constraints over a

sequence of images. The other important method is *Stereo vision matching* [6][7]. This technique performs depth prediction by matching the scene from two different viewpoints. This is inspired from the human eyes and the disparity maps of the two images captured using two cameras are calculated. This method overcomes the serious drawback of *SfM*, the scale information is included in depth estimation during stereo vision matching process which avoids the monocular scale ambiguity.

Sensor-based methods: We can get the depth information using depth sensors like RGB-D and LIDAR for the corresponsing images. RGB-D cameras produce pixel-level depth maps, they have limited measuremet range and outdoor sensitivity. LIDAR is most commonly used hardware in Autonomous systems for depth measurement but it can only generate the sparse 3D maps.

Deep learning-based methods: In the recent years due to rapid developments in deep learning, which led to an increase in performance of deep neural networks. Recent methods have shown that an end-to-end methods [8] are able to produce pixel-level depth maps from a single image. Many kinds of neural networks were able to effectively address the problem of Monocluar depth estimation, such as convolutinla nuera networks (CNN's) [9], recurrent nueral networks (RNN's) [10], variational auto-encoders (VAE's) [11] and generative adversarial networks (GAN's) [12].

The below sections describe various methods on Monocular depth estimation using deep learning techniques since our work is based on neural networks.

2.2 Deep learning based Monocular depth estimation

Deep neural networls are considered as a black box and they learn some structural information to do the depth prediction with the help of suopervised signals. However the biggest challenges of deep learning is to acquire large amount pof annotated data also called as groundtruth. Getting groundtruth data is very time consuming and expensive. In this section we will discuss about monocular depth estimation methods in terms of using groundtruth: supervised methods [13], unsupervised methods [14] and semi-supervised methods [15]. One major difference is even though the training processes of unsupervised and semi-supervised make use of monocular videos or stereo image pairs while training, they use only single images while inference.

2.2.1 Supervised monocular depth estimation

In a basic model of a supervised method, there exists a supervisory signal based on ground truth of depth maps, so that monocular depth estimation can be regarded as a regressive problem. The network is guided by a L_2 **Loss:**

$$L2\left(d, d^*\right) = \frac{1}{N} \sum i^N \left\| d - d^* \right\|_2^2$$

One of the earliest work that paved way for the current supervised techniques for monocular depth estimation is by Eigen et al. [16]. This paper introduced regressing for depth over pixels for the first time. The proposed method consists of a dual component structure namely the global coarse network and local fine network to predict depth from a

single image. They use a novel *scale-invariant* loss to account for scale dependent error. During the training process, they guide the network with a loss function set as below:

$$L\left(d, d^{\star}\right)=\frac{1}{N} \sum_{i}^{N} y_{i}^{2}-\frac{\lambda}{N^{2}}\left(\sum_{i}^{N} y_{i}\right)^{2}$$

where $y_i^2 = \log(d) - \log\left(d^*\right)$.λ refers to the balance factor. The global coarse network predicts the overall depth map and the coarse network performs the local refinements on previous output by utilizing vanishing points, object locations.

Xu et al. [17] proposed a new method that is derived from the multiscale predictions derived from CNN inner layers by structurally fusing them within a unified CNN-CRF framework. Similarly, Li et al. [18] introduced to fuse features from two different streams of CNN's, by extracting features at intermediate layers at different scales. There are two different techniques to fuse these intermediate layers:

Multi-scale CRF: Given an LN-dimensional vector $\hat{s}$ obtained by concatenating side output score maps $\{s_1, \ldots, s_L\}$ and an LN dimensional vector d, the CRF modelling the conditional distribution is defined as:

$$P(\mathrm{~d} \mid \hat{\mathrm{s}})=\frac{1}{Z(\hat{\mathrm{s}})} \exp\{-E(\mathrm{~d}, \hat{s})\}$$

Cascade CRFs: The cascade model is based on L CRF models, each one at a specific scale L which are progressively stacked to use only the previous scale as an input to define features at the next level.

Methods based on conditional random fields: Instead of using an additional network for different tasks in [16], Li et al [18] propose a refinement method using conditional random

fields which were already used for many tasks like semantic segmentation [19][20]. In these methods, the depth map is refined from a super-pixel level to pixel level using CRF, and the energy function is as follows:

$$\mathbf{E}(\mathbf{d}) = \sum_{i \in S} \phi_i \left(d_i\right) + \sum_{(i,j) \in \varepsilon_s} \phi_{ij} \left(d_i, d_j\right) + \sum_{c \in P} \phi_c \left(d_c\right)$$

where S stands for the set of super-pixels and refers to the set of pixels that share a common boundary. Similarly, another method is proposed after this by Liu et al. [21] to solve the problem of monocular depth estimation.

Xu et al. [22] proposed a very similar structure to that of Xu et al. [17], this method consists of the same kind of front end of the CNN architecture using the multi-scale features are extracted and passed to the next step through conditional random fields. One new significant change that they proposed in the new framework is to regulate how much information is exchanged between different features at various scales. They have achieved this using a structured attention model. The important idea of using an attention model is to manage the flow of information. More specifically, an attention model $A = \{A_s\}\, s = 1^{S-1}$ parameterized by binary variables $As = \left\{a_s^i\right\}\, i = 1^N, as^i \in \{0, 1\}$ is introduced. The attention variable a^i regulates information which is allowed to flow between intermediate scale s and final scale S for pixel I.

A CRF can be structured given the observed multi-scale feature maps **X** and the estimated latent multiscale representation **Y** as:

$$E(\mathbf{Y}, \mathbf{A}) = \Phi(\mathbf{Y}, \mathbf{A}) + \Xi(\mathbf{Y}, \mathbf{A}) + \Psi(\mathbf{A})$$

Here the first term represents the summation of unary potentials and correspondingly the second term represents the relation between features of the latest layers at the final scale

with the intermediate scale. They employed a ResNet-50 for the feature extractor frontend. The model is optimized using the backpropagation with stochastic gradient descent. Below I have summarized various methods in a table.

Table 2 : Comparision of supervised methods for monocular depth estiomation

Method	Network Architueeture	Multi-scale Features	CRF's	Learing Paradigm
Multi-Scale CNN	Deep CNN	Yes	No	Supervised
Multi-Scale CRFs	Deep CNN	Yes	Yes	Supervised
SAGCNF Xu et al.	Deep CNN	Yes	Yes	Supervised
DORN Fu et al.	DSE+SUM	Yes	Yes	Supervised

2.2.2 Unsupervised monocular depth estimation

Instead of using ground truth while training the models, unsupervised techniques make use of geometric constraints and the underlying epipolar geometry present in the training data. Acquiring labels or ground truth data is very expensive and time-consuming. To avoid such cumbersome processes, unsupervised methods are proposed and are gaining attention in the research community over recent years.

Monocular depth estimation with Left-Right consistency:

The first leap towards unsupervised methods for MDE is taken by Godard et al. [23]. The main proposal of this work is to make use of binocular stereo footage with left and right view images of the scene while training but only one view while testing. By exploiting the *epipolar geometric constraints* their models generate the disparity maps

using an image reconstruction loss. Since the model learns from stereo pair it doesn't require any ground truth labelled data to predict pixel-level disparity between pairs of rectified stereo images.

The novelty of this method is just not using stereo pairs of images, they also introduced depth estimation as image reconstruction and a novel loss to predict depth with high accuracy. The main idea is to be able to learn a function f such that we can predict per-pixel scene depth $\hat{d} = f(I)$ given a single image I,. Applying this function in a supervised setting is not practical considering the fact that making labels is expensive. Also, expensive hardware such as LIDAR can only produce sparse maps. Thus they propose to learn a function capable of reconstructing one image from another. During training, the network is provided with two images I^t and I^r which are left and right colour images from a calibrated stereo pair. The goal of the network is to find a depth relation d^r which should enable us to reconstruct the right image when it applied to the left image. The reconstructed image is denoted as $I'(d^r)$ or $\bar{I}^r$. In the case of a rectified images, d resembles the image disparity – where each pixel is a scalar value that the model has learned to predicted. Once we have the disparity maps and the static distance between the cameras b and their focal length f, we can get the depth $\hat{d}$ as $d = bf/d$.

The authors proposed a network that generates the depth predicted image using a *bilinear sampler*. With the help of a novel left-right consistency loss, Godard et al. [23] proposed to generate the disparity maps for both the left and right images instead of predicting depth pixel values individually for one and sampling on the other. They

employed an encoder and a decoder with skip connections. The authors have used a loss C_s for outputs individually at each scale which resulted in a total loss of:

$$C_s = a_{ap}\left(C_{ap}^l + C_{ap}^r + \alpha_{ds}\left(C_{ds}^l + C_{ds}^r\right) + \alpha_{lr}\left(C_{lr}^l + C_{lr}^r\right)\right)$$

where C_{ap} encourages similarity in the reconstructed image, C_{ds} enforces smoothness disparities, and C_{ir} prefers predicted left and right disparities to be consistent.

For the second part of sampling one stereo image from the other using the produced disparity map, they employed the image sampler from the spatial transformer networks (STN) [24] to sample any given input image using the disparity map. The C_{ap} loss is a combination of an $L1$ and single scale SSIM [] loss. This loss is defined as:

$$C_{ap}^l = \frac{1}{N}\alpha\frac{1 - \mathrm{SSIM}\left(I_{ij}^l, \tilde{I}ij^t\right)}{2} + (1-\alpha)\|Iij^l - \bar{I}_{ij}^l\,|$$

Disparity Smoothness Loss Disparities are encouraged to be locally smooth with an $L1$ penalty on disparity gradients ∂d. Because depth may have discontinuity, this loss is weight with an edge-aware term using image gradients ∂I ,

$$C_{ds}^t = \frac{1}{N}\left|\partial_x d'_{ij}\right| e - \left\|\partial_x I_{ij}^l\right\| + \left|\partial_y d_{ij}^l\right| e - \|\partial_y I_{j,\|}^l\,|$$

Self-supervised Training with Stereo Vision

To mitigate the issue mentioned above, Garg et al. and Godard et al. propose self-supervised training methods for monocular depth estimation. These approaches exploit the warping function to transfer the coordinates of the left image to the right image plane. In particular, design a photometric loss combining SSIM with L1 term and geometric warping. Moreover, casting depth estimation as an image reconstruction task represents a

very attractive way to overcome the need for expansive, ground truth labels by using a large amount of unsupervised imagery. During training, the network's predicted disparity is used to transform one image to the other image in a stereo pair. The transformed image is compared against the corresponding training example with a photometric loss term. Other optimizations are presented including disparity smoothness loss and consistency between disparities predicted on the left and right images.

Semantically guided monocular depth estimation

Our work is closely related to few recent works [49][50] in a way that the general framework that is used to enhance the depth features by using region aware object boundaries. Our work and theirs utilizes the semantic segmentation as a guidance network, but our work is novel in 4 different ways:

1. Work done in [49] uses two different encoders for the two tasks of semantic and depth estimation, whereas in our work we use a common encoder for both the tasks which enables our model to learn more accurate task-sharing features .
2. In [49], they propose to use a pretrained semantic network and freeze/fix the weights, but in our work we also train the semantic segmentation network simultaneously with depth estimation which enables us to perform attention guidance in a dynamic nature.
3. In [50] they propose to use the attention mechanism after the encoder and before the decoder. In our work we try to use the attention mechanism inbetween the two task-specific decoders to create the semantic aware depth features.

4. In [49], they use pixel adaptive convolutions two fuse the feature maps from separate decoders, whereas we utilize an attention mechanism which helps us to reduce the task-interference during feature fusing.

Chapter 3

Method Overview

Over the recent years, there has been increased interest in the self-supervised monocular depth estimation also commonly known as *structure-from-motion (SfM)* using deep neural networks since the self-supervision relaxes the hard requirement of having annotated ground truth which itself is cumbersome. This methodology has been gaining popularity among researchers for the above reasons. It all started with the seminal work of Zhou et al. [25]. Since then there has been a proliferation of interest in this area. Many new works have started to nail down and systematically tackle each component in this framework.

In this thesis, we exploit the geometric relationship between depth estimation and semantic segmentation for understanding a scene more precisely in the context of autonomous navigation. To be specific, we propose to enforce the semantic boundary information into the depth features by introducing a joint multitask learning framework that takes into consideration the geometric constraints which help in adaptively choosing the most significant features for monocular depth estimation. Understanding the spatial context of the scene will help the model to comprehend the individual characteristics of the pixels in that scene very well which in turn improve the performance of monocular depth estimation. Even after so many advances in the monocular depth estimation techniques suffer from some drawbacks. For example, they struggle with the cases where the texture

of the scenes is very similar they fail to distinguish and also they lack enough information of the occluded objects in the scene. We aim to address these issues by incorporating the global context of the scene into network with the help of semantic cues.

Our work is closely related to Meng et al. [26] in which the authors propose to use the fused input augmentation of semantic features and RGB images and fed them into the network. The major drawback of this method is that since the input features were just stacked together without considering any constraints there is a higher chance that the model tends to get confused and also the learned depth features may get corrupted due to the deeper nature of current neural networks.

3.1 Self-Supervised Monocular Depth Estimation

In a self-supervised setting the groundtruth comes from the input signal itself. In this case it is the RGB images. The pioneering work in this direction is set by work of Zhou et al.

Self-supervised learning framework: The defined goal is to resynthesize the images based on some input images. In our context, the images synthesized are making use of 3D scene geometry.

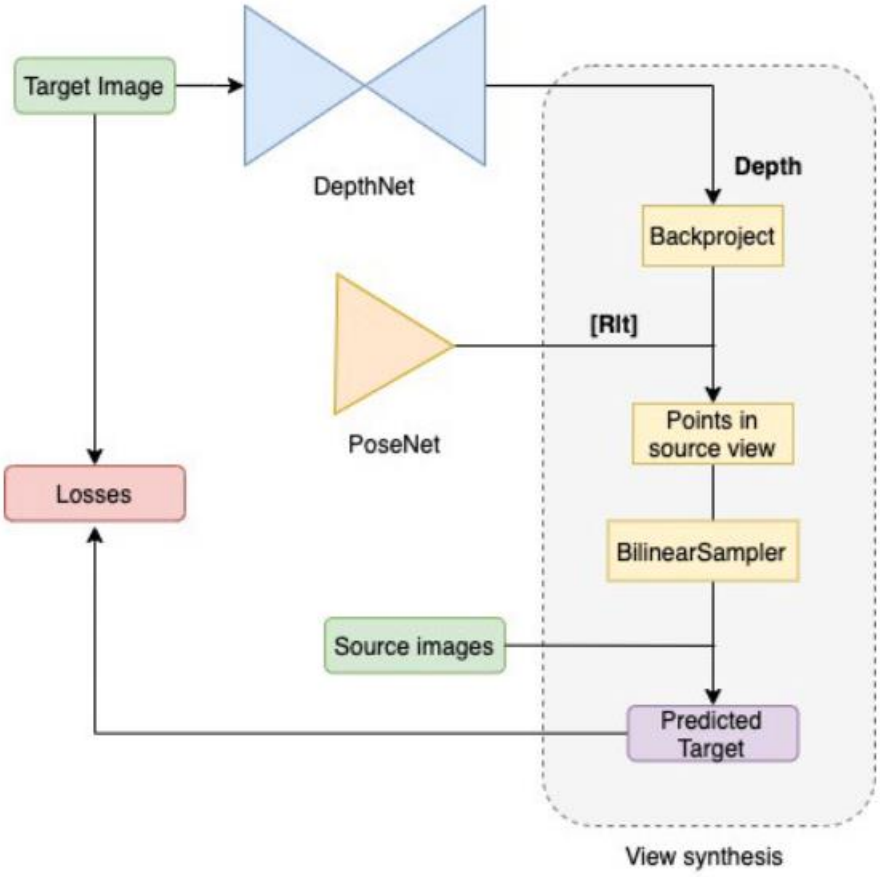

Figure 3: Self-supervised depth estimation pipeline. This pipeline shows the schematic of an unsupervised learning framework where we try to get the supervision from the input data itself. This has two blocks: The convolutional neural network for depth estimation, and view synthesis module to perform image reconstruction using Image geometry. Source[48]

When depth together with egomotion is provided, we can synthesize a new view (*target*) by applying a projective warping from the *source* camera point of view. From the figure above, the warping is achieved using a view synthesis module. One the main thing to notice here is that depth is an input to the module and in our case is predicted from a neural network. Learning in a self-supervised structure-from-motion setting requires two networks: a monocular depth model $f_D : I \rightarrow D$, that outputs a depth prediction

$\hat{D} = f_D(I(p))$ for every pixel p in the target image I; and a monocular ego-motion estimator $f_{\mathrm{x}} : (I_t, I_S) \rightarrow \mathrm{x}t \rightarrow S$, that predicts the 6 DoF transformations for all $s \in S$ given by

$$xt \rightarrow s = \begin{pmatrix} R & t \\ 0 & 1 \end{pmatrix} \in \mathrm{SE}(3)$$

between the target image I_t and a set of temporal context source images $I_s \in I_S$. In all reported experiments they use I_{t-1} and I_{t+1} as source images.

View synthesis module: As discussed above, in the view synthesis module using the monocular image sequences the projection is performed between neighbouring frames:

$$p_{n-1} \sim K T_{n \rightarrow n-1} D_n \left(p_n\right) K^{-1} p_n$$

where p_n stands for the pixel on image I_n, and p_{n-1} refers to the corresponding pixel of p_n on image I_{n-1}. K is the camera intrinsics matrix, which is known. $D_n \left(p_n\right)$ denotes the depth value at pixel p_n, and $T_{n \rightarrow n-1}$ represents the spatial transformation between I_n and I_{n-1}. Hence, if $D_n \left(p_n\right)$ and $T_{n \rightarrow n-1}$ are known, the correspondence between the pixels on different images $(I_n$ and $I_{n-1})$ are established by the projection function. Inspired by this constraint, Zhou et al. design a depth network to predict the depth map $\hat{D}n$ from a single image In, and a pose network to regress the transformation $\hat{T}n \rightarrow n-1$ between frames $(In$ and $I_{n-1})$. Based on the output of networks, the pixel correspondences between I_n and I_{n-1} are built up:

$$p_{n-1} \sim K \hat{T}n \rightarrow n-1 \hat{D}n \left(p_n\right) K^{-1} p_n$$

Then, the photometric error between the corresponding pixels is calculated as the geometric constraints. Zhou et al. are inspired by [27] to use a view synthesis as a metric. and the reconstruction loss is formulated as:

$$L\mathrm{Us} = \frac{1}{N}\sum p^{N}\left|I_n(p) - \hat{I}_n(p)\right|$$

where p indexes over pixel coordinates. $\hat{I}n(p)$ denotes the reconstructed frame. The structure similarity based on SSIM is also introduced into Lvs to quantify the differences between reconstructed and target images:

$$Lvs = \alpha\frac{1-\mathrm{SSIM}\left(In-\hat{I}n\right)}{2} + (1-\alpha)\left|In-\hat{I}_n\right|$$

where α is a balance factor. An edge-aware depth smoothness loss is adapted to encourage the local smooth of depth map:

$$L\mathrm{smooth} = \frac{1}{N}\sum p^{N}|\nabla D(p)| \cdot \left(e^{-|\nabla I(p)|}\right)^{T}$$

3.2 Multi-task depth estimation using global context

Multi Task Learning has been developed for a single CNN model to handle a multitude of tasks and yield better results in all of them. Previous MTL methods based on CNNs commonly utilize parameter sharing, which share some layers across all tasks and add task-specific layers on the top of the shared networks. These naive approaches have two limitations. First, since these methods combine all the task-specific losses without considering optimal weight parameters, the model cannot learn multiple objectives properly. Thus, some papers propose ways to assign the weights to balance each task.

Second, task-specific features may discourage the network from performing other tasks. Alternative studies are presented to learn task-shared features and task-specific features, respectively. In [22], task-specific attention modules allow the shared network to achieve this goal. Maninis et al. also apply the attention mechanisms, such as Squeeze and Excitation blocks [20] and Residual Adapters to calibrate intermediate features. These approaches enable the separate learning of task-specific and task-shared features.

Chapter 4

Experiments and Results

4.1 Implementation details

In this thesis, we have extended Zhou et al's [25] work to incorporate the global context-aware features to improve the performance of our depth estimation pipeline. Our network is based on an encoder-decoder atchitecture with skip connections. To compare the results we have used the same *ResNet* encoder which is pretrained on *ImageNet* dataset. Without a direct association between tasks, task interference can occur, which can corrupt each task-specific feature. We propose a network with the parameter sharing that two tasks share an encoder and have each decoder branch. Therefore, the task-specific schemes are designed to prevent corruption in single encoder.

Attention : To avoid interference between the tasks of depth estimation and segmentation, we build an attention mechanism in order to adaptively recalibrate and fuse feature maps from task-specific decoder networks, we design a new architectural unit for our network architecture. The goal of this block is to explicitly model the correlation between the two task-specific feature maps before passing them to the separate decoders so that the network

can exploit the complementary features by learning to selectively emphasize more informative features from one task, while suppressing the less informative features from the other. We construct the topology of this block in a fully-convolutional fashion which empowers the network with the ability to emphasize features from a task- specific for only certain spatial locations or object categories, while emphasizing features from the complementary task for other locations or object categories. Moreover, this dynamically recalibrates the feature maps based on the input scene context.

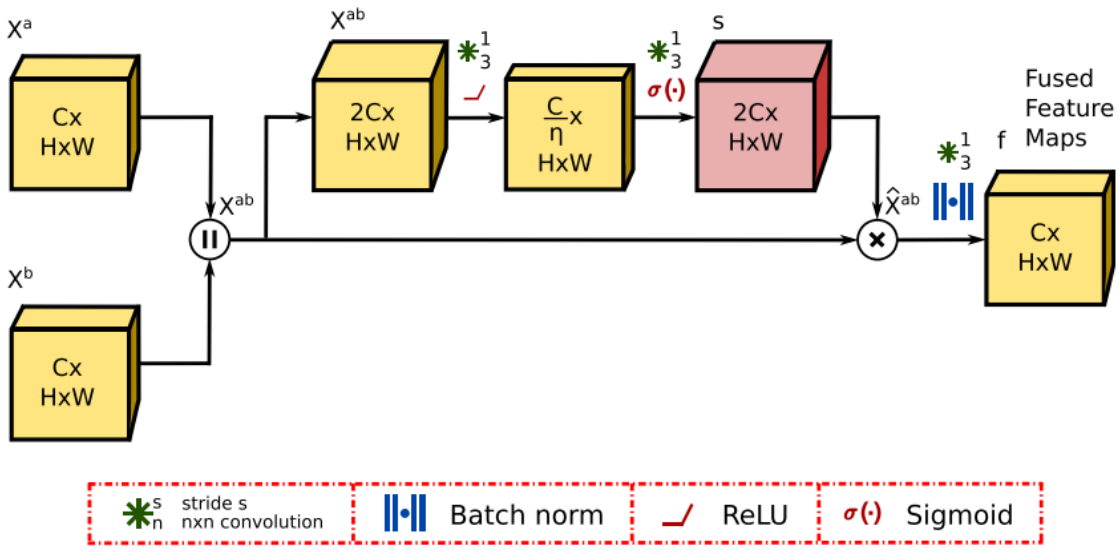

Figure 4: Our proposed Attention module for the depth network. The topology of our proposed Attention module that adaptively recalibrates and fuses task-specific feature maps based on the inputs in order to exploit the more informative features from the task-specific streams. η denotes the bottleneck compression rate.

The structure of the Attention block is shown in Figure 4. Let $\mathbf{X}^a \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{X}^b \in \mathbb{R}^{C\times H\times W}$ denote the task- specific feature maps from task A and task B respectively,

where C is the number of feature channels and $H \times W$ is the spatial dimension. First, we concatenate the task-specific feature maps to yield $\mathbf{X}^{ab} \in \mathbb{R}^{2 \cdot C \times H \times W}$. We then employ a recalibration technique to adapt the concatenated feature maps before fusion. In order to achieve this, we first pass the concatenated feature map $\mathbf{X}^{ab}$ through a bottleneck consisting of two 3×3 convolutional layers for dimensionality reduction and to improve the representational capacity of the concatenated features. The first convolution has weights $\mathcal{W}_1 \in \mathbb{R}^{\frac{1}{\eta} \cdot C \times H \times W}$ with a channel reduction ratio η and a non-linearity function $\delta(\cdot)$. We use ReLU for the non-linearity, similar to the other activations in the encoders.

The subsequent convolutional layer with weights $\mathcal{W}_2 \in \mathbb{R}^{2 \cdot C \times H \times W}$ increases the dimensionality of the feature channels back to concatenation dimension $2C$ and a sigmoid function $\sigma(\cdot)$ scales the dynamic range of the activations to the [0,1] interval. This can be represented as:

$$\mathbf{s} = F\left(\mathbf{X}^{ab}; \mathcal{W}\right) = \sigma\left(g\left(\mathbf{X}^{ab}; \mathcal{W}\right)\right)$$

The resulting output *s* is used to recalibrate or emphasize/de-emphasize regions as:

$$\hat{\mathbf{X}}^{ab} = F\left(\mathbf{X}^{ab}; \mathbf{s}\right) = \mathbf{s} \circ \mathbf{X}^{ab}$$

4.2 Training, Hyperparameter, and System details

For the training of the depth estimation, we resize all images to a resolution of 640×192 , if not mentioned otherwise, while for the semantic segmentation, the images are images to a resolution of 640×192, randomly cropped to the same resolution. We adopt the

zero-mean normalization for the RGB images used during training of the ResNet encoder. For input images we use augmentations including horizontal flipping, random brightness (±0.2), contrast (±0.2), saturation (±0.2) and hue (±0.1), while the photometric losses are calculated on images without color augmentations.

4.3 Evaluation criteria

To evaluate the depth estimation we follow other works [3,25] in computing four error metrics between predicted and ground truth depth as defined in [16], namely the absolute relative error (Abs Rel), the squared relative error (Sq Rel), the root mean squared error (RMSE), and the logarithmic root mean squared error (RMSE log). Additionally, we compute three accuracy metrics, which give the fraction δ of predicted depth values inside an image whose ratio and inverse ratio with the ground truth is below the thresholds 1.25, 1.25^2 and 1.25^3. We follow [25] by applying median scaling to the predicted depths. The semantic segmentation is evaluated using the mean intersection over union (mIoU).

The four error depth metrics used for evaluation on the Eigen and KITTI split are defined. The absolute relative error averages the error for all the predicted depth and the groundtruth depth pixels as shown below:

$$\text{Abs Rel} = \frac{1}{HW} \sum_{i \in \mathcal{I}} \frac{\left| d_i - \bar{d}i \right|}{\bar{d}i}$$

For the Squared relative error, one difference from absolute relative error is, before doing average over all the pixels in the image, the error between the predicted depth and the groundtruth depth is squared.

$$\text{Sq Rel } = \frac{1}{HW} \sum_{i \in \mathcal{I}} \frac{\left(d_i - \bar{di}\right)^2}{\bar{di}}$$

By measuring the Root mean square error metric on the test data, we will be able to understand the standard deviation of the residuals (Prediction errors). RMSE is a measure of how spread out these residuals are. Given the predicted depth pixel values and groundtruth values we can calculate it as follows:

$$\text{Rmse} = \sqrt{\frac{1}{HW} \sum_{i \in \mathcal{I}} \left(d_i - \bar{d}_i\right)^2}$$

Similarly, in case if we have an outlier depth values which is predicted by the network, the root mean square error may blow up into a bigger value, in order to overcome that we also use the root mean square error log while evaluating our model.

Let $\mathcal{I}$ being the set of all pixels and H and W being the width and height of the image, respectively. The accuracy metrics help us in finding how many of the predicted depth pixels are same as in the groundtruth data. In other words, when our model predicts depth on test images, the accuracy metrics tells us the percentage of pixels across all images that are correctly estimated. In order to calculate that we make use of Iversion bracket as shown below:

$$\frac{1}{HW} \sum_{i \in I} \left[\max\left(\frac{d_i}{\bar{di}}, \frac{\bar{di}}{d_i}\right) < 1.25 \right]$$

where [·] is defined as the Iverson bracket, which is 1 if the condition inside the bracket is true, and 0 if the condition is false. Similarly we also calcuarte accuracy metric at two other δ values as well. For thresholds of 1.25^2 and 1.25^3, we calculate a similar metric for all the test images.

4.4 Datasets used

For training and inference generation, following datasets have been used. We always utilize one dataset to train the semantic segmentation and another one for self-supervised training of the depth estimation of our model. For training the semantic segmentation we utilize the Cityscapes dataset [32] while at the same time we use different subsets of the KITTI dataset [33] for training the depth estimation. Similar to other state-of-the-art approaches we compare our depth estimation results by training and evaluating on the Eigen split [16] of the KITTI dataset, following [25] in removing static scenes from the training subset. We also train and evaluate on the single image depth prediction Benchmark split from KITTI [34]. To evaluate the joint prediction of depth and segmentation we utilize the KITTI split defined by [23] whose test set is the official training set of the KITTI Stereo 2015 dataset [35]. The number of training images deviates slightly from the original definitions, as we need a preceding and a succeeding image to train the depth estimation. The below table summarizes the datasets used:

Table 3: The table summarizes the amount of training data available for various dataset

Dataset	Subset	Number of Images
Eigen split	Train	21880
	Val	4187
	Test	697
Kitti split	Train	28937
	Val	1158
	Test	200
Cityscapes	Train	2975
	Val	500
	Test	1525

4.4.1 KITTI: Eigen split

The split of the KITTI dataset, which is most frequently used to compare depth estimation models, is the Eigen split [16], containing 697 images for testing. While the number of test images is constant throughout recent approaches, the number of training and validation images has been redefined by [71] to exclude static scenes. As shown in below table, there is a significant improvement in the RMSE values on the addition of the semantic supervision using attention mechanisms to our baseline model. It is to be noted that no post processing techniques have been used, demonstrating the sole effect of the attention guided semantic cues on an unsupervised monocular depth estimation model. In figure 4, we have shown qualitatively, how we obtain more continuous and crisp boundaries using our method.

Table 3: Evaluation of our self-supervised multitask attention guided depth estimation on the KITTI Eigen split.

Method	Resolution	Lower is better				Higher is better		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Zhou et al.[25]	416 × 128	0.198	1.836	6.565	0.275	0.718	0.901	0.960
Mahjourian et al.[3]	416 × 128	0.159	1.231	5.912	0.243	0.784	0.923	0.970
Yin and Shi et al.[36]	416 × 128	0.153	1.328	5.737	0.232	0.802	0.934	0.972
Wang et al.[10]	416 × 128	0.148	1.187	5.583	0.228	0.810	0.936	0.975
Casser et al.[37]	416 × 128	0.141	1.026	5.291	0.215	0.816	0.945	0.979
Meng et al.[26]	416 × 128	0.139	0.949	5.227	0.214	0.818	0.945	0.980
Godard et al.[23]	416 × 128	0.128	1.087	5.171	0.204	0.855	0.953	0.978
Our Method	416 × 128	0.124	1.140	4.960	0.198	0.860	0.949	0.982

4.4.2 KITTI: Kitti split

We train and evaluate on the KITTI split [18], whose test set are the official 200 training images from the KITTI 2015 Stereo dataset [40]. This test set has the advantage that it has available labels for both depth and semantic segmentation, which makes it suitable to observe the benefits of multi-task training for depth and semantic segmentation.

4.4.3 Cityscapes

The Cityscapes dataset has 2,975 labeled training images on which we train the semantic segmentation part of our network. Our evaluation on this dataset is conducted on the official validation set containing 500 labeled images.

Conclusion:

After evaluationg our proposed method on the KITTI test data, it is clearly evident that our method has gained a significant improvement in terms of the error metrics and accuracy. The proposed novel attention mechanism which helps in creating the semantic aware depth features has a positive impact on the overall performance of the model. From the baseline metrics, we can see that using a common encoder for both the tasks has helped our model to learn the task-sharing features without any features getting corrupted. To further enhance the results, we can try to implement our novel attention mechanism in a fully convolutional manner which would help to backpropagate the error through attention module.

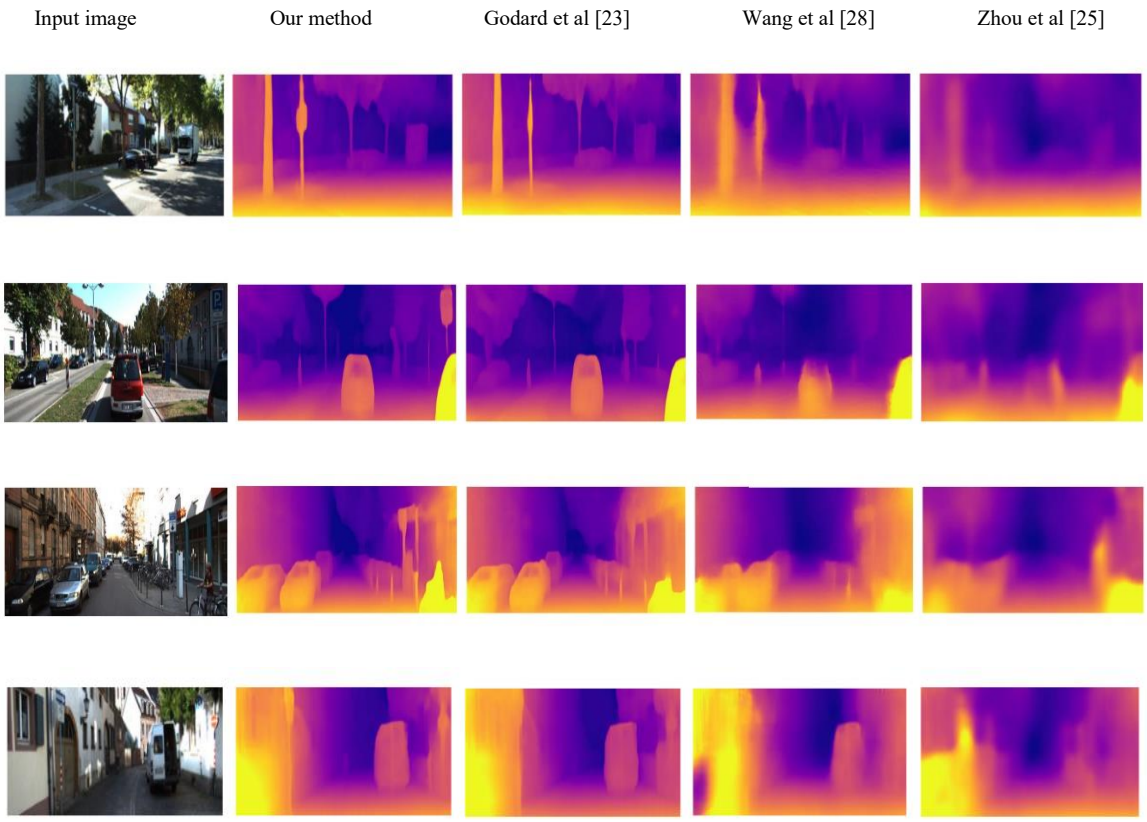

Figure 5: Qualitative examples of our proposed method in comparison to other recent methods.

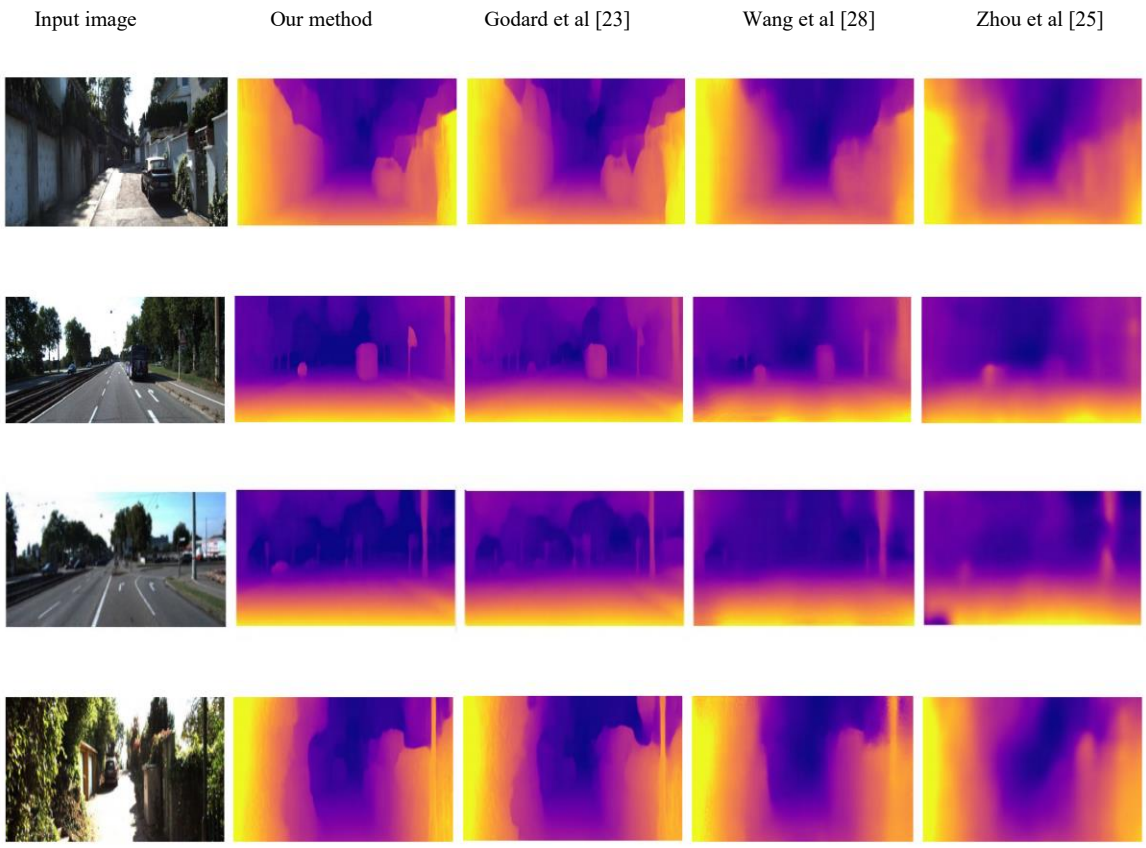

Figure 6: Additional examples of our proposed method in comparison to other recent methods.

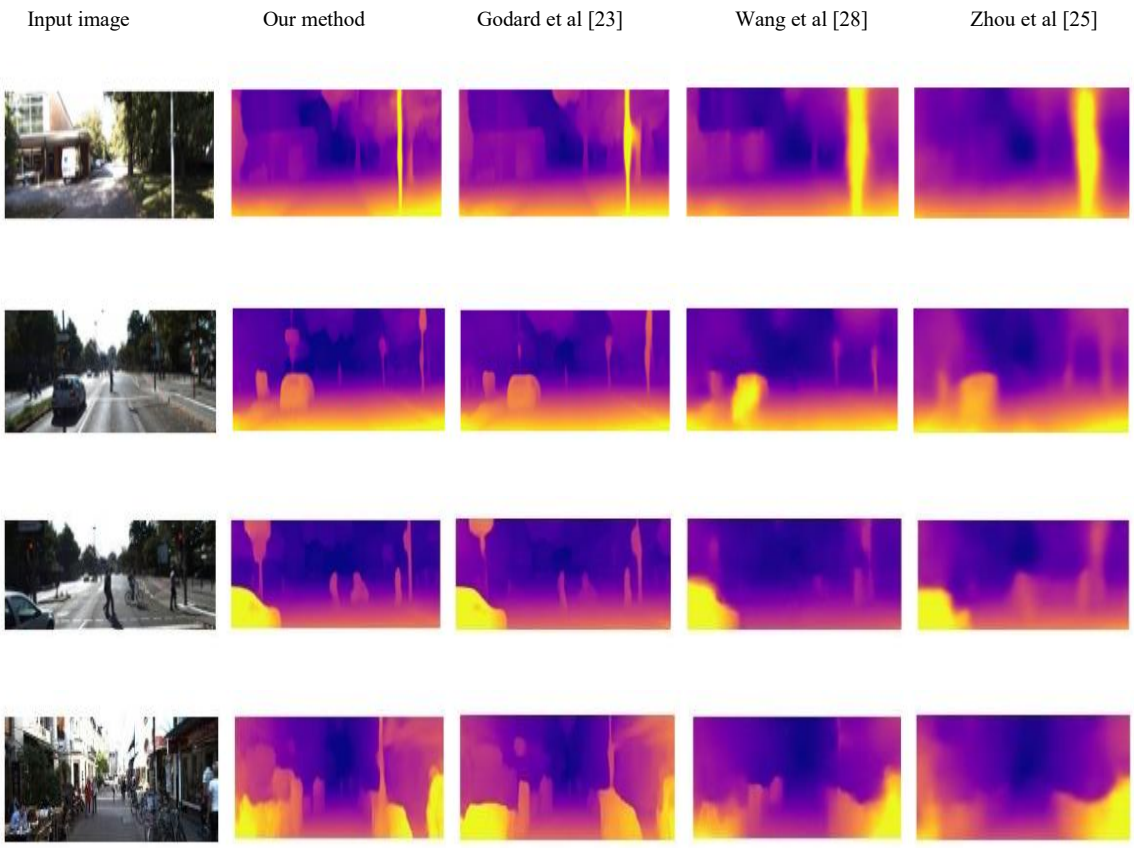

Figure 7: Additional examples of our proposed method in comparison to other recent methods.

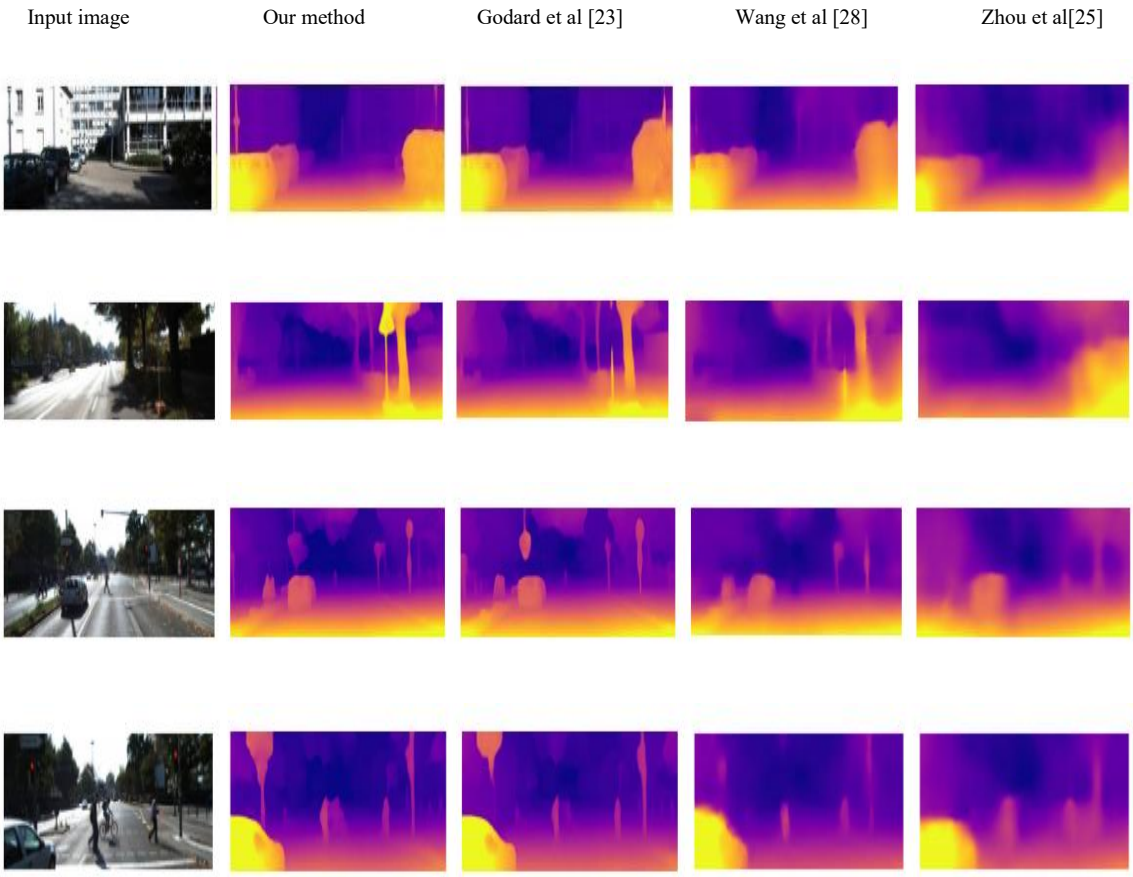

Figure 8: Additional examples of our proposed method in comparison to other recent methods.

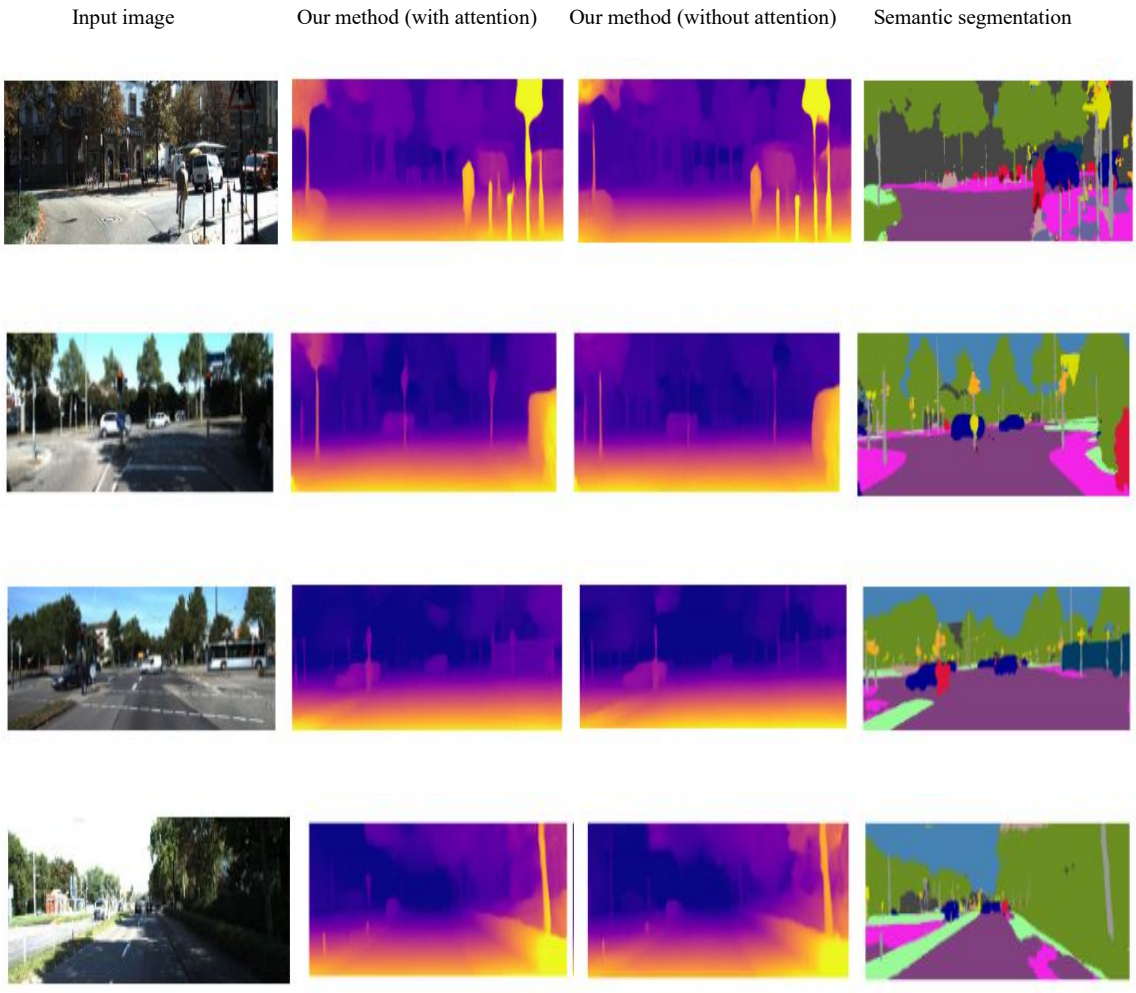

Figure 9: Examples of how our method compares to the models trained with and without attention guidance for the tasks of depth estimation and semantic segmentation.

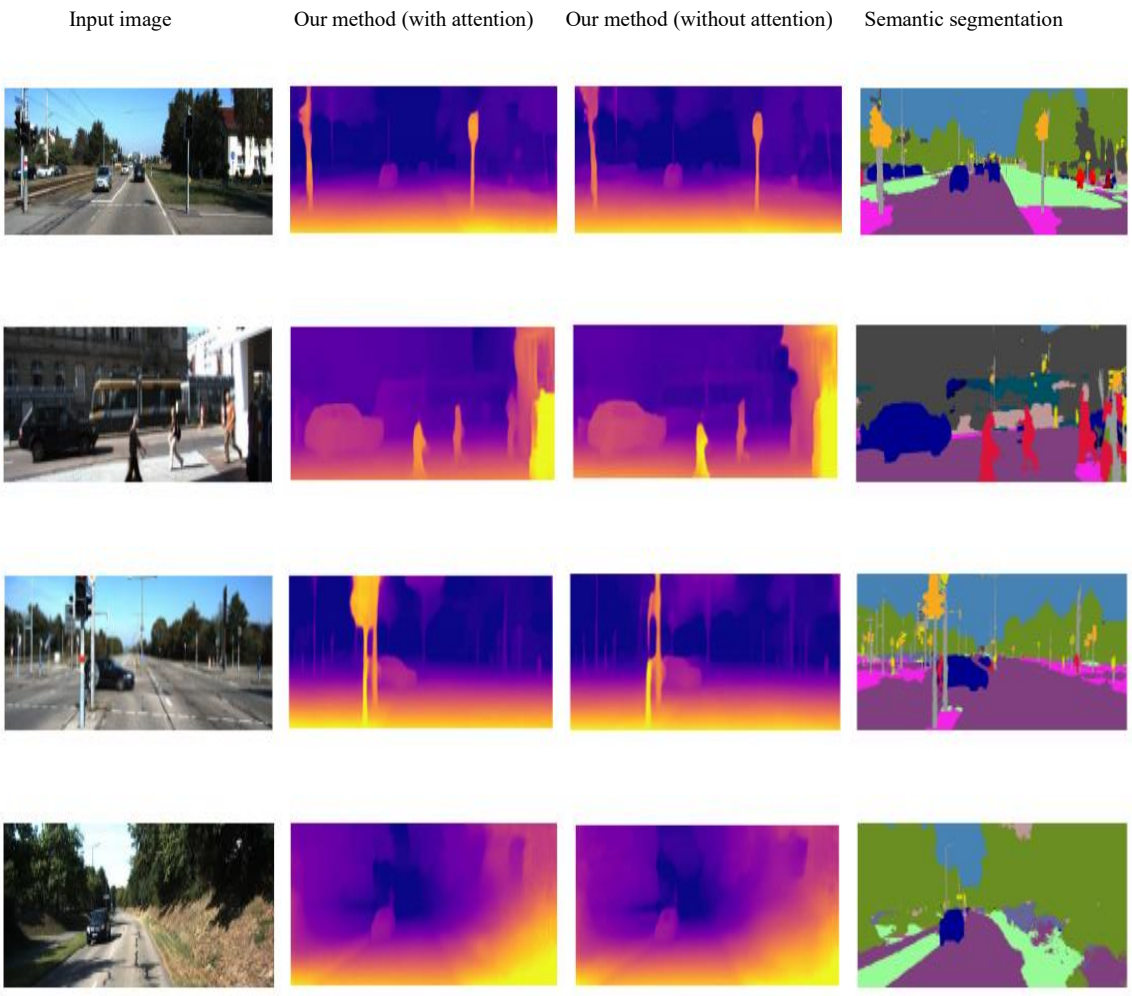

Figure 10: Additional examples of how our method compares to the models trained with and without attention guidance for the tasks of depth estimation and semantic segmentation.

References:

1. Visual Comfort of Binocular and 3-D Displays, Frank L. Kooi, Alexander Toet, *in* Proceedings of SPIE — The International Society for Optical Engineering 25(2):99–108 · August 2004

2. Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia. Every pixel counts: Unsupervised geometry learning with holistic 3d motion understanding. arxiv.org/pdf/1806.10556, 2018

3. R. Mahjourian, M. Wicke, and A. Angelova. 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, 2018.

4. D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo corre- spondence algorithms. IJCV, 47:7–42, 2002.

5. S. Ullman, "The interpretation of structure from motion," Proceedings ofthe Royal Society ofLondon. Series B. Biological Sciences, vol. 203, no. 1153, pp. 405–426, 1979.

6. L. Zou and Y. Li, "A method of stereo vision matching based on opencv," in 2010 International Conference on Audio, Language and Image Processing. IEEE, 2010, pp. 185–190

7. Z.-L. Cao, Z.-H. Yan, and H. Wang, "Summary of binocular stereo vision matching technology," Journal of Chongqing University of Technology (Natural Science), vol. 29, no. 2, pp. 70–75, 2015.

8. J. M. Facil, B. Ummenhofer, H. Zhou, L. Montesano, T. Brox, and J. Civera, "Cam-convs: camera-aware multi-scale convolutions for single-view depth," in Proceedings ofthe IEEE conference on computer vision and pattern recognition, 2019, pp. 11 826–11 835.

9. R. Garg, V. K. BG, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in European Conference on Computer Vision. Springer, 2016, pp. 740–756.

10. R. Wang, S. M. Pizer, and J.-M. Frahm, "Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5555–5564.

11. P. Chakravarty, P. Narayanan, and T. Roussel, “Gen-slam: Generative modeling for monocular simultaneous localization and mapping,” in 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019, pp. 147–153.

12. F. Aleotti, F. Tosi, M. Poggi, and S. Mattoccia, “Generative adversarial networks for unsupervised monocular depth prediction,” in European Conference on Computer Vision. Springer, 2018, pp. 337–354.

13. A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, “End-to-end learning of geometry and context for deep stereo regression,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 66–75

14. R. Mahjourian, M. Wicke, and A. Angelova, “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, 2018, pp. 5667–5675.

15. Y. Kuznietsov, J. Stuckler, and B. Leibe, “Semi-supervised deep learning for monocular depth map prediction,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6647–6655

16. D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in Advances in neural information processing systems, 2014, pp. 2366–2374

17. D. Xu, E. Ricci, W. Ouyang, X.Wang, and N. Sebe, “Monoc- ular depth estimation using multi-scale continuous crfs as sequential deep networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018

18. B. Li, C. Shen, Y. Dai, A. Van Den Hengel, and M. He, “Depth and surface normal estimation from monocular im- ages using regression on deep features and hierarchical CRFs,” in Proceedings of the IEEE Computer Society Con- ference on Computer Vision and Pattern Recognition, 2015.

19. Q. Huang, M. Han, B. Wu, and S. Ioffe, “A hierarchical conditional random field model for labeling and segmenting images of street scenes,” in CVPR 2011. IEEE, 2011, pp. 1953–1960.

20. L. Ladick`y, C. Russell, P. Kohli, and P. H. Torr, “Associative hierar- chical crfs for object class image segmentation,” in 2009 IEEE 12th International Conference on Computer Vision. IEEE, 2009, pp. 739– 746

21. F. Liu, C. Shen, G. Lin, and I. Reid, “Learning depth from single monocular images using deep convolutional neural fields,” IEEE trans- actions on pattern analysis and machine intelligence, vol. 38, no. 10, pp. 2024–2039, 2015

22. D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, “Structured attention guided convolutional neural fields for monocular depth estimation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3917–3925

23. C. Godard, O. Mac Aodha, and G. J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” in CVPR, vol. 2, no. 6, 2017, p. 7.

24. M. Jaderberg, K. Simonyan, A. Zisserman et al., “Spatial transformer networks,” in Advances in neural information processing systems, 2015, pp. 2017–2025

25. T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, “Unsupervised learning of depth and ego-motion from video,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1851–1858

26. Meng, Y., Lu, Y., Raj, A., Sunarjo, S., Guo, R., Javidi, T., Bansal, G., Bharadia, D.: SIGNet: Semantic Instance Aided Unsupervised 3D Geometry Perception. In: Proc. of CVPR. pp. 9810–9820. Long Beach, CA, USA (Jun 2019)

27. R. Szeliski, “Prediction error as a quality metric for motion and stereo,” in Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2. IEEE, 1999, pp. 781–788.

28. Wang, R., Pizer, S.M., Frahm, J.M.: Recurrent Neural Network for (Un-)Supervised Learning of Monocular Video Visual Odometry and Depth. In: Proc. of CVPR. pp. 5555–5564. Long Beach, CA, USA (Jun 2019)

29. Maninis, K.K., Radosavovic, I., Kokkinos, I.: Attentive single-tasking of multiple tasks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1851–1860 (2019)

30. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)

31. Rebuffi, S.A., Bilen, H., Vedaldi, A.: Efficient parametrization ofmulti-domain deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8119–8127 (2018)

32. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes Dataset for Semantic Urban Scene Understanding. In: Proc. of CVPR. pp. 3213–3223. Las Vegas, NV, USA (Jun 2016)

33. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision Meets Robotics: The KITTI Dataset. International Journal of Robotics Research (IJRR) 32(11), 1231–1237 (Aug 2013)

34. Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., Geiger, A.: Sparsity Invariant CNNs. In: Proc. of 3DV. pp. 11–20. Verona, Italy (Oct 2017)

35. Menze, M., Geiger, A.: Object Scene Flow for Autonomous Vehicles. In: Proc. of CVPR. pp. 3061–3070. Boston, MA, USA (Jun 2015)

36. Yin, Z., Shi, J.: GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose. In: Proc. of CVPR. pp. 1983–1992. Salt Lake City, UT, USA (Jun 2018)

37. Casser, V., Pirk, S., Mahjourian, R., Angelova, A.: Depth Prediction Without the Sensors: Leveraging Structure for Unsupervised Learning from Monocular Videos. In: Proc. of AAAI. pp. 8001–8008. Honolulu, HI, USA (Jan 2019)

38. https://www.cl.cam.ac.uk/teaching/1011/CompVision/Town%20-%20ComputerVision%20-%20L1.pdf

39. https://unsplash.com/photos/X5AhPIuKsrA?utm_source=unsplash&utm_medium=referral&utm_content=creditShareLink

40. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in neural information processing systems. pp. 2366–2374 (2014)

41. Garg, R., BG, V.K., Carneiro, G., Reid, I.: Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: European Conference on Computer Vision. pp. 740–756. Springer (2016)

42. 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)

43. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in neural information processing systems. pp. 2017–2025 (2015)

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

45. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A.: End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 66–75 (2017)

46. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4040–4048 (2016)

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

48. SFM Self Supervised Depth Estimation: Breaking Down The Ideas | by Daryl Tan | Towards Data Science

49. V. Guizilini, R. Hou, J. Li, R. Ambrus, and A. Gaidon. Semantically-guided representation learning for self-supervised monocular depth. In ICLR, 2020.

50. A. Johnston, G. Carneiro. Self-supervised Monocular Trained Depth Estimation using Self-attention and Discrete Disparity Volume, In CVPR 2020